Chapter 11
Automatic Speaker Localization and Tracking:
Using a Fusion of the Filtered Correlation with the Energy Differential

Siham Ouamour
USTHB University, Algeria

Halim Sayoud
USTHB University, Algeria

Salah Khennouf
USTHB University, Algeria

ABSTRACT

This paper presents a system of speaker localization for a purpose of speaker tracking by camera. The authors use the information given by the two microphones, placed in opposition, to determine the position of the active speaker in trying to supervise the audio-visual recording. To achieve the speaker localization task, the authors have proposed and employed two methods, which are called respectively: the filtered correlation method and the energy differential method. The principle of the first method is based on the calculation of the correlation between the two signals collected by the two microphones and a special filtering. The second is based on the computation of the logarithmic energy differential between these two signals. However, when different methods are used simultaneously to make a decision, it is often interesting to use a fusion technique combining those estimations or decisions in order to enhance the system performances. For that purpose, this paper proposes two fusion techniques operating at the decision level which are used to fuse the two estimations into one that should be more precise.

DOI: 10.4018/978-1-4666-0119-2.ch011

INTRODUCTION

The supervision of audiovisual recordings in multi-sensor smart-rooms (Neumann, Casas, Macho, & Ruiz Hidalgo, 2009) requires a combination of several localization methods by a special fusion technique, which will control the speaker tracking according to the information given by all the sensors.

Tracking technology is required both to keep the camera focused on the speaker and to display audience members when they talk. There are four general classes of tracking technology: sensor-based, motion-based, microphone-array-based and speaker-recognition-based. While all the four methods can be used for a single speaker, only the third and the last ones are normally used for multi-speaker audience (Liu, Rui, Gupta, & Cadiz, 2000).

In the context of automatic analysis of meetings, robust localization and tracking of active speakers is of fundamental importance, particularly for enhancement and recognition of speech in microphone-array based *ASR* (Automatic Speaker Recognition) systems. Microphone arrays provide hands-free and high-quality distant speech acquisition through beamforming techniques, which rely on speaker location for speech enhancement (Cox et al., 1987).

Furthermore, localization and tracking of active speakers from multiple far-field microphones are challenging tasks in smart room scenarios, where the speech signal is corrupted with noise from presentation devices and room reverberations (Maganti & Perez, 2006).

Sound source localization is defined as the determination of the coordinates of sound sources in relation to a point in space. It is achieved by using differences in the sound source received by different microphones to estimate the direction and if possible the actual location of the sound source. For example, human ears act as two different sound observation points, enabling humans to estimate the direction of source of the sound (Ui-Hyun, Jinsung, Doik, Hyogon, & Bum-Jae, 2008).

So how can these ears make an estimation of the speaker position?

To try to respond to the question, or at least simulate this faculty with two opposite cardioid microphones, we have done a thorough experimental investigation on two new proposed techniques based on the filtered correlation and the energy differential, which led us to several interesting results.

However, since we have implemented two different methods of speaker localization and since the two detection decisions of these methods are not necessarily similar, we have proposed and implemented two fusion techniques, in order to improve the precision of speaker localization and tracking.

SPEECH DATABASE

We have built four experimental databases with different scenarios, different speakers and different configurations:

- *DB8* database: the distance between the two microphones is 4.20 m.
- *DB9* database: the distance between the two microphones is 2 m.
- *DB10* database: the distance between the two microphones is 1 m.
- *DB11* database: the distance between the two microphones is 1 m.

In this paper, we will describe only the experiments done on *DB11* database, since the results got with long distances (*DB8* and *DB9*) are very affected by the echo effect, and those obtained on the *DB10* are insufficient.

The *DB11* database contains several scenarios with different speakers speaking alternatively in a natural manner and with different configurations. There are two general configurations: a stable