

BNAIC Demo: Online Speaker Detection by the iCat Robot

B. Terwijn

A. Noulas

*ISLA group, Informatics Institute, University of Amsterdam Kruislaan 403 1098 SJ
Amsterdam The Netherlands*

Abstract

The demonstration described in this document demonstrates it is possible to do robust real-time speaker detection using audio and video data on a single laptop computer.

1 Introduction

Automated speaker detection is the problem of detecting the speaking persons in a group of people at any particular time. Robustly solving this problem is useful in areas such as automated video analysis and human-machine interaction. For the detection we use a multimodal approach which learns the combined voice and appearance model of each person using a Dynamic Bayesian Network by correlating the audio and video data.

We demonstrate our speaker detection solution on the iCat robot platform (Figure 1) where the iCat looks at the person currently speaking to it. The iCat robot is developed by Philips Research [5] for research in human-machine interaction. It is equipped with a microphone and a camera mounted on its head which pans and tilts. The iCat can make many different facial expressions using thirteen servos to manipulate its face and body.



Figure 1: iCat in between two laptops

2 Purpose, User Groups and Projects

The purpose of the demonstration is to show it is possible to do online robust speaker detection on a dedicated standard laptop computer. We distinguish between offline and online speaker detection. With offline speaker detection all the data is available before the actual detection process to learn the model parameters. In contrast, with online detection the parameters need to be learned on the fly and the detection is done in real time, making this more difficult.

Speaker detection can be used in a variety of applications. Some examples are automated systems for making minutes of meeting or for video analysis where the system records who says what. More challenging are online systems such as an augmented video conferencing application where the speaker is automatically highlighted, or an automated ticket counter where people can buy tickets using a natural dialog.

Besides knowing who speaks the speaker detection can also benefit the speech recognition, especially in noisy environments and if there is more than one person speaking. Firstly, the recognition needs only be running when a person is indicated to speak which prevents processing background noise as speech. Secondly, because it is known who is speaking at what time, more accurate voice models can be learned. Thirdly, it is known at what time to use which person's voice model for recognition.

3 Technology

The technology background of this demo is based on ongoing research in the Intelligent Systems Laboratory Amsterdam on multi-modal speaker diarization. In this work, [3], [2], [4], information coming from the audio and video modality are fused using correlations in the joint-audiovisual space. This fusion is achieved under a probabilistic framework in the form of a dynamic Bayesian network. More information can be found in the full paper [1] submitted along with this demo.

4 Developers and Implementation

The speaker detection system described here is developed by the ISLA group of the University of Amsterdam. More specifically, Athanasios Noulas developed the speaker detection software and Bas Terwijn performed the integration with the iCat robot.

For controlling the iCat we use the OPPR2.0 library developed by Philips Research and for face detection we use the OpenCV1.0 face detector originally developed by Viola Jones.

5 Requirements and Duration

The speaker detection will run on a laptop with two 2.33Ghz processors and 2GB of RAM. The iCat has no onboard processing capabilities and is controlled by a separate laptop with 3Ghz processor and 512MB of RAM.

For our demonstration we require one table and three 230 Volt power outlets. Our demonstration is expected to take ten minutes excluding time for questions.

References

- [1] A. K. Noulas and B. J. A. Kröse. Learning in multimodal information streams. In *Belgian-Dutch Conference on Artificial Intelligence*, page submitted, 2007.
- [2] A. K. Noulas, N. Vlassis, and B. J. A. Kröse. Cross entropy for learning in multi-modal streams. In *Joint Workshop on MultiModal Interaction and Related Machine Learning Algorithms*, page to appear, 2007.
- [3] Athanasios K. Noulas and Ben J. A. Kröse. E.m. detection of common origin of multi-modal cues. In *International Conference on Multimodal Interfaces*, pages 201–208, 2006.
- [4] Athanasios K. Noulas and Ben J. A. Kröse. On-line multimodal speaker diarization. In *International Conference on Multimodal Interfaces*, page submitted, 2007.
- [5] A.J.N. van Breemen, Xue Yan, and B. Meerbeek. iCat: an animated user-interface robot with personality. In *The Fourth International Conference on Autonomous Agents and Multi Agent Systems*, pages 143–144, 2005.