euronoise

Front-end processing of a distant-talking speech interface for control of an interactive TV system

Maurizio Omologo

Fondazione Bruno Kessler - IRST, Via Sommarive, 18, Povo, 38050 Trento, Italy
omologo@fbk.eu

A user-friendly interface is being investigated for the access to a virtual smart assistant enabling the interaction with TV-related digital devices and infotainment services. In the given scenario, the users can speak in a natural and comfortable way, not encumbered by any hand-held or head-mounted microphone. The environment is typically a living room, equipped with digital TV, Hi-Fi audio devices, etc., and populated by a group of people (e.g., family members).
Among the most challenging issues involved in this scenario are a multi-microphone front-end for an effective processing of the given acoustic scene, an Acoustic Echo Cancellation (AEC) component to compensate the sound produced by loudspeakers, and eventually a multi-modal distant-talking spoken dialogue system.
As far as the front-end is concerned, multiple speaker localization, speech activity detection, speaker identification, and speech recognition will have to perform accurately even when AEC is applied to the given microphone array. The paper aims to present preliminary results of this research, which is being conducted under the European Project DICIT.

1 Introduction

An attractive future scenario consists in the development of new workspaces where the so-called "ambient intelligence" is realized through a wide usage of sensors (cameras, microphones, etc.) connected to computers that fade in the background, largely invisible and significantly less intrusive to humans. Given a microphone network distributed in the environment, the analysis of the resulting acoustic scenario is accomplished by a multi-channel processing aimed at extracting real-time information for speaker tracking, acoustic event classification, and distant-talking speech recognition.

During the last two years, the European project DICIT (Distant-talking Interfaces for Control of Interactive TV) has been conducted to focus on the development of advanced multi-microphone based technologies and on the related application to the smart-home environment. In this case, the main objective of the project is an automatic voice interaction system which operates in a complementary modality to the use of remote control allowing one to control an interactive TV system in a natural and flexible way. In fact, according to recent studies on human-machine interfaces, augmenting remote control devices with speech represents an important opportunity worth of investigation. During the last years, new devices have been introduced in the market, which integrate remote control and speech recognition to control a SetTopBox (STB) platform. However, the use of these devices is not easy as the set of admitted commands is quite restricted, the system is trained to work with a single specific user, and the maximum user-device distance to obtain acceptable recognition performance is rather short.

The purpose of DICIT is to progress in different technical areas in order both to extend these systems to a voice interaction under far-field conditions, namely at 2-3 meters (or more) from the TV set-up, and to manage voice input from multiple speakers.

Although quite significant advances have been so far achieved in distant-talking ASR, today high performance systems can be guaranteed only for small vocabularies, with a good match between training and testing conditions, with rather controlled speaker position, head orientation and speaking style. In DICIT, distant-talking ASR is being addressed to overcome most of the above mentioned constraints.

This paper aims to describe in general the acoustic front-end being developed under DICIT and to provide, in more detail, an overview of the research and the technological components being addressed at FBK-irst. The remainder of this paper is organized as follows: Section 2 will provide a brief description of the general objectives and the foreseen scenario; Section 3 will focus on the front-end architecture and on some of the related components; Section 4 will report on an acoustic WOZ data collection and Section 5 will outline some of the planned future work.

2 Objectives and foreseen scenario

DICIT objectives encompass a large variety of research and development topics rotating around the basic goal of showing concrete progresses in acoustic and speech processing for noisy and reverberant environments as well as in multimodal spoken dialogue interaction with TV and related devices. To this purpose, advances are needed in fields related to multi-microphone processing for acoustic echo cancellation, speaker localization, beamforming, blind source separation, dereverberation, speech enhancement [1,2], robust speech recognition and spoken dialogue [3,4].

