Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 34 (2015) 129–153

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Robust speaker localization for real-world robots☆

Georgios Athanasopoulos[a,*], Werner Verhelst[a,b], Hichem Sahli[a,c]

[a] *Vrije Universiteit Brussel (VUB), Department of Electronics and Informatics (ETRO), Pleinlaan 2, 1050 Brussels, Belgium*
[b] *iMinds, Department of Future Media and Imaging, Gaston Crommenlaan 8, 9050 Ghent, Belgium*
[c] *Interuniversity Microelectronics Center (IMEC), Kapeldreef 75, 3001 Leuven, Belgium*

Received 23 May 2014; received in revised form 15 February 2015; accepted 25 March 2015
Available online 2 April 2015

Abstract

Autonomous human–robot interaction ultimately requires an artificial audition module that allows the robot to process and interpret a combination of verbal and non-verbal auditory inputs. A key component of such a module is the acoustic localization. The acoustic localization not only enables the robot to simultaneously localize multiple persons and auditory events of interest in the environment, but also provides input to auditory tasks such as speech enhancement and speech recognition. The use of microphone arrays in robots is an efficient and commonly applied approach to the localization problem. In this paper, moving away from simulated environments, we look at the acoustic localization under real-world conditions and limitations. Our approach proposes a series of enhancements, taking into account the imperfect frequency response of the array microphones and addressing the influence of the robot's shape and surface material. Motivated by the importance of the signal's phase information, we introduce a novel pre-processing step for enhancing the acoustic localization. Results show that the proposed approach improves the localization performance in joint noisy and reverberant conditions and allows a humanoid robot to locate multiple speakers in a real-world environment.

Keywords: Microphone arrays; Acoustic localization; Time delay estimation; Steered response power; Phase spectrum enhancement

1. Introduction

In human–robot Interaction (HRI) auditory information can contribute to resolve complex problems such as the focus of attention, activity recognition, etc. It has been observed that both adults and children do not perceive humanoid robots as a mechatronic device but attribute to them characteristics similar to those attributed to living organisms (Reeves and Nass, 1998). HRI is therefore expected to employ mechanisms similar to those of humans interacting with each other. This renders acoustic localization a fundamental element of the context aware HRI.

Acoustic localization is defined as the determination of the direction of each active sound source of interest in relation to a reference point, usually the robot itself. Over the years, various general purpose acoustic localization methods have been proposed (Brandstein and Ward, 2001). In robotics, acoustic localization methods range from simulating the

☆ This paper has been recommended for acceptance by R.K. Moore.

[*] Corresponding author. Tel.: +32 2 629 3955.
E-mail addresses: gathanas@etro.vub.ac.be (G. Athanasopoulos), wverhels@etro.vub.ac.be (W. Verhelst), hsahli@etro.vub.ac.be (H. Sahli).

http://dx.doi.org/10.1016/j.csl.2015.03.009
0885-2308/© 2015 Elsevier Ltd. All rights reserved.

Fig. 1. NAO, a real-world humanoid robotic platform interacting with children under the ALIZ-E project.

binaural cues of human hearing (Ferreira et al., 2009; Dávila-Chacón et al., 2012) and developing artificial ears (pinnae) (Hwang et al., 2011; Kumon and Noda, 2011) to utilizing several spatially separated microphones (microphone array). Due to their efficiency and robustness, microphone arrays are being increasingly used. Typical microphone array sizes for robots vary from two-four (small) to eight or more sensors (large).

Acoustic localization using a microphone array usually relies on time delay estimation (TDE) of the audio signal at the different microphones of the array. Different techniques for estimating the time delay can be applied (Mumolo et al., 2002; Trifa et al., 2007). The Generalized Cross-Correlation (GCC) (Knapp and Carter, 1976) is most commonly used due to its robustness and computational efficiency. TDE, however, is generally limited to a single dominant sound source. In scenarios where multiple overlapping sources need to be localized, steered-beamforming methods are employed (Valin et al., 2007; Breuer et al., 2012). Recent frameworks for localizing multiple sources have been proposed in Oualil et al. (2013) and Brutti and Nesta (2013).

As robotics technology advances, robots are expected to operate in various types of environments such as houses, schools, hospitals, etc. Fig. 1 shows an out of the lab HRI, conducted with the commercially available robot NAO (Aldebaran's, 2015) in the context of the ALIZ-E[1] project. These various environments exhibit different acoustic properties, such as the presence of ambient noise or reverberation. These properties are not always known in advance and may change dynamically. Therefore, the acoustic localization system of robots must be capable of coping with these conditions. Typically, the effect of noise and reverberation is addressed through spectral weighting of the signals coming from the microphones of the array (Abutalebi and Momenzadeh, 2011; Knapp and Carter, 1976; Rui and Florencio, 2004; Valin et al., 2007).

On the other hand, in real-world robots such as NAO, the sound wave is diffracted along the surface of the robot before reaching the microphones of the array. Different methods have been proposed for addressing the influence of the robot's shape in the acoustic localization. The Head Related Transfer Function (HRTF) can be used for modeling the diffraction of sound waves by the robot (Keyrouz, 2008). However, measuring the HRTF is not a trivial process. Besides, the HRTF depends on the room acoustics, and therefore the measurements might have to be repeated each time the robot operates in a different environment. As a simpler alternative, the Auditory Epipolar Geometry (Nakadai et al., 2001; Kim et al., 2011) and Scattering Theory (Nakadai et al., 2003) have been used in different systems. Both methods assume that the array is installed on a spherically shaped robot head and therefore their accuracy decreases when this assumption is not satisfied. On a different approach, an appropriate geometry for the microphone array can be selected so that at least one microphone pair can be used to refine the location estimate (Kwon et al., 2007).

Moving away from controlled conditions, in this paper we present an acoustic localization system designed for real-world robots and environments. Although localizing multiple speakers is our primary goal, we also cover aspects related

[1] ALIZ-E aims at designing and developing long-term adaptive social interaction between robots and child users (8–11 years old). Under ALIZ-E, HRI is deployed in real-world environments and settings (Belpaeme et al., 2012).