ON CALIBRATION AND COHERENCE SIGNAL ANALYSIS OF THE CHIL MICROPHONE NETWORK AT IRST

L. Brayda

Institut Eurecom
2229 route des Cretes
06904, Sophia Antipolis, France
Email: Luca-Giulio.Brayda@eurecom.fr

C. Bertotti, L. Cristoforetti, M. Omologo, P. Svaizer
Istituto Trentino di Cultura (ITC)-irst
Via Sommarive, 18 - Povo
38100, Trento, Italy
Email: omologo@itc.it

Introduction

The purpose of this work is to describe the Microphone Network presently used at ITC-irst for multi-microphone data collection and prototype development, with the specific aim of conducting research inside the CHIL European Project.
In the project, we define a generic multi-sensor system which consists of two main components: a distributed multi-camera system for visual room observation, including several calibrated cameras, and a multi-microphone system for acoustic scene analysis, which consists of microphone arrays, microphone clusters, table top microphones and close-talking microphones allowing detection of multiple acoustic events, voice activity detection, ASR and speaker location and tracking [1]. The target scenario comprises seminars and meetings. The entire audio acquisition system is based on a common sampling rate of 44.1 kHz and a sample accuracy of 24 bit. Also for acoustic sensors, a detailed characterization process as well as a calibration step are necessary, according to the purpose of having a jointly consistent description of the audio-video sensor geometry.

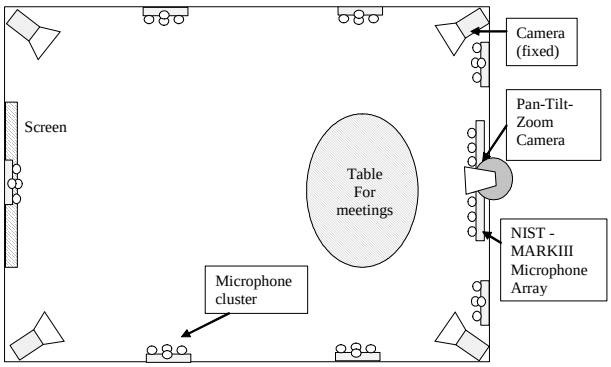

Figure 1: Map of the CHIL room at ITC-irst.

In the CHIL room at ITC-irst (See Figure 1), seven T-shaped microphone arrays, each consisting of four omnidirectional microphones, were installed in order to obtain an optimal coverage of the environment for speaker localization and tracking purposes. Moreover, a NIST-MarkIII array [2] of 64 microphones was installed on the wall facing the seminar speaker. Through its use the primary objective is far-field ASR: however, benefits are expected also for what concerns the use of MarkIII signals for speech activity detection and speaker localization.

Calibration

The accuracy of speaker localization as well as of beamforming-based enhancement systems is highly dependent both on the precision of input devices and on an accurate knowledge of sensor positions. In order to validate the geometry model of microphone arrangement, a semi-automatic procedure of calibration was defined together with a meticulous manual measurement of the coordinates of each microphone. This procedure can be accomplished by employing one or more loudspeakers in known positions and a test signal with appropriate characteristics (namely, a chirp-like signal). With the given waveforms and the signals acquired by the various microphones, an estimation of the relative delays can be derived. Consequently, a consistency check can be achieved and a calibration of the source localization system can be carried out on the test signals. Moreover, the possibility of estimating the room impulse response between a given source position and each microphone in the room is of great interest. In fact, the knowledge of the impulse responses is very useful to

characterize the multi-path propagation inside the room and to create realistic models for far-microphone signals acquired from real talkers. A more accurate modeling could be achieved by exploiting a talking head in place of the loudspeaker: however, this solution seems to be at the moment too complex to adopt in CHIL. Finally, note that a specific issue that should not be neglected in this calibration phase concerns the dependence of sound speed on temperature. If not accounted for, this could introduce a bias that would directly affect the results of the localization procedure.

