社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究資料
JSAI Technical Report
SIG-Challenge-A902-7

# Automatic Speech Recognition Under Ego-motion Noise of a Robot

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

[1]Honda Research Institute Japan Co., Ltd. 8-1 Honcho, Wako-shi, Saitama 351-0188, Japan,
{gokhan.ince, nakadai, yuji.hasegawa, tsujino}@jp.honda-ri.com

[2]Honda Research Institute Europe GmbH, Carl-Legien Strasse 30, 63073 Offenbach, Germany,
tobias.rodemann@honda-ri.de

[3]Dept. of Mech. and Env. Informatics, Graduate School of Information Science and Eng.,
Tokyo Institute of Technology 2-12-1-W8-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan,
imura@mei.titech.ac.jp

## Abstract

Active auditory perception related tasks like sound localization and speech recognition have to be performed with high accuracy even while the robot is moving. However, the joints of the robot inevitably generate noise because of the active motors, i.e. ego-motion noise. This problem is very critical, especially in humanoid robots, because they tend to have a lot of joints and the motors are located relatively closer to the microphones than the sound sources. In this work, we investigate methods for the prediction and suppression of the ego-motion noise. In the first part, we analyze the performance of different noise subtraction strategies, assuming that the noise prediction problem has been solved. In the second part, we present some results for a noise prediction scheme based on the current robot joint status. Performance is evaluated for a number of criteria, including Automatic Speech Recognition (ASR). We demonstrate that our method improves recognition performance during ego-motion considerably.

## 1 Introduction

An active auditory perception system is very essential for robots to be able to interact with their environment. Tasks like sound localization and speech recognition have to be performed with high accuracy even when the head (or whole robot) is moving. Unfortunately, the research done in the field of active audition suffers highly from this additive motor noise, which deteriorates the quality of the recorded sounds considerably. Therefore two restricting assumptions are made very often: Either the sounds are selected loud enough to ignore the motor noises generated during the body motion, or the sound processing is performed without movement at all [1]. An alternative method that overcomes the noise problem is utilization of a separate close-talk microphone [2], nevertheless it limits human-robot interaction.

In our research, the goal is to tackle the noise problem directly. We propose to utilize a biologically-inspired method for learning and suppressing the ego-noise that *weakly-electric fishes* exploit in the nature. They have evolved sensory systems that make use of copies of their self-generated dynamic electric wave patterns to decode the temporal characteristics of incoming sensory signals from the surrounding waves [3]. Localization and scene analysis procedures involve the computation of the spatial map of sensory expectations from recent inputs, and removal of the ego-motion effects, namely the spike events, from the total input image [4]. The ego-noise cancellation on a robot could be accomplished by autonomous mechanisms similar to the electrosensory system of the electric fishes, just like the way the animal learns what kind of noise template it has to subtract in case of the execution of a certain motor plan. In this paper, we first deal with fixed motion patterns that follow known trajectories. This approach is suitable for focusing on the noise suppression problem explicitly. Then, we generalize the ego-noise problem for freely moving robots by showing methods how the noise could be predicted. We demonstrate that the proposed methods can eliminate motor noise by evaluating them qualitatively in terms of ASR results.

### 1.1 Comparison to Related Work

In the field of "Robot Audition", noise suppression is mostly carried out using sound source seperation techniques with a microphone array [5]. However, in our case, the motors are located in the near field of the microphones and produce more like diffuse rather than directional sounds. In a standard task with robot motions where acoustic conditions such as power, frequencies and locations of the motor noise sources dynamically change at each time instance, the performance of sound source separation and ASR deteriorates drastically even when a microphone array is used. Nakadai et al [6] proposed a noise cancellation method with two pairs of microphones. One pair in the inner part of the shielding body records only internal motor noise and helps the sound localizer to distinguish between the spectral subbands that are

