Multi-talker Speech Recognition under Ego-motion Noise using Missing Feature Theory

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

2010

Preprint:

This is an accepted article published in IEEE-RSJ International Conference on Intelligent Robot and Systems (IROS 2010). The final authenticated version is available online at: https://doi.org/[DOI not available]

Multi-talker Speech Recognition under Ego-motion Noise using Missing Feature Theory

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

***Abstract*— This paper presents a system that gives a mobile robot the ability to recognize target speaker's speech, even if the robot performs an action and there are multiple speakers talking in the room. Associated problems to this system are twofold: (1) While the robot is moving, the joints inevitably generate ego-motion noise due to its motors. (2) Recognizing target speech against other interfering speech signals is a difficult task. Since typical solutions to (1) and (2), motor noise suppression and sound source separation, both introduce distortion to the processed signals, the performance of automatic speech recognition (ASR) deteriorates. Instead of removing the ego-motion noise with conventional noise suppression methods, in this work, we investigate methods to eliminate the unreliable parts of the audio features that are contaminated by the ego-motion noise. For this purpose, we model masks that filter unreliable speech features based on the ratio of speech and motor noise energies. We analyze the performance of the proposed technique under various test conditions by comparing it to the performance of existing Missing Feature Theory-based ASR implementations. Finally, we propose an integration framework for two different masks that are designed to eliminate ego noise and to filter the leakage energy of interfering sound sources. We demonstrate that the proposed methods achieve a high ASR accuracy.**

I. Introduction

Recently, robots are being equipped with audio signal processing techniques against environmental noises. However, the robot's own noise, so called ego noise, also poses a threat against accurate recognition of spoken words during an interaction with a human, even if there are no other interfering sound sources in the environment. One special type of ego noise, which is observed while the robot is executing some motions using its motors, is called *ego-motion noise*. As explained and quantified in [1], it causes drastic reduction in both speech recognition and sound localization accuracies. This type of noise, however, is so far either ignored or circumvented by a close-talk microphone, because the problem is rather challenging due to the complex characteristics of this particular noise. The complexity gets higher the more motors are in action, meaning that the noise is even more severe for a moving robot with a high number of degrees of freedom. Nevertheless, mobility is a necessary condition for improving perceptual capabilities of the robots, thus an autonomous robot requires very robust ego-motion noise suppression ability at any moment.

In this work, we incorporate Missing Feature Theory (MFT) to solve the ego-motion noise problem of a robot within the context of multiple speakers talking simultaneously. Primarily, a multi-channel audio processing framework that consists of Sound Source Localization (SSL), Sound Source Separation (SSS), and Speech Enhancement (SE) is adopted from already existing studies. These processes yield speech signals that are 1) uttered by each talker, 2) refined from background noises and reverberation, 3) but still contaminated by motor noise. To improve the performance of Automatic Speech Recognition (ASR), we propose to use MFT with a mask computation model that is based on the instantaneous ego-motion noise and speech estimations. We calculate the reliability of the speech features, which is represented by spectrotemporal masks.

In this respect, the main contribution of our work will be the design of an original missing feature mask (MFM) generation method based on the measure of a frequency bin's Signal-to-Noise Ratio (SNR), which is computed from the ratio of speech and estimated motor noise energies and called ego-motion noise MFM. Firstly, a hard mask, that uses either 0 or 1 to estimate the reliability of each acoustic feature, is suggested. We, later, enhance the proposed method further by using a soft mask represented as continuous values between 0 and 1. Furthermore, we focus on various integration methods for fusing the *ego-motion noise MFM* with the *multi-talker MFM* (originally introduced by Valin *et al.* [2]). The SNR-weighted integration mechanism of MFMs, another idea presented in this paper, prevents the robot from applying unnecessary suppression to speech features by exerting the ego-motion masks while the robot is moving slowly or comes to rest. Thus, it optimally balances the contribution of the two masks on the noise masking.

