DICIT acoustic WOZ data

FBK-irst has conducted Wizard of Oz experiments with the aim of collecting useful data for testing signal processing algorithms in the scenario foreseen by the DICIT project.

Introduction

This file describes the data collected under the Wizard of Oz experiments conducted at FBK-irst during the European DICIT project (http://dicit.fbk.eu). Usually, in a Wizard of Oz experiment, a subject is requested to complete specific tasks using an artificial system. The user is told that the system is fully functional and should try to use it in an intuitive way, while the system is operated by a person not visible to the subject. The operating person – called wizard – can react to user inputs in a more comprehensive way than any system could, because he/she is not confined by computer logic. In an effort to simulate as closely as possible the behaviour of a real system based on voice interaction, recognition errors are randomly simulated by the wizard.

The goal of our WOZ experiments is to create realistic usage scenarios for acoustic pre-processing purposes. As a consequence our interest in a high level analysis of the interaction is limited. This data can be useful for testing in a real and adverse scenario different types of signal processing algorithms: Acoustic Echo Cancellation (AEC), Speech Activity Detection (SAD), Sound Source Localization (SLoc), Speaker Identification.

In the following the experimental setup, the modalities of the data collection and the data format are described.

WOZ experimental setup

The WOZ experiments were conducted in a standard room located at FBK-irst whose reverberation time is about 700 ms. The television was simulated by means of a video beamer, projecting its output on a wall, and two high-quality loudspeakers that were placed on either side of the screen.

The participants sat on four seats, positioned at a fixed distance (about 2 meters) from the screen. It was observed that, even when allowed to move, participants rarely went closer than one meter from the arrays and the TV.

Both traditional television and teletext were simulated by using previously recorded TV video clips and teletext pages. The two channels of the TV audio output were decorrelated before playback in order to allow an effective implementation of stereo acoustic echo cancellation without impairing listening quality. The system was controlled by the wizard through a Windows PC station located in an adjacent room. A schema of the WOZ room can be seen in Figure 1.

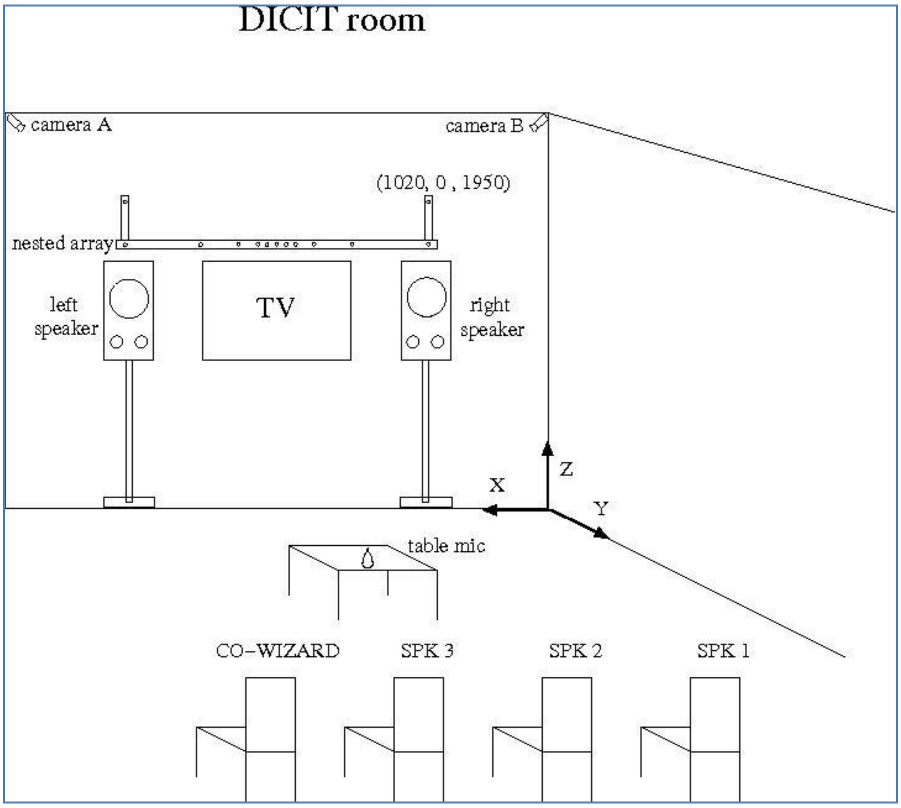

Figure 1: WOZ room setup. The picture includes also the Cartesian coordinate system and the coordinates of the microphone associated to the first channel of the array.