noisy and not noisy, and to ignore the ones where the noise is dominant. In contrary to our approach, this technique does not suppress the noise. Nishimura et al [7] estimates the ego-noise using robot's gestures and motions. With the help of the motion command, the pre-recorded correct noise template matching to the recent motion is selected from the template database and subtracted. Compared to their small set of noise template database of limited motions, we target to deal with the whole ego-noise space that is generated by any possible motor combination of the robot. Ito et al [8] developed a new approach of frame-by-frame based prediction with a neural network (NN) to cope with unstable walking noise. The trained network had to predict the noise spectrum from angular velocities of the joints of the robot. However, they concentrated on a small robot with limited degrees of freedom. For a huge dataset, NN will have a slow training speed and online adaptation is difficult to achieve. Therefore we rather propose the usage of a template database due to its efficiency and additionally enhance the accuracy of the templates further by incorporating more information related to the joints. Besides, both Nishimura [7] and Ito [8] based their research mainly on the estimation of templates for different motions, but neither focused on the possibility of quality improvement by utilizing spectral enhancement optimization factors nor evaluated the performance with any other criteria except ASR.

## 2 Blockwise Template Subtraction

This section gives an outline of the noise reduction strategy that we followed. Main point of investigation in this section is clearly not the prediction of the noise, but the suppression of it. Therefore, we concentrated especially on a single motion (quick horizontal motion of the neck) generated by the experimental robot head which looks like the head of ASIMO.

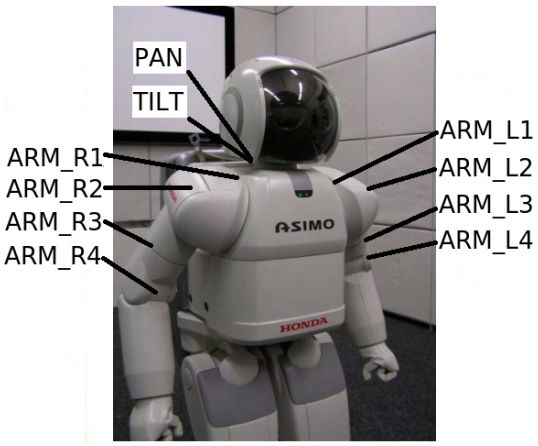


Figure 1: Hardware setup of ASIMO.

### 2.1 Template Generation

Spectro-temporal investigations conducted on the recorded ensemble of noise data for the same motion (same origin, target, velocity and onset time) revealed following results:

- The regions of the spectrum where noise power is densely distributed, correspond to the increased rotational velocity of the motor (see Fig. 2 for the case of one active joint). Most critical phases are acceleration and breaking.
- The energy distribution remains nearly the same during the constant velocity phase.
- The duration of the signals does not change by more than a few samples.
- Envelope shape does not deviate much from the mean envelope of the same motions.

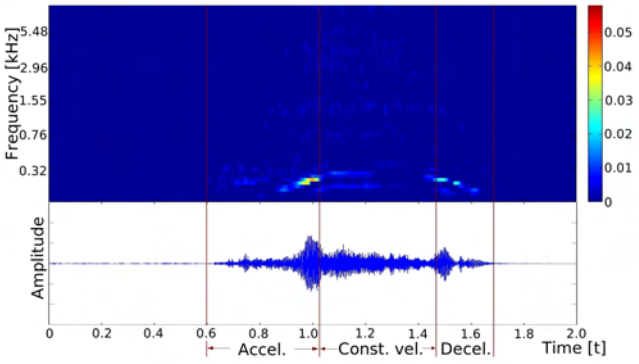


Figure 2: Envelope spectrum of the head motor noise for a rotation from -70° to 70° in the horizontal plane

The underlying notion for our first method, *blockwise template estimation*, relies on the idea that the motor noise can be predicted, if the motion performed by the robot has a pattern of a prior known duration and onset time. Noise spectra of different motions can be recorded by repeating the same motion $M$ times. An important preprocessing step after short-time spectral decomposition is the removal of stationary background noise, which involves an adapted version of Cohen's Minimum Controlled Recursive Averaging [9]. Furthermore, the electrical noise of the motors (static noise caused by the electrical circuits) is also suppressed by this background noise reduction scheme, so that only non-stationary mechanical noise remains as a final product of the processing chain. Template generation follows as the consequent stage. Time alignment of recorded motor noise is required before calculating the templates. The synchronization point regarding each element is determined at the sample number where the cross-correlation function of each spectrum and pilot (one specific pre-selected instance) spectrum gets its maximum value. Let $D(n, \Omega)$ be the short-time basis frequency spectrum of the distortion (motor noise), where $\Omega$ stands for the discrete frequency representation and $n$ for the current frame. A single template is represented by an average matrix $\overline{D}(n, \Omega)$ and a standard deviation matrix $\sigma_D(n, \Omega_i)$ such as follows:

