

WOZ Acoustic Data Collection For Interactive TV

A. Brutti*, L. Cristoforetti*, W. Kellermann+, L. Marquardt+, M. Omologo*

* Fondazione Bruno Kessler (FBK) - irst

Via Sommarive 18, 38050 Povo (TN), ITALY

+ Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg (FAU)

Cauerstr. 7, 91058 Erlangen, GERMANY

E-mail: {brutti|cristofo|omologo}@fbk.eu; {wk|marquardt}@LNT.de

Abstract

This paper describes a multichannel acoustic data collection recorded under the European Dicit project, during the Wizard of Oz (WOZ) experiments carried out at FAU and FBK-irst laboratories. The scenario is a distant-talking interface for interactive control of a TV. The experiments involve the acquisition of multichannel data for signal processing front-end and were carried out due to the need to collect a database for testing acoustic pre-processing algorithms. In this way, realistic scenarios can be simulated at a preliminary stage, instead of real-time implementations, allowing for repeatable experiments. To match the project requirements, the WOZ experiments were recorded in three languages: English, German and Italian. Besides the user inputs, the database also contains non-speech related acoustic events, room impulse response measurements and video data, the latter used to compute 3D labels. Sessions were manually transcribed and segmented at word level, introducing also specific labels for acoustic events.

1. Introduction

The goal of the European project Dicit¹ (Distant-talking Interfaces for Control of Interactive TV) is a user-friendly human-machine interface that enables a speech-based interaction with TV, related digital devices and infotainment services. In the foreseen scenario, the user is supposed to interact with the system in a natural and spontaneous way, without being encumbered by any head-mounted close-talk microphone, in a living room furnished with several digital devices, amongst others a TV equipped with a set-top box (STB). Different users, once at a time, will have access to the system, which offers information service from the Electronic Program Guide (EPG). Since the system is expected to operate in the presence of multiple acoustic sources, noisy background and TV audio, the project has to address some challenging and new technical issues: multichannel acoustic echo cancellation, blind source separation, acoustic event classification, multiple speaker localization, distant-talking automatic speech recognition and mixed initiative dialogue and multimodal integration. Multimodality is exploited to allow the user to choose between or combine two different interaction modalities: traditional TV remote control and spoken commands. Besides the design of robust dialogue strategies and the implementation of an accurate speech recognizer, one of the most critical aspects of the project is the development of a reliable acoustic front-end, capable of tackling all the difficulties of the given applicative scenario. In an effort to characterize the user behaviours in such a context and to better figure out the most crucial aspects of the project in terms of acoustic front-end processing, a set of WOZ experiments was conducted.

After a detailed description of the WOZ experiments, this paper presents the adopted multichannel setup and gives an overview of the annotation process. Final discussions conclude the paper.

2. Description Of The Wizard Of Oz Experiments

In a Wizard of Oz (WOZ) 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. From a WOZ study, interaction patterns can be extracted and applied to an actual prototype.

Our WOZ focused on the need for creating realistic usage scenarios for acoustic pre-processing purposes. The WOZ experiments were conducted in two standard rooms located at two different sites, without any expedient to improve their acoustic properties.

The WOZ model has been translated and the experiments have been recorded in three languages: English, Italian and German. The data collection is composed of twelve sessions as described in Table 1.

| Site | Language | Number of sessions |
|------|----------|--------------------|
| FBK | Italian | 6 |
| FAU | German | 5 |
| FAU | English | 1 |

Table 1: Number of recorded sessions at each site and for each language

Each session included three naïve users and one supervisor (co-wizard). The persons came from our work places, but not all of them were technology professionals. Some of them were administrative professionals and some were students. Before the experiments, all of them were given 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. We

¹ For further details see: <http://dicit.fbk.eu>.

chose to do recordings with a group of four people to simulate a typical home scenario, like 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. At the same time, the supervisor had to intentionally generate some acoustic events that were common in a real scenario. The above mentioned events were a subset of the ones exploited in previous data collections conducted under the European project CHIL² (Temko et al., 2006). In our WOZ experiments the following events were taken into consideration: door slamming, chair moving, phone ringing, coughing, laughing, objects falling, paper rustling.

Each session was split in two phases. At the beginning, all the participants were sitting in front of the television and read out some phonetically rich sentences³ that may be exploited to train algorithms for speaker identification and verification (Furui, 1997). During the second phase, each person interacted with the system trying to accomplish a list of predefined tasks. These included the typical actions to control a traditional television: channel switching, zapping, volume control and so on. In order to let users get familiar with the system, the first part of the interaction was conducted using only the remote control. After this warm-up stage the users were allowed to control the system via both remote control and voice-commands while sitting on their chairs. In the final part of the experiments the subjects were asked to find specific pages in the teletext using only voice-commands, while moving around in the room. This movement was especially intended for the testing of the source localization algorithms in later stages. In an effort to simulate as closely as possible the behaviour of a real system based on voice interaction, recognition errors were randomly introduced by the wizard. Since our focus was on data recording for technology testing rather than dialogue modelling, we did not care about the fact that the later users gained some experience observing the errors of the first users. It could be also interpreted as a simulation of different levels of expertise.

