Proceedings of the International Multiconference on Computer Science and Information Technology pp. 381–388
ISBN 978-83-60810-27-9
ISSN 1896-7094

Development of a Voice Control Interface for Navigating Robots and Evaluation in Outdoor Environments

Ravi Coote
Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE
Neuenahrer Str. 20, 53343 Wachtberg, Germany
ravi.coote@fkie.fraunhofer.de

***Abstract*—In this paper the development of a prototypic mobile voice control for navigating autonomous robots within a multi robot system is described. As basis for the voice control a hidden markov model based speech recognizer with a very low vocabulary of 30 words is utilized. It is investigated how many training samples for a markov model are required for a normal operation of speaker-dependent speech recognition. Therefore, hidden markov models were developed successively in parallel with an own training data corpus containing finally 2290 utterances from 12 speakers. Within the successive development of acoustical models and training corpus, the work revealed details about how many speakers are necessary to achieve an acceptable degree of speaker independence. We focused on an evaluation of the speech recognizer in adverse outdoor environments. The evaluation ranges from almost calm conditions of about 39 dB up to very adverse noise conditions of 120 dB. It is investigated whether a small vocabulary attenuates the noise vulnerability and in how far an increase of speaking volume can compensate noises of different intensity. The voice control was tested in outdoor environments and aspects of its usage are described.**

I. Introduction

IN HUMAN-MACHINE scenarios, e.g., where the user does not have his hands free to type in commands, or where the user is handicapped, the ability to control a system by voice can be considered. For those purposes usually small vocabularies are sufficient. In calm acoustical environments, e.g., in flats, low vocabulary speech recognition performs almost perfectly. Unfortunately in outside-scenarios or in in-vehicle-scenarios the acoustical environment can be very adverse.

It is not clear in how far a small vocabulary can compensate such bad acoustical conditions in order to maintain an acceptable word recognition rate (WRR) of 95%. Furthermore, it is an open question in which noise scenarios an increase of the speaking volume can maintain this recognition rate.

To this end, a low vocabulary speech recognizer was developed and evaluated under several adverse conditions. The evaluation ranged from almost calm conditions of about 39 dB up to very adverse noise conditions of 120 dB. For the training of models we constructed an acoustic corpus containing 2290 hand labeled German utterances from 12 people with different accents and relevant issues in corpus construction are described. For the unit of speech that has to be acoustically modeled by hidden markov models (HMMs), the word was chosen. Suitable numbers of gaussian mixture components were specified for speaker-dependent and speaker-independent training. The successive development of the acoustic models revealed insights in how many training samples are required per model and how many speakers are needed for speaker independent speech recognition. The core speech recognizer was connected to the software framework of the robots by means of suitable software libraries as will be explained in Section II. Finally, the speech recognizer was integrated into a voice control application for navigating robots within a multi robot system and issues in operating the voice control in outside environments are described.

A. Related work and Goals

In various studies experimental speech communication with robots has already been developed and successfully used (see, e.g., [1], [2]). However, in these works no studies were conducted regarding the performance of speech recognition when used in different noise scenarios. Therefore, we developed the voice control application with the aim to give answers to the following questions:

1) How strong is the effect of street noise, crowd noise, and in-vehicle noise of different degrees on the performance of speech recognition? Can an increase of speaking volume improve recognition rate and can a vocabulary with less than 50 words compensate such noise?
2) How many speakers are necessary to achieve an acceptable level of small-vocabulary speaker independence?
3) Does direct voice input suit to perform spatial navigation tasks?

The rest of the paper is structured as follows. Next, Section II describes groundwork and conceptual considerations for the voice control. The successive development of the application is illustrated in Section III. The speech recognizers performance is evaluated in Section IV and Section V concludes with a discussion.

II. Conception

This section describes basic conditions and conceptual considerations on which basis the voice control was developed.

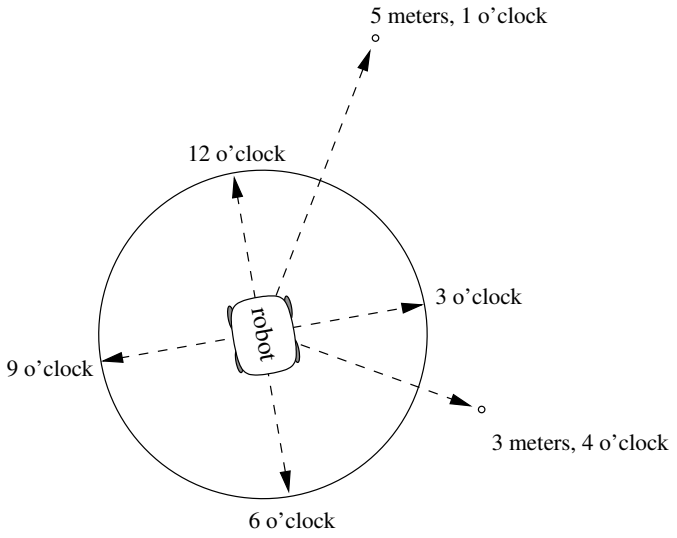

Fig. 1. Specifying target coordinates in a robot-centered 2-dimensional area

A. Navigation by the use of voice

The goal was to enable a navigation of robots onto a two-dimensional arbitrary ground, i.e., to move the robots backwards and forwards, and to rotate and stop. To allow for such a control, one single command was defined to consist of the specification of a direction and a distance. The values of distance and direction are given in *meters*[1] and *o'clock*. For instance, if the user commands robot 1 to go five meters forwards, the command to be uttered is formed as *"robot 1: 5 meters, 12 o'clock"*. If robot 1 should drive five meters to the right, the command is *"robot 1: 5 meters, 3 o'clock"*. The values for the parameters distance and direction are always specified relative to the robot (see also Figure 1).