$$\overline{D}(n,\Omega) = \frac{1}{M}\sum_{k=1}^{M} D(n,\Omega) \quad (1)$$

$$\sigma_D(n,\Omega) = \sqrt{\frac{1}{M}\sum_{k=1}^{M}\left(D_k(n,\Omega) - \overline{D}(n,\Omega)\right)^2} \quad (2)$$

### 2.2 Template Subtraction

Let $S(n,\Omega)$ and $D(n,\Omega)$ be the spectrum of useful signal and motor noise, respectively. Then the spectrum of the observed signal $Y(n,\Omega)$ is defined by

$$Y(n,\Omega) = S(n,\Omega) + D(n,\Omega). \quad (3)$$

The spectrum of the useful signal can be estimated by using the inverse operation:

$$Y_r(n,\Omega) = Y(n,\Omega) - \overline{D}(n,\Omega), \quad (4)$$

where $Y_r(n,\Omega)$ stands for the spectral magnitude comprising the magnitudes of useful sound and residual motor noise. The reason for the existence of this residual magnitude is that the original magnitudes of the motor noise $D(n,\Omega)$ deviate from their arithmetic mean $\hat{D}(n,\Omega)$. To compensate this error, we further suggest to use spectral subtraction approach that exploits *overestimation factor*, $\alpha$, and *spectral floor*, $\beta$. $\alpha$, also termed *aggressiveness factor*, allows a compromise between perceptual signal distortion and noise reduction level. On the other hand, $\beta$ is required to deal with the problem called *musical noise*. The cause of musical noise is a nonlinear mapping of the negative or small-valued spectral estimates, producing a metallic noise sounding like the sum of tone generators with random fundamental frequencies which are turned on and off constantly [10]. $\beta$ reduces the effect of the sharp valleys and peaks in the spectrum which is caused by the smaller attenuations of the frequencies compared to relatively larger attenuations of their neighboring frequencies due to the random fluctuations in the magnitude estimations. *Overestimated template subtraction* is introduced such as in the following formula:

$$\hat{H}_{SS}(n,\Omega) = max\left(1 - \alpha(n,\Omega)\frac{\hat{\sigma}_D(n,\Omega)}{Y_r(n,\Omega)}, \beta(n,\Omega)\right), \quad (5)$$

Finally, the template is conceptually 'subtracted', by weighting the signal $Y_r(n,\Omega)$ with the gain coefficients $\hat{H}_{SS}(n,\Omega)$:

$$\hat{S}(n,\Omega) = Y_r(n,\Omega)\cdot\hat{H}_{SS}(n,\Omega) \quad (6)$$

## 3 Parameterized Template Subtraction

In this section, we explain the techniques that are necessary to extend the proposed solution of the ego-noise reduction problem from a stereotyped motion level towards complicated motions with higher degrees of freedom. So far disregarded subjects like synchronization of templates, effect of increased number of motors and noise prediction are inspected further in this section.

Note that the *blockwise template subtraction* had several shortcomings, e.g. it could be performed properly only after the detection of the exact starting moment of the template, which is a very hard task to achieve. Another drawback was that it would require a large collection of signal representations consisting of the motor noise statistics like average values and standard deviations of the whole dataset of a given motion. Besides, it requires a huge amount of data for each possible motion. Considering the impossibility to collect and produce templates for each joint of different combinations of origin, target, position, velocity and acceleration parameters, the former approach was simply not feasible to be applied in a realistic scenario.

To overcome these deficits, a new technique is proposed that parameterizes a discrete audio segment under consideration using motor status and get a spectral energy vector to represent the ego-noise at that time instant. The experiments for parameterized template subtraction are conducted on Honda (humanoid robot) ASIMO (Fig. 1) due to the necessity of additional body joints beside the head motors. ASIMO has sensors that measure the angular positions of all of its joints separately.