In addition to the experiments, the users had to fill out questionnaires before and after the experiments. The former was on general statistical issues, like dialect and technical background; the latter focused on feedback about the experiments.

Each user interaction lasted about 10 minutes, which led to a total of 360 minutes of recordings.

3. Experimental Setup

Two standard rooms were equipped for the WOZ experiments. The objective was to simulate a typical living room, with similar dimensions and reverberation conditions.

Since it was necessary to hide the wizard from the users, adjacent rooms were prepared as well. All the hardware setup and computers were located in the wizard room and everything was wired to the experimental room.

² For further details see <http://chil.server.de>.

³ These sentences include a quasi balanced combination of all phonemes of the language in question leaving out all combinations that are invalid for that language.

3.1. WOZ Room Setup

The television was simulated by means of a video beamer, projecting its output on a wall, and two high-quality loudspeakers were placed on the sides of the screen. Both traditional television and teletext were simulated using some previously recorded TV video clips and teletext pages, provided in three languages. The two channels of the system audio output were decorrelated in order to allow an effective implementation of stereo acoustic echo cancellation without impairing listening quality (Huang and Benesty, 2004). The system was controlled by the wizard through a Windows PC station located in an adjacent room. “EB GUIDE Studio”, a tool suitably designed to manage the dialogue flow in the WOZ experiments, was adopted to record the dialogue sessions and control the system (Goronyz and Beringer, 2005). Additionally, a TV remote control was integrated into the system through an IR receiver in the experimental room which was connected to the serial port of the PC of the wizard. A schema of one of the WOZ rooms can be seen in Figure 1.

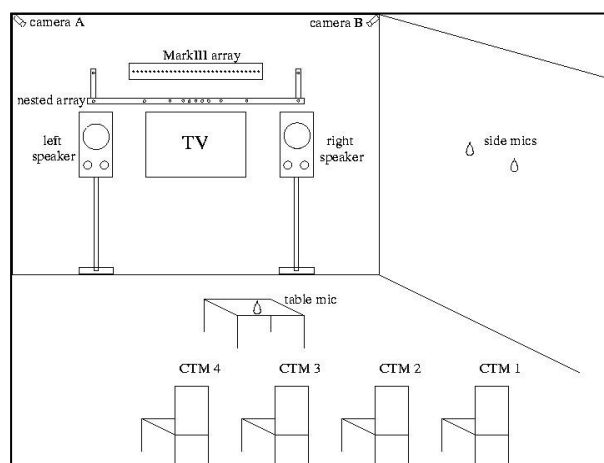


Figure 1: Schema of one of the rooms adopted in the WOZ experiments

3.2. Audio And Video Sensor Setup

A harmonic 15-electret-microphone array, which has been specifically developed, was located above the television and represented the acoustic setup that the DICIT consortium intends to exploit. It forms four linear sub-arrays composed by equidistant microphones, three of which consist of five microphones each and one consists of seven. Figure 2 shows the microphone arrangement in the harmonic array.

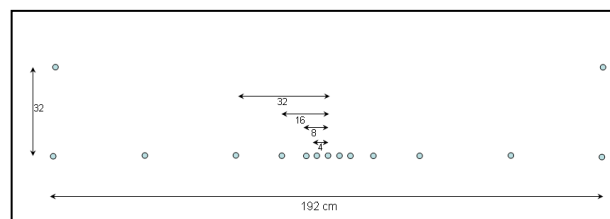


Figure 2: Layout of the harmonic microphone array

The array allows for the exploitation of different sub-arrays in order to meet the requirements of each of the

different acoustic pre-processing modules in terms of inter-microphone spacing.

For comparison purposes, the sessions were also acquired by a modified NIST MarkIII linear array (Brayda et al., 2005), placed just above the harmonic array. The MarkIII, depicted in Figure 3 is composed of 64 uniformly-spaced electret microphones, specifically developed for far-field voice recognition, speaker localization and audio processing. It records synchronous data at a sampling rate of 44.1 kHz or 22.05 kHz with a precision of 24 bits. The particularities of this array are its modularity, the digitalization stage and the data transmission via an Ethernet channel using the TCP/IP protocol.

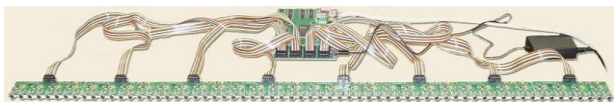


Figure 3: NIST MarkIII Microphone Array

The distances between the participant seats and the microphone arrays were about two meters. It was observed that, even when allowed to move, participants rarely went closer than one meter from the arrays.

