2007 IEEE International Conference on
Robotics and Automation
Roma, Italy, 10-14 April 2007

Adaptive Noise Reduction and Voice Activity Detection for improved Verbal Human-Robot Interaction using Binaural Data

Robert Brueckmann, Andrea Scheidig and Horst-Michael Gross

Abstract— **Speech has become an important part in Human Robot Interaction (HRI), e.g. for person detection systems by using localized sound sources or for applications in Automatic Speech Recognition (ASR) systems. By using speech in HRI in real world environments, we have to deal with mostly high and varying background noise, reverberation and also with different sound sources superimposing speech and other noises. Therefore, for real world scenarios a suitable signal preprocessing is essential.**

In this paper, we present a part of the artificial auditory system implemented on the mobile interaction robot HOROS **using only two low cost microphones. We combined neural Voice Activity Detection (VAD) and adaptive noise reduction which are essential aspects for HRI using mobile robot systems in changing and populated real-world environments.**

In the result, our system is able to robustly react on speech signals from its human interaction partner while ignoring other sound sources. Experiments show a significantly improved ASR performance in demanding environments making the system suitable for the use in real-world scenarios.

I. INTRODUCTION

When interacting with a mobile robot, speech plays an important role as natural interface between human and a robot. It is therefore desirable for a mobile robot to robustly recognize speech in real world scenarios. Unfortunately, real world environments often lead to reverberation, and besides the sound source of interest, there may be other interfering sound sources present, like fans, noises produced by the robot itself, etc.. In the consequence, the performance of most speech processing methods, e. g. speech detection, sound source localization, Automatic Speech Recognition (ASR), etc., will be significantly degraded. Hence, the preprocessing of the sound signals is necessary providing an improved input signal. Aspects that should be considered for the preprocessing of sound signals should integrate the detection of different sound sources, the classification of them as speech/non-speech, and signal enhancement by adaptive noise reduction.

Sound source localization and separation on mobile robots requires microphone arrays with at least two microphones. In our work, we focus on an implementation using only two low cost acoustic sensors. These provide the minimum hardware equipment to apply sound source separation and localization, although the accuracy of both aspects could be increased using more microphones. The robot BIRON [1] also uses two microphones for sound source localization and separation, but there is no Voice Activity Detection (VAD) used. Speech is assumed if other sensory cues detect further hypotheses of a human present at the direction of a detected sound source (eg. using face detection). Another robot using two microphones is ARMAR II [2] which implements VAD based on energy and zero-crossing-rate.

While it is basically possible to gather localization information from only two microphones, there are several implementations using more microphones. Both the robots SIG2 [3] and Spartacus [4] use eight microphones for source separation and speech recognition. In the result, Spartacus is able to track up to four sound sources simultaneously. Furthermore, the simultaneous voice of three speakers can be separated with this setup.

Speech/non-speech classification of sound sources is also known as Speech Segmentation or Voice Activity Detection (VAD). Using energy-based algorithms, one can detect speech and silence segments at high signal-to-noise ratios (SNR) [7]. Speech is assumed if the signal level exceeds a threshold value, even if this is caused by a non-speech sound. Methods based on spectral entropy have been proposed to detect signal segments in noisy conditions [8], [10], [11]. These permit the detection even when the SNR is low. But both energy and spectral entropy based algorithms do not guarantee that the detected signal really contains speech. Other sounds (music, hand claps, closing doors, etc.) might be classified as speech as well. Neural or statistical classification of speech and non-speech is able to distinguish these kinds of sounds from human speech. VAD based on neural networks using Multi-Layer-Perceptrons (MLP) and cepstral matrices as input is implemented in [12]. This approach needs to be trained with noisy input signals of several background noises, and there is the need to train multiple MLPs, one for each type of background noise. Our approach uses adaptive noise reduction to pre-process the input signal of the VAD, therefore improved input signals can be used for training and there is no need to train multiple MLPs. Recurrent neural networks considering temporal aspects, requiring a more complex training than MLPs, have also been proposed for VAD [13], showing only slightly better classification performance.

