A Robust Speech Recognition System against the Ego Noise of a Robot

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

2010

Preprint:

This is an accepted article published in Proceedings of Interspeech. The final authenticated version is available online at: https://doi.org/[DOI not available]

A Robust Speech Recognition System against the Ego Noise of a Robot

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

[1]Honda Research Institute Japan Co., Ltd., Japan
[2]Honda Research Institute Europe GmbH, Germany
[3]Dept. of Mechanical and Environmental Informatics, Tokyo Institute of Technology, Japan
gokhan.ince@jp.honda-ri.com

Abstract

This paper presents a speech recognition system for a mobile robot that attains a high recognition performance, even if the robot generates ego-motion noise. We investigate noise suppression and speech enhancement methods that are based on prediction of ego-motion and its noise. The estimation of ego-motion is used for superimposing white noise in a selective manner based on the ego-motion type. Moreover, instantaneous prediction of ego-motion noise is the core concept to establish the following techniques: ego-motion noise suppression by template subtraction and missing feature theory based masking of noisy speech features. We evaluate the proposed technique on a robot using speech recognition results. Adaptive superimposition of white noise achieves up to 20% improvement of word correct rates (WCR) and the spectrographic mask attains an additional improvement of up to 10% compared to the single channel recognition.

Index Terms: speech enhancement, noise reduction, ASR

1. Introduction

Robots with listening capabilities are being equipped with audio signal processing techniques against environmental noises [1], [2]. However, these methods are not effective against robot's own noise, in particular ego-motion noise, which arises when the robot performs a task, action or motion using its motors. Ego-motion noise is rather challenging due to its close proximity to the microphones and non-stationarity, therefore conventional noise reduction methods like spectral subtraction [3] do not work well in practice. A directional noise model such as assumed in case of interfering speakers [1] or a diffuse background noise model [2] does 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. Nishimura *et al.* [4] and Ito *et al.* [5] tackled this 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 limited number of gestures and motions with fixed trajectories. By exerting Missing Feature Theory (MFT), Yamamoto *et al.* [1] and Takahashi *et al.* [10] 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 this work, we target to eliminate the diminishing effects of ego-motion noise in the context of automatic speech recognition (ASR). In order to generate speech features we use an environmental noise robust feature extraction framework that consists of Sound Source Localization (SSL), Sound Source Separation (SSS), and Speech Enhancement (SE), which we have adopted from already existing studies. To enhance the acoustic features further, we propose to incorporate three methods into this framework that are based on instantaneous information extracted from ego-motion related processes: (1) selective white noise superimposition based on the motion type, (2) parameterized template subtraction, and (3) MFT-based masking on speech features. We demonstrate that the proposed system achieves a high noise cancellation performance and improves ASR accuracy.

2. Noise Robust Feature Extraction

In this section, we describe a standard multi-talker speech recognition system using a microphone array, which is robust to environmental noise and interfering speakers (see Fig. 1). 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 SSS stage, which is a linear separation algorithm called Geometric Source Separation (GSS) [1]. It is based on a hybrid algorithm that exerts Blind Source Separation (BSS) and beamforming. Current GSS implementation is an adaptive algorithm that can process the input data incrementally, and makes use of the locations of the sources explicitly. To estimate the separation matrix properly, GSS introduces cost functions that must be minimized in an iterative way [2].

After the separation process, a multi-channel post-filtering (PF) operation proposed by Cohen [6] is applied, which can cope with nonstationary interferences as well as stationary types of noise. This module treats the transient components in the spectrum as if they are caused by the leakage energies that may occasionally arise due to poor separation performance. For this purpose, noise variances of both stationary noise and source leakage are predicted. Whereas the former one is computed using the Minima Controlled Recursive Averaging (MCRA) [7], to estimate the latter the formulations proposed in [2] are used.

In order to achieve an optimum recognition performance, we create an acoustic model matched with a known noise. Therefore, a censequent additive white noise step applied after post filtering improves the speech recognition results by generating an artificial floor in the spectrum of speech signal. Details about the improvement of this module is subject to further discussion in Sec. 4.1. Finally, acoustic features are generated by calculating Mel-Scale Log Spectrum (MSLS) [8] that does not spread distortions to all coefficients of the cepstrum unlike Mel-Frequency Cepstral Coefficients (MFCC).