B. Reused tools, software and platforms

In the following, toolkits and robot platforms utilized for the implementation of the voice control application are described.

Robots software and hardware: For the operation of the voice control we used robots of ATRV series from Real World Interface[2]. An ATRV-robot is a four-wheeled mobile platform equipped with sonar sensors and wireless ethernet communications. The ATRV-robots employ the software robot framework *RoSe* [3], [4], which serves as framework for control and communication among roboters. A C++ application is embedded in this framework and is called a *RoSe service*. The framework provides methods that allow RoSe services to communicate with each other via wireless link. A relevant service for the voice control is the *collision avoidance service* [5]. To this service a target coordinate in a two-dimensional robot-centered coordinate system can be passed. The service computes a path to the target coordinate which prevents to collide with obstacles. In order to put the robot in motion the collision avoidance service instructs a further RoSe service which is responsible for control of the robot's motor.

[1]All commands have to be uttered in German. But for better readability they are written in English throughout the paper.

[2]formerly RWI, now iRobot, http://www.irobot.com

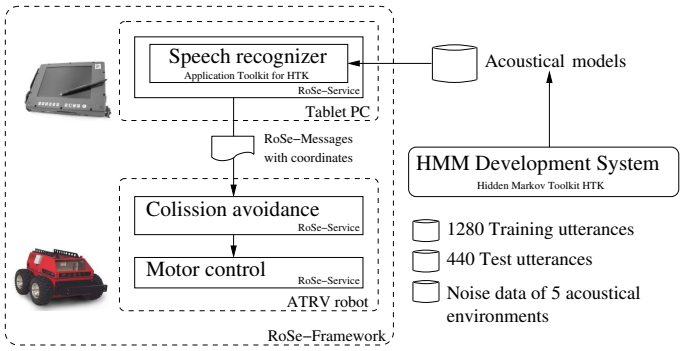

Fig. 2. Several software components arranged into an overall system for development, evaluation, and tests in outdoor environments

Toolkits for speech recognition: For development of acoustic models and evaluation under noise, the hidden markov Toolkit (HTK, [6]) was used. For embedding the developed acoustic models into a C++-application of the robots framework the *Application Toolkit for HTK* (ATK, [7]) was used. ATK is designed to create experimental applications based on HTK. It consists of a C++ layer which directly accesses the HTK library modules. ATK enables acoustic models that have been developed by HTK to be reused and integrate them into arbitrary C++-applications wrapped by ATK.

Composition of the voice control application: A combination of the software tools and hardware platforms described above composed the voice control application as follows. Acoustic models were trained with a development system based on HTK. These models were loaded into ATK. A RoSe service was written which utilizes the ATK libraries and, in particular, markov recognition algorithms and thus represents a RoSe service for voice control. Since the voice control appears as RoSe service, it is able to send messages with target coordinates to the collision avoidance service (see Figure 2).

C. Determining the speech recognizers application scope

This section describes the determination of adequate parameters for defining the speech recognizers application scope.

Vocabulary size: A small vocabulary of around 30 words was chosen to cover the required words for navigating the robot like, e.g., "robot", "meter", "o'clock" and several numbers ranging from "one" to "twelve".

Degree of speaker independence: Investigations dealing with speaker independence indicate different numbers of speakers required to achieve an acceptable degree of speaker independence [8], [9], [10]. According to *Lee* [9], at least 100 male speakers are in the training set as a minimum requirement for speaker independence. Furthermore, *Kubala* [10] shows that with 12 carefully selected speakers the same degree of speaker independence as a reference system can be obtained which was trained with 100 speakers. Thus, in this work utterances of 12 speakers of our research group were used as data basis. According to the statement of *Kubala* [10], it

is assumed that even with this little number of speakers an approximately similar degree of speaker independence can be achieved as with a system that is trained with great speaker numbers.

D. Conception of Acoustic Modeling

The parameters that define the structure of the hidden markov models and the feature extraction are as follows.

Unit of speech: As unit of speech that has to be modeled by hidden markov models the word was chosen.

Number of states: The number of states per word-model was chosen dependent from words number of phonems but an extra state was added for the closure phase of plosives to model their non-stationarity more adequately. The number of states of the HMM was chosen to be linear to the number of phonemes in the corresponding word. This was intended to ensure that the phonetic structure of words is identical with the states of the HMMs. Furthermore, Bakis models [11] were used in which each of the next state may be skipped. This was intended, to take care of articulatory phenomena like vowel reduction.

Number of gaussian mixture densities: In literature there is no way known how to calculate the correct number of gaussian mixture densities. Accordingly, the number of gaussian mixture densities was kept variable. In development of acoustic models various numbers of gaussian mixture densities were tested in order to determine a suitable value.

Feature extraction: Mel Frequency Cepstral Coefficients (MFCC) were used to simulate a frequency sensitivity that is similar to that of the human ear.

E. Conception of the Acoustic Corpus

The overall speech corpus was recorded with a *Sennheiser PC 30* microphone.