It is crucial for *adaptive noise reduction* algorithms to estimate the noise spectrum in order to apply spectral subtraction and to gather the original speech signal. Cohen [14] proposed the Minima Controlled Recursive Averaging (MCRA) technique based on minimum statistics. This approach adapts its noise spectrum estimation and is therefore able to track the

This work is partially supported by TMWFK-Grant #B509-03007 to H.-M. Gross and HWP-Grant to A. Scheidig

R.Brueckmann, A.Scheidig, and H.-M.Gross are with the Neuroinformatics and Cognitive Robotics Lab, Technical University of Ilmenau, Germany andrea.scheidig@tu-ilmenau.de

noise statistics even in non-stationary noisy environments. Rangachari [17] derived a similar algorithm yielding faster adaptation when the noise spectrum is rapidly changing. As we will discuss in Sec. III-C we extended this approach to use the output of a neural VAD to detect non-speech regions which further improves the noise suppression.

This paper is organized as follows. Sec. II describes the mobile robot HOROS on which the auditory system is implemented, and potential applications to the new auditory system are presented. Sec. III gives an overview of the auditory system itself and describes the integrated methods of source separation, voice activity detection, and adaptive noise reduction. Finally, in Sec. IV, we present experiments demonstrating the performance of the VAD and the adaptive noise reduction.

II. ROBOTIC PLATTFORM AND APPLICATION

For our experiments we use the mobile interaction robot HOROS (HOme RObot System)[1]. HOROS' hardware platform is an extended Pioneer robot from ActiveMedia. It integrates an on-board PC (Pentium M, 1.6 GHz, 512MB) and is equipped with a laser-range-finder (SICK) and sonar sensors. For the purpose of Human Robot Interaction (HRI), this platform was mounted with different interaction-oriented modalities such as front and omnidirectional cameras, a touchscreen, and a speaker.

Two low-cost microphones of type YOGA EMR-106 are mounted at both sides of the robot's head. The distance between them is 27 cm. Since we are only using two microphones, we are able to use the on-board sound card of the Pioneer's PC for audio recording, avoiding the need for additional multi-channel audio capture hardware.

Our proposed auditory system used on HOROS is designed for the following applications:

- ASR: By pre-processing the input sound signal of an ASR system, we expect lower word error rates.
- Localization of speakers: By combining sound source localization and VAD we will differentiate between localized speakers from other sounds, which is especially important for HRI.
- Recording of voice messages of users: The voice recorder feature is used as a sort of answering machine. The user can provide voice messages which will be stored by the robot for later use. By using VAD, the beginning and the end of the user's speech can be detected to automatically start and stop the recording. Thus, there is no user interaction required, e.g. by pressing start/stop buttons on the robot's touch screen.

III. INTEGRATED METHODS

The auditory system consists of several technical and methodical aspects, whereby the steps of audio signal processing are depicted in Fig. 1. The input for the auditory system is provided by the raw stereo signal of the two microphones, sampled at 44.1 kHz. Additionally, the sound localization and the people tracker [18] provide information on the direction of possible interaction partners currently speaking. This information is used to initialize delay-and-sum beamformers in the respective directions [6]. The beam patterns of these beamformers are used by the Geometric Source Separation algorithm (see Sec. III-A).

The adaptive noise reduction uses a minimum statistics approach [19] to estimate the noise spectrum and to improve signal quality by applying a Wiener-type gain filter (see Sec. III-C). The enhanced signal is used to detect speech using a neural voice activity detector (see Sec. III-B). The results of the VAD are used to further enhance the noise spectrum estimate, especially if no speech is present.

The respective parts of the auditory system will be described in the following sections, mainly focussing on the aspects of the VAD and adaptive noise reduction.

A. Source Separation

For many processing tasks, it is often desired to only process one of the captured sound sources, e.g. the desired speaker's voice as input to the ASR system. Sound source separation techniques can be used to gather the sound source of interest out of an audio signal mixture recorded by spatially separated microphones.

