Proposal for a Students' Project (Project Group)

1 Topic

Active Localization and Tracking of Sound Sources Using an Artificial Head

2 Duration

Summer Term 2007, 8 hours weekly

3 Organizers

Gernot A. Fink, Informatik XII / IRF[1], Department for *Intelligent Systems*
OH 16, Raum E23, Tel.: 6151, *Gernot.Fink@udo.edu*

Thomas Plötz, IRF[1], Department for *Intelligent Systems*
OH 8, Raum 103, Tel.: 4645, *Thomas.Ploetz@udo.edu*

Alfred Hypki, IRF[1], Department for *Industrial Robotics and Handling Systems*
OH 8, Raum 203, Tel.: 5623, *Alfred.Hypki@udo.edu*

[1]Robotics Research Institute

4 Task

4.1 Motivation

Acoustic signals – especially spoken language – are an attractive means for interacting with intelligent environments or mobile robots. In the simplest case a user might snap his fingers in order to attract the attention of an intelligent agent. In more advanced scenarios a robot could be instructed by giving spoken language commands. A challenging problem in such settings is to separate the relevant acoustic events from interfering sounds and capture the data with a quality sufficient for further processing. The most important pre-requisite for this is the use of multiple microphones in order to be able to localize acoustic events.

From the wide range of possible sensor configurations in the laboratory of the Intelligent Systems Group at the Robotics Research Institute (IRF) a setup with two microphones mounted in an artificial head is investigated, which closely mimics the acoustic sensing capabilities of a human. An important aspect of such an anthropomorphic recording system is the possibility to actively orient the artificial head with the integrated sensors towards the signal source currently of interest. This leads to improved quality of the captured audio data. Another point of interest is the realization of a human-like motion behaviour for the head to maximize the impression of having an attentive and listening partner vis-a-vis.

4.2 Goal

The goal of the proposed project group is the development of a system for automatic detection, localization and tracking of an acoustic source, e.g.. an active speaker, within a conference room. Two microphones, i.e. all investigations and developments are based on stereo signals, which are integrated (fixed) in an artifical head, which itself is mounted on an active "pan-tilt unit" (PTU), should actively be oriented towards the acoustic signal of interest. Technically this implies, that after some detection process, the particular speaker has to be localized by analyzing acoustic data only. The result of this detection and localization process represents the basis for the active adjustment of the PTU towards the acoustic signal of interest.

Generally, the overall task can be subdivided into two main parts: **acoustic processing** and **control of the PTU**, i.e. reactively (re-)adjusting it's orientation according to the localization hypothesis.

First, the existence of a *relevant* acoustic signal to be localized and tracked in further steps needs to be detected. Sound detection also includes the discrimination between relevant and irrelevant signals can be either based on actual speech or on specific initiation signals (e.g. a snap of the fingers etc.). Basically, this corresponds to the development of some (simple) attention system.

In order to allow for proper (re-)adjustment of the microphones towards the sound source the initial coordinates of the active speaker's position need to be estimated. Subsequently, an automatic tracking procedure continuously updates the position of the sound source.

The results of the localization and tracking procedure, respectively, are then transferred to a motion control unit, which is responsible for the physical re-adjustment of the artificial head and thus the microphones mounted. The methods to implement for the positioning of the PTU should be based on an analysis of the movement of the heads of real human beings, e.g. by using videos cameras to record human head movements and a subsequent motion tracking. Eventually, the microphones are adjusted towards the spatial position which is optimal for e.g. robust speech recognition as mentioned above. A goal is the development and implementation of motion control algorithms resulting in close-to-human movements of the head mounted on the PTU.

All parts outlined need to be integrated into an overall system which as an example can be used to track the position of an active speaker in a conference room.

4.3 Foundations

For the detection of speech signals the automatic discrimination between general acoustic noise and actual speech data – so-called voice activity detection (cf. e.g. [ASMG05] and references therein) – needs to be realized. Usually, certain statistical features (zero-crossing rate, energy, etc.) computed directly on the acoustic signal can be fed into the binary decision process. For robust speech detection integration over certain time steps needs to be performed.