Utterances to be spoken: It was determined to use speech samples out of continuously spoken utterances. Phenomena of coarticulation as they will occur in the operation of the speech are aimed to be included into the models. For instance, a sentence that had to be recorded looked like the following.

```
robot one drive ten meters twelve o'clock
```

Amount of utterances to be spoken: As requirement for HMM training for each model that should be trained at least 10, but better 50 or 100 samples should be available [8]. With a vocabulary of about 30 words, it should be sufficient to take 60 training utterances as basis to achieve a set of 20 references per word. Thus, in this work 60 training utterances were specified as minimum.

Manual annotation of utterances: Training procedures for hidden markov models require model specific pre-annotated audio data. For good results, at least an initial annotation for the models should be provided [12], [8]. For that reason parts of the corpus were manually segmented on word level. For the utterances for which no segmentation has taken place, a complete orthographic annotation was required instead.

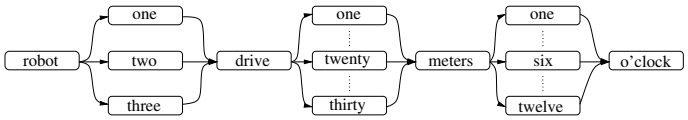

Fig. 3. Grammar network to cover navigation commands

III. Development and Implementation

In this section the development of the speech recognizer, its connection to the robots framework, and its integration onto a Tablet PC is described.

A. Realization of the Grammar

To allow for the requested operation the following grammar has been constructed. A command basically consists of a citation of the robot which is to execute a command, e.g., "robot two", and the actual statement, e.g., "drive ten meters towards twelve o'clock". Figure 3 depicts the grammar where silence models are omitted for better clarity.

B. Development of the Acoustic Corpus

In order to create the corpus, speaker utterances for training and for testing were recorded, afterwords post-processed, and finally partly annotated on word level with *Praat* [13].

Recording: The recordings have taken place in a carpeted large room with curtains. In total 220 training and 50 test utterances were recorded from the author. From each of the 12 speakers, 72 training and 50 test utterances were recorded. The whole corpus consists of 2290 utterances which includes the training part of 1850 utterances and the test part of 440 utterances. Issues related to the recording procedure were as follows. In order to maintain an adequate recording level and to avoid overmodulation, the distance to the mouth was re-adjusted for each speaker. For very loud or very soft voices the recording volume of the sound device had to be adjusted. Increasing the recording volume had to be done carefully in order to avoid too much inclusion of ambient noise into the signal. Sometimes it was difficult to maintain the same mean energy due to a movement of the microphone or a change of speaking volume. To ensure a flawless corpus, it was necessary to review the recorded utterances and, if some utterances were faulty, to capture those again.

Post-processing: The recorded utterances had needed to be post-processed such that only those audio data was included in the speech signals that were specified by our orthographic annotation. Thus, the utterances were freed from previous and following silence with standard sound-editing software. Random reviewing of temporal and spectral variation has taken place at approximately one third of the statements. Attention in inspecting the utterances was paid to modulation issues like insufficient modulation or overmodulation.

Manual annotation: All 240 utterances of the author were entirely manually annotated considering to use them for speaker-depended training. 10 of the 70 utterances of remaining speakers have been annotated manually aiming to use them for speaker-independent training. An automatic conversion

TABLE I
NUMBER OF UTTERANCES FOR TRAINING AND EVALUATION

Subject	Sex	Accent	Training	Segmented	Test
AK	male	Franconian	72	10	50
AT	male	Russian	72	10	10
BB	male	High German	72	10	50
DS	male	High German	72	10	50
FH	male	High German	72	10	50
FS	male	High German	0	0	10
HLW	male	High German	72	10	50
HM	male	Arabic	0	0	10
HN	male	High German	72	10	50
MS	male	High German	0	0	10
RC	male	High German	220	220	50
SR	male	High German	72	10	50
TB	male	High German	72	10	50
TR	male	High German	72	10	50
			1850	320	440

TABLE II
MODEL VERSIONS FOR SPEAKER-DEPENDENT TRAINING

Mixtures	References per model	WRR
4	≥ 5	97.56
4	≥ 10	98.00
4	≥ 21	94.17
4	≥ 21	99.04
4	≥ 23	100.00

TABLE III
MODEL VERSIONS FOR SPEAKER-INDEPENDENT SPEECH RECOGNITION

Mixtures	Training Speakers	Test Speakers	WRR
4	RC	SR	75.00
4	RC,SR	HN,DS	65.00
4	RC,SR,HN,DS	BB,TR,FS,MS,HLW	76.00
2	RC,SR,HN,DS,HLW	BB,TR,FS,MS	52.00
3	RC,SR,HN,DS,HLW	BB,TR,FS,MS	90.00
3	RC,SR,HN,DS,HLW,TB	BB,TR,FS,MS	90.00
3	RC,SR,HN,DS,HLW,TB,FH	BB,TR,FS,MS,HM,AT	93.34

of annotations from the Praat format into HTK format has been realized with linux scripts. Naturally appearing speaking sounds like those of smacking and exhalation were separately annotated intending to train particular models for them later on. When smacking sounds merged with subsequent verbal sounds those were included into the annotation boundaries of the whole word. Small pauses between words were handled such that the word boundary was set in the middle of the pause. Regarding the annotation of silence, it had to be ensured that the duration of the intervals was consistently the same. Therefore, we chose an interval of 110 to 130 ms for silence annotations.

