Speech Operated Smart-Home Control System for Users with Special Needs

A. Vovos 1, B. Kladis 2, N. Fakotakis 1

1 Wire Communication Laboratory (WCL), University of Patras, Greece
2 Knowledge SA (LogicDIS Group), Patras, Greece

Abstract
This paper presents a speech operated smart-home control system with a focus on users with special needs – elderly people, people suffering from blindness or low vision and people with mobility impairments. The proposed system simplifies the use of various home appliances by providing a unified speech-controlled interface and thus improves the quality of their everyday life. The paper presents the implementation scheme and decisions taken in order to have a real-time operational system as well as real users’ experimental results.

1. Introduction
Nowadays, modern houses are equipped with a set of household appliances ranging from simple lights to feature-rich hi-fi systems, DVD players and HDTV television sets. Usually these devices are controlled by buttons placed on them, on wall switches or remote controls. A major subject of research for these interfaces is how to simplify and design them in a user-friendly way. Thus, the design of the interface is considered successful if the average user can operate the device easily through it. Although this design approach is widely used, it does not guarantee that the device would be acceptable or even usable by all the possible users. People with special needs are not considered average users and usually face major difficulties operating household devices. These users often are not able to move easily or move at all to the vicinity of the device’s interface. Also, they may not be able to press the interface’s buttons or, in the case of users suffering from sight impairment, they may not be able to differentiate the controls. Additionally, the complexity of these devices might be overwhelming for users not accustomed to modern technology, users that just want to perform simple tasks. A lot of elderly people belong to this category of users. On top of all that, there are no common standards agreed between the various manufacturers on how these interfaces should be designed and built, so they present considerable differences even between appliances of the same kind and the same manufacturer. So, even if a user becomes accustomed to use one of these interfaces it is not guaranteed that he/she will be able to operate a different one.

This paper presents a speech operated smart-home control system with a focus on people with special needs that serves as a common interface between them and their home appliances. Since speech is the fundamental communicative and interacting mean for humans, it was considered as the user-friendliest and most natural interface: The user just has to say what he/she desires and the system will orchestrate the home appliances in order to fulfill the specific wish, just like a human assistant would. Using wireless microphones or microphone arrays to acquire the user’s spoken commands, the system is operable regardless of any mobility disabilities the user may be facing. The system does not require the user to press any buttons since the microphones are always active and the command spotting task is handled strictly by software using various filtering techniques. If a home appliance is replaced or a new one is added, as soon as it is installed, the user can start immediately to use it because there is no need for button interface familiarization. In addition to that, the system is responsible for handling and operating properly the home appliances. Thus the user can utilize them without the stress of misusing and rendering them inoperable. This feeling of assurance combined with the user-friendliness that a speech interface offers, can help people with special needs to include to their everyday life home appliances that seemed complex and out of reach till now. Finally, the system is designed taking into account that people with special needs should not be confined in any way to the vicinity of their homes, so a telephony interface was implemented enabling full remote access to the system through the telephony network.

In order to create a user-friendly system that would receive a widespread use, a number of restrictions were imposed. The user of the system wants to provide simple verbal commands to the system without having to remember or memorize any specific list of predefined phrases. In case the user does not provide to the system all the information required to execute a task, it should be flexible enough to form an appropriate question to the user requesting the crucial missing data. All the dialogues with the system should be unconstrained and should create the impression that the system is a friendly helpful entity ready to accommodate the user’s needs and desires. High recognition results are required in order to achieve user satisfaction and avoid the system’s rejection. Audio feedback should be provided during system interactions to enhance usability especially for users with sight impairment. Last, but certainly not least, the system must be real-time and respond in fractions of time so small that would be unnoticed by the user.

The present research is based on modifications and enhancements to the EC project INSPIRE [1]: INfotainment management with SPEech Interaction via REMote-microphones and telephone interfaces IST-2001-32746.

2. System Overview

Fig. 1 shows the proposed system architecture. The system is divided into components each assigned with specific tasks. The components implementations and interconnections were based upon detailed design descriptions and guidelines.
2.1. System Components Description

A description of each of the system’s modules follows:

- Two different microphone arrays were tested during the experiments. The first one was a custom made array of eight microphones that formed a frequency domain delay-and-sum beamformer [2]. Other algorithms were also tested but this was the only one fast enough to perform real-time beamforming. The second array was a commercial one [3]. Both beamformers provided comparable recognition results under the same conditions and input; 91.56% and 93.4% Word Accuracy Rate (WAR) respectively. Because of this slightly better performance but mostly because of its faster response time the commercial array was selected.

- During the experiments the Sound Speakers were placed outside of the beamformer’s scan range in order to minimize their interference to the system.

- The Sound Acquisition and Generation module is responsible for sound acquisition (sound card) and encoding to a compatible to the system format. It also plays back the prompts created by the System’s Responses module.

- The Computer Telephony Interface performs the same task with the Sound Acquisition module with the difference that it uses the Telephony Card instead of the sound card. It is also responsible for the full control of the Telephony Card and tasks such as incoming calls answering, hang-up detection and dialing.

- The Voice Activity Detection (VAD) module is an Energy-based VAD in the time domain [4]. The energy of the signal is compared with a threshold depending on the noise level. Speech is detected when the energy exceeds the threshold. In order to avoid clipping, a hang-over is added to compensate for small energy gaps in the speech and a number of frames preceding the detection are also included. The specific algorithm was selected because it is less computationally intensive and because of its advantages as mentioned in [5].