We use the „Geometric Source Separation" (GSS) technique described in [5] and [6] for separation of the speaker's voice and one interfering sound source from a different direction. The sound source localization (see Fig. 1) is used to detect new sound sources and to initialize the beam patterns of the GSS algorithm accordingly with delay-and-sum beamformers [6]. Additionally, our people tracker provides direction information on where speakers are to be expected, e. g. by detecting legs and skin color using other sensory cues. This information is used to select the desired GSS output channel containing the speech of the user for further processing. Although the attenuation using the two microphones is only 1-2 dB for typical noisy real-world recordings, the preprocessing provides a better signal than using only one microphone.

B. Voice Activity Detection

Distinguishing speech from other sound sources gives the robot a powerful new opportunity to detect potential interaction partners. We use a neural network approach for detecting speech in the surrounding of the robot. An MLP network is trained to classify short periods of the audio signal into speech or non-speech segments.

Mel Frequency Cepstral Coefficients (MFCC) [9] are used for dimensionality reduction of the network's input data and because of their ability to represent audio signals according to human perception. We use 12 MFCC as features for speech/non-speech classification, excluding the first coefficient mostly representing the overall signal energy. To include information from previous frames, the differences between the last and the current coefficients are calculated along with the change of these delta values. Additionally, we

[1] http://wcms1.rz.tu-ilmenau.de/fakia/HOROS-Homepage.horos_project.0.html?&L=1

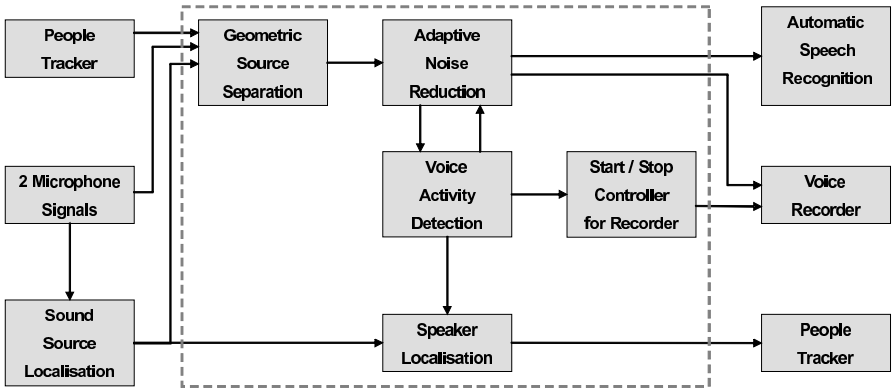

Fig. 1. Overview of the proposed Auditory System. The methods discussed here are displayed in the dashed box in the middle.

added spectral entropy as a classification feature, which is calculated according to [10]

$$H = -\sum_{k=1}^{N} p_k \log p_k \tag{1}$$

where p_k is the probability density function (PDF) of the current frame's input spectrum $S(k)$, which can be calculated using:

$$p_k = \frac{S(k)}{\sum_{i=1}^{N} S(i)}. \tag{2}$$

1) Network Architecture: The MLP gets its input from the MFCC with their delta values and the entropy of the signal. Thus, the input of the net consists of 37 values. The hidden layer of the network provides 15 hidden neurons with $\tanh$ activation function. We also tested networks with two hidden layers, which however did not show improved classification performance. The output of the network consists of one neuron with $\tanh$ activation function. Its value is trained to be in the range of -1...+1 where -1 means non-speech and +1 means speech frame.

2) Training data: The training data recordings for the neural net were divided into frames of 1,024 samples (23 ms at 44.1 kHz samplerate) with 50 % overlap. The frames need to be short enough to contain a stationary audio signal. Voice recordings of 9 speakers (5 female, 4 male) were used, which were recorded in a silent environment without significant noise and reverberation. Therefore, we could label the frames as speech or non-speech just by using an energy threshold that we empirically set to -24 dB below the maximum amplitude of the recorded signal. Afterwards, the resulting speech frames were included in the data set for training and testing. We used a total set of 13,521 frames consisting of 6,316 speech frames and 7,205 non-speech frames. The non-speech samples were gathered from three different kinds of sounds: noises (eg. PC fans), music, and environmental sounds such as shutting doors, clicks, coughs, etc.. The target vector used to calculate the training and test error contained binary teacher values (-1 for non-speech frames and +1 for speech frames). The entire data set was divided into training data (50 %), validation data (25 %), and test data (25 %).