C. Development of acoustic models

In literature there was no precise information found about how much data is necessary for training of acoustic models. For this reason, the development of acoustic models was keeping pace with the creation of the acoustic body. In each development step, a model version was created. Below the structure and development of acoustic models is described.

Creating the HMM structure: The number of model states N_s was initially selected for each model by $N_s = N \cdot (P+3)$ with the word's number of phonemes P, and a weight factor N. A small number of 3 states were provided for the word's beginning and the word's end. After the results for $N > 1$ were significantly worse, N was set to 1. In determining the number of phonemes of a word an extra state was provided for closure phases of plosives. For instance, for the German word 'roboter' 11 states were provided where two states are considered to model the closure phases of /b/ and /t/.

Training of HMMs: The training of models was carried out in two main steps. First, a speaker-dependent model was developed and optimized for high detection rates. Second, the speaker-dependent model was gradually extended to other speakers to achieve a high degree of speaker independence. To keep the development cycle as small as possible, the entire exercise was automated with linux scripts for which HMM parameters could be specified. The varied parameters of the training were *number of gaussian mixture components* and *number of training reestimations*. Tables II and III show the development of the speaker-dependent and speaker-independent models. As expected, it was observed that the degree of speaker independence increased for each additional speaker in the training set.

D. Creating the link to the robot framework

After development of the acoustic models was finished, a RoSe service has been written in C++ in which ATK connects the acoustic models with the RoSe-framework (see Fig. 2). ATK provides methods for starting the speech recognition process and returns the word chain that is assumed to be uttered. From the recognized word chain values for distance and direction are extracted by regular expressions. If the distance r in meters and the angle α is given in clock the target coordinate (x, y) was determined by $x = r \cdot sin(\frac{\alpha \cdot \pi}{6})$ and $y = r \cdot cos(\frac{\alpha \cdot \pi}{6})$. The target coordinate is then sent via a RoSe-message to the RoSe service for collision avoidance which is responsible for further activation of the robot's motor.

E. Integration onto a Tablet PC and outdoor tests

In the integration and test of the speech recognizer on a Tablet PC, it was observed that the sensitivity of the microphone had to be adjusted such that less ambient noise was included in the signal. If the recording level was set too high, the detection rate fell off dramatically. This was noted especially for operation in a outdoor environments when the microphone was adjusted too sensitive because even little noise was included in the signal.

IV. EVALUATION

The overall development of the voice control described above has already revealed some details about the degree of

TABLE IV
RESULTS OF THE LEAVE-ONE-OUT-TEST FOR MEASURING THE DEGREE OF SPEAKER-INDEPENDENCE, 1 MIXTURE, REESTIMATIONS 1-10

Training data	Test data	WRR
HN,HLW,AK,BB,FH,TB,SR,AT,DS,TR	RA	98%
HLW,AK,BB,FH,TB,SR,AT,DS,TR,RA	HN	100%
AK,BB,FH,TB,SR,AT,DS,TR,RA,HN	HLW	100%
BB,FH,TB,SR,AT,DS,TR,RA,HN,HLW	AK	84%
FH,TB,SR,AT,DS,TR,RA,HN,HLW,AK	BB	100%
TB,SR,AT,DS,TR,RA,HN,HLW,AK,BB	FH	98%
SR,AT,DS,TR,RA,HN,HLW,AK,BB,FH	TB	98%
AT,DS,TR,RA,HN,HLW,AK,BB,FH,TB	SR	96%
DS,TR,RA,HN,HLW,AK,BB,FH,TB,SR	AT	100%
RA,HN,HLW,AK,BB,FH,TB,SR,AT,DS,TR	FS	90%
RA,HN,HLW,AK,BB,FH,TB,SR,AT,DS,TR	HM	100%
RA,HN,HLW,AK,BB,FH,TB,SR,AT,DS,TR	MS	100%
	Mean	97%

TABLE V
AVERAGE SOUND LEVELS OF MALE SPEAKERS IN 1 M DISTANCE OF SPEAKERS MOUTH FOR SPECIFIED SPEAKING STYLES; P = PRIVATE FIELD, FROM LAZARUS [16], SUPPLEMENTED WITH SPECIFICATIONS AT SPEAKERS MOUTH

Speaking style	1 m distance	3,125 cm distance
whispering	36 dB	66 dB
softly speaking	42 dB	72 dB
relaxed speaking (p)	48 dB	78 dB
relaxed, normal (p) speaking	54 dB	84 dB
normal, raised (p) speaking	60 dB	90 dB
raised speaking	66 dB	96 dB
speaking loudly	72 dB	102 dB
speaking very loudly	78 dB	108 dB
screaming	84 dB	114 dB
screaming maximally	90 dB	120 dB
screaming maximally (single cases)	96 dB	126 dB

speaker independence and vulnerability to noise. In this section these features are examined in more detail.

A. Speaker independent recognition

The speaker-independent training has shown that the degree of speaker independence increased with the number of speakers in the training set. For evaluation of the degree of speaker independence, only those speakers had to be used who were not in the training set of acoustic models. By repeatedly leaving out a speaker in the respective training set and testing only this specific speaker, i.e., performing a *leave-one-out test*, one measurement of the degree of speaker independence for this specific speaker could be achieved. Since a total number of 11 speakers were available, the exhaust test was repeated several times for each available speaker. Depending on the mean value of the results, the general degree of speaker independence was estimated. The test results are shown in Table IV.