A table microphone and two lateral microphones (located on a side wall of the room) were also used for recording. The microphones were all omnidirectional. The table microphone was placed one meter from the users and was meant to simulate a remote control equipped with a microphone. It may also be an alternative solution in case the quality of the signals acquired by the array is considerably lower due to a bigger distance. As to the lateral microphones, they will be exploited only for experimental analyses. Finally, each participant was also recorded by a close-talk microphone whose signals are used to guarantee robust segmentations and accurate transcriptions. In total, 24 input signals (26 at FAU) + 64 channels from the MarkIII were recorded. The 24/26 signals were recorded at 48 kHz sampling frequency with 16 bits precision, in a synchronous way and aligned at sample level. The MarkIII array was equipped with its own acquisition board at 44.1kHz and 24 bits precision.

Figure 1 also shows the positions of the acoustic sensors, besides the acoustic channels the room was furnished with 3 video cameras: one placed on the ceiling (not shown in the picture) and two on the upper corners. Video data were used both to monitor the experiments during the annotation process and to derive 3D reference positions for each participant. Notice that video and audio signals were manually aligned taking advantage of some impulsive events present in the recordings, for instance a door slam.

3.3. Recording Hardware Setup

To simulate the prototype of the WOZ and to record in parallel all the acoustic data, three PCs had to be employed. Two Linux machines (PC1 and PC2) were used for the data recording, while a Windows machine (PC3) was used to run the EB GUIDE Studio simulator tool. Hardware setups at FAU and FBK laboratories were similar, with only minor differences. The FBK setup is presented exemplarily and depicted in Figure 4.

PC1 was equipped with a digital board, a RME

HDSP9652, connected via optical ADAT ports to three RME OctaMicII preamplifiers with integrated AD converters. Sample synchronization was guaranteed to all the boards via a BNC cable connected to the word clock input. As recording software, we used a tool called Ecasound⁴, able to interface with the ALSA drivers.

PC2 was connected to the MarkIII array by a dedicated network interface card and a LAN cross cable. A specific software developed by NIST was used to record the data on the hard disk. The amount of recorded data on PC2 was about 480MB/minute.

PC3 was partially described in Section 3.1. The WOZ tool displayed the TV clips on the beamer through a dual-head graphic card, while the TV audio was connected to both the acquisition board of PC1 and to the loudspeakers.

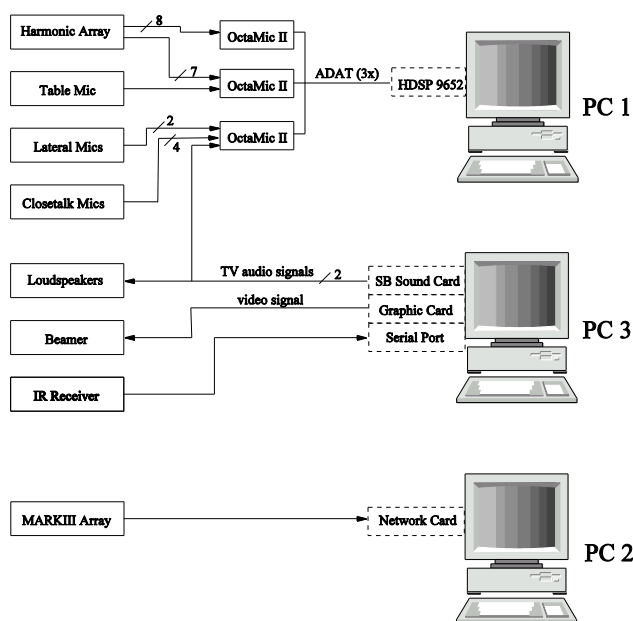


Figure 4: FBK recording setup

3.4. Room Impulse Response Measurements

Room impulse response measurements were carried out in order to provide data which could be used later for purposes such as speech contamination. Measurements were made at FAU in the same room used for the WOZ experiments utilizing the Maximum Length Sequence (MLS). A single loudspeaker was used to play the MLS sequence while the harmonic array and five separate microphones recorded the output in a synchronous way. The loudspeaker was moved to 12 different positions within the room and the measurements were repeated. At FBK, impulse response was measured in the WOZ room using a chirp sequence played by a loudspeaker positioned on the seat of each participant. The two microphone arrays recorded the output and 4 different positions were investigated.

