

# Towards Robust Speech Recognition for Human-Robot Interaction

Stefan Heinrich and Stefan Wermter

Knowledge Technology Group, Department of Informatics, University of Hamburg, Hamburg, Germany

Email: {heinrich,wermter}@informatik.uni-hamburg.de

**Abstract**—Robust speech recognition under noisy conditions like in human-robot interaction (HRI) in a natural environment often can only be achieved by relying on a headset and restricting the available set of utterances or the set of different speakers. Current automatic speech recognition (ASR) systems are commonly based on finite-state grammars (FSG) or statistical language models like Tri-grams, which achieve good recognition rates but have specific limitations such as a high rate of false positives or insufficient rates for the sentence accuracy. In this paper we present an investigation of comparing different forms of spoken human-robot interaction including a ceiling boundary microphone and microphones of the humanoid robot NAO with a headset. We describe and evaluate an ASR system using a multi-pass decoder – which combines the advantages of an FSG and a Tri-gram decoder – and show its usefulness in HRI.

## I. INTRODUCTION

With current speech recognition systems it is possible to reach an acceptable word recognition rate if the system has been adapted to a user, or if the system works under low-noise conditions. However, on the one hand in *human-robot interaction* (HRI) or in *ambient intelligence environments* (AmIE), the need for robust and automatic speech recognition is still immanent [1], [2]. On the other hand research in *cognitive neuroscience robotics* (CNR) and multimodal communication benefits from a robust and functioning speech recognition as a basis [3]. Headsets and other user-bound microphones are not convenient in a natural environment in which, for instance, a robot is supposed to interact with an elderly person. A microphone built into the robot or placed at the ceiling, a wall, or a table allows for free movement but reduces the quality of speech signals substantially because of larger distances to the person and therefore more background noise.

One method to deal with the additional problems is of course a further adaptation of the speech recogniser towards a domain-specific vocabulary and grammar. Enhancing recognised speech with a grammar-based decoder (*finite state grammar*, FSG) can lead to improved results in terms of recognised sentences, but it also leads to a high rate of false positives, since an FSG decoder tries to map the recognised utterances to legal sentences. To deal with this problem, one can combine the FSG with the classical Tri-gram decoder to reject unlikely results. Such a multi-pass decoder can be applied also to noisy sound sources like a ceiling boundary microphone or microphones, installed on a robot.

In the past research has been done on combining FSG and  $N$ -grams decoding processes: In 1997 Lin et. al. used an FSG and an  $N$ -gram decoder for spotting key-phrases in longer sentences [4]. Based on the assumption that sentences of interest are usually surrounded by carrier phrases,

they employed  $N$ -gram decoding to cover those surrounding phrases on the one hand and FSG decoding on the other hand if a start word of the grammar was found by the  $N$ -gram decoder. Furthermore, with their approach they rejected FSG-hypotheses if the average word score exceeded a preset threshold. However, this approach combined FSG and  $N$ -grams while modifying and fine-tuning the decoding processes on a very low-level, preventing to switch to another FSG or  $N$ -gram model easily. Therefore it would be interesting to exploit the dynamical result of an  $N$ -gram hypotheses list for the rating of an FSG-hypothesis instead of a fixed threshold.

Levit et. al. combined 2009 an FSG decoder and a second different decoder in a complimentary manner for the use in small devices [5]. In their approach they used an FSG decoder as a fast and efficient baseline recogniser, capable of recognising only a limited number of utterances. The second decoder, used for augmenting the first decoder, was also FSG-based but according to the authors could be replaced by a statistical language model like  $N$ -grams, too. An augmentation for the first decoder could be a 'decoy', which is a sentence with a similar meaning, similar to an already included sentence. However, those decoys can only be trained off-line. In this approach the result of the first decoder was not rated or rejected afterwards, but the search space was shaped to avoid the appearance of false positives.