### 3.1 Template Generation

For that purpose, joint status information provided by the sensors on the motors will be utilized, with the following assumptions:

- Current motor noise depends on position, velocity and acceleration of that specific motor.
- Similar combinations of joint status will result in similar motor noise spectral vectors at any time instance.
- The superposition of single joint motor noises at any arbitrary time equals to the whole body noise at that specific time instance.

Figure 3 illustrates the proposed template generation scheme. During the motion of the robot, actual position ($\theta$) 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 $N$ joints are active, feature vectors consisting of $3N$ 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 $F = [\theta_1, \dot{\theta}_1, \ddot{\theta}_1, \theta_2, \dot{\theta}_2, \ddot{\theta}_2, \ldots, \theta_N, \dot{\theta}_N, \ddot{\theta}_N]$. At the same time, motor noise is recorded and spectrum of the motor noise is calculated by the sound processing branch running in parallel. Both feature vectors and spectra are continuously labeled with time tags so that templates are generated when their time tags match. Finally, a large noise template database that consists of short noise templates for many joint configurations is created.

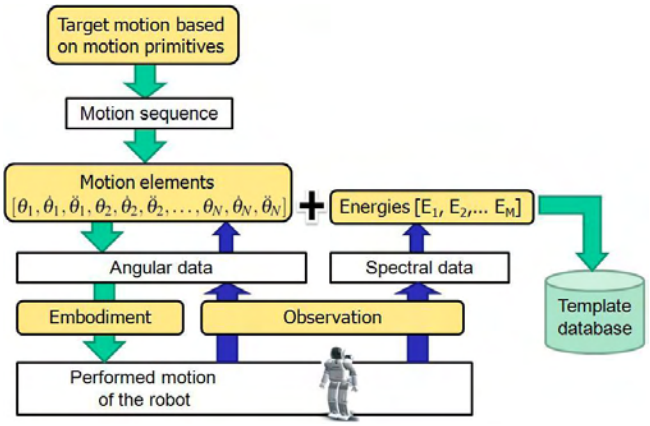


Figure 3: Flowchart of the proposed template generation and database creation.

### 3.2 Template Prediction and Selection

The prediction phase starts with a search in the database for the best matching template of motor noise for the current time instance. Finding the correct template involves a search among all the templates in the database for most similar joint configuration. We implemented a nearest neighbor (1-NN) search to accomplish this task. The spectral vector associated with the point in the database that has the shortest distance to the query point is used as the template. The prediction process is applied for every frame. In that sense, the block template for an arbitrary motion (e.g. neck motion template in Sec. 2) can be regarded as the concatenation of smaller templates that are predicted according to the abovementioned approach on a frame-by-frame basis with the following setting:

For a given database $\mathbf{S}$ of templates in $3N$-dimensional feature space $V$ and a query point $\mathbf{q} \in V$, find the closest point in $\mathbf{S}$ to $\mathbf{q}$. $V$ is taken to be the $3N$-dimensional Euclidean space and the distance is measured by the Euclidean distance between two points $\mathbf{q} = (q_1, q_2, ..., q_{3N})$ and $\mathbf{s} = (s_1, s_2, ..., s_{3N})$, where $\mathbf{s}$ is an element of the set $\mathbf{S}$.

$$d(\mathbf{q}, \mathbf{s}) = ||\mathbf{q} - \mathbf{s}|| = \sqrt{\sum_{i=1}^{3N} (q_i - s_i)^2} \qquad (7)$$

The spectral vector associated with $\mathbf{s}$ having the shortest distance to $\mathbf{q}$ is used as the template.

### 3.3 Template Subtraction

On contrary to blockwise template subtraction, there is no ready-to-use average template for parameterized template subtraction. Occasionally, the prediction accuracy could even become very low. In this respect, we employ a slightly changed version of weight calculation formula for spectral subtraction:

$$\hat{H}_{SS}(n, \Omega) = max\left(1 - \alpha(n, \Omega)\frac{D_{pr}(n, \Omega)}{Y(n, \Omega)}, \beta(n, \Omega)\right), \qquad (8)$$

where $D_{pr}(n, \Omega)$ stands for predicted template. This operation is followed by Eq. 6 to finish the noise reduction operation.