Coherence among microphone pair signals

A relevant effort was devoted to characterize any possible bias in the coherence between signals of the same acquisition system (that is based on the same clock and hardware). This analysis was conducted on all the microphones of the network. In particular, the most interesting insights came out from the use of the NIST Mark III array, a very complex and effective acquisition system able to acquire 64 signals at 44.1 kHz and with a 24 bit accuracy. This platform was conceived to allow interested laboratories access to a relatively cheap and reliable means of acquiring multi channel speech signals suitable for phased array processing research. It represents an ideal solution for the purposes of CHIL project, in order to study new techniques of speaker localization and distant-talking automatic speech recognition. According to the preliminary experiments conducted in an insulated room of ITC-irst laboratories, it was found that the potential of the device could be improved through a hardware intervention aimed to eliminate some residual noise components. This might be neglected in the general case, but in this specific context it does introduce a significant bias in a generic time delay estimation process.

The left part of Figure 2 shows the coherence between two non adjacent channels of the array, through a bidimensional representation deriving from a Cross-power Spectrum Phase (CSP) analysis [3]. One can note a constant peak that is centered at 0 samples delay. This effect is evident for any pair of microphones belonging to any microboard of the array. It is worth noting that any localization technique as well as beamforming algorithms would be affected by the artificial coherence at zero samples delay, leading to the hypothesis of a source in front of the array (at an "infinite" distance) any time there is not a more dominant speech source (e.g. a speaker having a relatively loud voice).

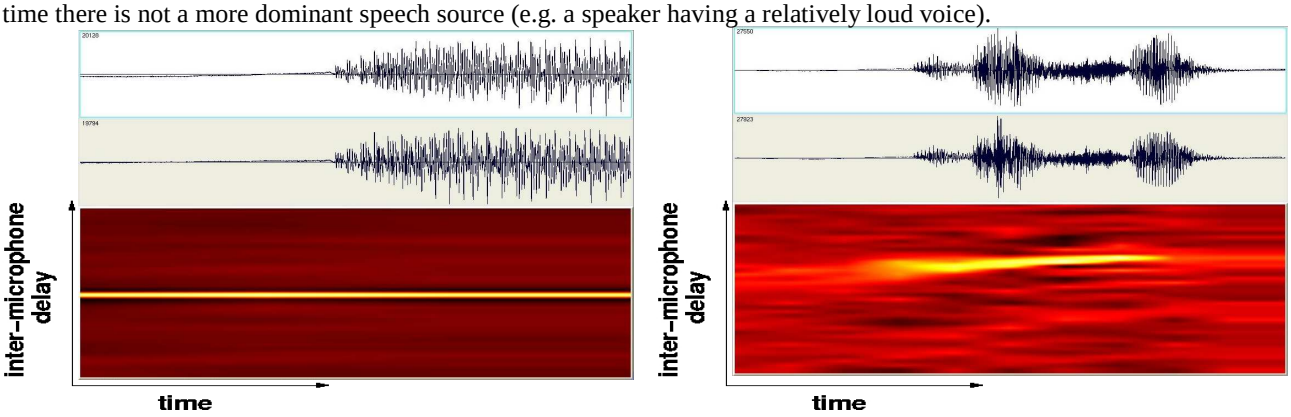

Figure 2: Signals extracted from Channel 1 and Channel 8. The peak of the CSP function (centered at 0 samples delay) reported in the left part of the figure shows a strong coherence between the device noise sequences. On the contrary, the peak of the CSP function reported in the right part of the figure, obtained from two signals acquired by the modified array, shows a strong coherence only when the speaker is talking (and moving to the right of the array).

The problem was solved by replacing some electronic components as well as introducing a set of rechargeable batteries to power only the amplification stages, in order to eliminate the above mentioned common-mode noise. More details on this analysis and the related intervention can be found in [4]. The right part of Figure 2 shows the coherence between the same channels, after the hardware intervention. The latter result derived from the use of the array in an environment with one active speaker pronouncing an isolated word.
Note that in the intervals without speech it is evident that the coherence bias disappeared.

References

[1] Dusan Macho et al., "First experiments of automatic speech activity detection, source localization and speech recognition in the CHIL project," submitted to HSCMA 2005.
[2] Cedrick Rochet., URL: http://www.nist.gov/smartspace/toolChest/cmaiii/userg/Microphone_Array_Mark_III.pdf
[3] M. Omologo, P. Svaizer "Acoustic event localization using a Cross-power Spectrum Phase based technique", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 1994*.
[4] C. Bertotti et al., URL: http://www.eurecom.fr/~brayda/MarkIII-modified-at-IRST.pdf