AN INTEGRATED SYSTEM FOR SMART-HOME CONTROL OF APPLIANCES BASED ON REMOTE SPEECH INTERACTION

I. Potamitis, K. Georgila, N. Fakotakis, G. Kokkinakis

Wire Communications Laboratory, Electrical and Computer Engineering Dept.,
University of Patras, 261 10 Rion, Patras, Greece, Tel:+30 2610 991722, Fax:+30 2610 991855
e-mail: potamitis@wcl.ee.upatras.gr

Abstract

We present an integrated system that uses speech as a natural input modality to provide user-friendly access to information and entertainment devices installed in a real home environment. The system is based on a combination of beamforming techniques and speech recognition. The general problem addressed in this work is that of hands-free speech recognition in a reverberant room where users walk while engaged in conversation in the presence of different types of house-specific noisy conditions (e.g. TV/radio broadcast, interfering speakers, ventilator/air-condition noise, etc.). The paper focuses on implementation details and practical considerations concerning the integration of diverse technologies into a working system.

1. Introduction

The ever-increasing complexity of home appliances and services, combined with the difficulties encountered by a great portion of the population to handle complex equipment including the inability of elderly and disabled people to use it, renders the creation of intelligent, intuitive and flexible interfaces that facilitate the human-machine interaction, an endeavor of great importance. As humans predominantly communicate via spoken language, a natural interface should allow the user to interact with the home environment directly using his/her voice requesting an agent to perform some action on behalf of the user. Natural access to the appliances should be ensured by the use of spoken dialogue that will allow inquiry regarding the current status of the appliances and their control, request of assistance on their use, etc.

This paper focuses on a speech-operated electronic-assistant installed in a PC that can handle the interaction between a speaker and home appliances. The challenges faced by the front-end of the system are related to the well-known acoustic adversity of home environments (with respect to ASR systems) using distant microphones. The assistant should be able to receive commands and interact with the speaker while s/he walks in the presence of competing speaker and music interference as well as stationary noises originating in home appliances.

Microphone arrays have received considerable attention lately due to cost effective DSP advances and the increasing need for hands-free applications [1]. Their ability to provide spatially selective speech acquisition by restricting their receptive field to desired talkers while simultaneously suppressing ambient noise and interfering talkers makes them a viable solution for the problem at hand. Beamforming techniques need an estimation of the direction of arrival (DOA). The DOA can be estimated by high resolution spectral estimation techniques like ESPRIT or MUSIC [2], by finding the maximum power of a steered response over a range of angles [3] and implicitly by first estimating the time delay of arrival (TDOA) between a set of microphones and then the DOAs [4].

The prime factors that place tight constraints on the design of the system are the limitations of current algorithms associated with DOA and TDOA estimation. The design of the system was carried out under some restrictive conditions primarily because of the many potential speakers, the fact that speakers can be moving while talking, and that the main source of interference comes from non-desired speakers and music. This kind of non-stationary interference shares the same statistical characteristics with the signal of the desired user. We present the limitations of beamforming in realistic room-environments in order to decide which restrictions can be relaxed while maintaining a good level of robustness.

a) The speaking scenarios and room layout should be as general as possible. Moreover, the system should not be dependent on specific environmental scenarios regarding the nature of noise and reverberation. The latter limitation entails that the vast majority of one-channel speech enhancement algorithms cannot be of much help when the background interference is a competing speaker or music because in general, the noise estimation algorithms and voice activity detectors cannot discern the useful signal from background music or speech.

b) The ability to resolve closely spaced sources is reduced in multipath environments either by using DOA or TDOA estimation. High resolution techniques are sensitive to modeling errors and reverberation [1, 2]. As regards TDOA estimation, under medium reverberation accurate techniques require long data segments to perform some kind of ensemble averaging (since this would require that the speakers remain at a fixed location while being active) making them unsuitable for moving speakers (see [3] and DiBiase in [1]).

c) The system must be able to respond in a robust way to many moving speakers. In the case of multiple moving speakers, repeated application of DOA or TDOA estimation does not yield tracking of speakers due to source ambiguity association (i.e. unresolved correspondences between consecutive measurements and particular speakers). The major problem though, with most localization techniques is that they are generally unsuitable for a multi-speaker environment [1, 3]. A TDOA estimation technique capable of fine resolution at a high update rate [4], in a real room over a distance of 3 meters returns many spurious estimates because there are no sufficient data to perform some kind of averaging. These observations can be smoothed by using Kalman filters and data association techniques [5, 6] but this approach is computationally very demanding and a still developing research area.