3) Training: The network was trained using classic back-propagation [15]. An increase of the validation error indicates overlearning of the network so the training was stopped in that case. The test data was used to measure the classification performance on data which was not included during training. The output of the network for each test input pattern was assigned to contain speech/non-speech using a threshold parameter set to 0.

4) Smoothing: Although the classification result of the neural network can be used to segment a recording into speech and non-speech by using the threshold value of 0, the output of that VAD is still very noisy. This is because of outliers in the classification output. Their amount can be reduced by postprocessing the output of the neural network (see Fig. 2, top row). In a first step, the fixed threshold of 0 is replaced by a state model using hysteresis-like thresholds to switch between the states "speech" and "non-speech". This state model forbids transitions between the classes speech and non-speech if the network output is between -0.5 and +0.5. Therefore, weak classification results close to 0 cause the state model to remain in its current state. As can be seen in the middle row of Fig. 2, this state model can eliminate many outliers of the final classification result. Further improvement can be obtained if the network outputs are averaged prior to applying the state model. The lower row of Fig. 2 shows the network output after using a mean of 5 adjacent frames. To avoid the loss of weak speech components at the beginning and the end of a speech segment, we propose to use 5 future frames for smoothing if the current state is non-speech and 5 frames of the past otherwise. To have access to the needed future frames, a delay of the classification result by 4 frames ($\sim$46 ms) has been introduced.

C. Adaptive Noise Reduction

The recordings made with the microphones of the mobile robot are disturbed significantly by additive noise. Parts of that noise are produced by the robot itself, mainly by the fans of the PCs and noise produced by the sonar controllers. Additionally, non-stationary background noises might appear dependent on the environment of the robot. The noise markedly degrades the performance of typical speech

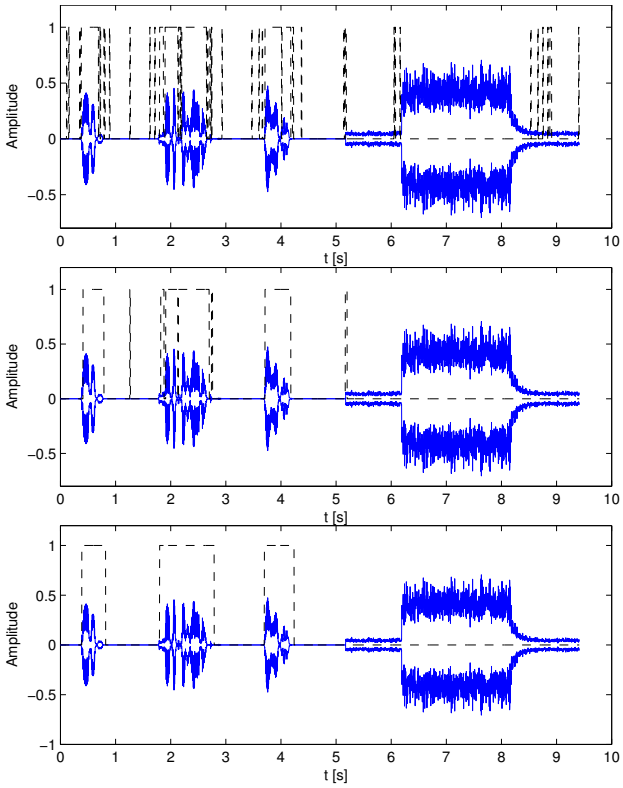

Fig. 2. Classification of a sound input consisting of three speech utterances (first 4 sec.) and one non-speech noise signal (between 6^{th} and 8^{th} sec.) using different post-processing. **Top:** Speech segmentation using a fixed threshold of 0.5. **Middle:** Speech segmentation using a state model ignoring weak classification results close to zero. **Bottom:** Speech segmentation by averaging over 5 frames of the neural network output and using the state model.

processing methods such as VAD and ASR. To improve the performance, we use a modified version of the adaptive noise reduction proposed by Rangachari et al. [17]. It enables the robot to operate in different noisy environments without manually collecting noise segments. The method is based on minimum statistics of the noisy input signal introduced by Martin [19].