Audio signals are recorded by means of a nested microphone array, four close-talk microphones worn by the three participants and the co-wizard, and a table microphone positioned in front of them. The 15-electret-microphone array, which allows for various configurations has been specifically developed for the project. It was located above the television screen and represented the acoustic sensor setup that the DICIT consortium intends to exploit. It forms four linear sub-arrays composed of equidistant microphones, three of which consist of five microphones each and one that consists of seven. Figure 2 shows the microphone arrangement within the nested array.

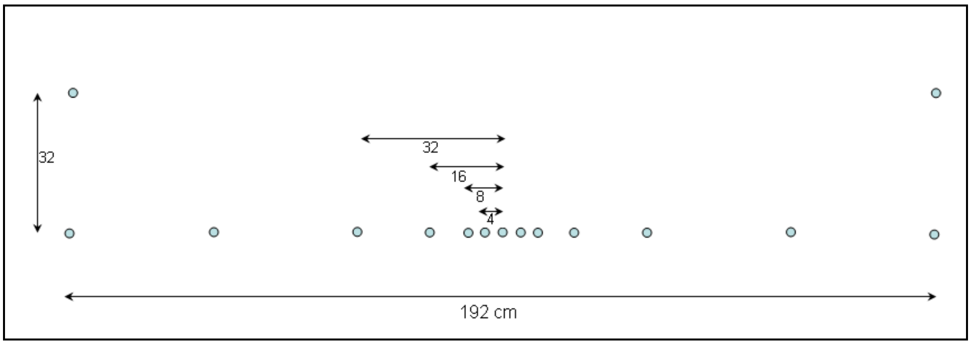

Figure 2: the nested array structure.

Description Of The Wizard Of Oz Experiments

The data collection is composed of six sessions. Three naïve users and one supervisor (co-wizard) participated in each session. The subjects were recruited from the staff at FBK-irst, therefore the sample was composed not only of technology professionals but also of subjects from other fields of work (administration etc.). Before the experiments, all subjects received an instruction sheet describing the tasks and the expected behaviour. Although all four participants were simultaneously present in the room, only one person at a time was allowed to interact with the system. In any case, other participants unconsciously made occasional noises that were recorded by the system.

We chose to do recordings with a group of four people to simulate a typical home scenario (e.g., a family watching TV). The supervisor had a double role: First of all, he/she had to help naïve users in navigating the dialogue system, to ensure the accomplishment of the experiment's objective. At the same time, the supervisor had to generate a number of acoustic events that were found to be typical for a real domestic scenario. In our WOZ experiments the following events were specified as relevant for the DICIT scenario: slamming door, chairs being moved, ringing phones, coughing, laughing, falling objects and rustling paper.

Each session was split in three phases. At the beginning, all participants sat in front of the television and read out a set of phonetically rich sentences that may be exploited to train algorithms for speaker identification and verification. During the second phase, each person kept on sitting and interacted with the system trying to accomplish a list of predefined tasks. These included the typical actions to control a traditional television: channel switching, volume control etc. In the third part of the experiment the subjects were asked to find specific pages in the teletext using voice-commands, while moving around in the room. This movement was especially intended for testing the source localization algorithms.

Data structure

Root directory includes directories 210507, 220507, 230507, 240507_1515, 240507_1630, 280507 representing the six WOZ sessions. Each session directory contains the following directory list:

- sp1_ses1
- sp1_ses2
- sp1_ses3
- sp2_ses1
- sp2_ses2
- sp2_ses3
- sp3_ses1
- sp3_ses2
- sp3_ses3

the corresponding reference files:

