EARS@HOME
Team Description Paper for RoCKIn@HOME Challenge 2015

Heinrich W. Löllmann[1], Hendrik Barfuss[1], Alastair Moore[2], Vladimir Tourbabin[3], Patrick Naylor[2], Boaz Rafaely[3], Christine Evers[2], Antoine Deleforge[1], Saša Bodiroža[4], Guido Schillaci[4], Claas-Norman Ritter[4], Gregory Rump[6], Verena Hafner[4], Fabien Badeig[5], Radu Horaud[5], Rodolphe Gelin[6], and Walter Kellermann[1]

[1] Multimedia Communications and Signal Processing, Friedrich-Alexander University Erlangen-Nürnberg, Germany

[2] Dept. of Electrical and Electronic Engineering, Imperial College London, UK

[3] Dept. of Electrical and Computational Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel

[4] Institut für Informatik, Humboldt-Universität zu Berlin, Germany

[5] INRIA Grenoble Rhône-Alpes, France

[6] Aldebaran Robotics SA, Paris, France

Website: `http://robot-ears.eu`

***Abstract*—The EARS@HOME team consists of members of the research project Embodied Audition for RobotS (EARS) and aims to take part at the FBM3 "Speech Understanding" of the RoCKIn@HOME Challenge 2015. For this, the NAO H25 robot of the manufacturer Aldebaran Robotics SA with a new prototype head having 12 microphones will be employed. This paper provides a description of the team, the robot platform and the algorithms that should be employed for the challenge.**

I. Introduction

The EARS@HOME team[1] consists of members of the EU-funded research project Embodied Audition for RobotS (EARS). This project explores new algorithms for enhancing the auditory capabilities of humanoid robots. The main focus of this project is on the development of algorithms which facilitate a natural Human-Robot-Interaction (HRI) in adverse acoustical environments as found in typical real-world application scenarios (e.g., a hotel lobby). The Functionality Benchmark *Speech Understanding* of the RoCKIn@HOME challenge offers an opportunity to benchmark a new robot prototype system developed within the EARS project, which motivates the application of the EARS team for this challenge.

The remainder of this Team Description Paper (TDP) is structured as follows: In Sec. II, the objectives of the EARS project are outlined (as far as they are related to the challenge). In Sec. III and Sec. IV, the robot platform and new head array design that should be used for the challenge are described. The software and computing platform is outlined in Sec. V, and Sec. VI includes a description of the signal enhancement algorithms as well as the ASR system that should be employed for the challenge.

II. Objectives of the EARS project

Achieving a HRI that is perceived as natural and intuitive to the human will critically depend on how responsive the robot will be to all forms of human expressions and how well it will be aware of its environment.

With acoustic signals distinctively characterizing physical environments and speech being one of the most effective means of communication among humans, humanoid robots must be able to extract fully the rich auditory information from their environment and to use voice communication as much as humans do. While vision-based HRI is well developed, current limitations in robot audition do not allow for such an effective and natural acoustic human-robot communication in real-world environments. This is mainly because of the severe degradation of the signals captured by the robot's microphones due to noise, interference and reverberation.

To overcome these limitations, the EARS project aims at providing intelligent 'ears' with high performance and use it for HRI in complex real-world environments. Novel microphone arrays and powerful signal processing algorithms will be designed to localize and track multiple sound sources of interest and to extract and recognize the desired signals. After fusion with robot vision, embodied robot cognition will then derive HRI actions and provide knowledge of the entire scenario, and feed this back to the acoustic interface for further auditory scene analysis. As a prototypical application, EARS will consider a welcoming robot in a hotel lobby experiencing the above challenges. Representing a large class of generic applications, this scenario is of key interest to industry. The robot manufacturer Aldebaran Robotics SA will integrate the results achieved by EARS into a robot platform for the consumer market and validate the performance.

III. Robot Platform

The platform of the EARS project is the commercial humanoid robot NAO [1] manufactured by Aldebaran Robotics SA (Paris, France) as shown in Fig. 1. This robot has a height of $57.4\,\mathrm{cm}$ and a weight of $5.4\,\mathrm{kg}$.

