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. Gallese, 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. Lööf, 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.