B. Recognition in noise

In the following the speech recognizers evaluation under noise adversity of low and high degrees is described.

Acquisition of noise: For the environmental conditions *calm outdoor environment* and *busy street*, the noise was recorded with the same microphone used for creating the acoustic corpus. The adjustment of the recording volume was done manually such that it was initially set to zero and continuously increased up to a good modulation of amplitudes between values of 0.5 and -0.5. For every recording, the sound pressure level was measured in dB with a sound level meter. Noise of the high adversity environments *babble* and *track vehicles* was taken from the corpus *Noisex* [14].

Overlaying procedure: Both the speaker-dependent and speaker-independent models were evaluated. For speaker-dependent evaluation in noise, test utterances of speaker RC were used. For speaker-independent evaluation in noise, test utterances of all speakers were used. Furthermore, noise adversity was simulated by artificially superimposing the clear commands from the test corpus with the software tool *FaNT* (Filtering and Noise Adding Tool, see [15]). The operation of FaNT requires SNR values to be specified which represents the intensity with which the noise superimposes the speech signal. The ratio of signal to noise depends on the sound level of speech and on the sound level of the noise. Sound levels of noise were measured in case of recording and are specified by *Noisex* [14] in case of noise corpus usage. Sound levels of speaking styles are taken from Lazarus et al. [16]. They provide average sound levels of different speaking styles, i.e., whispering, speaking softly, and relaxed speaking. Therefore, men produce by whispering at a distance of one meter a sound pressure level of 36 dB. By screaming, up to 96 dB can be achieved in some cases. For a summary of the data collected see Table V. Thus, utterances were superimposed at several SNRs in the range of small SNRs where the noise was barely noticeable up to large SNRs where the speech was hardly intelligible. In particular, superimposition ranged from -5 dB to 50 dB SNRs in 1dB steps. For the various noise scenarios, the test corpus was superimposed several times and new modified testing corpuses were created. The speech recognizer was then scheduled on the created corpora and word recognition rates were logged. This overlaying procedure has been used for noise evaluation of speech recognition in several works, e.g., [17], [18].

Overlaying Results: The sounds of a *calm outdoor environment* were obtained by recording at an average sound pressure of 39 dB. It can be seen that in a quite acoustic environment with a noise level of 40 dB, it is sufficient to speak softly to achieve very good recognition rates of at least 95% (see Figures 4 and 5). The sounds of a *busy street* were obtained by recording in a distance of 5 meters at an average sound pressure of 61 dB. The experiment showed that no good word recognition rates were possible when speaking relaxed. But by increasing the speaking volume good detection rates above 90% were achieved (see Figures 6 and 7). The *babble sounds* came from a large hall in which 100 people spoke to each other producing an average sound level of 88 dB. Hence, by

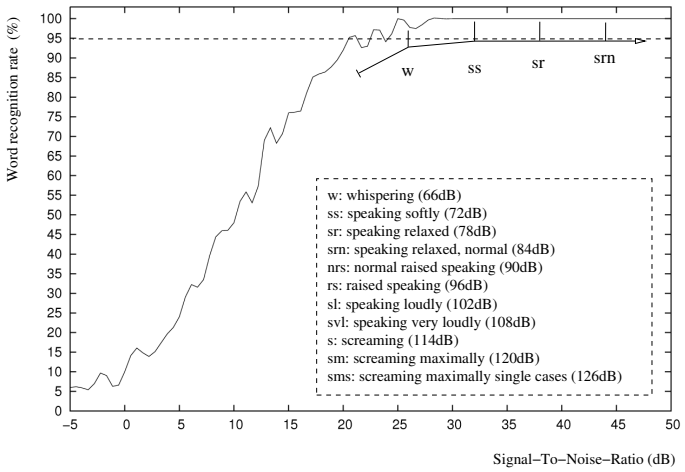

Fig. 4. Speaker-dependent recognition in 40 dB noise of a calm outdoor environment

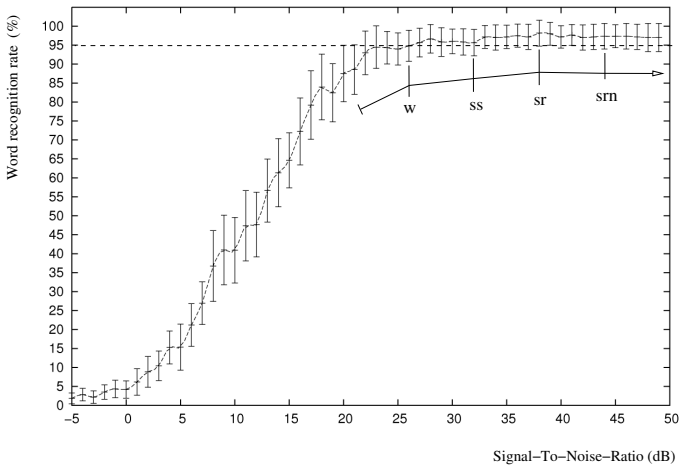

Fig. 5. Speaker-independent recognition in 40 dB noise of a calm outdoor environment

speaking very loud with 108 dB effecting an SNR of 20 dB, it could still be possible to achieve acceptable word recognition rates of around 90% (see Figures 8 and 9). The in-vehicle sounds came from *track vehicle 1* driving at a speed of 30 km/h producing an in-vehicle sound level of 100 dB. Good detection rates were virtually not able to be achieved. With a maximum achievable speaking volume of 120 dB, it would be theoretically possible even to achieve 80% recognition rate (see Figures 10 and 11). The in-vehicle sounds came from *track vehicle 2* driving at a speed of 70 km/h. The sound power level was specified with 114 dB. The results of the experiment may be taken from Figures 12 and 13. For those background noises virtually no satisfactory recognition rates were achieved.