[1]Team leader: Heinrich W. Löllmann, email: `loellmann@lnt.de`

Fig. 1. Humanoid robot NAO used as platform for the RoCKIn@HOME Challenge.

The current version of the NAO robot utilizes an array with four hypercardioid microphones mounted on the robot head. Two cameras are also installed which can stream videos at a resolution of 1280×960 pixels at a frame rate of 30 frames per second.

IV. New microphone array design

The design of the microphone array greatly affects the performance of any spatial filtering algorithm employed. The microphone positions should be chosen such that the array captures as much information as possible from the sound field. Within the project, a new prototype head, in which 12 omnidirectional microphones are integrated, has been developed and manufactured. The new microphone array facilitates the application of spatial filtering algorithms in the spherical harmonics domain, see, e.g., [2], [3], which are computationally efficient and powerful.

The optimal microphone positions were determined by minimizing the spatial aliasing level for an array design in the spherical harmonics domain. Regions where microphones cannot be mounted due to mechanical constraints were excluded in the design process. The optimal 12 microphone positions were finally chosen from a set of 327 possible microphone placements.

The microphones are attached to the surface of the head from the inside using a custom-designed plastic holder assembly. The microphones are placed underneath the head surface and the conduit that connects the surface of the head and each of the microphones has a diameter of approximately 1 mm and a length of approximately 3 mm. Due to its short length relative to the wavelength of sound even at high frequencies, no significant scattering effect due to the conduit is expected on the signals picked up by the microphones in the frequency range of interest. The complete head assembly with all internal parts and the 12 microphones is illustrated in Fig. 2.

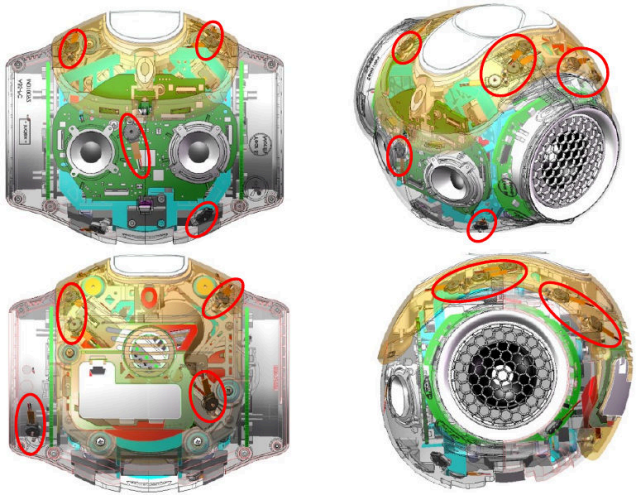

Fig. 2. Prototype head for the NAO robot with 12 integrated microphones.

Each microphone of the head is attached to a pre-amplifier which outputs standard line level. The output circuitry is designed to provide a differential and balanced transmission line and is terminated with a $1/4$" Tip Ring Sleeve (TRS) connector. As shown in Fig. 3, the microphone outputs are connected to an external standard Analogue-to-Digital Converter (ADC) like the RME, M-32AD with 32 input channels. The prototype head will be mounted on the NAO robot for the challenge, but since the new robot head is a prototype, the microphone cables are connected with an external ADC and laptop computer for the signal processing.

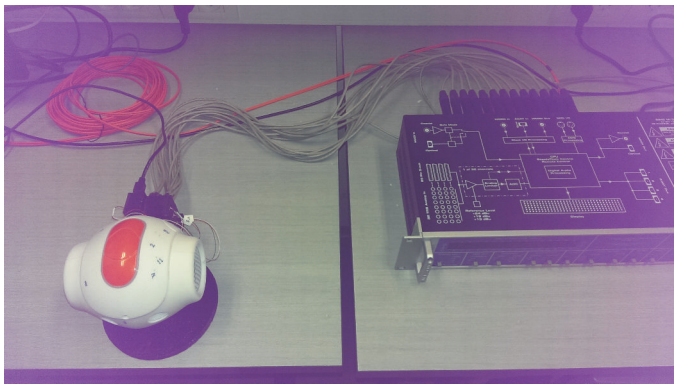

Fig. 3. New prototype head with 12 mics to be used by the EARS@HOME team. The analog microphone signals are fed into an external multichannel A/D converter.