## 4 Results

For the first part of our experiments, we evaluated the blockwise template subtraction. Tests are done on the robot head which is a close derivative of the actual ASIMO head. It is equipped with Sennheiser DPA 4060-BM omni-directional microphones for recording. We used only one microphone on the left side. For more information regarding the ears and pinnae refer to [11]. The head motor is an Amtec Robotics PowerCube070. Data was recorded in a noisy, very echoic room ($T_{60} = 1100ms$). The tests for blockwise template subtraction are focused on motor noise signals generated by a horizontal motion of 140° with a very high angular velocity ($v_{max} = 200°/sec$). Sampling rate was set to 48kHz. We used a Gammatone filterbank with 60 channels where center frequencies are increasing quasi-logarithmically from 100Hz to 10 kHz.

The obtained noise signals are added to clean male speech. Not only the signal-to-noise ratio (SNR) is very low (nominalSNR=-5.7dB and segSNR=-2.8dB), but also the frequency bins with high energy content of both speech and noise are overlapping. These signals and their spectrally enhanced versions after noise reduction are evaluated using Perceptual Evaluation of Speech Quality (PESQ, ITU-T P.862 Standard). It is designed to calculate an index value of quality that correlates to a mean opinion score (MOS) given by human subjects in evaluation sessions. It predicts subjective opinion scores of a degraded audio sample in a range from 4.5 to -0.5, with higher scores indicating better quality. Results in relation with $\alpha$ and $\beta$ are given in Tab. 1. When the aspect of *intelligibility* is considered, overestimated subtraction with low spectral floor is not appropriate for speech enhancement, because the human ear is especially sensitive to musical noise. Therefore, high spectral floor values ($\beta > 0.3$) are desirable. The best score is achieved when mean template subtraction is applied.

Table 1: PESQ results for Magnitude Spectral Subtraction

| MOS for noisy signal: 0.361 | | MOS after mean template subtraction: 2.681 | | | | | |
|---|---|---|---|---|---|---|---|
| MOS values | | $\beta$ | | | | | |
| | | 0.0 | 0.2 | 0.3 | 0.5 | 0.8 | 0.9 |
| $\alpha$ | 1 | 0.329 | 0.309 | 0.277 | 0.221 | 1.526 | 2.429 |
| | 1.5 | 0.322 | 0.291 | 0.216 | 0.249 | 1.535 | 2.606 |
| | 2 | 0.313 | 0.312 | 0.262 | 1.448 | 1.541 | 2.592 |
| | 2.5 | 0.331 | 0.248 | 0.244 | 1.464 | 1.545 | 2.194 |
| | 3 | 0.35 | 0.241 | 0.291 | 1.476 | 1.546 | 2.619 |
| | 4 | 0.255 | 0.338 | 1.371 | 1.439 | 1.547 | 2.62 |

The second evaluation criteria we utilize, exploits the *Precedence Effect* [12], which makes localization in echoic

environments possible for humans. Using this model, the detection of noise and sound signals is to be verified on their onset points. Onsets are the points where a position measurement for sound localization is done. They are frames where the signal amplitude increases and the effect of echoes is still small. Therefore we assume, the larger the energy of the onset, the larger the impact on localization. They are used particularly for sound localization in order to suppress the onsets caused by the echoes of the same sound source, by introducing the inhibition of the local echo onset points other than these particular desired signal onsets (See [13]).

Provided that the noisy signal consists of the superposition of the noise and speech signals, the onsets of both signals can be extracted separately by giving only the interested signal to the input. That way, the energies and positions of the onsets are saved individually. A likelihood method is introduced so that the onsets of the degraded signal can be compared with the onsets of its noise and speech components assessed before. Given a certain confidence area (explained below), it should return an objective measure how likely the onsets of the degraded signal are to its nearest onset belonging to either one of those classes.

Considering that the onsets are computed for each channel, two parameters are selected to tune the confidence area, namely the *timing* and the *energy* of the onsets. An optimized *timing* confidence interval of 60 ms defines the limit of interest for the corresponding onsets. The onsets beyond the limits are considered as completely dissimilar onsets. The second parameter in the confidence area is the *energy* level of the onset. The onset of a class whose energy seems to be reduced far more than the other class is rewarded more. The total confidence value (product of the position and energy confidence) acts as an indicator for the competition between the noise and speech onsets in the reference onset set. The candidate which has the greatest confidence value is selected as the winner and the onset is assigned to belong to either speech, motor noise, or indecisive category. This method gives out a measurement bench how many onsets from the noise are suppressed, how much energy has remained in the onsets of the noise (see Fig. 4).