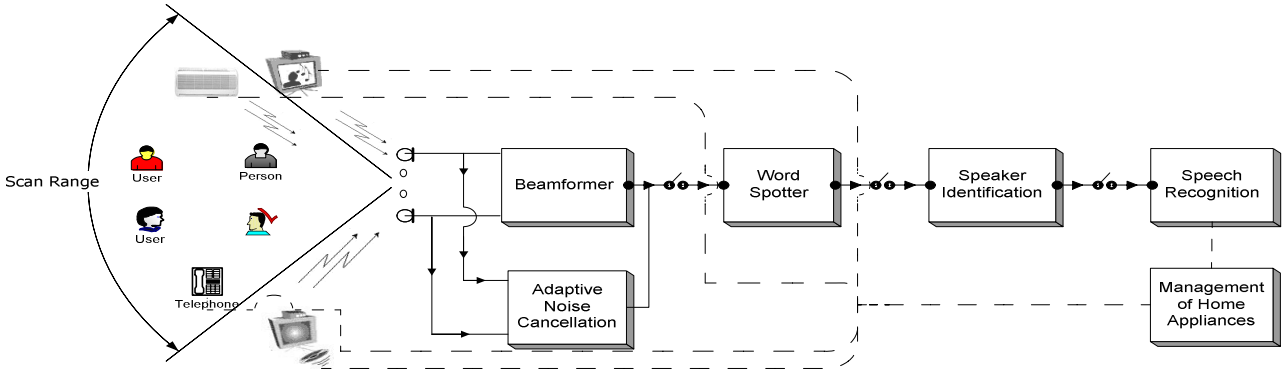

Fig. 1: Schematic diagram of the system.

d) The number of active speakers is usually assumed known or is calculated by employing variations of Akaike's information criteria or minimum description length [3]. However, in practice, the number of active speakers is not known and is also time-varying, while a series of experiments we performed in a typical room with 0.3 sec reverberation time has demonstrated a very poor performance of the algorithms that estimate the number of speakers from a limited number of speech frames.

e) Last (but certainly not least), the whole set-up must not be computationally demanding in order to induce a short processing latency to make it practical for real-time operation.

The present research is supported by the EC project **INSPIRE: IN**fotainment management with **SP**eech **I**nteraction via **RE**mote-microphones and telephone interfaces IST-2001-32746.

2. System Overview

2.1. Selecting the Voice Separation Strategy

The greatest part of interference inside a house that disrupts the message conveyed between a speaker and an ASR machine is directional and non-stationary. In our implementation, the basic noise and reverberation suppression function is supplied by the spatial selectivity of sound provided by a beamformer. We selected a Generalized Sidelobe Canceller (GSC)-based adaptive beamformer since it has a demonstrated efficacy in canceling directional interference even from non-stationary noise-types. The gain has been shown to be high for a variety of linear array configurations [9] and under reverberant conditions [8]. The greatest problem has been to deal with the multiple speakers case. The decision on how to design the system is mainly derived by studying the conversational attitude of people instead of applying voice separation techniques, namely:

- In a polite conversation the speakers participating are not simultaneously active, that is, most of the time a single speaker is active (though not the same speaker).
- In conversational speech, less than 20% of the overall frames include overlapping of speech [4, 7].
- In cases of overlapping speech the system should resolve the overlapping voices only when someone is claiming access to the smart-home agent otherwise the agent is indifferent to the overlapping of speech.

These observations led us to face the problem as a situation where the receptive beam switches to the speaker possessing the strongest SNR ratio and the effort is to correct the problems arising when one speaker is assumed when in fact there are more.

2.2. Composition of the System

The nature of the application entails that the recognition engine is active continuously. To prevent accidental activation/de-activation of appliances and a robust operation of the home automation system, a multiple-stage 'gating' of incoming signals (each stage functioning as a switch) is included in the design of the system as described in Fig. 1.