4. Data Utilization

4.1. Data Annotation

In order to be used for later algorithm testing and speech

⁴ See <http://eca.cx/ecasound/index.php> for details.

recognition, the six FBK sessions (in Italian) have been manually transcribed and segmented at word level, introducing also specific labels for acoustic events. An annotation guideline, modified from a previous work (Cristoforetti et al., 2000), was used in order to ensure as much consistency as possible between the two annotators. The data were annotated using Transcriber⁵, a free graphic annotation tool which permits multichannel view. To ease the effort in understanding the dialogues between users and the system, stereo audio files were created putting the signal coming from the table microphone on the left channel and the sum of the close-talk microphones on the right channel. In this way, the annotators could listen in a selectable way to the environmental noises or to the uttered sentences.

Annotators were provided with a preliminary automatic segmentation based on the energy of the close-talk signals. Even if not reliable due to cross-talk effects and non-speech human sounds, this segmentation turned out to be a very useful starting point. It was also possible to display the automatic segmentation for each speaker, to help in understanding which user was speaking or producing some noise. Markers were inherited from the automatic segmentation and adjusted manually in order to have some silence around the usable signal.

Annotation information comprises the name (ID) of the speaker, the transcription of the uttered sentences and any noise included in the acoustic event list. Seven classes of noises were identified and annotated with square brackets (e.g., [pap] standing for paper rustling). Two other classes were created to label speaker's or unknown noises. Noises and their associated labels are listed in Table 2.

| 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 |
| [unk] | other unknown noises |

Table 2: Noise event classes

Annotators were also instructed to properly annotate those sentences that were personal comments and were not intended for the system.

As to the video data, a set of 3D coordinates for the head of each participant was created with a video tracker based on a generative approach (Lanz, 2006). Given the 3D labels, for each session a reference was derived, which includes the ID of the active speaker, his/her coordinates and some information about the presence of noises. The reference files were obtained as a combination of the raw 3D labels generated by the video tracker and the manual acoustic annotation, with a rate of 5 labels per second.

⁵ See <http://trans.sourceforge.net/en/presentation.php> for details.

4.2. Data Exploitation / Testing

The data collected with the WOZ experiments have been exploited for a preliminary evaluation of some FBK algorithms.

The main goal of the evaluation was to understand the peculiarities of the DICIT scenario and verify their influence on localization techniques in order to correctly handle them in the development of the first prototype. For instance we observed that user sentences were usually very short and silence was predominant.

The WOZ data were used to test the speaker verification and identification system: the system was applied to the signals of the close-talk, the single central microphone of the array and the beamformer, using matched model condition and different training material quantity. The results showed that beamforming yields benefits to the system performance when compared to the single microphone case, but the results are still inferior to the close-talk microphone case. The WOZ data were also used to test the acoustic event detection system. The test data were composed of 682 speech segments and 108 non-speech segments extracted from the continuous audio stream exploiting the manual annotation. The results are promising and highlight that the most confusable events are speech, cough and laugh.

5. Conclusions

This collection of data has been the first of its kind and is of significant benefit to acoustics front-end algorithms and dialogue strategies. Some on-going experimental activities on speaker localization, speech activity detection, distant speaker ID and others are being conducted on this corpus.

From these studies, it was observed that participants tended to use very short sentences, in a command-like fashion, to control the television. This aspect leads to a series of technical difficulties that must be taken into account correctly in order to ensure satisfactory performance. It was also observed that users prefer in general the speech based modality even with the presence of recognition errors. However, it is possible that this behaviour may be partly due to the novelty of such human-machine interface to the naïve users.

6. Acknowledgements

This work was partially funded by the Commission of the EC, Information Society Technologies (IST), FP6 IST-034624, under DICIT.

7. References

- Brayda, L., Bertotti, C., Cristoforetti, L., Omologo, M., and Svaizer, P. (2005). Modifications on NIST MarkIII array to improve coherence properties among input signals. In *Proceedings of AES, 118th Audio Engineering Society Convention*. Barcelona, Spain.
- Cristoforetti, L., Omologo, M., Matassoni, M., Svaizer, P., and Zovato E. (2000). Annotation of a multichannel noisy speech corpus. In *Proceedings of LREC 2000*. Athens, Greece.
- Furui, S. (1997). Recent Advances in Speaker Recognition. *Pattern Recognition Letters*, pp. 859–872

Goronzy, S., and Beringer, N. (2005). Integrated Development and on-the-Fly Simulation of Multimodal Dialogs. In *Proceedings of Interspeech 2005*. Lisbon, Portugal, pp. 2477--2480.

Huang, Y., and Benesty, J. (2004). *Audio Signal Processing for Next-Generation Multimedia Communication Systems*. Boston: Kluwer Academic.

Lanz, O. (2006). Approximate bayesian multibody tracking. *IEEE transaction on Pattern Analysis and Machine Intelligence*, pp. 1436--1449.

Temko, A., Malkin, R., Nadieu, C., Zieger, C., Macho, D., and Omologo, M. (2006). CLEAR Evaluation of Acoustic Event Detection and Classification systems. *CLEAR'06 Evaluation Campaign and Workshop*. Southampton, UK: Springer.