In the addressed Set-Top-Box (STB) scenario, a microphone array is placed nearby the TV and the related devices. Due to the interference of other coexisting active sound sources (e.g. loudspeakers, other talkers or noise sources) and to the effect of room acoustics, both the processing and the understanding of vocal messages become more problematic with respect to the ideal situation encountered when using a close-talking microphone. The state of the art in this area is still poor of effective solutions. For example, one of tasks of DICIT concerns the so-called "acoustic scene analysis": the goal is to get a system capable of automatically detecting how many speakers (or in general how many acoustic sources) are active at a given instant, who is speaking and where, what is he/she saying, etc. All this should work even when the TV itself is diffusing its audio in the room, independently of the volume and the content of the sound reproduced by the loudspeakers.

In practice, advanced techniques for acoustic echo cancellation, detection and classification of acoustic events are also necessary in order to prevent the system from reacting to other input than user's utterances. Since no push-to-talk button activation is foreseen during the interaction, a challenging aspect is to ensure a high performance in terms of event detection together with a very limited false alarm rate.

Moreover, functionalities and procedures essential for an effective voice interaction between user and TV platform are required. One of the ambitious aspects of the project concerns integration between spoken commands and manual remote control: all functionalities will be available in both modalities, thus giving to the user a feeling of flexible and unconstrained use of the system.

The system should be able to understand (at a semantic level) the user's request and to carry out the corresponding operation or, alternatively, to interact in a cooperative manner with the user (e.g. by asking more details about the user's intention, or by providing hints about the possible options).

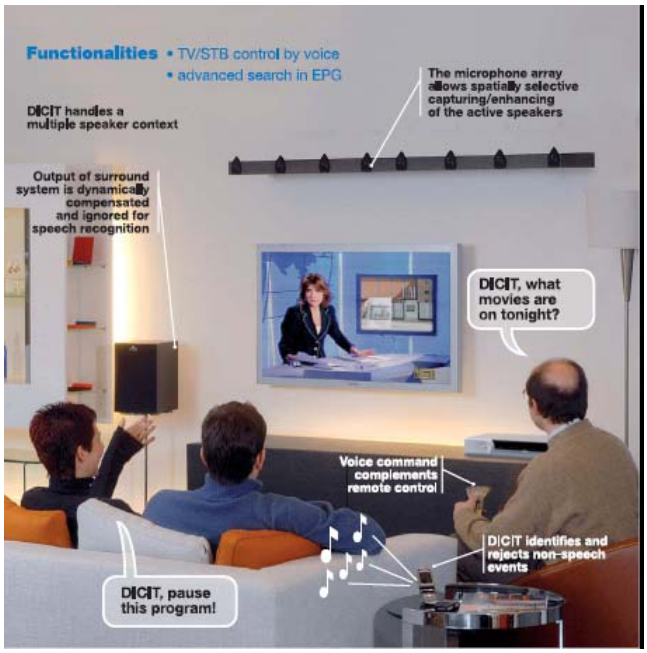

Fig.1 Interactive TV scenario of the DICIT project.

The dialogue with the system will be carried out allowing interchangeability between spoken commands and use of the remote control. A progressively increasing complexity of language will be adopted as the project advances (with larger vocabulary and variety of sentences recognized by the system). At the final stage one can think of a system capable of understanding and managing as complex requests as "Tell me what cartoon movies are scheduled for tomorrow afternoon" or "Please give me a list of sport shows being transmitted now" or "*This* is fine".

3 The multi-microphone front-end

A general block diagram of the DICIT prototype front-end system is shown in Figure 2. Describing the entire system architecture goes beyond the purpose of this paper; more details and information about the project can be found at http://dicit.fbk.eu.

The input to the front-end consists of the TV audio channels as well as of the acoustic signals (coming from a harmonically nested microphone array) which are acquired, pre-amplified and A-to-D converted by a multichannel soundcard.

The first step of the front-end processing is beamforming the microphone array input signals. At the same time, the acquired input data are made available to the speaker localization module that processes them in parallel.