The first stage is a band-pass filter applied to each microphone channel restricting the frequency range from 300 to 3400 Hz removing the high-frequency spectral content of music signals and most of the energy of low-frequency noise signals emitted by appliances functioning at steady state (e.g. refrigerator, ventilator, computer fans, etc.). The band-pass filter is followed by a pre-emphasis filter $H(z)=1-0.97z^{-1}$ that greatly improves DOA estimation in reverberant environments. The band-pass filter and pre-emphasis are placed in the beginning of the process to assist the estimation process of the following stages. These filters do not induce any degradation to the speech recognition system since the far-end speech recognition system includes such filters in its feature extraction front-end.

The second stage is composed of an adaptive spatial filtering array with a prefixed scan range (i.e. $\pm 60^{o}$). DOA estimation schemes based on MUSIC and ESPRIT have a high computational cost due to the eigenvalue decomposition of the spatial correlation matrix. In our framework the DOA estimation is based on the minimum variance technique [2] from five 512-sample frames overlapping by 50% at 8 kHz sampling rate. If a source is detected outside the scan range of the array the microphone output is blocked, that is, a software switch turns off the microphones and the path to the recognizer is blocked. At this phase a general constraint is imposed on the array positioning so that the array is placed inside the room in a way that the loudspeakers of the TV set or audio equipment are placed outside its scan range. This practical procedure though simple in concept, greatly reduces the misclassifications of the recognition engine since the interference originating in TV or radio does not reach the recognizer as long as there is no speaker active inside the scan range. If however, the voice of a speaker is detected inside the scan limits by the DOA estimation and these appliances are also active then the interference cancellation load is transferred to the Adaptive Noise Cancellation (ANC) part of the GSC (see section 3).

The Word Spotter excludes accidental commanding of the appliances by constantly seeking a specific word (that of the

agent). As long as the keyword is not detected the propagation of the signal to the speech recognition module is obstructed. Characteristic examples of the permissible commands of the appliances are the following:

Agent open the light please.
Agent turn on the TV to channel four please.
Agent is the ventilator working?

One should have in mind that commands without the name of the agent preceding are ignored. This is due to robustness reasons. An appliance should not receive commands just because someone is talking about them. If the keyword is spotted, the waveform samples following the keyword are recorded and passed on to the speaker verification stage.

The Personalization Module consists of a Speaker Verification component and a User Modeling component. A speaker verification module cannot respond on a per frame basis, therefore, it is placed after the word spotting module and its decision is based on the chunk of the speech waveform recorded after the keyword of the agent is spotted. The speaker recognition module further restricts unauthorized access to the system and excludes part of the input signal that might have been skipped by the previous stages (e.g. music parts). In case of a positive decision, the signal is forwarded to the speech recognition module.

The User Modeling component decides which one of a certain group of registered users is currently interacting with the system, in order to respond properly (e.g. adjust the volume of the speaker output in the case of elderly people or ask a refining question in a different way depending on the idiosyncrasy of the speaker).

The Speech Recognition Module consists of a Continuous Speech Recognition component, which performs the main recognition process, using a Language Modeling sub-module and the appropriate dynamically-built lexicons. In this work the task is considered completed when the recognition engine returns the right transcription of the speaker's intention. In a larger scale implementation the appliances and the interface will all be integrated in a network that controls the home appliances through hardware (see e.g. [10)].

3. Implementation Details and Discussion

The GSC structure achieves greater voice separation than stand-alone beamforming since it includes an adaptive noise canceling configuration. The upper path of the GSC forms a fixed beamformer by filtering the array data through weights, which are designed to steer the directivity to desired directions. The lower path of the GSC contains a weight vector (blocking matrix) that forms a deep null in the direction of the sound source so that only the noise component passes which subsequently serves as a noise reference in the ANC stage. The GSC needs a reliable angle estimate otherwise part of the target signal leaks to the interference cancellation part leading to signal degradation. In a reverberant environment and due to the limitations presented in the introduction the angle estimation from few frames can vary significantly. We tested several implementations of the GSC like GSC with an Adaptive Blocking Matrix configuration where adaptive noise cancellation takes place before the blocking in order to face the leakage of the desired signal to the adaptive noise cancellation part. We also tried out various implementations in the frequency and time domain as well as block implementations [2]. Although these variations demonstrate an advantage under controlled situations (see Hoshuyama et al. in [1]) and small distance of the speaker from the microphones we did not observe any distinct advantage over a typical distance of 3 meters compared to a Recursive Least Square (RLS) of 14 taps in each frequency bin. However, we observed a distinct advantage of GSC compared to standard beamforming especially in the cases where speakers were standing still while talking, a thing that we attribute to the fact that the filters of the ANC had enough time to converge.