A. Comparison to Related Work

The ego-motion noise is more difficult to cope with compared to background noise or static fan-noise of the robot, because it is non-stationary and, to a certain extent, similar to the signals of interest. Therefore, conventional noise reduction methods like spectral subtraction [3] do not work well in practice. Several researchers tackled this problem by predicting and subtracting ego-motion noise using templates recorded in advance: Ito *et al.* [4] proposed a frame-by-frame based prediction technique using a neural

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

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, 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

network to cope with unstable walking noise of a robot. The trained network had to predict the noise spectrum from angular velocities of the joints of the robot. Ince *et al.* [1] proposed to use parameterized template subtraction which incorporates tunable parameters to cope with noise template representations that do not match to the instantaneous noise due to the deviations in the noise spectra. However, these methods suffer from the distorting effects of *musical noise* [5], a phenomenon that occurs when noise estimation fails. Thus, nonlinear single-channel based noise reduction techniques generally degrade the intelligibility and quality of the audio signal. Linear SSS techniques are also very popular in the field of robot audition, where noise suppression is mostly carried out using SSS techniques with microphone arrays [6],[8]. A directional noise model such as assumed in case of interfering speakers [7] or a diffuse background noise model[2] does not hold entirely for the ego-motion noise. Especially because the motors are located in the near field of the microphones, they produce sounds that have both diffuse and directional characteristics.

In the speech recognition literature, it was shown that unreliable speech features degrade recognition performance severely [9]. MFT, which can be basically described as filtering of the missing or damaged acoustic features, has already found useful applications. For example, speech corrupted by music and several types of background noise can be recognized with MFT more robustly (refer to [9] for a comprehensive study). For a simultaneous speech recognition task of several speakers, Yamamoto *et al.* [7] and Takahashi *et al.* [10] proposed a model for mask generation based on the disturbing effect of leakage noise over speech, because an imperfect source separation causes distorting elements, however their model is unable to deal with ego-motion noise. Nishimura *et al.* [11] estimated ego noises of distinct gestures and motions of the robot. Using motion commands, the pre-recorded correct noise template matching to the recent motion was selected from the template database and the acoustic features of the aligned template are used for MFT weight calculation. However, their mask model is based on a simple energy thresholding, therefore it is not feasible to use this method in a real-world scenario, where the SNR of speech can change depending on the loudness and distance of the speaker from the robot. Secondly, they utilized blockwise templates which cannot cope with dynamic changes of the motion trajectories in time. Our approach overcomes the former problem by introducing an SNR-based weighting of mask generation, whereas the latter drawback is tackled by a parameterized template prediction method as in [1].

In our previous work [12], we showed a multi-channel noise suppression scheme, where SSS is applied to separate speech of interest from the fan noise of a robot and the directional portion of ego-motion noise. Diffuse portion of the noise, however, is suppressed by a consequent post filtering operation. In a parallel work, we demonstrated the effectiveness of simultaneous usage of (1) single-channel template subtraction and (2) multi-channel noise suppression. To improve the ASR accuracy, we exerted MFT and designed a mask based on the similarity of the refined signals at the output of the two noise suppression methods ((1) and (2)) [13]. However, the mask was vulnerable to the distortions caused by the residuals of motor noise due to incorrect/overestimated noise estimations. Besides, the system was only able to deal with one speaker at a time. In this work, we aim to eliminate the artifacts of musical noise by disposing template subtraction from our system and to extend the speech recognition system so that it can recognize multiple speakers at the same time while the robot is moving.

