Active Hearing and Automatic Speech Recognition for the MiRo Robot

Saeid Mokaram[1], Samuel Fernando[1,2,3], Hamideh Kerdegari[3],
Heidi Christensen[1,2,3], Jon Barker[1,2,3], Tony J. Prescott[1,2,3]

I. Introduction

The performance of Automatic Speech Recognition (ASR) systems can drop with even moderate levels of background noise such as when multiple speakers talk simultaneously. Human listeners appear to have little difficulty in such environments and can follow one voice amongst a mixture. Various factors have been shown to contribute to this ability, including the use of dynamic head movements. This active strategy to improve hearing has been extensively investigated for human sound source localization and separation in multi-source, reverberated environments [1], and has stimulated developments in the area of *Robot Audition* [2]. The purpose of our research is to investigate the benefits of augmenting a mobile robot platform with an active hearing system to improve its speech recognition and perception abilities. This paper presents a milestone in our research on providing tools for developing active hearing on the MiRo robot platform.

II. The MiRo robot

The MiRo robot is a programmable mobile developer platform for companion and social robotics that resembles a pet animal whilst being clearly a robot [3]. Developed by Consequential Robotics [1], MiRo has a unique biomimetic design. It has two physically directable ears that have the potential to deliver an animal-like ability to localize and track sound sources in an active manner and therefore support speech recognition capabilities.

III. Modelling Robot-Specific Acoustic Factors

Building an active hearing system for a robot requires the understanding of two sets of robot-specific factors: the spatial filtering properties of the robots ears and the robot's self-noise. The spatial filtering properties of the MiRo robot was evaluated by placing the robot in an anechoic chamber and measuring the responses of the microphones to carefully controlled sounds played from a grid of radial directions. This provided a set of impulse response recordings and the process was repeated for a range of orientations of the robots ears. The robot self-noise was also recorded as heard through MiRos own ears, enabling us to explore noise cancellation techniques in order to mitigate the effect of motors being placed close to the robots microphones.

We aim to develop a MiRo hearing simulator to be added to the existing MiRo motion simulator. This will enable the evaluation of hearing algorithms at an early stage in a simulated environment. It will also enable the filtering of standard speech and noise datasets, so they appear as if they have been recorded on the robot reflecting a variety of directions and environments. Such data augmentation is standard practice for creating the needed large multi-conditional datasets for training of Deep Neural Network (DNN) based systems in state-of-the-art ASR.

IV. ASR for the MiRo robot

A baseline ASR system was developed on the robot platform using recordings from the robot's microphones. A subset of *The Wall Street Journal (WSJ)* read speech corpus [4] (a *de facto* benchmark in ASR research) was used as training data. As well as the clean data, separate training sets were created which augmented the data with speed perturbations of 0.9 and 1.1, and also added background noise from the *CHIME* speech corpus [5]. The training portion of this data was used to train acoustic models for ASR using the Kaldi toolkit.

We also tested on recordings of live speech made by eight speakers each saying short phrases such as 'Hello MiRo', 'Stay There' and 'Go to sleep'. For the WSJ test, an unrestricted tri-gram model was used for decoding, but for the live speech a smaller phrase-based grammar model was deployed. Using a DNN approach, we obtained word error rates of 5.04% for the baseline WSJ system and of 8.87% for the short phrases, when using multi-conditional training incorporating data augmented with speed and noise. For future work, we will explore the use of active hearing strategies for improving MiRo's ASR in challenging environments.

References

[1] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," *Annu. Rev. Psychol.*, vol. 42, no. 1, pp. 135–159, 1991.

[2] S. Argentieri, A. Portello, M. Bernard, P. Danes, and B. Gas, "Binaural systems in robotics," in *The technology of binaural listening*. Springer, 2013, pp. 225–253.

[3] B. Mitchinson and T. J. Prescott, "Miro: A robot mammal with a biomimetic brain-based control system," in *Conference on Biomimetic and Biohybrid Systems*. Springer, 2016, pp. 179–191.

[4] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSRI (WSJ0) Complete," Linguistic Data Consortium, Philadelphia, Tech. Rep., 2007.

[5] H. Christensen, J. Barker, N. Ma, and P. D. Green, "The chime corpus: a resource and a challenge for computational hearing in multisource environments." in *Interspeech*, 2010.

[1]Department of Computer Science, University of Sheffield, 211 Portobello, Sheffield S1 4DP, U.K. {s.mokaram, s.fernando, h.kerdegari, j.p.barker, heidi.christensen, t.prescott}@sheffield.ac.uk

[2]Centre for Assistive Technology and Connected Healthcare (CATCH), 217 Portobello, Sheffield S1 4DP, U.K.

[3]Sheffield Robotics, Mappin Street , Sheffield, S1 3JD, U.K.

[1]http://consequentialrobotics.com/miro/