Doostdar et. al. proposed 2008 an approach where an FSG and a Tri-gram decoder processed speech data independently based on a common acoustic model [6]. The best hypothesis of the FSG decoder was compared with the  $n$ -best list of hypotheses of the Tri-gram decoder. Without modifying essential parts of the underlying system they achieved a high false positive reduction and overall a good recognition rate, while they restricted the domain to 36 words and a command grammar. Although aiming for applying their system on service robots, they limited their investigation to the use of a headset. Yet it would be interesting to test such an approach far-field in a real environment using the service robots' microphones or other user-independent microphones.

In contrast, Sasaki et. al. investigated 2008 the usability of a command recognition system using a ceiling microphone array [7]. After detecting and separating a sound source an extracted sound was fed to a speech recogniser. The used open source speech recognition engine was configured for the use of 30 words and a very simple grammar allowing only 4 different sentence types like GO TO X or COME HERE. With their experiments, the authors have shown that using a ceiling microphone in combination with a limited dictionary leads to a moderate word accuracy rate. Also they claim that their

approach is applicable to a robot, which uses an embedded microphone array. A crucial open question is the effect on the sentence accuracy if a more natural interaction and therefore a larger vocabulary and grammar is being used. Based on the presented moderate word accuracy the sentence accuracy is likely to be small for sentences with more than three words, leading to many false positives.

In this paper we present a speech recognition approach with a multi-pass decoder in a home environment addressing the research question of the effect of the decoder in the far-field. We test the usability of HRI and investigate the effect of different microphones, including the microphones of the NAO humanoid robot and a boundary microphone, placed at the ceiling, compared to a standard headset. After analysing the background of speech recognition we will detail the description of a multi-pass decoder in section 2. Then we will describe the scenario for the empirical evaluation in section 3, present the results of our experiments in section 4, and draw a conclusion in section 5.

## II. THE APPROACH

Before explaining the multi-pass decoder in detail, we first outline some relevant fundamentals of a statistical speech recognition system and the architecture of a common single-pass decoder (see also [8]).

### A. Speech Recognition Background

The input of a speech recogniser is a complex series of changes in air pressure, which through sampling and quantisation can be digitalised to a pulse-code-modulated audio stream. From an audio stream the features or the characteristics of specific phones can be extracted. A statistical speech recogniser, which uses a Hidden Markov Model (HMM), can determine the likelihoods of those acoustic observations.

With a finite grammar or a statistical language model, a search space can be constructed, which consists of HMMs determined by the acoustic model. Both, grammar and language model, are based on a dictionary, defining which sequence of phones constitute which words. The grammar defines a state automaton of predefined transitions between words, including the transition probabilities. Language models in contrast are trained statistically based on the measured frequency of a word preceding another word. With so-called  $N$ -grams, dependencies between a word and the  $(N - 1)$  preceding words can be determined. Since  $N$ -grams of higher order need substantially more training data Bi-Grams or Tri-grams are often used in current *automatic speech recognition* (ASR) systems.

During processing of an utterance, a statistical speech recogniser searches the generated graph for the best fitting hypothesis. In every time frame, the possible hypotheses are scored. With a best-first search, or a specialised search algorithm like the Viterbi Algorithm, hypotheses with bad scores are pruned.

In principle it is possible to adapt ASR for improving the recognition rate with two different approaches:

- 1) The acoustic model is trained for a single specific speaker. This method leads to precise HMM's for phones, which allows for a larger vocabulary.
- 2) The domain is restricted in terms of a limited vocabulary. This restricted approach reaches good recognition rates even with an acoustical model trained for many different speakers.

### B. Multi-Pass Decoder

Both introduced methods, the *finite state grammar* (FSG) based decoder as well as the Tri-gram decoder, have specific advantages and limitations.

