1 Publishable summary

1.1 Introduction

The DIRHA (Distant-speech Interaction for Robust Home Applications) project was launched as STREP project FP7-288121 in the Commission's Seventh Framework Programme on January 1st, 2012, with a duration of 36 months. The Action Line of the related call was FP7-ICT-2011-7 Objective 4.2 – Language Technologies.

During the three years, the project achieved several results and progressed consistently both at scientific and at application-oriented level, if compared to the state-of-the-art from which the consortium started in January 2012. The remainder of this document aims to highlight the main goals and achievements.

It is also worth noting that at the end of the first and second years the project was evaluated at the EC review meeting as in good progress, and it was confirmed to continue just with very minor modifications to the original work plan.

1.1.1 General goals

The DIRHA project addressed the challenge of natural spontaneous speech interaction with distant microphones in a home environment.

The main fields on which research was conducted, and for which suitable solutions can be identified and embedded in real-time prototypes, are: multichannel acoustic processing, distant speech recognition and understanding, speaker identification/verification, and spoken dialogue management.

The project also aimed to investigate the use of a new type of acquisition device consisting of MEMS (Micro Electrical-Mechanical System) digital microphone arrays.

The project addressed four target languages: Italian, Greek, Portuguese and German.

An intermediate prototype was foreseen during the second year, to install in a real automated home, located in Trento, and operate in the Italian language.

The final prototype was conceived to interact in the given four languages, and to be integrated and evaluated by real end-users in their automated homes.

1.1.2 Facts and Figures

Project details	
Project Reference: 288121	**Contract Type:** Specific Targeted Research Project
Start Date: 2012-01-01	**End Date:** 2014-12-31
Duration: 36 months	**Project Status:** Execution
Project Cost: 4.85 million euro	**Project Funding:** 3.45 million euro

Coordinator	
Contact Person: Name: OMOLOGO MAURIZIO Tel: +39-0461-314563 Fax: +39-0461-314591 Email: omologo@fbk.eu	**Organisation:** Fondazione Bruno Kessler (FBK) VIA SANTA CROCE 77 38100 – TRENTO, ITALY

Other Beneficiaries	**Country**
ATHENA RC-IAMU	GREECE
DOMOTICAREA	ITALY
INESC-ID	PORTUGAL
NEWAMUSER	ITALY
ST Italy	ITALY
TU Graz	AUSTRIA

1.2 Scientific, technical, and application oriented objectives

Several scientific/technological challenges were tackled within the DIRHA project. Most of them related with the problem of modifications encountered by acoustic waves propagating from a sound source (e.g. the speaker) to a set of far-field microphones. Such modifications have a negative effect on the performance of a speech recognition system or, more generally, of a system for acoustic event interpretation.

One of the most challenging and innovative aspects of the DIRHA project is the development of a distant speech interaction system, robust to speaker position, even in a noisy and reverberant environment and eventually in a multi-speaker context. In the past, other projects addressed this concept, and tried to realize some early solutions. However, DIRHA investigated on a novel approach and on techniques for distant-speech interaction in a multi-room environment, and possibly with multiple users.

Among the most relevant innovative aspects, it deserves to be mentioned that acoustic scene analysis is performed in an "always listening" mode (i.e., without the need of any push-to-talk button), with the goal of interpreting acoustic/speech activities concurring in the given environment, and eventually delivering speech chunks to the recognition and understanding components. To this end, one needs to realize robust technologies able to tackle unforeseen acoustic environments and noisy conditions. Such goals were new and far beyond the state of the art, not only for an application in the home scenario but also for other domains.

The targeted application included voice-enabled interaction with appliances and other automatic services available in a household. Although in some cases users could simply try to speak close to the microphone and in a rather controlled way, the expectation is that in the future they would require being able to interact at four-five meters from microphones in a crowded room, with music playing, and other possible active sound sources. For some individuals (e.g. motor impaired), this is a strong immediate requirement, which is the main reason for addressing firstly this category of users under the DIRHA project. To this purpose, a group of possible end-users was involved from the beginning of the project, in order to define concrete and realistic user requirements. A target was to integrate the most advanced technologies resulting from the project in a real-time prototype installed in automated homes, and daily used by the end-users for evaluation purposes. This task is going on, with real installations almost ready to be experimented by end-users.