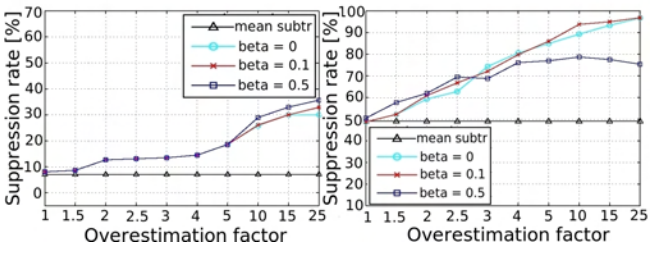


Figure 4: Onset based results using magnitude spectral subtraction for (a) voice onset energy suppression rates (b) noise onset energy suppression rates

The results demonstrated that the higher the overestimation factor is selected, the more the noise reduction is achieved. Template reduction can suppress 76% of the total energy of noise onsets, while keeping voice suppression in low rates like 15%. (see Tab. 1 and Fig. 4 for $\alpha = 4$ and $\beta = 0.5$)

We also evaluated the speech recognition results with Sphinx-4 to inspect the *qualitative* aspects of our noise suppression scheme. Totally 200 evaluation word sequences (Resource Management Speech Corpus) are selected each comprising of 5 to 12 words chosen randomly. Utterances belong to both male and female speakers. The recognition is performed speaker- and gender- independent. No grammar is used in the tests. The results will be evaluated for 7 different SNR values between approximately -10 and 40 dB.

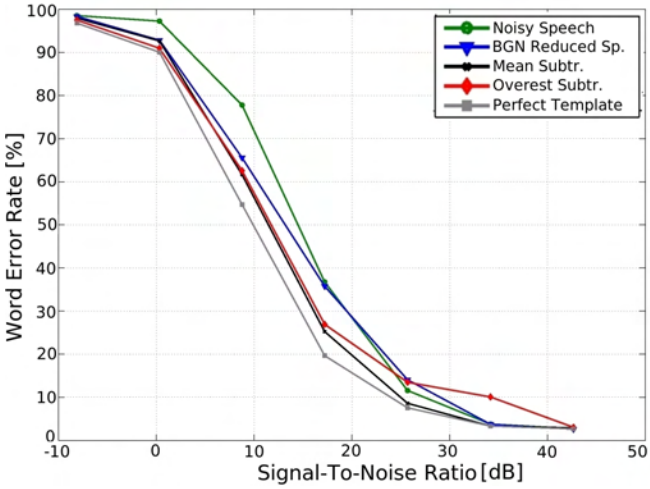


Figure 5: ASR results

The experiments carried out with ASR show that the word error rate after mean template subtraction decreases substantially in the sensible region between -5dB and 30dB compared to both the reference recognition results with noisy signals and to the results after applying stationary background noise reduction scheme as shown in Fig. 5. For an SNR value of 17dB, the improvement is 12% and for 8dB case 16% improvement is achieved. For an additional test bench, the recognition performance of an perfect template is introduced as well. This perfect template is in fact nothing but the identical spectrogram that the motor noise has. This defines the upper limit of performance and defines a benchmark for the comparison of all methods by providing a best case scenario.

It is also clear that the templates generated by variance weighting (Overestimated Temp. Subtraction) are not suitable to be applied to the signals with high SNR. They worsen the Word Error Rate (WER), which is an expected consequence coinciding with the results obtained from the previous PESQ and onset measurement tests. However, recognition for low SNRs (below 0dB) yields better performance if an overestimation of the noise variance is used (within a certain range). For moderate SNR levels, usage of variance weighting techniques reduces WERs by up to 10%.