- The FSG decoder can be very strict, allowing valid sentences without fillers only. Unfortunately, such an FSG decoder maps every input to a path in the search space, which is spanned from all valid starting words to all valid finishing words. For example if the speaker is using a sentence like NAO \*EHM\* PICK PHONE, the decoder may map it to a most likely sentence like NAO WHERE IS PHONE. Even if the speaker is just randomly putting words together, the decoder may often produce a valid sentence and therefore – very often – a false positive.
- With a Tri-Gram decoder an ASR system is more flexible and can get decent results if the quality of the audio signal is high and the data set for training the language model is sufficiently large. However, since Tri-grams mainly take into account the last two most probable words, they cannot deal with long-range dependencies. Therefore even if the word accuracy is reasonably high, the sentence accuracy as a cumulative product is fairly moderate [8].

To overcome the limitations of both single decoders, we can combine them to a multi-pass decoder. First, we use the FSG decoder to produce the most likely hypothesis. Second, we use the Tri-gram decoder – which is able to backoff to Bi-grams or Uni-grams – to produce a reasonably large list of best hypotheses. Even if the best hypothesis of the Tri-gram decoder is not appropriate there is a good chance that one of the similar sentences is. In the next step, we compare the best hypothesis of the FSG decoder with the list of  $n$ -best hypotheses of the Tri-gram decoder. If we find a match we can accept this sentence, otherwise we reject the sentence. Figure 1 illustrates the HMM-based ASR system using the multi-pass decoder.

### C. Speech Recogniser and its Adaptation

In this study, we use the ASR framework *Pocketsphinx*, because it is open source and has been ported and optimised for hand-held devices [9]. In comparison to other promising systems [10], [11] it provides the advantage of being an effective research tool on the one hand and being applicable to devices and robots with moderate computing power on the other hand. *Pocketsphinx* comes with a speaker-independent acoustic-model 'HUB4' based on English broadcast news. Also available is a language model trained on the same data.

Since it is our aim to keep the system speaker independent, we decided to limit the vocabulary and to reduce the format

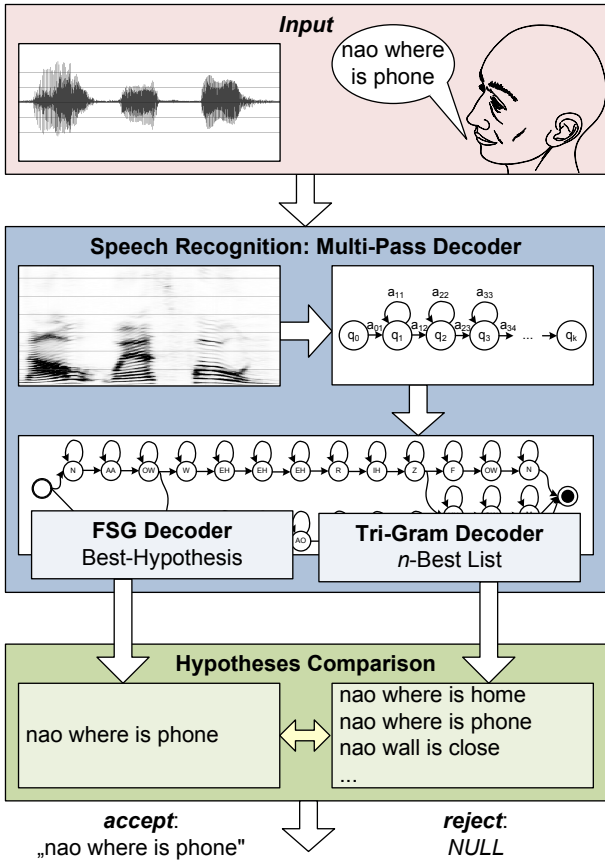


Fig. 1. Architecture of a multi-pass decoder

of a sentence to a simpler situated grammar or command grammar, as it can be useful in HRI. Devices and robots in our AmIE are supposed to be used for a specific set of tasks, while the scenario can have different human interactors. The acoustic-model HUB4 was trained over a very large set of data (140 hours) including different English speakers [12]. With a vocabulary reduction to 100 words and the new grammar, as outlined in figure 2, we generated an own FSG automaton on the one hand and trained an own language model on the other hand. For the training of the language model, we used the complete set of sentences which can be produced with our grammar. The grammar allows for short answers like YES or INCORRECT as well as for more complex descriptions of the environment like NAO BANANA HAS COLOR YELLOW.