1.3 Expected results

The DIRHA project aimed both to make advances at research level in the given scientific fields and to progress at technological level, with the development of a proof-of-concept system, which can represent the starting point for a next exploitation action to be addressed by the involved industrial partners.

Research activities also included the creation of experimental tasks and corpora, which might also enable future initiatives of dissemination and benchmarking at international level.

The main target at the end of the project was defined as a final prototype that runs based on microphone devices installed in different rooms in order to monitor selectively acoustic and speech activities observable inside any space of the household. In the targeted scenario, the user can speak from any position in space, i.e. any point in any room of a house given any background noise and acoustic conditions typical of a household, and no matter of where the closest microphones are. A spoken dialogue session can be activated based on a user request, for instance in order to switch on

the light, open a door, have access to appliances and devices, or to services regarding emergency situations.

1.4 Impact

The final objective of the project targeted application of automatic speech recognition in four languages with common multi-microphone front-end, spoken dialogue management, and user interface. This had a relevant impact in terms of synergetic approach to the development of spoken language interaction systems and to the immediate evidence of a possibly easy portability to other languages. The project represents a milestone for developers and integrators of home automation systems, since the targeted prototype can be seen as a first proof-of-concept realization in a real world context, based on concrete and realistic user requirements and operational constraints.

The DIRHA consortium aimed to examine the impact of its novel technologies primarily with collaborative users. In other words, the DIRHA system was conceived for subjects who have, in principle, no difficulty in understanding the way to access the system in order to obtain the highest satisfaction (e.g., based on a high completion rate in the proposed tasks) and who have a very good attitude towards this experimentation. Once the basic technology has been established and evaluated as reliable, other categories of users (e.g., elderly people) may be addressed in future projects.

Another impact of the project regards the portability of the foreseen solutions to other possible domains. In fact, the DIRHA approach and the resulting technologies could eventually be applied to several application contexts characterized by noisy environment and by the need of talking far from the microphone as, for instance, robotics, surveillance, telepresence, gaming, industry sector and manufactory.

1.5 Achievements of the project

1.5.1 General achievements during Year 1

- A **User study** was conducted during the first months of the project in order to analyze requirements of a set of motor-impaired end-users who are available to experiment the DIRHA prototype. Following this activity, a set of scenarios and related functionalities were outlined, which also represented a guideline for the definition of the user interface to adopt. In particular, it emerged that functionalities as doors, windows and rolling shutter management and lights, temperature and entry-phone/interphone control are the most important features to be supported by voice for motor-impaired people. Based on this, during the last months of Year 1 Wizard-Of-Oz experiments were conducted to test the usability of a voice interface aimed at managing the above mentioned devices and other ones chosen by the users. In the meantime, the WoZ experiment allowed us to collect spontaneous speech of 11 subjects while giving voice commands to the simulated system.
- A first set of **Experimental tasks** was defined to support all the experimental activities which are being conducted under scientific and technological work packages. Both simulated and real acoustic and speech corpora were created. Real data includes collection of distant-speech material as well as acoustic measurements in home environments. Simulated data were produced thanks to a technique that reconstructs, in a very realistic manner, multi-microphone front-end observations of typical scenes occurring in a domestic

environment. To this purpose, an ITEA (Istituto Trentino per l'Edilizia Abitativa S.p.A.) apartment is available in Trento; this apartment will also represent the site where DIRHA prototypes will initially be developed and tested.