II. System Overview

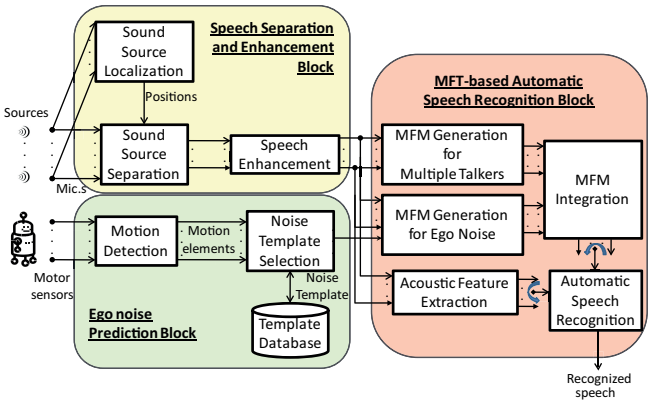


Fig. 1. Proposed multi-talker speech recognition system

The overall architecture of the proposed system is shown in Fig. 1. The first block of our processing chain, composed of elements for performing SSL, extracts the location of the most dominant sources in the environment. The estimated locations of the sources are used by a SSS. The next stage after SSS is a speech enhancement module that attenuates stationary and non-stationary noises. These three main modules constitute the **speech separation and enhancement** block (Sec. II-A), whereas the second block performs **ego noise prediction** (Sec. II-B). Former block is responsible for the extraction of *audio features* for speech recognition, while both blocks produce *spectrograms* to be processed further in the MFM generation stages. Finally, a third block, **MFT-based speech recognition**, uses both the features and the spectrograms created in the pre-processing stages in order to extract the most suitable features to achieve a more robust ASR. This part will be discussed in Sec. III in detail. Note that by using a switching arrow the number of inputs of ASR sub-block is reduced to one for the acoustic features and MFMs, which indicates that any talker's utterance can be recognized. The talker selection can be triggered by a selective attention system that is out of scope of this paper.

A. Speech Separation and Enhancement Block

We used linear separation algorithm called Geometric Source Separation (GSS) [7] for SSS. It is based on a hybrid algorithm that exerts Blind Source Separation (BSS) [14] 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 [15] is applied, which can cope with nonstationary interferences as well as stationary noise. Transient components in the spectrum are treated as if they are caused by the leakage energies that may 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) [5], to estimate the latter the formulations proposed in [2] are used.

B. Ego Noise Prediction Block

During the motion of the robot, actual position (θ) information regarding each motor is acquired regularly in the template generation (database creation) phase. Additionally, using the difference between consecutive sensor outputs, velocities ($\dot{\theta}$) are calculated. Considering that N joints are active, feature vectors with a size of $2N$ are generated. The resulting feature vector has the form of $F = [\theta_1, \dot{\theta}_1, \theta_2, \dot{\theta}_2 \ldots, \theta_N, \dot{\theta}_N]$. At the same time, motor noise is recorded by a single microphone and spectrum of the motor noise is computed by the sound pre-processing that runs simultaneously with motion element acquisition. MCRA [3] and removal is applied consequently. Both feature vectors and spectra are continuously labeled with time tags. The templates are created at those time points, when their time tags match with each other. Finally, a large noise template database that consists of short noise templates for the desired joint configurations is created (Please refer to [1] for details).

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 the feature (joint-status) vectors. Please note that in contrary to the approaches used in [4] and [1], the templates are not subtracted from the noisy signal.

III. MFT-based Automatic Speech Recognition Block

As stated in [9], MFT-ASR is a very promising Hidden Markov Model based speech recognition technique that basically applies a mask to decrease the contribution of unreliable parts of distorted speech. By keeping the reliable parameters that are essential for speech recognition, a substantial increase in recognition accuracy is achieved. In the following sections, we will discuss two reliability estimation techniques, one designed especially against speaker separation artifacts (Sec. III-A) and one for the purpose of eliminating ego-motion noise (Sec. III-B). They are followed by mask generation algorithms(Sec. III-C) and proposed method for the integration of the two masks (Sec. III-D). In general, the masking operation can be considered as a confidence-based weighting of the time-frequency representation of audio signals, therefore the masks and acoustic features must be provided to MFT-ASR simultaneously as depicted in Fig. 1. Detailed explanation about acoustic feature extraction can be found in [13].

A. Reliability Estimation for Multiple Talkers

As mentioned in Sec. II-A the noise estimate in post-filtering is decomposed into stationary (background noise) and transient (leakage energies of interfering sources) components for each source of interest. In order to predict the amount of noise present at a certain time in a certain frequency, Yamamoto *et al.* proposed a computation method for measuring the reliability as given like following [8]:

$$m_m(f,k) = \frac{\hat{S}_{out}(f,k) + \hat{B}(f,k)}{\hat{S}_{in}(f,k)}, \tag{1}$$

where $\hat{S}_{in}$ and $\hat{S}_{out}$ are respectively the post-filter input and output energy estimates for frame k and Mel-frequency band f. $\hat{B}(f,k)$ denotes the background noise estimate and $m_m(f,k)$ gives a measure for the reliability based on multi-talker (m) effects.