In summary we adapted Pocketsphinx to recognise instruction, information, and question sentences in English.

### III. OUR SCENARIO

The scenario of this study is an ambient intelligent home environment. To investigate opportunities and chances of technical devices and humanoid robots in home environments, those scenarios are of increasing relevance [13], [14]. In particular EU research projects like KSERA aim to develop a socially assistive robot that helps elderly people [15]. Such a scenario consists of a home environment including interactive devices and a humanoid robot.

```

public <utterance> = <confirmation> | (nao <communication>);

<communication> = <information> | <instruction> | <question>;
<instruction> = <command> | <action>;
<information> = ((<object> | <agent>) close to (<object>
| <agent> | <place>))
| (<object> can be <affordance>)
| (<object> has color <color>);
<question> = (what can <object>)
| (which color has <object>)
| (where is (<object> | <agent>));
<confirmation> = yes | correct | right | (well done) | no
| wrong | incorrect;
<command> = abort | help | reset | (shut down) | stop;
<action> = <head_action> | <hand_action> | <body_action>;
<hand_action> = (<affordance> <object>)
| (show (<object> | <agent>));
<body_action> = (turn body <direction>) | (sit down)
| (walk <number>) | (bring <object>)
| (go to (<agent> | <object>)) | (come here);
<head_action> = (turn head <direction>)
| ((find | look at) (<object> | <agent>))
| (follow <agent>);
<agent> = nao | i | patient;
<object> = apple | banana | ball
| dice | phone | oximeter;
<direction> = left | straight | right;
<number> = one | two | three;
<affordance> = pick | drop | push;
<color> = yellow | orange | red | purple | blue | green;
<place> = home | desk | sofa | chair | floor | wall;

```

Fig. 2. Grammar for the scenario

#### A. Environment

Our AmIE is a lab room of 7x4 meters, which is furnished like a standard home without specific equipment to reduce noise or echoes, and is equipped with technical devices like a ceiling boundary microphone and a NAO H25 humanoid robot. A human user is supposed to interact with the environment and the NAO robot and therefore should be able to communicate in natural language. For this study the human user is wearing a headset as a reference microphone. The scenario is presented in detail in figure 3. The details of the used microphones are as follows:

a) *Ceiling Microphone*: The ceiling boundary microphone is a condenser microphone of 85 mm width, placed three meter above the ground. It is using an omni-directional polar pattern and has a frequency response of 30Hz - 18kHz.

b) *NAO*: The NAO robot is a 58 cm tall robot with 25 degrees of freedom (DOF), two VGA cameras, and four microphones, developed for academic purposes [16]. Besides his physical robustness, the robot provides some basic integrated functionalities like an initial set of prepared movements, a detection system for visual markers, and a text-to-speech module. Controllable over WLAN with a mounted C++ API namely NaoQi, the NAO can be used as a completely autonomously agent or as a remotely controlled machine. The microphones are placed around the head and have an electrical bandpass of 300Hz - 8kHz. In its current version the NAO uses a basic noise reduction technique to improve the quality of processed sounds.

c) *Headset*: The used headset is a mid-segment headset specialised for communication. The frequency response of the microphone is between 100Hz - 10kHz.

To allow reliable comparison, the location of the speaker is at a distance of 2m meter to the ceiling microphone as well as to the NAO robot.



Fig. 3. Scenario environment

### B. Dataset

The set of data to test the approach was collected under natural conditions within our AmIE. Different non-native English mixed male and female test subjects were asked to read a random sentence, produced from our grammar. All sentences were recorded in parallel with the headset, the ceiling microphone and the NAO robot in a 16 bit format and a sample rate of 48.000 Hz. In summary we collected 592 recorded sentences each, which led to 1776 audio files.