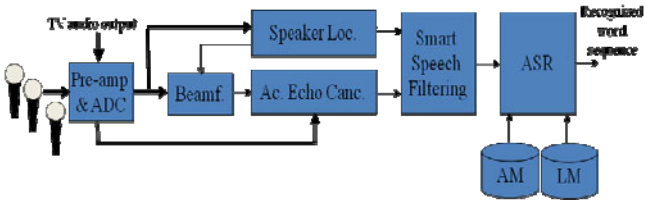

Fig.2 Multi-microphone front-end and ASR.

The speaker localization module (see the next section) is in charge with determining the position of the currently active speakers. The module receives as input the signals acquired by the nested array. The output includes the azimuth and the coordinates estimations, but it can be easily extended with any other information that may be useful to the other modules, as for instance the level of plausibility of the estimation.

The beamforming unit exploits the spatial distribution of sources and interferers in order to attenuate the latter. By means of filtering and adding up the given microphone signals, a beam with an increased sensitivity in the 'look direction' is formed. This 'look direction' depends on the information of the source localization unit, so that it is possible to track movements of the source.

A multichannel AEC real-time system is then used, which was developed in the past at FAU-Erlangen Nuremberg University (Germany) and is suitable for typical interactive home entertainment systems [5]. The AEC component uses the information from the available loudspeaker signals to suppress the residual acoustic feedback from the loudspeakers within the beamformer output signal. The AEC filters therefore have to adaptively model the system consisting of the loudspeaker-enclosure-microphone combination and the following time-varying beamforming. The output of the echo canceller should contain no echo anymore and serves as the major input for the Smart Speech Filtering (SSF) unit.

3.1 Speaker localization

Acoustic source localization based on Time Difference Of Arrival (TDOA) and triangulation represents one of the most common method to locate the position of a given speaker. GCC-PHAT is a widely used technique [1,3] to derive a TDOA estimate for each microphone pair. However, this approach is often not robust enough in adverse acoustic situations, characterized by reverberation, reflections and occlusions of the direct path between source and microphones.

On the other hand, spatial maps in the form of Global Coherence Field (GCF) [3] (also known as SRP-PHAT [1]) and Oriented GCF (OGCF) [6], are very effective representations for the given target. Both GCF and OGCF are composed by exploiting not only the maximum peak of generalized cross-correlation, but the whole GCC-PHAT based coherence measure at any time lag. GCF and OGCF based techniques have been widely adopted to tackle the SLOC problem when limited to a single source.

When two, or more, sources are simultaneously active, it was observed that most of the time the coherence map presents two, or more, evident peaks in correspondence of

the sources. However, searching for two local maxima may fail in the given context.
In fact, depending on the spectral contents of the involved speech signals, the main peak jumps from one source to the other while the second peak may be considerably lower than the main one and may be overtaken by spurious peaks. Figure 3a) shows an example of a map when two sources are active. In this case the two sources, denoted by the circles, are on the left and on the right of a linear microphone array that is placed in the upper part of the picture. It can be observed that most of the coherence concentrates around the speaker on the right, while the peak on the left is quite smooth.

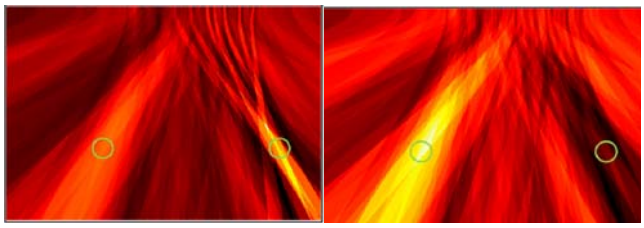

Fig.3 a)-b) Example of GCF map in presence of two sources. Bright colors represent high value, while dark colors identify low values. Notice how the presence of the main peak in the left map tends to compress the contrast of the remaining part of the map. The right map shows how the secondary source has gained evidence thanks to the de-emphasis of OGCF.