- Preliminary research activities on **Multi-microphone front-end processing** were conducted. Some algorithms have been selected and tested to assess their performance in the real application scenario and to determine their most suitable combination for a prototype implementation. The target is to derive, starting from the signals of the available multiple microphones, an input of sufficient quality for an effective speech recognition process. The possibility of using novel solutions based on MEMS digital microphone arrays is also explored.
- As far as **Distant-speech and speaker recognition** is concerned, preliminary activities were conducted on the development and evaluation of baseline recognition components in the four languages of interest in DIRHA and of the baseline distant speaker identification and verification. In addition to baselines description and experimental results, some novel research activities have been started.
- As for **Speech Understanding**, two approaches have been investigated, the more traditional one based on the use of hand-crafted grammars and the other based on a statistical framework. In particular, the latter topic is being investigated towards the development of a robust system for speech understanding in order to overcome the limitations offered by the former approach. A first set of related tools has been developed to this purpose.
- Activities on **Spoken Dialogue Management** regarded the definition of the project-wide strategy for the handling of Concurrent dialogues, each one taking place in separated spaces of the house, the design and development of a reusable Concurrent Dialogue Manager based on the StateCharts paradigm, and the Modelling of the User+House State and Profile.
- A significant effort was devoted to define the hardware setup to be used and the **System architecture** for the next development of the intermediate DIRHA prototype which will run in a real environment fully connected to an automated home (i.e., ITEA apartment).
- Nine **Showcases** were realized, which represent proofs of concept for what regards the state-of-the-art technologies on multi-microphone front-end processing, ASR and dialogue management, available at DIRHA partner sites. Some of the showcases refer to targeted functionalities for the intermediate prototype. In particular, a showcase was implemented in the ITEA apartment and does already support real-time execution of speech input requests for command-and-control of some devices.

1.5.2 General achievements during Year 2

- In March 2013, the **Year 1 Review** meeting was held in Luxembourg. The decision of the EC Commission was to continue the project with minor modifications. After the review, the project web site was updated including all the **Public deliverables** produced during the first year, which are available for download.
- Based on the user study conducted during Year 1, and on the following activities characterizing the application scenario, the **User interface** and the dialogue flow of the intermediate prototype have been defined.
- A second campaign of acoustic measurements and speech collection has been conducted at different partner sites. The following activities led to the creation of a database of multi-channel impulse responses referred to different environments, and to the realization of a **Multi-language multi-microphone DIRHA corpus** of real and simulated data multi-channel impulse responses for different rooms of different environments. A portion of the

latter corpus is available free upon request, and has been recently distributed to labs working on speaker localization and speech detection.

- An **Experimental Matrix** framework has been defined by the consortium in order to enable the process of evaluation of a technique, and of the related algorithm, when combined with other ones in a multi-microphone processing chain that ranges from early multi-channel signal processing steps to speech recognition and understanding. A set of experimental tasks has been defined, together with a standard data exchange format, in order to compare different algorithms aimed to address a given problem (e.g., speech enhancement). The Experimental Matrix framework can represent another output of the project that may be disseminated in the related scientific community together with the multi-microphone corpora mentioned above.
- Research activities on **Multi-microphone front-end processing** of the second year led to significant achievements on the different topics addressed by the scientific partners. In particular, the most promising techniques selected during the first year for acoustic source localization, source enhancement, event detection and classification were evaluated on data representative of the addressed scenarios (DIRHA-II simulated corpus). Algorithms for acoustic source detection, localization, and for the cancellation of known sources were also tested on real multi-channel data acquired in the real apartment. Performance of beamforming plus postfiltering was assessed both on simulated data and on real data acquired with a MEMS array. Acoustic feature enhancement was further investigated using different compensation techniques. Concerning speech activity detection, efforts concentrated in the utilization of multi-microphone input for better detection of speech segments within the reverberant and noisy DIRHA environment. A first set of real-time components (acoustic event detection, speech-non-speech discrimination, localization) has been already integrated in the intermediate prototype and is being tested directly in the real apartment. It is also worth noting that during Year 2 a new MEMS array based platform has been delivered and used to collect acoustic data and embed early solutions of multi-microphone front-end processing.
- Research activities on **Distant-speech recognition** led to consolidate and improve baseline algorithmic components that were developed for the various addressed tasks during the first year. Novel robust techniques were developed that exploit the multi-channel network available in the DIRHA scenario for speech and speaker recognition purposes. It is worth noting that a significant effort was devoted by the consortium in order to create a common framework for the four targeted languages, based on the use of the above-mentioned DIRHA multi-language corpora, for comparable analysis and evaluation of the effectiveness of the proposed methods. This is an important progress, compared to what was done during Year 1, and makes it possible to also derive an experimental matrix framework, in which the performance of a given front-end processing technique can be measured, in a consistent way, in terms of recognition accuracy in the four languages. This activity allowed comparing baseline systems in the four languages of interest and analysing the impact of the different characteristics of the training data.
- As for **Distant-speaker recognition**, besides robustness to the environmental noise and reverberation, a second aspect that was addressed during Year 2 regarded the assessment of the impact of similar voices (e.g., from relatives). Based on the experiments so far conducted, the latter issue does not seem to be critical, at least in the application scenario addressed by this project.
- **Speech understanding** represents another issue on which important efforts were devoted during Year 2. A part of the activities focused on refining hand-crafted grammars to use for the development of the intermediate prototype, which represent the traditional approach. Based on the integration of the FBK recognition engine, it was possible to handle recursive