### C. Evaluation Method

For the empirical validation, we converted all files to the monaural, little-endian, unheadered 16-bit signed PCM audio format sampled at 16000 Hz, which is the standard audio input stream for Pocketsphinx.

With Pocketsphinx we run a speech recognition test on every recorded sentence. Since it is not the focus of this study to test for false negatives and true negatives, we did not include incorrect sentences or empty recordings in the test. The result of the speech recogniser was compared with the whole desired sentence to check for the sentence accuracy as means of comparability. If the sentence was completely correct

it was counted as true positive, otherwise a false positive. For example if the correct sentence is NAO WHAT COLOR HAS BALL, then NAO WHAT COLOR HAS WALL as well as NAO WHAT COLOR IS BALL are incorrect.

To test for statistical significance of the false positive reduction with the multi-pass decoder, we calculated the *chi-square* ( $\chi^2$ ) score over the true-positives/false-positives ratios. If, for example, the  $\chi^2$  score over the tp/fp ratio of the multi-pass against the tp/fp ratio of the FSG decoder is very high, then we have evidence for a high degree of dissimilarity [17].

## IV. EMPIRICAL RESULTS

The empirical investigation of our approach consists of two parts. First, we analysed the overall rate of true and false positives of the multi-pass decoder in comparison to specific single-pass decoders. Second, we determined the influence of the size  $n$  of the list of best hypotheses. Every investigation has been carried out in parallel for every microphone type as described above.

### A. Effect of Different Decoders

With the 592 recorded sentences we tested the speech recognition using the FSG-decoder and the Tri-gram decoder in a single-pass fashion and combined them in a multi-pass fashion, using  $n$ -best list size of 10. In table I the results are presented where every row contains the number of correctly recognised sentences (true positives) and incorrectly recognised sentences (false positives).

TABLE I  
COMPARISON OF DIFFERENT DECODERS

(a) FSG decoder			
	True positives	False positives	Tp/fp ratio
Headset	458 (77.4%)	101 (17.1%)	81.93%
Ceiling mic.	251 (42.4%)	251 (50.3%)	45.72%
NAO robot	39 (6.6%)	447 (75.5%)	8.02%

(b) Tri-gram decoder			
	True positives	False positives	Tp/fp ratio
Headset	380 (64.2%)	212 (35.8%)	64.19%
Ceiling mic.	133 (22.5%)	459 (77.5%)	22.47%
NAO robot	14 (2.4%)	322 (54.4%)	4.17%

(c) Multi-pass decoder, $n = 10$			
	True positives	False positives	Tp/fp ratio
Headset	378 (63.9%)	24 (4.1%)	94.03%
Ceiling mic.	160 (27.0%)	76 (12.8%)	67.80%
NAO robot	31 (5.2%)	130 (22.0%)	19.25%

$$\text{tp/fp ratio} = \text{tp} / (\text{tp} + \text{fp}) * 100$$

The data shows that for a headset every decoder led to a relatively high rate of correct sentences, counting 458 (77.4%) with the FSG, 380 (64.2%) with the Tri-gram, and 378 (63.9%) with the multi-pass decoder. The single-pass decoder produced 101 false positives (tp/fp ratio of 81.93%) with FSG and 212 false positives (tp/fp ratio of 64.19%) with Tri-gram, while the multi-pass decoder produced 24 false positives (tp/fp ratio of 94.03%).

For the ceiling microphone the rate of correct sentences was fairly moderate, reaching 251 (42.4%) with the FSG, 133 (22.5%) with the Tri-gram, and 160 (27.0%) with the multi-pass decoder. The number of produced false positives was relatively high for the single-pass decoder reaching 298 (tp/fp ratio of 45.72%) with FSG and 459 false positives (tp/fp ratio of 22.47%) with Tri-gram, whereas the multi-pass decoder produced 76 false positives (tp/fp ratio of 67.80%).