A microphone signal containing additive noise can be expressed as

$$y(t) = s(t) + n(t) \tag{3}$$

where $s(t)$ is the clean input signal and $n(t)$ denotes the additive noise component. To apply noise reduction to the recorded noisy signal $y(t)$, it is necessary to obtain an estimate of the noise spectrum. Subsequently, spectral subtraction can be used to retrieve an estimate of the clean speech signal. Therefore, the noise reduction is applied in the frequency domain using a Wiener-type gain function $G(k)$ resulting in an estimate of the clean input signal $\tilde{S}(k) = G(k)Y(k)$. The gain function is described by [17]

$$G(k) = \frac{\phi_s(k)}{\phi_s(k) + \mu\phi_n(k)} \tag{4}$$

where $\phi_s(k)$ is an estimate of the clean signal PDF and $\phi_n(k)$ is the PDF of the current noise spectrum estimate. The oversubtraction factor $\mu \geq 1$ can be used for stronger attenuation if the segmental signal-to-noise ratio is very low, assuming there is currently only noise present [20]. The gain function is used to attenuate the input signal by the amount of the estimated noise component. Therefore, its values are close to 0 if there seems to be only noise currently present. If speech without noise is assumed, the values of $G(k)$ are close to 1, leaving the input signal nearly unprocessed.

The noise reduction method by Rangachari et al. [17] uses a signal-detection to determine which frequency bins are likely to contain speech components. The presence of speech at a specific frequency is assumed if there appears a sudden increase of the energy level in the respective frequency bin. In the consequence, the adaptation of the noise power spectrum estimate is inhibited not to contain the detected speech components.

Since the given approach does not differentiate between speech and other non-stationary sound sources, an extension with a VAD can reduce the error of the adapted noise power spectrum estimate. We propose to use the neural VAD presented in Sec. III-B for this purpose. The method by Rangachari provides a frequency dependent "speech presence probability" $I(\lambda, k)$ at time frame λ. We use the output $y(\lambda) \in [-1 \ldots +1]$ of the VAD to scale down the speech presence probability if a non-speech sound has been detected at the respective time frame.

$$I^*(\lambda, k) = \begin{cases} \frac{y(\lambda)+1}{2} \cdot I(\lambda, k) & \text{if } y(\lambda) < -0,5 \\ I(\lambda, k) & \text{otherwise} \end{cases} \tag{5}$$

Since the use of the noisy microphone signal as input data lead to weak classification results of the VAD, we propose a two-pass noise reduction to enhance the classification performance. In a first step, the method of Rangachari is applied using its original signal decision value $I(\lambda, k)$. The resulting enhanced audio signal is used as the input of the VAD. A second noise reduction with the same input data as the first one is applied using the modified speech presence probability $I^*(\lambda, k)$ and providing the final noise reduced audio signal.

IV. Experimental Results

Our proposed auditory system will be used for the applications described in section II. Fig. 3 shows an example for the output of the auditory system integrating the aspects discussed. The adaptive noise reduction reduces the underlying noise automatically providing better input signals to the VAD and ASR systems. Originally, the sound source localization responded to any sound source present. By using the VAD, the non-speech detections can be ignored providing the robot with a speaker localization.

The adaptive noise reduction decreases the influence of the noise automatically within the first few seconds (see Fig. 3). It is even capable of adapting to sudden changes in the underlying noise spectrum, as can be seen in Fig. 4. Therefore, the VAD receives a signal containing much less

noise which improves the speech/non-speech classification performance.