- sp1_ses1.ref, sp1_ses1.loc
- sp1_ses2.ref, sp1_ses2.loc
- sp1_ses3.ref, sp1_ses3.loc
- sp2_ses1.ref, sp2_ses1.loc
- sp2_ses2.ref, sp2_ses2.loc
- sp2_ses3.ref, sp2_ses3.loc
- sp3_ses1.ref, sp3_ses1.loc
- sp3_ses2.ref, sp3_ses2.loc
- sp3_ses3.ref, sp3_ses3.loc

sp[1-3] represents the speaker label, while ses[1-3] represents the sub-session label of each speaker. In particular, ses1 is the session part in which the participant was asked to pronounce four Italian phonetically rich sentences: audio files in this directory can be used to train the speaker model for the speaker recognition task. ses2 is the session part in which the speaker was asked to interact with the system with simple commands: during this session all speakers were sitting on chairs. Finally, ses3 is the session in which the speaker was asked to interact with the system in order to browse teletext: speakers were free to move around the room and meanwhile noises were produced by the co-wizard.

Each directory sp*_ses* includes the following audio files compressed with NIST sphere standard:

- ctm[1-4].sph: 4 close-talk signals belonging to speaker 1,2,3,4
- left.sph and right.sph: left and right channels of TV output to be used as reference signals for Acoustic Echo Cancellation.
- nested_[1-15].sph: 15 signals belonging to the nested array. The first channel refers to the rightmost top microphone of Figure 1 and Figure 2.
- table.sph: signal acquired by a table microphone placed close to the speakers.

Audio data are acquired at a sampling rate of 48 kHz with a precision of 16 bits.

The sp*_ses*.ref are the annotation files relative to session sp*_ses*. In the following the annotation format is explained by means of some examples.

The first example refers to an utterance pronounced by a speaker:

96000 234816 sp2 il giovane gnomo ha emesso un urlo agghiacciante

- "96000" and "234816" are the speech time markers expressed in sample number
- "sp2" is the speaker label
- "il giovane gnomo ha emesso un urlo agghiacciante" is the transcription of the spoken utterance

The second example refers to a noise produced by the co-wizard:

6800495 6855744 noise [sla]

- "6800495" and "6855744" are the noise time markers expressed in sample number
- "noise" is the label for noise
- [sla] is the label for the noise type

The third example refers to a speech utterance overlapping with a sound produced by the co-wizard:

2053584 2202672 sp1 [pap-] meteo meteo amsterdam [-pap]

- "2053584" and "2202672" are the speech time markers expressed in sample number
- "sp1" is the speaker label
- "meteo meteo amsterdam" is the transcription of the spoken utterance while [pap] is the label of the produced noise

In Table 1 the noise classes with their respective labels are reported.

Label	**Acoustic Event**
[sla]	door slamming
[cha]	chair moving
[pho]	phone ringing (various rings)
[cou]	coughing
[lau]	laughing
[fal]	objects falling (water bottle, book)
[pap]	paper rustling (newspaper, magazine)
[spk]	noises from speaker mouth
[fil]	hesitation expression (uh, eh, ah...)

Table 1: Noise classes and labels.

In a similar way, sp*_ses*.loc files contain the coordinates of the active speaker. Moreover the files include information on: the speaker label, number of active noise sources and status, either *standing* or *sitting*, of the speaker. The time resolution of the localization references is 0.5 s. Table 2 shows an excerpt of a *.loc file.

Time	Speaker	# noises	X coor	Y coor	Z coor	Status
2.070000	1	0	2142.000000	2001.000000	1560.000000	standing
2.570000	1	0	2155.000000	2005.000000	1560.000000	standing
3.070000	1	0	2191.000000	1986.000000	1560.000000	standing
3.570000	1	0	2214.000000	2031.000000	1560.000000	standing
4.070000	1	0	2148.000000	2017.000000	1560.000000	standing
4.570000	1	0	2119.000000	2016.000000	1560.000000	standing

Table 2: Excerpt of a *.loc file

In case nobody is speaking a conventional set of values is adopted as reported in Table 3. The speaker label is 0, coordinates are all set to -1 and the status is *none*.

14.070000	0	0	-1.000000	-1.000000	-1.000000	none

Table 3: *.loc file line when nobody is speaking.

As the localization references were obtained automatically through a video tracker that is not flawless, they are not reliable for all the sessions. Thus we suggest to omit session 210507. Moreover, as far as the remaining sessions concern it was possible to manually verify the reliability of the references only for the segments where participants are sitting.