The rate of correct sentences for the NAO robot microphones was very low, getting only 39 (6.6%) with the FSG, 14 (2.4%) with the Tri-gram, and 31 (5.2%) with the multi-pass decoder. However, the single-pass decoder produced 447 false positives (tp/fp ratio of 8.02%) with the FSG and 322 false positives (tp/fp ratio of 4.17%) with the Tri-gram, while the multi-pass decoder produced 130 false positives (tp/fp ratio of 19.25%).

In table II some examples for the the recognition results with different decoder and microphones are presented. The results indicate that in many cases where sentences could not be recognised correctly, some specific single words like APPLE were recognised incorrectly. In some cases valid but incorrect sentences were recognised by both decoders, but were successfully rejected by the multi-pass decoder. Furthermore, with the NAO robot often only single words were recognised.

TABLE II  
EXAMPLES OF RECOGNISED SENTENCES

	True positive	Rejected	False positive
(a) "NAO GO TO OXIMETER"			
	FSG decoder	Tri-gram dec.	Multi-pass dec.
Headset	NAO GO TO OXIMETER	NAO WHAT COLOR OXIMETER	NAO GO TO OXIMETER
Ceiling mic.	NAO SIT DOWN	NAO SIT DOWN	NAO SIT DOWN
NAO robot	NAO GO TO OXIMETER	NAO BE	
(b) "NAO APPLE CLOSE TO PATIENT"			
	FSG decoder	Tri-gram dec.	Multi-pass dec.
Headset		NAO APPLE HAS CLOSE TO PATIENT	
Ceiling mic.	NAO I CLOSE TO PATIENT	NAO HEAD CLOSE TO PATIENT	
NAO robot	NAO PATIENT FIND	NAO TO PATIENT	
(c) "NAO WHICH COLOR HAS BALL"			
	FSG decoder	Tri-gram dec.	Multi-pass dec.
Headset	NAO WHICH COLOR HAS BALL	NAO WHICH COLOR HAS BALL	NAO WHICH COLOR HAS BALL
Ceiling mic.	NAO WHERE IS PHONE	NAO WHERE IS HEAD AT PHONE	
NAO robot	NO		
(d) "WELL DONE"			
	FSG decoder	Tri-gram dec.	Multi-pass dec.
Headset	WELL DONE	WELL DONE	WELL DONE
Ceiling mic.	WELL DONE	WELL DONE	WELL DONE
NAO robot	YES		

## B. Influence of Parameter $n$

To determine the influence of the size of the  $n$ -best list, we varied  $n$  over  $\{1, 2, 5, 10, 20, 50, 100\}$ . Figure 4 displays the ratio of true positives and false positives in comparison to the rate of correctly recognised sentences for every microphone type as described above.

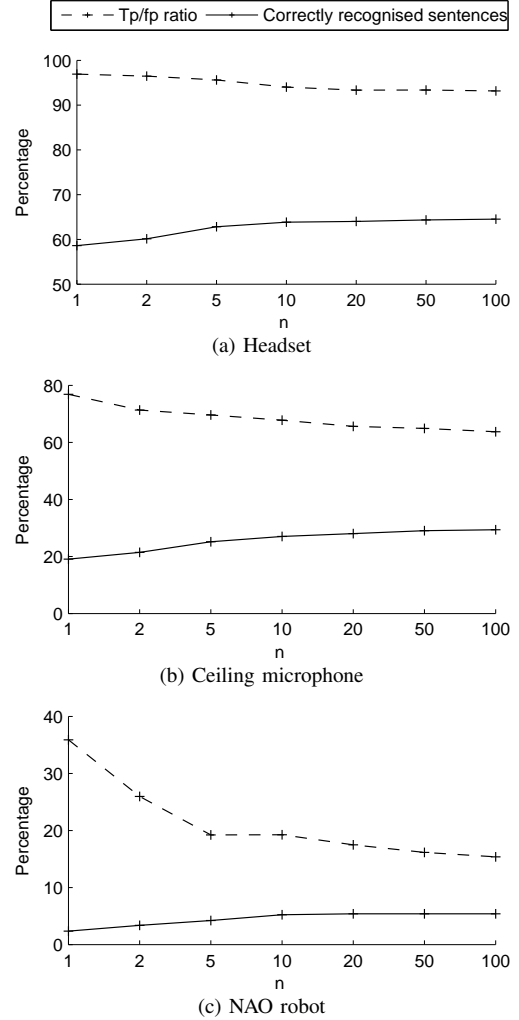


Fig. 4. Comparison of true positives/false positives ratio and correctly recognised sentences