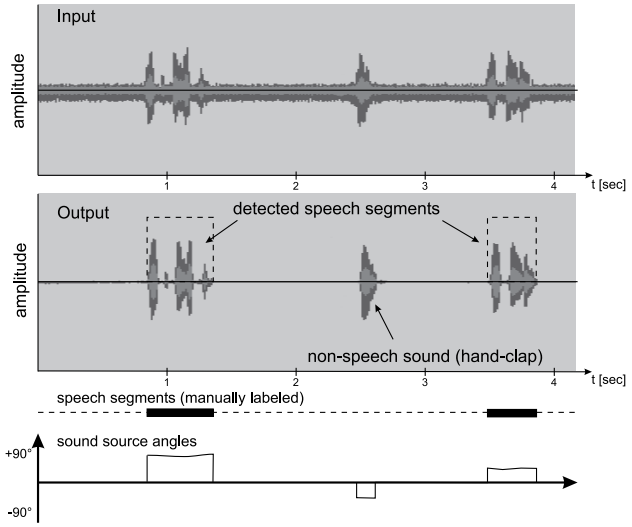

Fig. 3. **Top:** The noisy input signal containing two speech utterances and one non-speech sound (hand-clap). **Middle:** The output of the auditory system. The noise is reduced and the speech and non-speech segments have been classified. **Bottom:** The detected angle of the sound source localization, which can be combined with the classification results to ignore non-speech sounds.

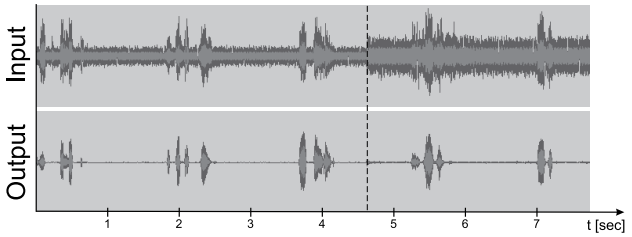

Fig. 4. The wave form of one microphone signal (top) and the output of the proposed auditory system (bottom). The noise reduction reduces the amount of noise over time, even after a sudden change in the noise characteristics (dotted line).

The classification performance of the neural VAD was evaluated using the voice of 30 different speakers recorded in an office environment at distances of 75 cm and 150 cm to the robot. These voice recordings were manually labeled as speech and non-speech segments. Additionally, non-speech sound sources were recorded such as hand claps, clicks, noises produced by computer-fans, etc.. In total, there were 369,377 frames processed by the VAD (∼70 min. at a window length of 1024 samples and 50 % overlap). Tab. I shows the resulting classification rates: 83,68 % of the overall time frames were correctly classified as speech/non-speech. Additionally, the voice recordings were grouped into recordings containing keywords of the ASR system and recordings of a read out text passage. As can be seen in Fig. 5, the classification rates of the keywords are slightly higher because these were articulated better, whereas the read out text contained more weak speech components. As can also be seen, the classification performance is degrading with larger distances. This is because the signal-to-noise ratio gets lower and the influence of reverberation is increasing in that case. The best classification performance (87,23 %) is achieved with ASR keywords at a distance of 75 cm. Since the robot HOROS is a service robot aiming at dialog-based interaction, this is a typical use case. Most of the time the speaking user will be located right in front of the robot at interaction distance.

	frames	correct
speech	184,136	152,699 (82,93 %)
non-speech	185,241	156,382 (84,42 %)
overall	369,377	309,081 (83,68 %)

TABLE I

THE TOTAL NUMBER OF FRAMES TESTED AND THE NUMBER OF CORRECTLY CLASSIFIED FRAMES USING THE PROPOSED NEURAL VAD.

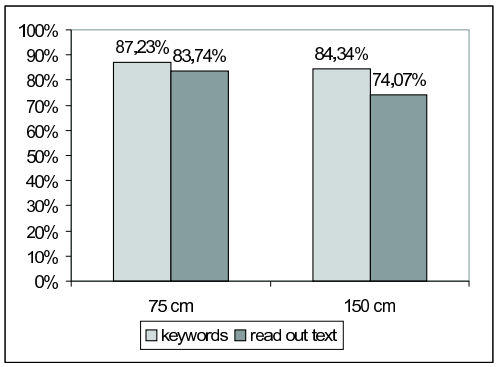

Fig. 5. The speech/non-speech classification rates of the VAD at speaker distances of 75 cm and 150 cm for keywords of the used ASR system and a read out text paragraph.