B. Reliability Estimation for Ego Noise

There are several problems associated with GSS and PF based ego-motion noise reduction. First of all, GSS lacks the ability to catch motor noise originating from the same direction of the speaker and separate it, because the noise is considered as part of the speech in that case. 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. Apart from the directional portion, motor noise also has a diffuse portion that is caused by the highly reverberant propagation patterns of motor noise waves inside the body covers of the robot. Although diffuse noises are tackled by the post filter, the nonstationarity of the motor noise makes PF ineffective against ego-motion noise. Based on those drawbacks, we claim that it is impossible to remove the motor noise completely just by applying source separation or speech enhancement. However this multi-channel noise reduction chain (GSS+PF) is optimally designed for "simultaneous multiple speakers" scenarios with background noise and demonstrates a very good performance when no motor noise is present. Therefore, we plan to support our ASR with a probabilistic framework designed for reliability estimation, MFT-ASR.

We base our reliability measurements for ego noise on the retrieved templates during the motion. In contrary to speech separation and enhancement block (cf. Fig. 1), ego noise prediction block does not make any assumption about the directivity or diffuseness of the sound source and can correlate the incoming noisy signal to the retrieved template in a frame-by-frame basis. In this work, we make the simplifying assumption that the additive motor noise is distributed uniformly among the existing sound sources. Therefore, we divide the noise energy by the number of sources. The ratio of the template (estimated noise) energy and the noisy signal energy of interest yields a reliability measurement about whether the corresponding frequency

bin is strongly contaminated by ego-motion noise or not. A continuous measurement for the reliability based on ego noise (e) effects, $m_e(f,k)$, is calculated like following:

$$m_e(f,k)=1-min\left(1,\frac{\hat{S}_e(f,k)}{l\cdot\hat{S}_{out}(f,k)}\right), \tag{2}$$

where $\hat{S}_e(f,k)$ is the noise template and l represents the number of speakers. To make $m_e(f,k)$ and $m_m(f,k)$ value ranges consistent, the possible values that it can take are limited between 0 and 1. This formula suggests that if high motor noise ($\hat{S}_e(f,k)$) is estimated, the reliability is zero, whereas low motor noise sets $m_e(f,k)$ close to 1.

C. Missing Feature Mask Generation

The reliability of features is computed for each frame and for each mel-frequency band. If continuous values between 0 and 1 constitute the mask, it is called a *soft mask*. On the other hand, a *hard mask* contains only discrete values, either 0 or 1. We adopted two mask generation mechanisms that 1) meet our needs and 2) are generic in the sense that they generate masks for both ego noise (M_e) and for multi-talker (M_m) by substituting x with e or m in Eq. (3) and (4).

1) Hard (binary) mask:

$$M_x(f,k)=\begin{cases}1, & \text{if } m_x(f,k)\geq T_x\\ 0, & \text{if } m_x(f,k)<T_x\end{cases}. \tag{3}$$

2) Soft mask [10]:

$$M_x(f,k)=\begin{cases}\dfrac{1}{1+\exp(-\sigma_x(m_x(f,k)-T_x))}, & \text{if } m_x(f,k)\geq T_x\\ 0, & \text{if } m_x(f,k)<T_x\end{cases}, \tag{4}$$

where σ_x is the tilt value of a sigmoid weighting function and T_x is a predefined threshold. A speech feature is considered unreliable, if the reliability measure is below T_x. In this paper, we used both masks to assess their performance.

Additionally, we introduced a new heuristics-based concept called *minimum energy criterion (mec)* as in Eq. (5). It is used to override the decision of the above formulae, if the energy of the noisy signal is smaller than a given threshold, T_{mec}. It is used to avoid wrong estimations caused by computations performed with very low-energy signals, e.g. during pauses or silent moments.

$$M_x(f,k)=0, \qquad \text{if } \hat{S}_{out}(f,k)<T_{mec}. \tag{5}$$

D. MFM Integration

Even if some portion of the motor noise is removed by the pre-processing stages (SSS and SE), the real ego-motion noise suppression is performed in the masking stage. As we have stated, the two masks in Sec. III-A and Sec. III-B serve two different purposes. Nevertheless, they can be used in a complementary fashion within the context of multi-talker speech recognition under ego-motion noise.

$$M_{tot}(f,k)=w_mM_m(f,k)\dotplus w_eM_e(f,k), \tag{6}$$

where $M_{tot}(f,k)$ is the total mask and w_x is the weight of the corresponding mask. This is a generic framework that allows the system designer 1) to weight each single mask individually and 2) to perform either an addition (*OR*) or a multiplication (*AND*) operation with soft (*hard*) masks for integration.