On the one hand, for small  $n$  the percentage of false positives is smaller for every microphone type. On the other hand, a small  $n$  results in a more frequent rejection of sentences.

Finding an optimal  $n$  seems to strongly depend on the microphone used and therefore on the expected quality of the speech signals. In our scenario a larger  $n$  around 20 is sufficient for the use of headsets, in terms of getting a good true positives to false positives ratio while not rejecting too many good candidates. For a moderate microphone like the ceiling microphone, a smaller  $n$  around 5 is sufficient. With low-quality microphones like in the NAO robot the variance of  $n$  does not point to an optimal configuration. Smaller  $n$  result in very few correctly recognised sentences, while larger  $n$  result in a very low tp/fs rate.

### C. Result Summary

In summary, we observed that using a multi-pass decoder reduced the number of produced false positives significantly. For a low-noise headset as well as for boundary microphones and inexpensive microphones installed on a mobile robot, the experiment has shown that reducing the false positives to a good degree does not lead to a substantial reduction of true positives. The overall recognition rates with the NAO were insufficient, while the ceiling microphone worked with a reasonable rate using the multi-pass decoder. A good value for  $n$  depends on the hypotheses space and the microphone used. For our scenario, overall using  $n = 10$  best hypotheses was sufficient. If the expected quality is moderate and the number of different words and possible sentences are high, then a larger value for  $n$  is likely to lead to better results.

### V. CONCLUSION

In this paper we presented a study of speech recognition using a multi-pass FSG and Tri-gram decoder comparing a ceiling microphone and the microphones of a humanoid robot with a standard headset. The results of our approach are in line with [6], showing that a multi-pass decoder can successfully be used to reduce false positives and to obtain robust speech recognition. Furthermore we can state that using a multi-pass decoder in combination with a ceiling boundary microphone is useful for HRI: Adapting to domain-specific vocabulary and grammar on the one hand and combining the advantages of an FSG and a Tri-gram decoder leads to acceptable speech recognition rates. The size of the  $n$ -best list is not very crucial and depends on the search space to some extent. Build-in microphones of humanoid robots such as the NAO still come with a low SRN due to noisy fans or motors, and need intensive preprocessing to allow for speech recognition.

In the future the proposed method can be improved in various ways. First, one could improve the quality of the speech recorded by a (ceiling) microphone itself. Using for example a sophisticated noise filter or integrating a large number of microphones could lead to a more reliable result [18]. Second, one could not only integrate different decoding methods but also the context information into one ASR system to accept or reject recognised utterances. For example vision could provide information about lip movement and therefore provide probabilities for silence or a specific phoneme [19]. Speech recognition serves as a starting ground for research in HRI and CNR and as a driving force for a better understanding of language itself. In this context we have shown that using a multi-pass decoder and environmental microphones is a viable approach.

### ACKNOWLEDGMENT

The authors would like to thank Arne Köhn, Carolin Mönter, and Sebastian Schneegans for the support in automatically collecting a large set of data. We also thank our collaborating partners of the KSERA project funded by the European Commission under n° 2010-248085 and of the RobotDoC project funded by Marie Curie ITN under 235065.