- The Speech Recognition Engine (SRE) is responsible for converting the spoken utterance into a written form. Based on grammars (Speech Recognition Grammar Specification as defined by the W3C [6]) and various search processes and probability calculations, it produces a word graph which represents the words that most probably represent the user's spoken utterance. A commercial SRE [7] was used for the recognition task. The specific one was selected because of its robustness, performance, response speed and configurability. The acoustic models were trained with Greek SpeechDat-II database of utterances and their associated transcriptions [8]. This database is a collection of Greek annotated speech data from 5000 speakers (each individual having a 12-minute session).

- The Speech Understanding component re-evaluates the word graph from speech recognition and determines which sequence of words represented in the graph is the most probable. This is done by searching for best matches according to a related grammar. In addition, the semantic meaning of the sentence is extracted. During the development phase the system was tested periodically by real users and the grammars were modified to incorporate requests that were not anticipated. Users’ utterances were annotated and new language models were created. Language models contain statistical information about which words or word sequences are used more often by the users of an application. Two different kinds of language models were used. The first kind of models reflects the fact that a given word is typically followed by a particular word (word bigrams). The second kind contains information about concept bigram probabilities (a given concept is typically followed by another certain concept) and rule probabilities (certain rule alternatives are typically used more frequently than other rule alternatives). The joint acoustical and language model probability is measured to find the best sequence of words for the recognition result. Using language models the recognition performance is improved and the recognition process itself is speeded-up. Results of the SRE and Speech Understanding combination can be seen at [9].

- The Speaker Recognition component detects if the current command origins from an authorized user. In
The Dialogue Control is the central decision-making system that processes user commands and generates system responses and questions. The communication module handles the connection between the system and the Appliances Control Server in order to control home-appliances and retrieve information concerning their status. Finally, it generates the system responses in textual form and forwards them to the System’s Responses module.

- The Appliances Control Server and the home appliances controlled were simulated. Users evaluating the system received visual feedback about the system’s actions through an animated room simulation that can be seen at Fig. 2. The appliances simulated and their functionalities can be seen at Table 1.

2.2. User commands Identification

One of the main tasks that the system should perform is to filter out all signals that are not valid user commands. A first stage of filtering is performed by the beamformer, restricting its scan range to the area where users usually are and relocating all the noise sources out of that area if possible. The second stage of filtering is performed to the VAD module and the third to the SR module where phrases not belonging to valid users are rejected. The SRE module also includes noise filters that have been proven quite effective especially against non-speech noise [9]. Finally, the Dialogue module is constantly monitoring the recognition results of all the phrases to check if the system was addressed. This is done by checking if the name of the system (“Socrates”) was uttered anywhere inside the phrase probed. The system was named “Socrates”, a rare name nowadays that makes a strong impression to the users in order not to be confused with real persons, to be remembered and to create a positive attitude towards the system. Also, it is a word that the SRE module can easily identify and spot out among others utterances.

3. Real Users Experiments

In order to test the usability and acceptability of the system a number of experiments were conducted. The experiments took place in a 6.75m x 4.9m x 3m room with reverberation time of 0.3 sec. The microphone array was placed at the middle of the long wall of the room at a height of 1.6m. Two groups of users were used, one with users with special needs and one with ordinary users, in order to study the different impact that the system might have. The special needs group consisted of 15 members where 5 were suffering from mobility impairment, 3 were blind and 7 were elderly people, while the second group consisted of 20 members aged 25 to 38. The only instruction given prior to using the system was to address the system by its name every time they wanted to perform a task. Each user interacted with the system two times, one based on a set of general scenario task-cards and one without. They were asked to rate 14 Statements (No. 1 to No 14) on a 7-point Likert scale with 1 meaning “strongly disagree” and 7 meaning “strongly agree”. The mean results and their standard deviation can be seen in Table 2.

4. Experimental results analysis

Users generally were pleased by their interaction with the system (statements No. 3, 4, 6, 8, 9, 10, 11 and 14) that most
of the times surpassed their expectations (statements No. 1 and 2) although it wasn’t too human-like (statement No. 7). Their majority would be keen to use it again (statement No. 13). The experimental results indicate that people with special needs are more willing to accept such a system in their everyday life since in most statements they exhibited a positive reaction exceeding the respective one expressed by the ordinary users. This observation does not apply to statement No. 5, mainly because elderly people were hesitant especially during their first interactions with the system. Ordinary users are somehow reluctant to use the specific system (statement No. 12). Their age (mean of 29.9 years) indicates that they are accustomed to modern technology and some even stated that they prefer remote controls because they grant faster interaction. On the other hand, users with special needs confront the system almost as a one-way solution.

5. Conclusions

We have presented a complete speech operated smart-home control system with a focus on users with special needs. The system could be easily manufactured and distributed as a commercial low-cost product since nowadays a personal computer can incorporate a lot of household devices (TV/radio, DVD, hi-fi, answering machine, etc) and control various home appliances using affordable home networking solutions. Also, the experimental results indicate that the system is quite usable and well received so its widespread use is quite probable. Considering these, e-inclusion for people with special needs concerning their everyday life at home seems closer than ever. According to the experimental results, the system needs further improvements in the direction of a more human-like interface. This could be done by incorporating an animated agent or a robotic assistant and it will be the target of future work.

6. Acknowledgements

The authors would like to thank all the partners involved in INSPIRE and dedicate this paper to the recently deceased Paula Smeee that contributed utmost to the fulfillment of the project. She will be dearly missed by all of us.

7. References