IV. Evaluation

In this section, we present comparative results for hard & soft masking for ASR and the performance of an integrated mask, after describing the experimental settings.

A. Experimental Settings

To evaluate the performance of the proposed techniques, we use a humanoid robot developed by Honda. The robot is equipped with an 8-ch microphone array on top of its head. Of the robots many degrees of freedom, we use only a vertical head motion (tilt), and 4 motors for the motion of each arm with altogether 9 degrees of freedom. We recorded random motions performed by the given set of limbs by storing a training database of 30 minutes and a test database 10 minutes long. 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. We recorded clean speech utterances and converted them to 8 ch. data by convoluting with a transfer function of the microphone array. This Japanese word dataset includes 236 words for 1 female and 2 male speakers that are used in a typical human-robot interaction dialog. After normalizing the energies of the utterances to yield an SNR of $-6dB$ (noise:two other interfering speakers), the noise signal consisting of ego noise (incl. ego-motion noise and fan noise) and environmental background noise is mixed with clean speech utterances. Acoustic models are 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-open test. We used 13 static MSLS, 13 delta MSLS and 1 delta power as acoustic features. Speech recognition results are given as average Word Correct Rates (WCR). The position of the speakers are kept fixed at three position configurations throughout the experiments: $[-80^\circ, 0^\circ, 80^\circ]$, $[-20^\circ, 0^\circ, 20^\circ]$. To avoid the mis-recognition due to localization errors and evaluate the performance of the proposed method, we set the locations manually by by-passing SSL module. The recording environment is a room with the dimensions of $4.0\,\mathrm{m}\times 7.0\,\mathrm{m}\times 3.0\,\mathrm{m}$ with a reverberation time (RT_{20}) of 0.2s. We evaluated MFMs with the following heuristically selected parameters T_e=0, T_m=0.2, σ_e=2, σ_m=0.003 (energy interval:[0 1]).

B. Spectrograms and Masks

Fig. 2 gives a general overview about the effect of each processing stage until the masks are generated. In Fig. 2f), we see a tightly overlapped speech (Fig. 2a)+b)+c), SNR_m=$-6dB$) and motor noise (Fig. 2d), SNR_e=$-5dB$) mixture. GSS+PF as in Fig. 2g)–l) reduces only a minor part of the motor noise while sustaining the speech.

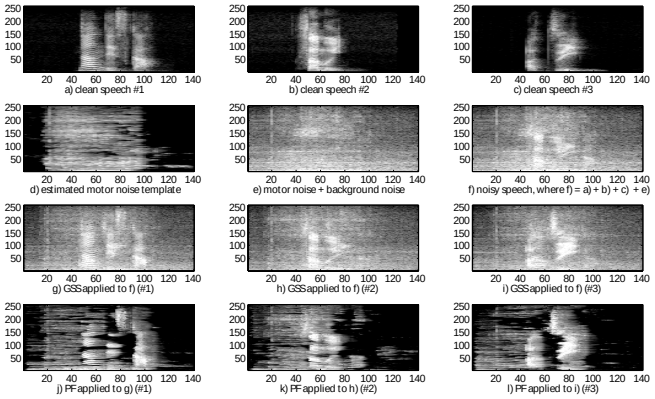


Fig. 2. Spectrograms for preprocessing. y-axis represents frequency bins between 0 and 8kHz. x-axis represents the index of frames.

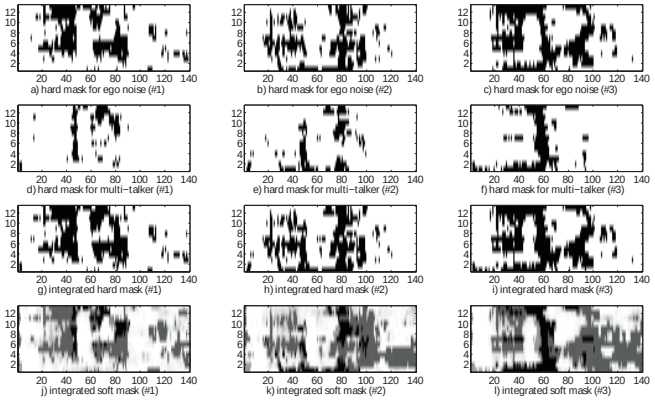


Fig. 3. Masks generated with a mild threshold value (T_e=0.1). y-axis represents 13 static mel-features. x-axis represents the index of frames.