Recent research activities under DICIT led to the development of a new acoustic map based technique to address the problem of localizing multiple speakers. It operates in two steps and attempts to highlight the weaker source by masking the main peak in the initial acoustic map, as shown in Figure 3b). More details on the technique can be found in [7,8].
The algorithm has proved to provide satisfactory results in the given DICIT scenario. Experiments were restricted to the case of two simultaneous speakers. It is worth noting that in case the number of speakers is not known, the amplitude of the secondary peak may be exploited to check whether a second source is possibly active or not. Potentially, the same solution can be extended to the case of three or more simultaneous speakers, even if the discriminative power of acoustic maps decreases as the number of sources increases.

3.2 Smart Speech Filtering

The aim of the SSF component is to process the continuous audio stream as well as other information coming from Speaker Localization and AEC modules in order to provide the ASR with speech chunks to be processed. Moreover, the role of SSF is to discard any non-speech event including background noise and other possible interferences. To this purpose, the SSF component comprises an activity detector aimed at providing the boundaries of a given speech utterance (or acoustic event) with a high accuracy. To reinforce the tracking of the primary speaker (e.g. the user) both a speaker ID and a speaker verification components are deployed. The latter ones represent tough problems in a difficult context as that of distant-talking interaction, due to the loss of speaker voice characteristics which can be observed in far-field speech recordings. Speaker verification is expected to be useful in authentication before having access to the STB-related services. Speaker ID can be usefully exploited in the TV-related scenario together with speaker localization and acoustic event classification in order to provide an accurate acoustic scene analysis.
Any speech sequence is eventually associated to one of the family members, or even rejected by classifying in this way the speaker as a possible interferer.
Once the acoustic event detection as well as the classification and speaker identification processes have been accomplished, the SSF eventually sends a speech signal segment to the ASR, whose role is that of providing the next module of natural language understanding with one or more recognized word sequences.
Some experiments on the use of the above mentioned techniques of acoustic event classification and speaker identification/verification are reported in [8].

3.3 Distant-talking ASR

The distant-talking ASR component is based on Hidden Markov Models (HMM). To train acoustic models, contaminated corpora are produced for the given environmental acoustics and for the addressed three languages (namely English, German, Italian), by following a procedure similar to that described in [9].
It consists of the following steps:

1. measurement of the impulse responses from the positions of the active speakers and the TV loudspeakers to each microphone;
2. recordings of the typical background noise with the actual microphone setup;
3. convolution of clean corpus signals (e.g., TIMIT for American English) with the impulse responses of the various speaker-microphones pairs;
4. application of multichannel pre-processing (e.g. delay-and-sum beamforming);
5. application of the echo canceller on the resulting simulated noisy speech sentence with the simulated TV audio as reference signal.

In other words, AEC processing is also applied to contaminated speech in order to take into account possible effects due to front-end pre-processing. Finally, acoustic model adaptation is applied to further reduce mismatch between training and test conditions.
More details and preliminary experimental results on these issues are reported in [8].

4 Acoustic WOZ data collection

In order to develop a reliable acoustic multi-microphone front-end and in an effort to characterize the user behaviours in such a context, a set of Wizard of Oz (WOZ) experiments was also conducted.

Figure 4 shows the layout of one of the two rooms used for the given experiments.

Each WOZ session included three naïve users and one supervisor (co-wizard). Although all four participants were simultaneously present in the room, only one person at a time was allowed to interact with the system in order to complete specific tasks. A hidden human operator, called wizard, provided the speech recognition and interpretation capabilities and reacted to user inputs.

Each of the twelve given sessions was split in two phases.