V. Discussion and Conclusions

This section describes finally which conclusions were made and how this work can be used as a basis for further developments.

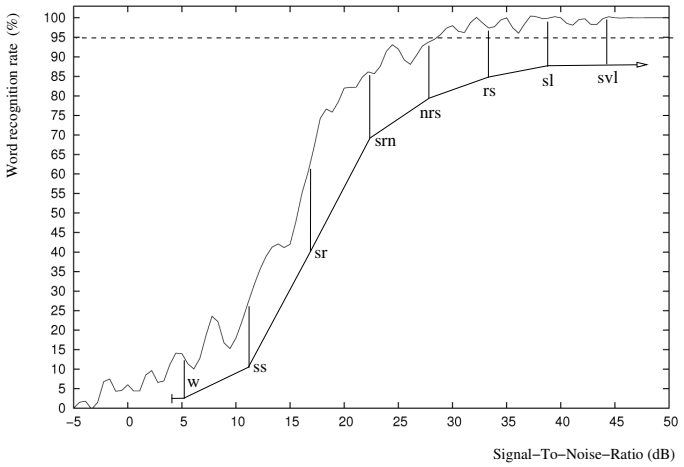

Fig. 6. Speaker-dependent recognition in 61 dB noise of a busy street in 5 meters distance

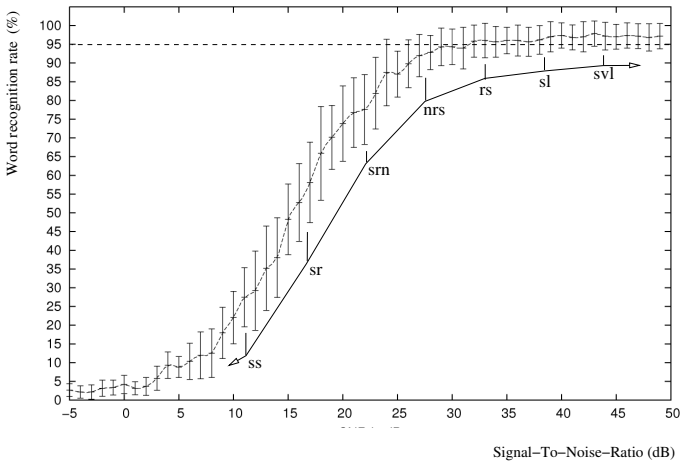

Fig. 7. Speaker-independent recognition in 61 dB noise of a busy street in 5 meters distance

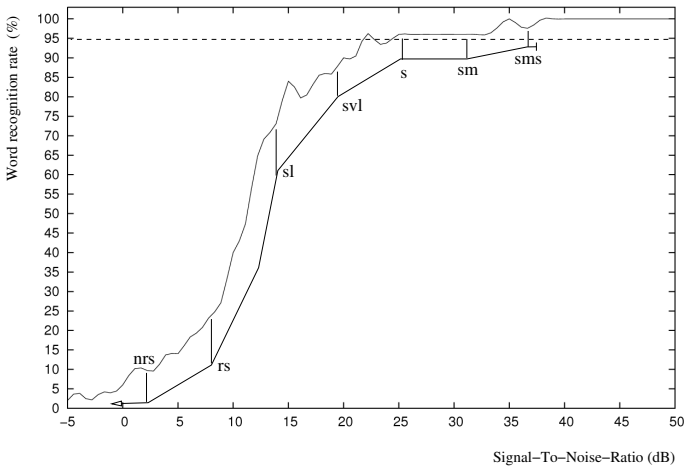

Fig. 8. Speaker-dependent recognition 88 dB noise of a crowd of 100 people in a large room

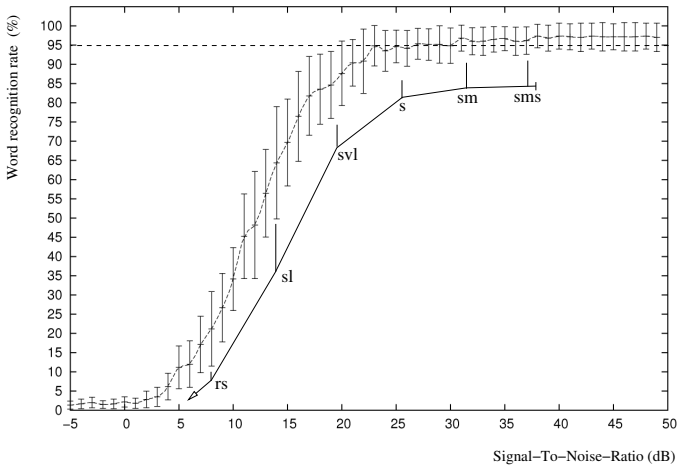

Fig. 9. Speaker-independent recognition 88 dB noise of a crowd of 100 people in a large room