3.1. The Recognition Task

For the development of the word spotter and the recognition module that performs the main recognition task we used the HTK Hidden Markov Models toolkit [11] running simultaneously in different modes of operation. The basic recognition units are tied-state, context-dependent triphones of five states each. In order to train the acoustic models we used the Greek SpeechDat-II database of utterances and their associated transcriptions [12]. This database is a collection of Greek annotated speech data from 5000 speakers (each individual having a 12-minute session). We made use of utterances taken from 3000 speakers in order to train our system. Each frame processed so far is already in the Fourier domain due to the DOA estimation stage and is subsequently passed through a set of 20 Mel-spaced triangular band-pass filter-bank channels. Thirteen-dimension feature vectors are formed after applying DCT to log-filter-bank output which reduces the 20 output channels into 12-dimension Mel Frequency Cepstral Coefficients plus a log-energy value. Cepstral mean normalization is applied to deal with the linear channel assumption. The 13 aforementioned coefficients and their temporal regression coefficients of first and second order form the final 39-dimension observation vector.

As regards the operation of the word spotter, the language modeling is based on word networks defined using the HTK Standard Lattice Format (SLF). An SLF word network is a text file, which contains a list of nodes representing words and a list of arcs representing the transitions between words. Thus the SLF file describes the network depicted in Fig. 2. The words '!ENTER' and '!EXIT' constitute the start and end nodes of the network. The node 'garbage model' is a phoneme pool that points to a sub-lattice, which allows every combination of the language phonemes. Every possible transition among the agent's name, silence ('sil') and the phoneme sub-lattice is permitted. Given the set of HMMs and the SLF file along with the corresponding dictionary, the HTK produces the best path of the word network using the Frame Synchronous Viterbi Beam Search algorithm. That is, the output of the HTK recognition unit will be any combination of phonemes and/or silence and/or the agent's name. The word

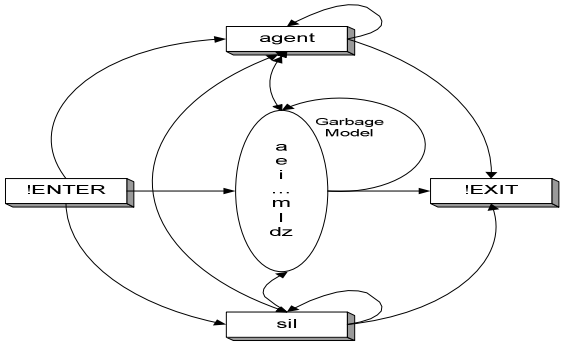

Fig. 2: The lattice of the word spotter.

spotter will propagate the signal to the speaker verification and speech recognition modules only if the agent's name is included in the output of the recognition unit.

Regarding the recognition module the language models are also SLF word networks, the paths of which cover the most probable speaker utterances for the specific task. This was due to the lack of training data in the initial phases of development in order to create robust statistical models. However, in the following stages of system's development it is anticipated that user utterances will be recorded and transcribed to be used for extending and improving the current language models.

The Speaker Verification component of the *Personalization Module* has as a main task to infer the identity of the speaker out of a small number of pre-registered voices. Therefore, due to the nature of the application the workload of the speaker verification is rather light compared to the number of speakers contemporary speaker recognition systems can identify (in our setting only two persons have the right to access the system). Although we have employed a widely accepted identification technique [13] any standard technique can be employed.