networks and finally produce a semantic parsing during the decoding phase. The other activities concerned an alternative approach based on semantic role labelling, which was addressed by developing a new semantic parser for the Italian language that solves the issues due to the lack of robustness of the syntactic processing when applied to the output of the ASR system. Preliminary results showed the advantage of this novel approach in terms of robustness to ASR errors. Some activities are under way to eventually localize the system on the other three languages of interest under DIRHA.

- The activities on **Concurrent Dialogue Management** were devoted to implement powerful data modelling, needed to support house (and user) profile and state, and other sophisticated mechanisms that could simplify the implementation of advanced services. The component has been integrated in the intermediate prototype, and is combined with a user interface handling 5 devices (i.e. doors, windows, light, shutters and heaters) of the ITEA apartment, even supporting interactions that can occur in more than one room. House profiling has been developed in a house-independent way, exploiting the data modelling and navigation capabilities of the CDM engine.
- During Year 2, the **Architecture** of the **Intermediate prototype** has been finalized and the prototype has been completed and integrated with the home automation of the ITEA apartment. The system runs in a satisfactory way, supporting for the moment only Italian language with a very good recognition accuracy, and interacting with users in a prompt and efficient way for the control of the above-mentioned devices.
- Three **Showcases** have also been realized, which regard the following technologies: speech understanding; close-talk speech recognition in Greek; joint acoustic map computation and speaker identification.

1.5.3 General achievements during Year 3

- In March 2014, the **Year 2 Review** meeting was held in Trento. The decision of the EC Commission was to continue the project with minor modifications.
- The **Intermediate prototype**, completed at the end of Year 2, was evaluated at the ITEA apartment, involving 12 motor-impaired end-users who experimented the system in realistic conditions, and provided an overall positive feedback and many useful suggestions.
- During Year 3, the **Collection and annotation of speech corpora** represented a particularly intense and productive activity, with high quality data sets created at different partner sites, under different conditions (e.g., for studies on acoustic echo cancellation and barge-in based iteraction), in different languages (e.g., now including UK and US English). Some portions of these corpora have already been made available publicly (e.g., DIRHA-GRID), and other ones will soon be distributed to the international scientific community.
- A new **MEMS array** prototype was realized by ST-Italy, that allows acquisition, decoding and streaming of up to 16 MEMS microphones. The device was also used, arranged in a harmonic geometry, for the English data recordings.
- Research activities on **Multi-microphone front-end processing** of the third year led to significant achievements on the different topics addressed by the scientific partners. New multi-channel front-end processing techniques were developed for what concerns acoustic source localization, echo cancellation, source enhancement, event detection and classification tasks. Most of the activities were conducted evaluating the techniques on both real and simulated DIRHA corpora, and based on simple combination tasks as addressed during the previous years through the experimental matrix framework. Moreover, the work on acoustic echo cancellation led to select the most suitable real-time algorithm then embedded in the final prototype, for barge-in based interaction in situations of known sound propagated by the multimedia player.