By extending the *adaptive noise reduction* method by Rangachari et al. [17] with the proposed neural VAD, the detection of non-speech regions in the recorded audio signal can be improved significantly. Fig. 6 shows the mean square errors of the noise power spectrum estimates using Rangachari's method and our proposed method. The recordings contained highly non-stationary signals of noise sources so the noise reduction algorithms needed to adapt their noise power spectrum over time. The proposed method is able to adapt faster to the original spectrum leading to significantly lower noise estimation errors. This is achieved by improving the detection of non-speech regions with the neural VAD.

Finally, we show the improvement of the *word recognition rate* using the commercial, speaker-independent, and keyword-based ASR system "Novotech GPMSC" [16]. For this purpose the recognition rates were evaluated using a total of 120 clean utterances provided by four speakers (2 male, 2 female) out of a vocabulary containing 93 utterances. These were mixed with noise recordings to provide different signal-to-noise ratios in the range from -5 to 30 dB in steps of 5 dB. The recognition rates of the ASR system were evaluated using the enhanced speech signal of the auditory system and compared to the results where the unprocessed microphone

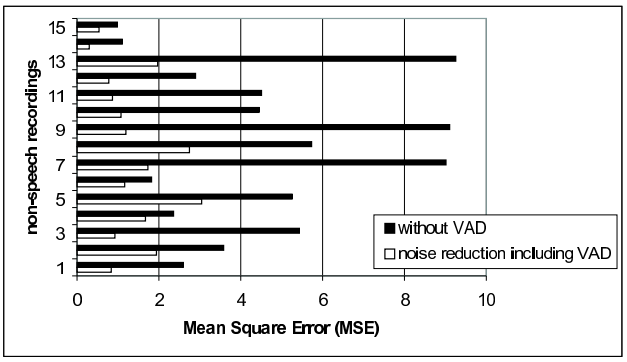

Fig. 6. Mean square error of the noise-power spectrum estimation of highly non-stationary noise sources using the method by Rangachari et al. [17] and the proposed extension with the neural VAD.

signals were used. The results are shown in Fig. 7. As can be seen, the recognition performance can be increased significantly in the presence of strong background noise. On recordings with an SNR of 0 dB, the ASR performance could be improved by 37.5%. As expected, for high signal-to-noise ratios the improvement in terms of recognition rates gets smaller. The SNR of typical voice recordings presented to the robot in real-world scenarios is app. 5-20 dB. It depends on the sound level of the speaker and the respective environment noise. The increase of the recognition rate for this range of SNR was 1-28%.

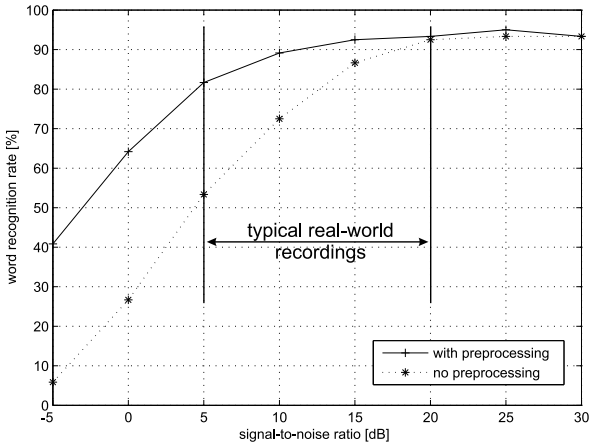

Fig. 7. Recognition rates of the ASR system at different signal-to-noise ratios.

V. Conclusions

We presented a new adaptive auditory preprocessing for a mobile interaction robot. The system was designed to improve automatic speech recognition by using adaptive noise reduction. Experimental results showed a noticeable increase in word recognition performance of up to 37.5% if the proposed auditory system was used. A voice activity detector has been implemented using a neural classifier approach. This allows the robot to detect speakers by combining its result with a sound source localization. A voice recorder has been enhanced with an automatic start/stop controller reacting on the user's speech. The system proved its ability to detect the user's speech even under noisy conditions making it suitable for the use in real-world environments.