The second part of the experiments is carried out on ASIMO. Experiment involves random motions of 10 different joints simultaneously. We rotated the head of

ASIMO (elevation = [-30° 30°], azimuth = [-90° 90°]) randomly, while the arms were performing a random grasp motion in the reaching space of the body without moving its torso or hip. Status information of the motors are gathered from the joints with an average acquisition rate of 7.3 ms. ASIMO also has a circular array consisting of 8 microphones mounted on the head. We made evaluations using the data recorded from the third microphone that corresponds to a spatial position of 90° counterclockwise with respect to the front. The training data was a joint database consisting of 30 minutes of motor noise and the corresponding feature vectors stored during this time span. The probability was very high that no similar motions could be generated for this scenario with another arbitrarily generated random trajectory. In that case, the performance of the experiments would be biased by the inappropriately selected test set. Therefore, we followed a similar trajectory used in the training session but with a sequence of slightly different destination points as before that deviated in their final positions by a certain random displacement. This distance is determined by a Gaussian distribution with a variance of $\sigma$=0.1. We stored a test database of 10 minutes long. Data is recorded in a noisy and echoic room (reverberation time (RT20) was about 0.2 seconds). Data was sampled on 16kHz and frame shift was 12.5 ms. Hamming window of 16 ms was used.

We evaluated the effectiveness of the proposed approach using Julius which is an open-sourced ASR. For this experiment, we created 35 different motor noise patterns from the test set. The length of each test set is kept flexible so that it matches the duration of the utterances used in the wordset. As speech corpus, ATR phonemically-balanced wordset (ATR-PB) was used. This word-set includes 216 Japanese words and average word correct rate was calculated as depicted in Fig. 6. Please note that the signals were formerly subject to stationary background noise reduction. Hence, SNR values are given for signal-to-motor noise ratio (background noise of the room was approx. 5dB).

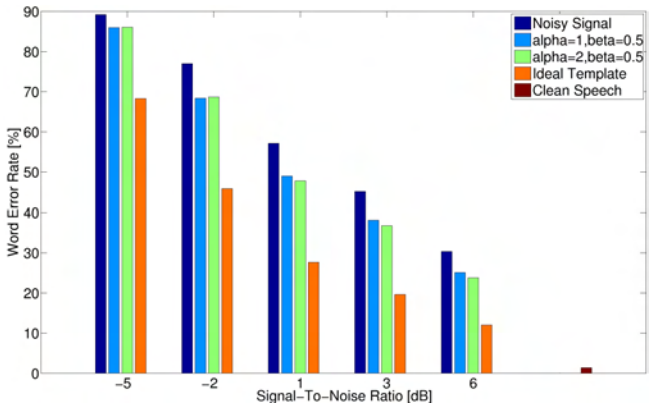


Figure 6: Recognition performance for different spectral subtraction settings

The graph shows that a template subtraction with $\alpha = 2$ and $\beta = 0.5$ is slightly better than a subtraction with $\alpha = 1$ and $\beta = 0.5$ for high SNR values. Latter set of parameters allow us to obtain improvement rates of up to 10% for SNR conditions that can be observed in a realistic human-robot interaction. Fig. 7 illustrates the ASR performance distribution on 35 different noise test sets. With the exception of two test cases (#12 and #29), high improvement rates are achieved. We also depicted the recognition rates for an ideal template subtraction for comparison. Ideal template represents the template that is constructed for the current test motor noise using the predictions from the test set. The reason of the gap between ideal template subtraction performance and the results for overestimated template subtraction with optimal settings is due to the incorrect predictions of the template. Nearest neighbor search does not make a decision on whether the final prediction is a reasonably correct template, that is why it is called a *lazy learning algorithm*. The errors are mostly caused by the absence of similar templates that are available for the current motor status combination.

Because the feature set has big impact on the prediction accuracy, we also tested the influence of the feature vector selection. For that purpose, we reduced the number of features from 30 to 20 by excluding the acceleration values. In the second condition, we eliminated the angular velocities and provided only the position and acceleration features (20 in total) for the prediction. We found out (See Tab. 2) that angular velocity and acceleration information do not provide independent features. The combination of $(\theta,\dot{\theta})$ has outperformed the other feature combinations. Additional benefit of this feature reduction is that the search algorithm works now considerably faster and is less affected by the curse of dimensionality.