- Research activities on **Distant-speech/speaker recognition** led to further improve baseline algorithmic components developed for the various addressed tasks during the first and second years. Acoustic models for the four addressed languages were created in a coherent way, which were then deployed in the corresponding real-time prototypes. Different combinations of processing components were investigated, as planned, also for what concerns the keyword-spotting task. A specific focus of the activities was devoted to the robustness of the proposed solutions, which was also explored studying new front-end features and related normalization. Moreover, different techniques for system adaptation to the speaker and the environment as well as multi-microphone based speaker recognition were further developed.
- During Year 3, **Speech Understanding** for the final prototype has been completed, localizing the hand-crafted grammars for Austrian German, Greek, and Portuguese languages.
- For each of these three languages, a **Prototype** is now available in the ITEA apartment, derived from the intermediate prototype formerly running only in the Italian language.
- The **Dialogue Manager** component has been re-implemented starting from a pre-existing technology available at FBK. This work led to a new concurrent dialogue management component suitable to be integrated in the DIRHA systems. Based on it, additional functionalities (e.g., barge-in and audio player management) to the intermediate prototype were developed and installed in the final prototype.
- This **Final Prototype**, operating in the Italian language, is now available in the ITEA apartment. It will also be experimented and tested soon by motor-impaired end-users in their automated homes. This prototype, together with other prototypes and showcases (e.g., a MEMS array based distant-speech recognizer working in Austrian-German and European Portuguese) realized at different partner sites, represent the most relevant achievements of the DIRHA project from the technological point of view. They are proof-of-concepts that confirm the appropriateness of the vision outlined four years ago, which characterized the original project proposal, as well as the concreteness and exploitability of the project results from an application oriented perspective.

1.5.4 Dissemination during the three years of the project

- During the first year, an official and publicly available project web site was established (see http://dirha.fbk.eu), which represented the main means for presentation and dissemination of project achievements inside and outside the consortium. It includes the most important information regarding the project allowing the download of publicly available documents (ex. the DIRHA brochure, state-of-the-art documents, etc.). The DIRHA website has been continuously updated in order to give always a global and timely overview of the activities and the results during the project. Furthermore, it has received a large number of visits (from about 1140 different visitors during the third year).
- General information about the project and promotional material (e.g., the DIRHA brochure) were made available to a wide range of specialists with different disciplines both inside scientific-technical communities and in the home automation related industrial field. In particular, the project was presented at LREC 2012 - EU Village in Istanbul, at META-FORUM 2012 in Brussels, at trade shows such as CES 2014, MWC 2014, Embedded World 2014, Electronics 2014, and at various other events during 2013 and 2014.
- The project was presented with invited lectures, talks and other scientific events (e.g., DAGA 2013, PanEuropean Researcher's Night 2013, PanEuropean Researcher's Night 2014, Athens Science Festival 2014). DIRHA was also disseminated to other EU project consortiums. Some papers have already been published and presented at international conferences and workshops in the field of signal processing and speech technologies (e.g. ICASSP, Interspeech, EUSIPCO, HSCMA, LREC, MEDIAEVAL), and journals, while other ones have been submitted and are currently under review. During the whole duration

of the project, three (3) journal publications were accepted and published, one (1) patent was filed and sixty-three (63) conference publications were accepted and presented in major conferences.

- DIRHA has also been described outside the related scientific community, with articles (e.g. see Speech Technology magazine, Spring 2013 issue), news websites, TV interviews, YouTube videos, and to associations related to industry and to disabled people.
- Public deliverables D7.1, D7.2, and D7.3 were produced, which provide further information about the project and its dissemination actions.
- The DIRHA consortium is finally very active in promoting the project through the participation to challenges (e.g., 2nd *CHiME* Speech Separation and Recognition Challenge) and through the dissemination of experimental tasks and corpora in the scientific community (e.g., DIRHA-GRID, data set for HSCMA 2014). During the last year, dissemination actions led to organize different DIRHA events, e.g., special sessions, tasks and workshops, related to EUSIPCO 2014, HSCMA 2014 and EVALITA 2014. A subset of the DIRHA corpora is currently available for free access to the community, and has already been downloaded by several labs outside the consortium. Other material will be made available publicly during 2015, with the aim of becoming a reference for the related international scientific community.