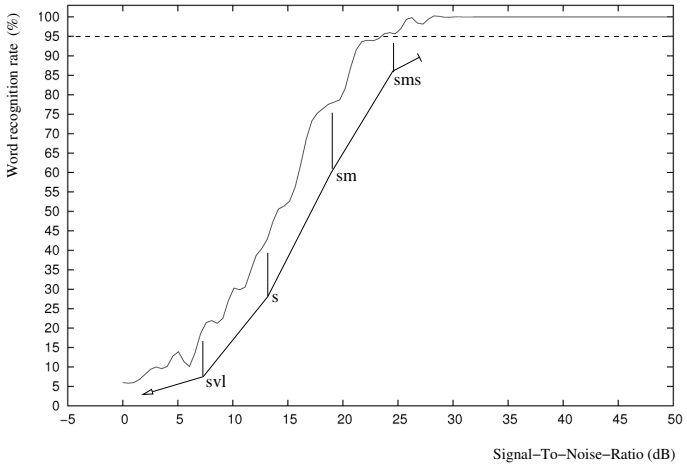

Fig. 10. Speaker-dependent recognition in 100 dB noise from within a track vehicle driving with 30km/h

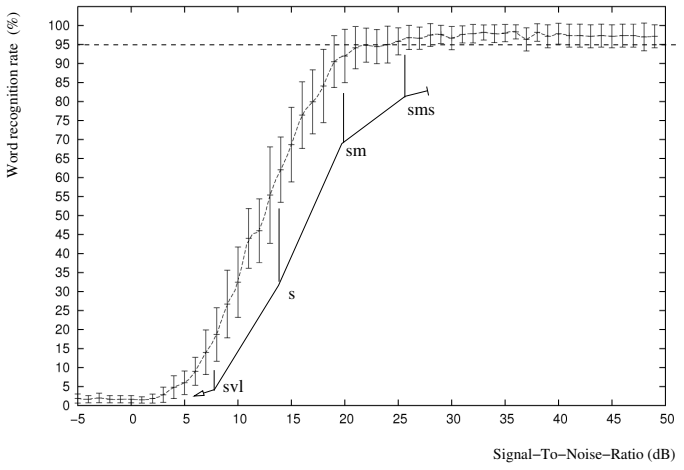

Fig. 11. Speaker-independent recognition in 100 dB noise from within a track vehicle driving with 30km/h

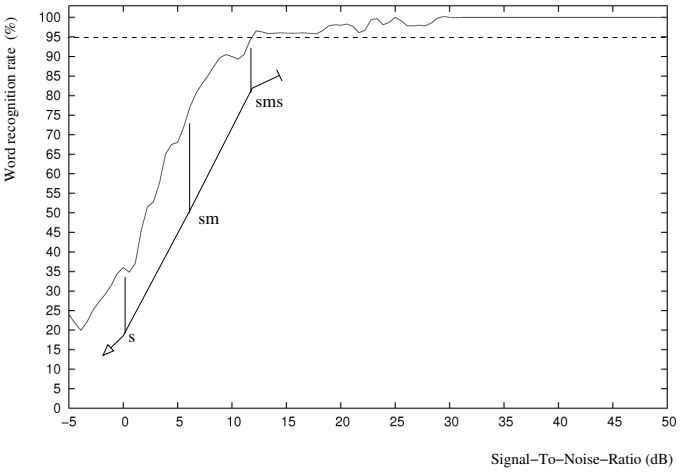

Fig. 12. Speaker-dependent recognition in 114 dB noise from within a track vehicle driving with 70km/h

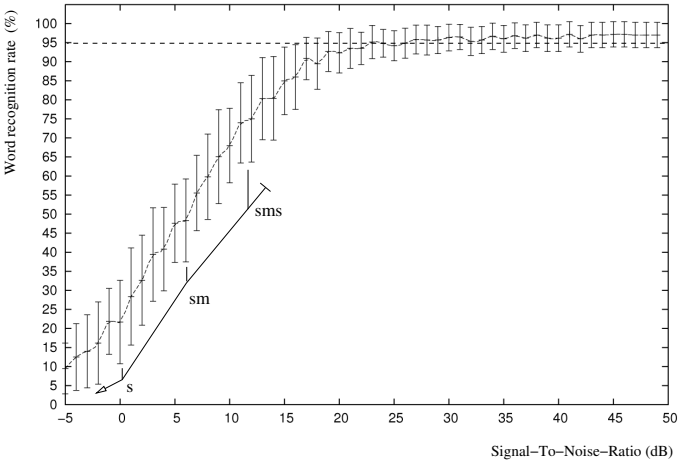

Fig. 13. Speaker-independent recognition in 114 dB noise from within a track vehicle driving with 70km/h

A. Acoustic modeling and degree of speaker independence

The training of acoustic models in Section III-C proved, that a number of around 20 samples per hidden markov model can be sufficient for normal operation of a HMM based speech recognizer. In the evaluation of speaker independence, it was shown that in a speech recognizer with a small vocabulary and a small number of speakers of around 10, a relatively high degree of speaker independence can be achieved.

B. Noise vulnerability

The evaluation had shown that when using a speech recognizer with vocabulary of solely 30 words in adverse environments of around 60 dB noise can be tackled by speaking with a raised voice in order to achieve a recognition rate of about 95%. In environments with extreme noise conditions from 80 dB up to 150 dB, the speech recognizer can no longer be used satisfactorily due to detection rates of around 80%. Here, a method for compensation of noise should be taken into account in conception of the speech recognizer. It could

be summarized that keeping a speech recognizers vocabulary small does not compensate adverse noise of high sound levels above 100 dB.