### REFERENCES

- [1] T. Kanda, M. Shiomi, Z. Miyashita, H. Ishiguro, and N. Hagita, "A communication robot in a shopping mall," *IEEE Robotics and Automation Society*, vol. 26, no. 5, pp. 897–913, 2010.
- [2] K. K. Paliwal and K. Yao, "Robust speech recognition under noisy ambient conditions," in *Human-Centric Interfaces for Ambient Intelligence*. Academic Press, Elsevier, 2009, ch. 6.
- [3] S. Wermter, M. Page, M. Knowles, V. Gallesse, F. Pulvermüller, and J. G. Taylor, "Multimodal communication in animals, humans and robots: An introduction to perspectives in brain-inspired informatics," *Neural Networks*, vol. 22, no. 2, pp. 111–115, 2009.
- [4] Q. Lin, D. Lubensky, M. Picheny, and P. S. Rao, "Key-phrase spotting using an integrated language model of n-grams and finite-state grammar," in *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*. Rhodes, Greece: ISCA Archive, Sep. 1997, pp. 255–258.
- [5] M. Levit, S. Chang, and B. Buntschuh, "Garbage modeling with decoys for a sequential recognition scenario," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU 2009)*. Merano, Italy: IEEE Xplore, Dec. 2009, pp. 468–473.
- [6] M. Doostdar, S. Schiffer, and G. Lakemeyer, "Robust speech recognition for service robotics applications," in *Proceedings of the Int. RoboCup Symposium 2008 (RoboCup 2008)*, ser. Lecture Notes in Computer Science, vol. 5399. Suzhou, China: Springer, Jul. 2008, pp. 1–12.
- [7] Y. Sasaki, S. Kagami, H. Mizoguchi, and T. Enomoto, "A predefined command recognition system using a ceiling microphone array in noisy housing environments," in *Proceedings of the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2008)*. Nice, France: IEEE Xplore, Sep. 2008, pp. 2178–2184.
- [8] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd ed. Prentice Hall, 2009.
- [9] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. (ICASSP 2006)*. Toulouse, France: IEEE Xplore, May 2006.
- [10] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius," in *Proceedings of the 2009 APSIPA Annual Summit and Conference (APSIPA ASC 2009)*. Sapporo, Japan: APSIPA, Oct. 2009, pp. 131–137.
- [11] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Löff, R. Schlüter, and H. Ney, "The RWTH Aachen University open source speech recognition system," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, Brighton, U.K., Sep. 2009, pp. 2111–2114.
- [12] J. Fiscus, J. Garofolo, M. Przybocki, W. Fisher, and D. Pallett, "English broadcast news speech (HUB4)," Linguistic Data Consortium, Philadelphia, 1997.
- [13] S. Wermter, G. Palm, and M. Elshaw, *Biomimetic Neural Learning for Intelligent Robots*. Springer, Heidelberg, 2005.
- [14] H. Nakashima, H. Aghajan, and J. C. Augusto, *Handbook of Ambient Intelligence and Smart Environments*. Springer Publishing Company, Incorporated, 2009.
- [15] D. van der Pol, J. Juola, L. Meesters, C. Weber, A. Yan, and S. Wermter, "Knowledgeable service robots for aging: Human robot interaction," KSERA consortium, Deliverable D3.1, October 2010.
- [16] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, "The NAO humanoid: A combination of performance and affordability," *CoRR*, 2008. [Online]. Available: <http://arxiv.org/abs/0807.3223>
- [17] C. D. Manning and H. Schuetze, *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.
- [18] H. Nakajima, K. Kikuchi, T. Daigo, Y. Kaneda, K. Nakadai, and Y. Hasegawa, "Real-time sound source orientation estimation using a 96 channel microphone array," in *Proceedings of the 2009 IEEE/RSJ Int. Conference on Intelligent Robots and Systems (IROS 2009)*. St. Louis, USA: IEEE Xplore, October 11-15 2009, pp. 676–683.
- [19] T. Yoshida, K. Nakadai, and H. G. Okuno, "Two-layered audio-visual speech recognition for robots in noisy environments," in *Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010)*. Taipei, Taiwan: IEEE Xplore, October 18-22 2010, pp. 988–993.