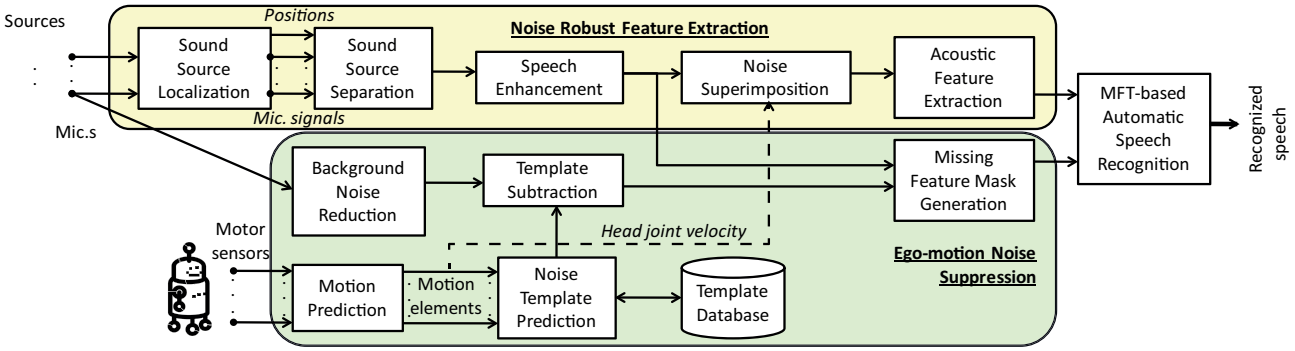


Figure 1: Proposed multi-talker speech recognition system

3. 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 does not change drastically for the same motions when the motion is performed again. However, a conventional *blockwise template prediction* [4] 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.

3.1. Motion Prediction and Template Database Generation

We make the following assumptions:

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

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. Resulting feature vector has the form of $[\theta_1(k), \dot{\theta}_1(k), \ddot{\theta}_1(k), \ldots, \theta_J(k), \dot{\theta}_J(k), \ddot{\theta}_J(k)]$, where k stands for the time-frame. At the same time, motor noise is recorded and background noise is removed from the recordings. The spectrum of the motor noise is given by $[D(1,k), D(2,k), \ldots, D(F,k)]$, where F represents number of frequency bins. Both feature vectors and spectra are continuously labeled with time tags so that corresponding templates are generated when their time tags match.

3.2. 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 (Fig. 2). We implemented a Nearest Neighbor search to find the correct template with most similar joint configuration among all templates in the database. The prediction process is applied for 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.

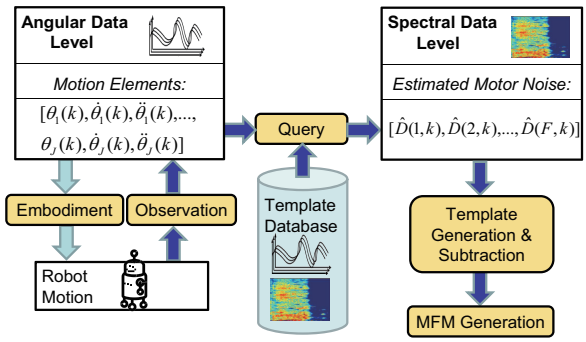


Figure 2: Parameterized template prediction method and its applications for ego-motion noise robust speech recognition

4. Ego-motion Noise Robust Speech Recognition

4.1. White Noise Superimposition

Since there are many different types of noise in a real-world environment, it is impractical to create matched models for each type of noise, especially for each ego-motion. Therefore, during the training phase we add white noise with a fixed amplitude value as a known noise source, where $\rho\,[dB]$ represents its magnitude relative to clean speech magnitude. The second advantage of using white noise is that it blurs the musical noise distortions caused by the spectral subtraction of the post-filter. Because the artifacts of high motor noise (i.e. head motion noise) is more harmful compared to the artifacts of arm-motion noise, leg motion noise, robot fan or background noise, we propose a switching mechanism for white noise level adjustment inside the noise superimposition module. The mechanism performes a decision between two white noise levels, which is triggered by the motion predictor. The motion predictor is able to discriminate, which joints are actively involved in the motion by calculating and checking their velocities. By doing that it determines, which motion is being performed at that moment. This method is applicable to all robotic systems and is scalable by the physical conditions regarding microphones, motors, their distances

and properties. Based on our preliminary experiments with our robot, we propose to implement the following rule-based routing in the switch:

$$\rho(k)=\begin{cases}-20dB, & \text{if any } |\dot{\theta}_{HeadJoint}(k)| > \epsilon \\ -40dB, & \text{otherwise}\end{cases}, \quad (1)$$

where $|\dot{\theta}_{HeadJoint}(k)|$ denotes absolute velocity of the pan or tilt motion of the head and ϵ is a certain speed value. ϵ, instead of zero, is used to prevent the activation of the switch during the *tail motion* of the head. It is used as a countermeasure to the situation where the motion has stopped, but the joint sensors still send very small position differences. Please note that the additive white noise will be cancelled out in the spectral mean normalization module of ASR.