Table 2: ASR performances for three feature sets ($\alpha = 2$ and $\beta = 0.5$)

| | $(\theta,\dot{\theta},\ddot{\theta})$ | $(\theta,\dot{\theta})$ | $(\theta,\ddot{\theta})$ |
|---|---|---|---|
| SNR = 1dB | 48.0 | **47.8** | 47.9 |
| SNR = 3dB | 37.2 | **36.8** | 37.1 |
| SNR = 6dB | 24.2 | **23.8** | 24.2 |

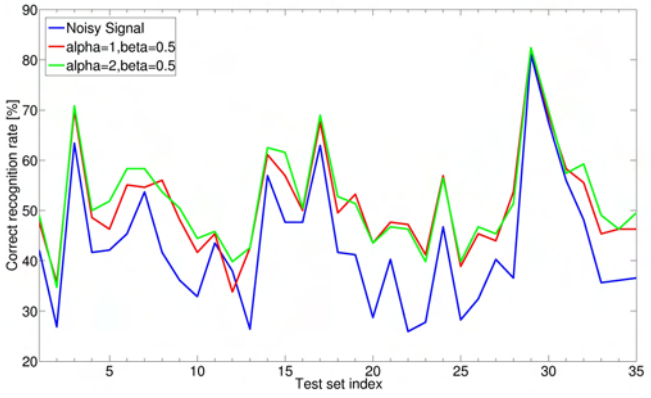


Figure 7: Distribution of the recognition performance over the noise set data

## 5 Summary and Outlook

In this paper, we have presented methods for removing ego-motion noise from sound signals. We showed that there is a trade-off between quality and intelligibility of the speech. Results are very promising in the sense that high suppression rates are achieved while keeping the speech as untouched as possible. We also demonstrated methods to maintain the same intelligibility while improving the quality of the speech by tuning the spectral subtraction parameters, $\alpha$ and $\beta$. We suggest to choose these parameters depending on one's purpose in using the enhancement algorithm: If the aim is sound localization, template subtraction can be used aggressively to remove the onsets originating from motor noise. For speech recognition, however, no harm to the speech signal can be tolerated, hence only milder suppression is recommended. We have also investigated methods for noise prediction based on joint status information. Results are preliminary, but they show that described concept works.

In its current form, our system has difficulties in achieving precise prediction of templates. Therefore, additional features that utilize cues about time series expansion of consecutive motion elements and incorporate information on motion primitives and motion-sequences would improve the reliability and performance of the predictions. Next steps involve an online implementation of the template subtraction scheme on ASIMO that performs motions using more joints. Besides, more sophisticated online compatible learning and indexing techniques will increase the speed of our approach and endow the system with a capability of online adaptation. An important advantage of parameterized approach would be that it can update the database on the fly making the prediction more adaptive and accurate in case any change in the characteristics of the motor noise (e.g. due to heating or alterations in the material) occurs at any time. Moreover, it can run online on the background while the robot is performing its duties and tasks. In order to improve the robustness, we plan to embed the current single-channel ego-noise reduction stage into a general multi-channel microphone array processing framework for speech recognition that utilizes geometric source separation and post filtering.

## 6 Acknowledgements

We thank Dr. Björn Schölling for his fruitful discussions.

## References

[1] T. Rodemann, M. Heckmann, B. Schölling, F. Joublin and C. Goerick "Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping", *Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, 2006.

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

[3] B. Rasnow and J. M. Bower, "Imaging with electricity: how weakly electric fish might perceive objects", *Proceedings of the annual conference on Computational neuroscience : trends in research*, 1997.

[4] P. D. Roberts, "Modeling Inhibitory Plasticity in the Electrosensory System of Mormyrid Electric Fish", *The Journal of Neurophysiology*, vol. 84, No.4, 2000.

[5] 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.

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

[7] 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.

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

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

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

[11] T. Rodemann, G. Ince, F. Joublin and C. Goerick "Using Binaural and Spectral Cues for Azimuth and Elevation Localization", *Proceedings of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, 2008.

[12] B.C.J. Moore, *An introduction to the psychology of hearing*, 5th ed. London: Academic Press, 2003.

[13] M. Heckmann, T. Rodemann, B. Schölling, F. Joublin and C. Goerick "Auditory Inspired Binaural Robust Sound Source Localization in Echoic and Noisy Environments", *Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, 2006.