If the attention of the agent is attracted by special initiation signals, like a snap with the fingers, such signals need to be detected robustly. In these premises, as a first attempt techniques based on (normalized) template matching can be applied [KE00]. Certainly, the effectiveness of more advanced signal detection methods needs to be analyzed (cf. e.g. [Cal02, HK02]).

For speech localization the time delay between channels in the recorded stereo signals have to be analyzed and exploited for the estimation of the most probable position of the acoustic source. According to the literature certain short-time cross correlation techniques can be applied (cf. e.g. [GRT03]).

In order to continuously update the relevant speaker's position acoustical tracking needs to be realized. Therefore, certain prediction filters like Kalman or (variants of) particle filters (cf. e.g. [WLW03, TBF05]) which are also used in robotics for tracking objects or the robots own position, seem to be very promising. Alternatively, the effectiveness of certain histogram based tracking techniques can be evaluated.

The PTU can be regarded as a kind of robot and the head can be seen as the workpiece handled. Therefore standard kinematical approaches and motion control algorithms [Pau82, Kor85] can be examined and eventually taken as the base for further developments. The movements of the employed PTU are controlled using a standardized hardware (RS232) and the software interface Sony VISCA. To connect the motion control system to the PTU the features of the software package libVISCA [Dou06] can be checked for suitability and possibly be adopted for the given task.

5 Pre-requisites

At the Robotics Research Institute an intelligent house – the "FINCA"[1] – including a smart conference room has been created serving as laboratory for the investigation of next generation man-machine interaction techniques.

Within the conference room, an artificial head including two microphones, and mounted on a pan-tilt unit (PTU) is available for the project. Furthermore, all hard- and software necessary for the capturing and processing acoustic signals is available and can be used during project work.

Participants of the project group should have basic knowledge in general signal processing techniques and in robotics (including the related mathematical foundations), programming skills, and high motivation. They should be able to work in teams with strong goal-orientation.

[1] FINCA is an apronym for *F*lexible *I*ntelligent e*n*vironment with *C*omputational *A*ugmentation.

6 References

[ASMG05] M.Y. Appiah, M. Sasikath, R. Makrickaite, and M. Gusaite. Robust voice activity detection and noise reduction mechanism. Technical report, Institute of Electronics Systems, Aalborg University, 2005.

[Cal02] Laurent Calmes. A binaural sound source localization system for a mobile robot. Master's thesis, Faculty of Mathematics, Computer Science and Natural Science at the Rheinisch Westfaelische Technische Hochschule Aachen, 2002.

[Dou06] Damien Douxchamps. libVISCA, 2006. `http://damien.douxchamps.net/libvisca/index.php`.

[GRT03] T. Gustafsson, B.D. Rao, and M. Trivedi. Source localization in reverberant environments: modeling and statistical analysis. *IEEE Trans. on Speech and Audio Processing*, 11(6):791–803, 2003.

[HK02] A.A. Handzel and P.S. Krishnaprasad. Biomimetic sound-source localization. *IEEE Sensors Journal*, 2(6):607–616, 2002.

[KE00] Martin Kermit and A. J. Eide. Audio signal identification via pattern capture and template matching. *Pattern Recognition Letters*, 21(3):269–275, 2000.

[Kor85] Y. Koren. *Robotics for Engineers*. McGraw-Hill, 1985.

[Pau82] R. P. Paul. *Robot Manipulators: Mathematics, Programming and Control*. MIT Press, 1982.

[TBF05] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics*. MIT Press, 2005.

[WLW03] D.B. Ward, E.A. Lehmann, and R.C. Williamson. Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Trans. on Speech and Audio Processing*, 11(6):826–836, 2003.

7 Legal Issues

The results of all project work incl. the software developed will be made available without any restrictions to the department of computer science of Dortmund University, and the Robotics Research Institute, respectively. Furthermore, any constraints or restrictions as well as non-disclosure agreements regarding the use and exploitation of the results of the project group are null and void.