Further improvement of the system might be the implementation of an improved sound source localization algorithm for detecting multiple sound sources simultaneously.

References

[1] S. Li, M. Kleinehagenbrock, J. Fritsch, B. Wrede, and G. Sagerer, "BIRON, let me show you something": Evaluating the Interaction with a Robot Companion. *Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC), Special Session on Human-Robot Interaction*, pp. 2827–2834, The Hague, The Netherlands, October 2004.

[2] R. Stiefelhagen, C. Fügen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel, Natural Human-Robot Interaction using Speech, Head Pose and Gestures, *Proc. IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, vol. 3, pp. 2422–2427, 2004.

[3] S. Yamamoto, K. Nakadai, J-M. Valin, J. Rouat, F. Michaud, K. Komatani, T. Ogata, and H.G. Okuno, Making a robot recognize three simultaneous sentences in real-time, *Proc. IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, pp. 4040–4045, August 2005.

[4] F. Michaud, Y. Brosseau, C. Cote, D. Letourneau, P. Moisan, A. Ponchon, C. Raievsky, J-M. Valin, E. Beaudry, and F. Kabanza, Modularity and integration in the design of a socially interactive robot, *IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN)*, pp. 172–177, August 2005.

[5] L. C. Parra and C. V. Alvino, Geometric Source Separation: Merging Convolutive Source Separation With Geometric Beamforming, *IEEE Trans. Speech and Audio Processing*, vol. 10, pp. 352–362, 2002.

[6] J-M. Valin, J. Rouat, F. Michaud, Microphone Array Post-Filter for Separation of Simultaneous Non-Stationary Sources, *IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 221–224, May 2004.

[7] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition, *IEEE Trans. Speech and Audio Processing*, vol. 10, pp. 146-157, 2002.

[8] A. Ouzounov, Robust Features for Speech Detection - A Comparative Study, *Int. Conf. Computer Systems and Techologies*, 2005

[9] L.R. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[10] B-F. Wu and K-C. Wang, Robust Endpoint Detection Algorithm Based on the Adaptive Band-Partitioning Spectral Entropy in Adverse Environments, *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 762-775, September 2005.

[11] J. Shen, J. Hung, and L. Lee, Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments, *Int. Conf. Spoken Language Processing*, 1998.

[12] J. Kacur, G. Rozinaj, and S. Herrera-Garcia, Speech Signal Detection in a Noisy Environment using Neural Networks and Cepstral Matrices, *Electrical Engineering*, 2004.

[13] R. Gemello, F. Mana, and R. De Mori, Non-linear estimation of voice activity to improve automatic recognition of noisy speech, *Interspeech*, 2005.

[14] I. Cohen, Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging, *IEEE Trans. Speech and Audio Processing*, vol. 11, pp. 466-475, September 2003.

[15] C. M. Bishop, Neural Networks for Pattern Recognition, *Oxford University Press*, 1995.

[16] Novotech, GPMSC (General Purpose Machines' Speech Control), *http://www.novotech-gmbh.de/speech control.htm*, 2006.

[17] S. Rangachari and P. C. Loizou, A noise-estimation algorithm for highly non-stationary environments, *Speech Communication*, pp. 220-231, 2006.

[18] C. Martin, E. Schaffernicht, A. Scheidig, and H.-M. Gross, Sensor Fusion using a Probabilistic Aggregation Scheme for People Detection and People Tracking, *Robotics and Autonomous Systems 54*, Elsevier Science, pp. 721-728, 2006.

[19] R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics, *IEEE Trans. Speech and Audio Processing*, vol. 9 (5), pp. 504-512, 2001.

[20] Y. Hu and P. Loizou, Speech enhancement based on wavelet thresholding the multitaper spectrum, *IEEE Trans. Speech Audio Processing*, vol. 12 (1), pp. 59-67, 2004.