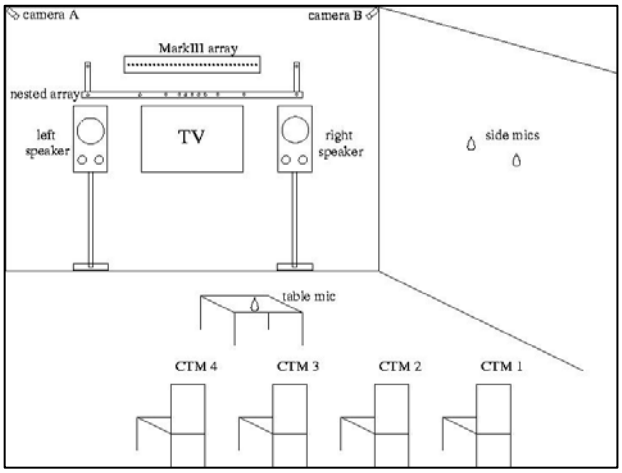

Fig.4 Layout of the FBK room used for WOZ experiments.

At the beginning, all the participants were sitting in front of the television and read out some phonetically rich sentences that may be exploited to train algorithms for speaker identification and verification. 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 controlling, finding specific pages in the teletext using only voice-commands, and so on. During the session, the subjects were allowed to move around in the room.

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.

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

The resulting database, comprising three languages (English, Italian and German), is being used for the performance evaluation of front-end components as speaker localization, speaker verification and identification, acoustic event detection, as well as of distant-talking ASR components described in the previous sections. In particular, six sessions (in Italian) were manually transcribed and segmented at word level, introducing also specific labels for acoustic events as well as information about ground truth of speaker positions.

More details about this corpus can be found in [10].

It is also worth noting that another WOZ data collection was devoted to the study of user-interface interaction as for the dialogue system design is concerned.

5 Future work

The purpose of this paper was to provide a partial overview of the activities being conducted under the DICIT project. Thanks to all the efforts so far devoted to this project and to the work that is under way as to the development of other components (natural language understanding, spoken dialogue, etc.), first DICIT prototypes have been integrated recently, which allow real-time TV control. One can find some video-clips with examples of that interaction in the project web-site.

Next activities will regard improvements of the given state of the art as far as all the technological components are concerned. The forthcoming prototype evaluation campaign will better show the directions to take as to increasing of system robustness.

In particular, 5+1 acoustic echo cancellation, blind source separation (also for multiple speaker localization), speaker identification and localization joint algorithms, joint optimization of the multi-microphone front-end and the ASR unit, self-calibration of the microphone network are some examples of acoustic information processing techniques which will be explored during the next phase of the project.

Acknowledgments

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

References

[1] M. Brandstein and D. Ward, “Microphone Arrays”, Springer-Verlag, 2001.

[2] J. Benesty, S. Makino, and J. Chen, “Speech Enhancement”, Springer Verlag, 2005.

[3] R. De Mori, “Spoken Dialogues with Computers”, Academic Press, 1998.

[4] X. Huang, A. Acero, and H.-W. Hon, “Spoken Language Processing: A Guide to Theory, Algorithm and System Development”, Prentice Hall, 2001.

[5] H. Buchner and W. Kellermann, “Improved Kalman Gain Computation for Multichannel Frequency-Domain Adaptive Filtering and Application to Acoustic Echo Cancellation”, Proc. of ICASSP 2002, pp. 1909-1912.

[6] A. Brutti, M. Omologo, and P. Svaizer, “Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays”, Proc. of Interspeech, pp. 2337-2340, 2005.

[7] A. Brutti, N. Omologo, P. Svaizer, “Localization of Multiple Speakers based on a two-step Acoustic Map Analysis”, Proc. of ICASSP, pp. 4349-4352, 2008.

[8] http://dicit.fbk.eu , Public Deliverables.

[9] M. Matassoni, M. Omologo, D. Giuliani, and P. Svaizer, “Hidden Markov model training with contaminated speech material for distant-talking speech recognition”, Computer Speech & Language, Volume 16, Number 2, April 2002, pp. 205-223(19).

[10] A. Brutti, L. Cristoforetti, W. Kellermann, L. Marquardt, M. Omologo, “WOZ Acoustic Data Collection for Interactive TV”, Proc. of LREC, 2008.