4. Real Room Experiments

The experiments took place in a 6.75m × 4.9m × 3m room with reverberation time of 0.3 sec. Our passive acoustic array consisted of 8 omni electret condenser Panasonic capsules WM54BT Ø 9.7 mm and an Aardvark Q10 Pro soundcard with internal pre-amps, 8-channel A/D converter for the signal acquisition and 0.05 m spacing between microphones. The signals were sampled at 44.1 kHz with 16 bits per sample and downsampled to 8 kHz. Experiments in a real room have the advantage of testing a system in a situation close to the final operational conditions. However, the unconstrained number of speakers, their movement and the variability of the environment make quantitative analysis problematic in terms of recognition accuracy. We are currently seeking a satisfying usability and acceptability assessment methodology, which will provide valid and reliable estimations of system quality over a variety of different application functionalities and system conditions, and to identify system weaknesses. Preliminary results focusing on acceptability evaluation based on the experience of a pre-selected group of ten potential users is based on the Task Completion Rate. Three scenarios were investigated. Firstly, four people discussing while seated in the room (quiet conditions). Secondly, three people walking engaged in conversation (quiet conditions) and, finally, three people walking engaged in conversation while music was on. One person had access to the system and commanded the appliances using phrases on the basis of twenty commands of the form: <AGENT – unconstrained words - COMMAND unconstrained words>. While real-time performance is feasible (real time performance of the system was actually achieved using a commercial beamformer [14], the word spotter and the main recognition engine), our implementation was about five times over real time. The users were informed on the succession or not of the task by the projection of the computer screen on the wall. Each user performed 3 sessions.

SCENARIOS	Task Completion Rate (%)		
	1st Trial	2nd Trial	3rd Trial
1st	96	100	-
2nd	85	92	96
3rd	79	84	90

Table 1: Mean Task Completion Rate (%) with respect to trials no. Each trial is composed of a full set of 20 commands.

5. Conclusions

We presented a robust combination of a number of diverse technologies, to construct a speech activated portal to complex devices to assist users who are disabled or not technically-inclined at home or in their workplace. We sought to frame the problem in a manner that departed from the highly constrained laboratory settings towards real-life operational conditions. The system integrated a linear microphone array to supply spatial sound selectivity, adaptive noise cancellation to deal with appliances that emit noise with the same statistical behaviour as the desired speech signal, a word spotter that provides a robust gate to the far-end speaker recognition system, a speaker recognition module that allows interaction of only registered users and finally a speech recognition system. Future work on the problem will focus on robust stochastic models based on back-off n-grams. Moreover, all appliances will be connected in a network and the line-out of TV and audio equipment will serve as an accurate noise reference in a ANC scheme that will provide more robustness than the integrated ANC of the beamformer. The possibility of wireless arrays is also to be investigated. Furthermore, the user should be able to employ the connection with the public network, use existing web-services (i.e. search engines, indexes, etc.), obtain specific information (e.g. on-line manuals, TV-programs, etc.), and access existing telecommunication voice services.

6. References

[1] Brandstein, M. and Ward, D., (Eds.), "Microphone Arrays Signal Processing Techniques and Applications", *Springer-Verlag,* 2001.

[2] Krim, H. and Vibeg, M., "Two Decades of Array Signal Processing Research", *IEEE Sig. Processing Mag.*, 1996.

[3] Johnson, D. and Dudgeon, D., "Array Signal Processing: Concepts and Techniques", *Prentice Hall,* 1993.

[4] Brandstein, M., "A Framework for Speech Source Localization Using Sensor Arrays", *PhD thesis, Brown University*, Providence, RI, 1995.

[5] Sturim, D., Brandstein, M., and Silverman, H., "Tracking Multiple Talkers using Microphone Array Measurements", *IEEE Proc. ICASSP*, pp. 371-374, 1997.

[6] Potamitis, I., Tremoulis, G., Fakotakis, N., "Multi-speaker DOA tracking using interactive multiple models and data association", submitted to *Eurospeech*, 2003.

[7] Brady, P., "A technique for investigating on-off patterns of speech", *Bell Syst., Technical Journal,* 44:1-22, 1965.

[8] Griffiths, L. and Jim, C., "An alternative approach to linearly constrained adaptive beamforming", *IEEE Trans. Antennas and Propagation*, 30(1):24-27, 1982.

[9] Hoffman, M., Pinkelman, C., Lu, X., and Li, Z., "Real-time and off-line comparisons of standard array configurations containing three and four microphones", *Journal of the Acoustical Society of America*, 2000.

[10] http://home-automation.org/Home_Networking/

[11] Young, S., et al., "The HTK Book", Entropic Cambridge Research Laboratory, Cambridge, 1997.

[12] Van den Heuvel H., Moreno, A., Omologo, M., Richard, G., Sanders, E., "Annotation in the SpeechDat projects", *Intern. Journal of Speech Tech.*, 4(2):127-143, 2001.

[13] Reynolds, D., "Speaker identification and verification using Guassian mixture speaker models," *Speech Communication*, 17:91-108, 1995.

[14] http://www.acousticmagic.com/