4.2. Template Subtraction

Let us start by defining $S(\omega, k)$ and $D(\omega, k)$ as the short-time basis frequency 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:

$$X(\omega, k) = S(\omega, k) + D(\omega, k). \quad (2)$$

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

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

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 this error, we further suggest to use spectral subtraction approach that exploits *overestimation factor*, α, and *spectral floor*, β. α, allows a compromise between perceptual signal distortion and noise reduction level, whereas β is required to deal with *musical noise* [3]. Finally, we calculate the gain coefficients, $\hat{H}_{SS}(\omega, k)$, and multiply them with the signal $X(\omega, k)$ as in Eq. (5):

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

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

4.3. Missing Feature Mask Generation

GSS lacks the ability to catch motor noise originating from the same direction of the speaker and suppress it, because the noise is considered as part of the speech. Moreover, when the position of the noise source is not detected precisely, GSS cannot separate the sound in the spatial domain. As a consequence, motor noise can be spread to the separated sound sources in small portions. However, it is optimally designed for "simultaneous multiple 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, the characteristics of predicted and actual noise can differ to a certain extent.

As stated above, the strengths and weaknesses of both approaches are distinct. Thus, they can be integrated into an MFT-based mask in a complementary fashion. A speech feature is considered unreliable, if the difference between the energies of refined speech signals generated by multi-channel and single-channel noise reduction systems is above a threshold T. Computation of the masks is performed for each frame, k, and for each mel-frequency band, f. First, a continuous mask is calculated like following:

$$m(f, k)=\frac{|\hat{S}_m(f, k) - \hat{S}_s(f, k)|}{\hat{S}_m(f, k) + \hat{S}_s(f, k)}, \quad (6)$$

where $\hat{S}_m(f, k)$ and $\hat{S}_s(f, k)$ are the estimated energy 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) [10] is used in the MFT-ASR:

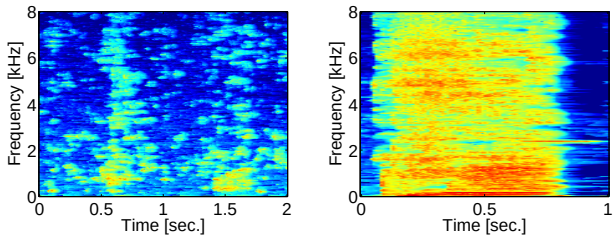
$$M(f, k)=\begin{cases}\dfrac{1}{1+\exp(-\sigma(m(f, k) - T))}, & \text{if } m(f, k) < T \\ 0, & \text{if } m(f, k) \geq T\end{cases}, \quad (7)$$

where σ is the tilt value of a sigmoid weighting function.

5. Evaluation

5.1. Experimental Settings

We used 8 microphones located on top of the head of the robot. We recorded (1) random whole-arm pointing behavior in the reaching space of the body 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 in average and show the spectral characteristics as in Fig. 3. Sensors give the angle of the joints every 5 ms and the length of the audio frames is 10 ms. We used constant values for α=1 and β=0.5 as template subtraction parameters, because in our previous study we observed that an increase in β improves ASR accuracy considerably compared for the case of β=0. For detailed evaluations regarding the α and β parameters, please refer to [9]. MFM parameters are selected as follows: T=0.75 and σ=10.



(a) Arm motion noise (b) Head motion noise

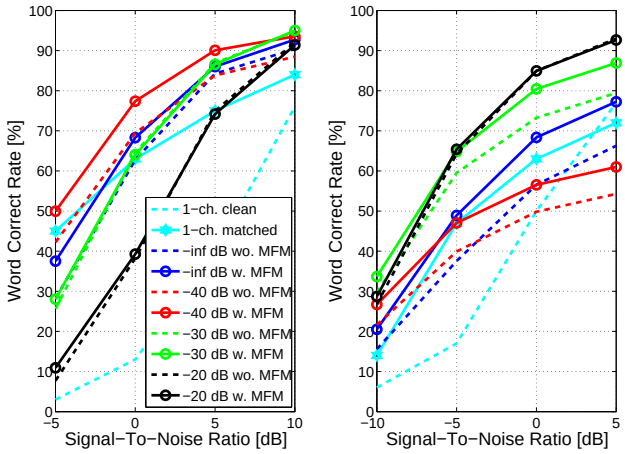
Figure 3: Spectrograms of arm and head ego-motion noise