V. Signal Processing and Computing Platform

The current NAO configuration has two drawbacks. Firstly, the on-board computing resources of the commercial version of the NAO robot consist of a 1.6 GHz ATOM Z530 CPU,

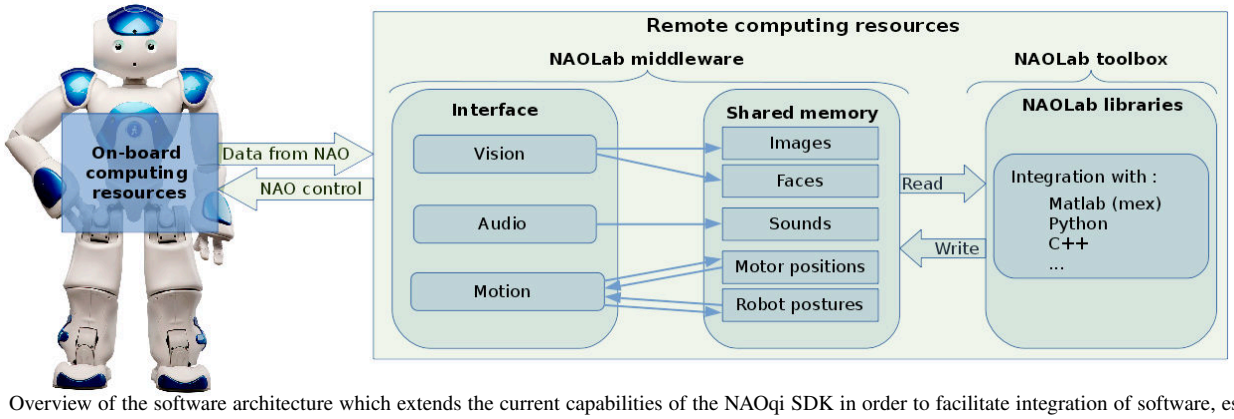

Fig. 4. Overview of the software architecture which extends the current capabilities of the NAOqi SDK in order to facilitate integration of software, especially in Matlab.

1 GB of RAM. These resources are not sufficient to implement sophisticated audio and video processing algorithms. Secondly, the current signal processing software is based on the NAOqi middleware, thus requiring sufficient expertise with the restrictive environment of the NAOqi Software Development Kit (SDK). Consequently, it is complicated to carry out thorough experimental validations and design proof-of-concept demonstrators which can dissuade the HRI practitioners.

The proposed solution is to combine a new middleware, named NAOLab middleware, with a functional toolbox, named NAOLab toolbox, to remotely access and control the robot NAO. The NAOLab toolbox provides two main features. Firstly, the NAOLab middleware complexity is assessable for the user. Secondly, a user-friendly interface is provided through C++ and Python libraries extended with mex functions for Matlab. This solution allows to deploy sophisticated audio and video processing algorithms without the constraints imposed by the NAOqi SDK, and NAO on-board computing resources are augmented with external computing resources.

An overview of the proposed architecture is shown in Fig. 4. The same modular approach as NAOqi is kept where the NAO capabilities are divided into three groups: vision, audio and motion. For each NAOqi module, an interface (vision, audio, motion) is associated and each interface needs to deal with the sensor access and the control of the actuators. The role of these interfaces is twofold: (i) to feed the sensor data into a memory space that is subsequently shared with existing software or with software under development, and (ii) to send back to the robot NAO commands that are generated by the external software. The NAOLab toolbox provides C++ and Python libraries (see Table I), i.e., a set of functions to interact with NAO. To use the famous signal processing toolboxes of Matlab, a mex library is also available.

NAOLab C++ library	Vision	Audio	Motion
Access:	getImage() getImages() getFaces	getSound()	getMotorInfo getAllMotorInfo() getPosture()
Control:	setResolution() setCameras()	textToSpeech() playAudioFile()	moveMotor() setPosture() moveHead() moveToPoint()

TABLE I
NAOLAB C++ LIBRARY.