The masks presented in Fig. 3 are all applied to the refined signals in Fig. 2j)–l). The ego noise masks in Fig. 3a)–c), show that ego noise located between the 20-45th frames can be detected and the features can be suppressed. Besides, as in Fig. 3c), the residuals of ego noise between 80-100th frames can be masked, as well. The multi-talker masks, on the other hand, are active especially where the source separation leaves residual energies that belong to interfering speakers. In the integrated masks (in Fig. 3g–i)), we see the contribution of ego noise masks is more dominant compared to multi-talker masks, as they seem to extend the contours of the ego noise masks. The soft masks in Fig. 3)j)–l), in addition, provide more detailed information about the reliability degree of each feature so that the noise-free features are weighted more than the noise-containing parts in the MFT-ASR.

C. ASR Accuracy using MFMs

Fig. 4 and 5 illustrate the ASR accuracies for a speaker setting with wide, resp. narrow, separation intervals and for all methods under consideration. The results are evaluated using an acoustic model trained with motor noise data. Single channel recognition is between 0-2 percent for all SNRs. Because the task is a multi-talker recognition, GSS+PF is considered as the baseline. There is only little improvement gained from minimum energy criterion (mec), as the results of the comparison for the hard masks presents in both figures. In overall it contributes only up to 1-3% to WCR. We see three general trends:

1) Soft masks outperform hard masks for almost every condition. This improvement is attained due to the improved probabilistic representation of the reliability of each feature.
2) We observe that the ego noise masks perform well for low SNRs, however WCRs deteriorate for high SNRs. The reason resides in the fact that faulty predictions of ego-motion noise degrade the quality of the mask, thus ASR accuracy, of clean speech more compared to noisy speech. On the other hands, in high SNRs (inferring no robotic motion or very loud speech) multi-talker masks improve the outcomes significantly, but their contribution suffers in lower SNRs instead.
3) As the separation interval gets narrower, the WCRs tend to reduce drastically. The presented WCRs in Fig. 4 and 5 are consistent with a multi-talker recognition study of [10]. We further observe a slight increase in the accuracy provided by M_m compared to M_e in -5dB for narrow separation angles (Fig. 5), because the artifacts caused by SSS for very close talkers become very dominant.

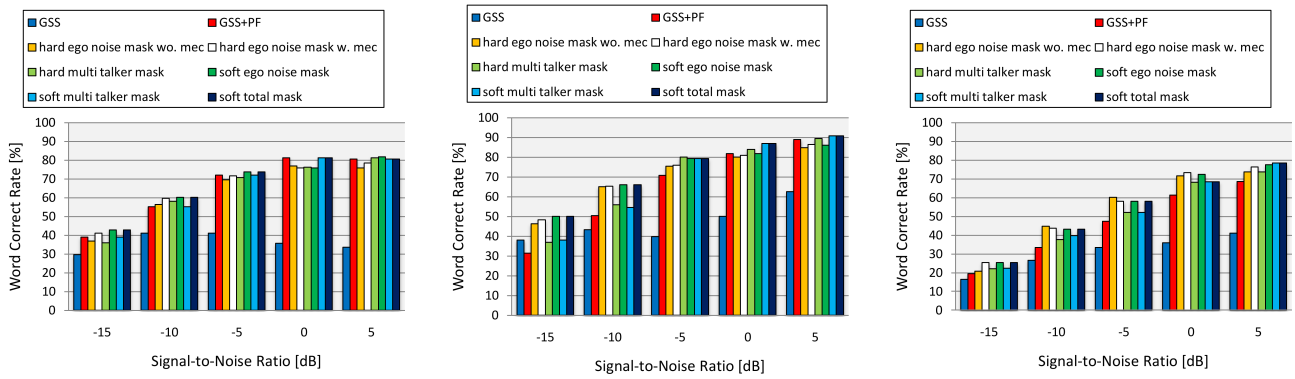
We evaluated both integration techniques for hard masks: AND and OR-based integration of M_e and M_m. However, the WCRs were all worse compared to the individual recognition performances of single masks. We applied a simple binary weighting based on the assessment of trend 1) and 2) above. We set the weights according the following conditional statements and apply an addition operation as in Eq. 6:

$$\{w_e, w_m\} = \begin{cases} \{1,0\} & \text{if } SNR < 0 \\ \{0,1\} & \text{if } SNR \geq 0 \end{cases}$$

Corresponding results are displayed in the final bin of each SNR segment, which indicate that fused masks work best for this problem. Our method demonstrated significant WCR improvement for soft masking (up to 20% compared to GSS+PF).