Because the noise recordings are comparatively longer than the utterances used in the isolated word recognition, we selected

those segments, in which all joints of the corresponding limb contribute to the noise. To generate precise SNR conditions before mixing, we amplified clean speech based on its segmental SNR. The noise signal consisting of ego noise (incl. ego-motion noise) and environmental background noise is mixed with clean speech utterances used in a typical human-robot interaction dialog and recorded by us. The audio data is converted to 8 ch. data by convoluting with a transfer function of the microphone array. This Japanese word dataset includes 236 words for 4 female and 4 male speakers. Acoustic model is triphone HMM (phonetic tied mixture) with 32 mix/state and 2000 mixtures. It is 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. The position of the speaker is kept fixed at 0° throughout the experiments. The recording environment is a room with the dimensions of 4.0 m$\times$7.0 m$\times$3.0 m with a reverberation time (RT_{20}) of 0.2s.

5.2. Results

We superimpose white noise of various SNR's ($-20, -30, -40, -\text{inf}dB$) and evaluate WCRs with and without MFMs. Fig. 4 illustrates 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. We found out that the white noise level plays a crucial role in the final results. In case of arm motion, which is considered as a relatively weaker noise, white noise of the same intensity level used in the acoustic model training has shown the best performance. On the other hand, best ASR accuracy during a head motion with high noise intensity is achieved with an additive white noise of $-20dB$. This proves also that the musical noise effects after noise suppression can be tackled with the white noise addition.



(a) Under arm motion noise (b) Under head motion noise

Figure 4: Recognition performance for different types of ego-motion noise

We also observe that the MFT-ASR outperforms standard ASR without MFMs. Although there is little gain of using MFM for the $-20dB$ white noise (See Fig 4(a) and Fig 4(b)), it is very beneficial to use the masks for all other cases. While the masks eliminate unreliable speech features contaminated with motor noise, they also compensate the erroneous effects of voice activity detection due to additive motor noise that contains a large portion of energy. They prevent misdetection of motor noise as speech, even though the speech has not started yet, or is already over. The reader should also note that in a real-time, real-world scenario with a robot, where the SNR is $[0\ 5]dB$ for the arm motion and $[-5\ 0]dB$ for the head motion noise depending on the distance and loudness of the speaker, 15% and 18% average WCR improvement is attained compared to the WCRs obtained by ego-motion noise matched single-channel speech recognition.

6. Conclusion

In this paper, we presented methods for eliminating ego-motion noise from speech signals. The system we proposed utilizes (1) a selective white noise superimposition scheme based on the motion type, (2) a template subtraction technique to remove the ego-noise, and finally (3) a masking stage to improve speech recognition accuracy. The experimental results indicate that WCRs based on soft masking are improved up to 10% compared to conventional recognition system. Furthermore, we have shown that selective white noise superimposition method contributes to 20% improvement for the problematic head-motion noise.

7. References

[1] 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", in *Proc. IEEE/RSJ IROS*, 2006.

[2] J.-M. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai and H. G. Okuno, "Robust Recognition of Simultaneous Speech By a Mobile Robot", in *IEEE Trans. on Robotics*, Vol. 23, No. 4, pp. 742-752, 2007.

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

[4] Y. Nishimura, M. Nakano, K. Nakadai, H. Tsujino and M. Ishizuka, "Speech Recognition for a Robot under its Motor Noises by Selective Application of Missing Feature Theory and MLLR", in *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition*, 2006.

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

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

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

[8] Y. Nishimura, T. Shinozaki, K. Iwano, S. Furui, "Noise-robust speech recognition using multi-band spectral features", *Proc. of 148th Acoustical Society of America Meetings*, 1aSC7, 2004

[9] G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, and J. Imura, "Ego Noise Suppression of a Robot Using Template Subtraction", in *Proc. IEEE/RSJ IROS*, pp.199-204, 2009.

[10] T. Takahashi, S. Yamamoto, K. Nakadai, K. Komatani, T. Ogata, H. G. Okuno, "Soft Missing-Feature Mask Generation for Simultaneous Speech Recognition System in Robots" in *Proc. Interspeech*, 992-997, 2008.