VI. SIGNAL ENHANCEMENT AND ASR

The planned signal processing pipeline for the challenge is illustrated in Fig. 5.

As a first step, the desired source (e.g., the loudspeaker in the case of the challenge) needs to be localized to determine the look direction for the multichannel speech enhancement algorithm. For the RoCKIn@HOME Challenge, two classes of algorithms are considered for this. The first class, GCC-PHAT [4], employs a weighted cross-correlation function of pairs of microphones to estimate the Time Difference of Arrival (TDOA) of a wavefront at the two microphones. This will be used for the first phase of the evaluation (stereo audio files provided by OC). The second class, a pseudo-intensity (PIV) based method [5]–[7], is targeted at the novel microphone array with 12 mics developed by the EARS consortium. It relies on a spherical harmonics-domain representation of the sound field and provides the estimated direction of arrival in azimuth and elevation. The approach to localize the speaker (loudspeaker) within the challenge might be switched off, if the array beam is broad enough such that it is sufficient to steer the robot head manually towards the direction of the source.

In a second step, spatial filtering will be employed in order to extract the desired source from the mixture of the desired source and undesired noise and interference which is captured by the robot's microphones. As spatial filtering algorithm, the so-called Minimum Variance Distortionless Response (MVDR) beamformer should be used, realized in the spherical harmonics domain [3]. The output of the beamformer is processed by a single-channel postfilter which aims to suppress residual noise components which the spatial filtering algorithm could not suppress sufficiently (see, e.g., [8]). The postfilter will be realized as a Wiener filter [9]. Instead of the Signal-to-Noise Ratio (SNR), the Coherent-to-Diffuse Power Ratio (CDR), which is the ratio between direct and diffuse

Fig. 5. Illustration of the signal processing chain for the RoCKIn@HOME challenge.

signal components, will be used to calculate the postfilter weights as proposed in [10].

The enhanced signal is then fed into the Automatic Speech Recognition (ASR) system. A well-known commercial engine will be used for the ASR with a general-purpose acoustic model. A grammar-based language-model will be specifically built for the task in order to better fit the closed-vocabulary nature and the rigid linguistic structure of the benchmark. This structure being implicit, the aforementioned grammar requires to be clearly defined to cope with any instruction given to the robot. This results in a reduced language model perplexity and, by this, to improved recognition results. Following the same purpose of minimizing perplexity, the language model of the ASR and the model used for the Natural Language Understanding (NLU) that deals with the conversion to the Command Frame Representation (CFR) will be closely matched.

In order to map the recognized sentence into a required structure of the CFR, the sentence is parsed using a library for natural language processing (NLP). The output of the parser is a parse tree, which represents the syntactic structure of the sentence. During the tree traversal, each occurred verb phrase is mapped into one of the defined actions along with its associated arguments. The output of this parser is finally written on a memory stick in the format as required by the rules of the challenge.

ACKNOWLEGMENT

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 609465.

REFERENCES

[1] *NAO NEXT Gen H25 Datasheet*, Aldebaran Robotics, December 2011.

[2] J. Meyer and G. W. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*. IEEE, 2002, pp. 1781–1784.

[3] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, and E. Fisher, "Spherical microphone array beamforming," in *Speech Processing in Modern Communication*, ser. Springer Topics in Signal Processing, I. Cohen, J. Benesty, and S. Gannot, Eds. Springer Berlin Heidelberg, 2010, vol. 3, pp. 281–305.

[4] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[5] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, Aug. 2010, pp. 442–446.

[6] C. Evers, A. H. Moore, and P. A. Naylor, "Multiple source localisation in the spherical harmonic domain," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Nice, France, July 2014.

[7] A. H. Moore, C. Evers, P. A. Naylor, D. L. Alon, and B. Rafaely, "Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2015.

[8] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, ser. Digital Signal Processing. Springer Berlin Heidelberg, Jan. 2001, pp. 39–60.

[9] E. Haensler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*. Wiley-Interscience, 2004.

[10] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE Trans. Audio, Speech and Lang. Process. (ASLP)*, Apr. 2015.