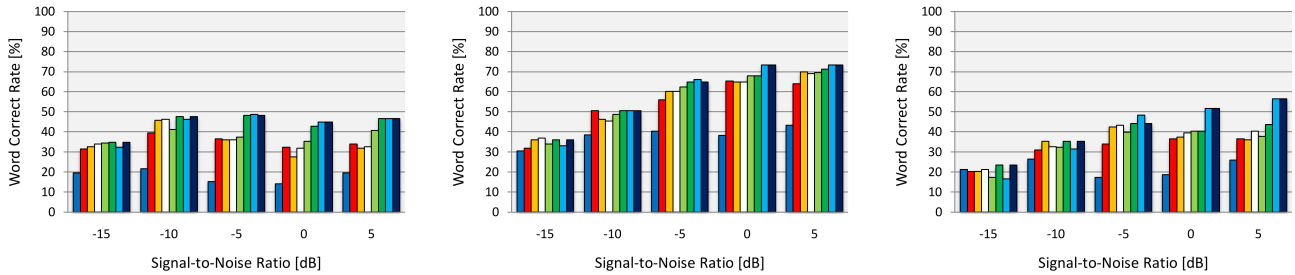
V. Summary and Outlook

In this paper we presented a method for eliminating ego-motion noise from simultaneous speech signals of multiple talkers. The system we proposed utilized (1) a multi-channel noise reduction module, (2) an ego-motion noise prediction module, and finally (3) a masking module to improve speech recognition accuracy. We used an MFM model, which is based on the the ratio of speech and motor noise energies. Furthermore, we proposed an integration framework for two masks that are designed to eliminate ego noise and to filter the leakage energy of interfering sound sources. We showed



(a) Speaker #1 (b) Speaker #2 (c) Speaker #3

Fig. 4. Recognition performance for the speaker located at $[-80°, 0°, 80°]$



(a) Speaker #1 (b) Speaker #2 (c) Speaker #3

Fig. 5. Recognition performance for the speakers located at $[-20°, 0°, 20°]$

that our integration method achieves a high ASR accuracy for any arbitrary separation angle between the talkers and any SNR value.

In future work, we plan to weight the masks not binary, but in a continuous way for a large SNR value interval. Besides, equal distribution of total ego-motion noise to all talkers is not a good representation, e.g. even if the ego-motion noise comes only from right arm, current ego noise masks do not distinguish the direction of the noise and the mask of each talker is affected by the same amount of noise. So, we plan to calculate the noise energy contributions based on the direction of the motors in relation with the directions of speech. Finally, the negative correlation between nearest neighbor search and database size will be tackled by interpolating missing templates appropriately.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

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

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

[6] K. Nakadai, H. Nakajima, Y. Hasegawa and H. Tsujino, "Sound source separation of moving speakers for robot audition", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.3685-3688, 2009.

[7] 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.

[8] S. Yamamoto, J. M. Valin, K. Nakadai, J. Rouat, F. Michaud, T. Ogata, and H. G. Okuno, "Enhanced Robot Speech Recognition Based on Microphone Array Source Separation and Missing Feature Theory", *Proc. of the IEEE/RSJ International Conference on Robotics and Automation (ICRA)*, 2005.

[9] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition", *IEEE Signal Processing Magazine*, vol. 22, pp. 101-116, 2005.

[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", *International Conference on Spoken Language Processing (Interspeech)*, pp.992-997, 2008.

[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] G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, and J. Imura "A Hybrid Framework for Ego Noise Cancellation of a Robot", *Proc. of the IEEE/RSJ International Conference on Robotics and Automation (ICRA)*, pp.3623-3628, 2010.

[13] G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, and J. Imura "Robust Ego Noise Suppression of a Robot", *International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA-AIE)*, LNAI 6096, pp.62-71, 2010.

[14] 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.

[15] I. Cohen and B. Berdugo, "Microphone array post-filtering for non-stationary noise suppression", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.901-904, 2002.