However, it should be noted that when speaking louder, not only SNR increases but also the vocal tract changes its shape and produces frequencies for which the acoustic models were not trained. The results obtained here are subject to some inaccuracy since the identified speech sound levels were calculated directly at the mouth and thus only represent an estimate. Furthermore, many microphones feature polar patterns so that the noise received on the side of the microphone is attenuated more than those which enter head-on. It can be concluded that slightly higher SNRs and better recognition rates could be achieved. The comparison of the speaker-specific and speaker-independent results suggests that models that were trained by much more data and therefore having a greater generality, do not necessarily offer worse results than models that were only trained by a speaker and are much more specific.

Regarding possible further developments noise sensitivity can be reduced based on several methods. Besides conventional signal filtering methods human strategies for speech understanding in noise can be employed. A recent survey about findings of human strategies for noise compensation can be found in *Loizou* [17].

C. Voice as input mode for navigation tasks

The usage of speech as an input mode for the control of robots must be carefully planned. In this work, speech which represents a verbal mean is used to perform a spatial continuous operation task. The objective called to enable a discrete navigation of the robots on an arbitrary two-dimensional ground. In considering how this task could be completed by means of speech, the continuous task was transformed into a discrete task by instructing the user to specify a target coordinate in a two-dimensional system that the user must consider first. The spatial thinking user has to transform his intention to move first into a verbal command which causes an additional cognitive load and costs time.

Regarding possible further developments, a holistic designed interface could take the requirements of the operation task and the expectations of the operator into account. The development should take place through an iterative design-implementation approach and the human-robot interface in field evaluation should be kept at pace with development. The result should be an effective human-robot interface that allows the user, even under extreme conditions like stress and noise, for a consistent, effective, and fast control of the robot. For control by voice discrete operating activities are suitable. It would be conceivable to raise the navigation commands for navigation to a higher level of abstraction. For example, the user could navigate the robots as follows:

- "Drive back to command center", or
- "Drive to robot group A, drive to robot group B"

Furthermore, semi-autonomous functions of the robot could be controlled. Examples of such instructions are:

- "Follow robot A",
- "Explore area, radius 50 m", or
- "Search for intruders".

REFERENCES

[1] S. Yamamoto, K. Nakadai, J. Valin, J. Rouat, F. Michaud, K. Komatani, T. Ogata, and H. G. Okuno, "Making A Robot Recognize Three Simultaneous Sentences in Real-Time," *Proceeding of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005.

[2] O. Majdalawieh, J. Gu, and M. Meng, "An HTK-Developed Hidden Markov Model for a Voice-Controlled Robotic System," *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004.

[3] A. Tiderko and T. Bachran, "A Service Oriented Framework for Wireless Communication in Mobile Multi Robot Systems," in *ROBOCOMM 2007: First International Conference on Robot Communication and Coordination*, Athen, 2007.

[4] ——, "A Framework for Multicast Communication over Unreliable Networks in Multi Robot Systems," *Proceedings of Towards Autonomous Robotic Systems*, 2007.

[5] F. Höller, *Personenbegleitung mit mobilen Robotern: Lokale Navigation mit probabilistischen Roadmapverfahren, FKIE-Bericht. Nr. 134*. Forschungsgesellschaft für angewandte Naturwissenschaften, Wachtberg, FKIE, 2007.

[6] S. Young, "The HTK Hidden Markov Model Toolkit: Design and Philosophy," Cambridge University (UK), Department of Engineering, Tech. Rep. 153, 1993.

[7] ——, "ATK - Application Toolkit for HTK," University of Cambridge, 2007.

[8] E. Schukat-Talamazzini, *Automatische Spracherkennung: Grundlagen, statistische Modelle und effiziente Algorithmen*. F. Vieweg, 1995.

[9] K.-F. Lee, *Automatic Speech Recognition: The Development of the SPHINX Recognition System (The Springer International Series in Engineering and Computer Science)*, 1st ed. Springer, 10 1988.

[10] F. Kubala and R. Schwartz, "A New Paradigm For Speaker-independent Training," in *ICASSP '91: Proceedings of the Acoustics, Speech, and Signal Processing. ICASSP-91*. Washington, DC, USA: IEEE Computer Society, 1991, pp. 833–836.

[11] R. Bakis, "Continuous speech recognition via centisecond acoustic states," *The Journal of the Acoustical Society of America*, vol. 59, 1976.

[12] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[13] P. Boersma and D. Weenink. (2008) Praat: Doing Phonetics by Computer (Version 5.0.32). [Online]. Available: http://www.praat.org

[14] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. Vol. 12, pp. 247–251, 1993.

[15] H. G. Hirsch. (2005) F a N T - Filtering and Noise Adding Tool. Hochschule Niederrhein. [Online]. Available: http://dnt.kr.hs-niederrhein.de

[16] H. Lazarus, A. C. Sust, R. Steckel, and K. P. Kulka M., *Akustische Grundlagen sprachlicher Kommunikation*. Springer, 2007.

[17] P. C. Loizou, "Comparison of Speech Enhancement Algorithms," in *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*. CRC, 2007, ch. 11, pp. 545 – 555.

[18] H. Hirsch and D. Pearce, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," *ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium*, 2000.