MICROPHONE ARRAY FRONT-END INTERFACE FOR HOME AUTOMATION

Gustavo Esteves Coelho[(1)], António Joaquim Serralheiro[(1,3)], João Paulo Neto[(1,2)]

[(1)] L2F – Spoken Language System Laboratory / INESC-ID
[(2)] IST – Instituto Superior Técnico / Technical University of Lisbon
[(3)] Academia Militar
www.l2f.inesc-id.pt
Email: {gustavo.coelho, antonio.serralheiro, joao.neto} @ l2f.inesc-id.pt

ABSTRACT

In this paper we present a Microphone Array (MA) interface to a Spoken Dialog System. Our goal is to create a hands-free home automation system with a vocal interface to control home devices. The user establishes a dialog with a virtual butler that is able to control a plethora of home devices, such as ceiling lights, air-conditioner, windows shades, hi-fi and TV features. A MA is used for the speech acquisition front-end. The multi-channel audio acquisition is pre-processed in real-time, performing speech enhancement with Delay-and-Sum Beamforming algorithm. The Direction of Arrival is estimated with the Generalized Cross Correlation with Phase Transform algorithm, enabling us to track the user. The enhanced speech signal is then processed in order to recognize orally issued commands that will control the house appliances. This paper describes the complete system emphasizing the MA and its implications on command recognition performance.

Index Terms— Home automation, microphone arrays, speech recognition, beamforming, source localization.

1. INTRODUCTION

Since speech is the most natural way of interaction between humans, it is reasonable to foresee that, in a near future, human-machine communication will comprise voice as well as the usual non-vocal forms. One of the several impairments to that desideratum is the need to adequately capture the speech signal in any place in a house. One way to avoid the nuisance of wearing close-captioning microphones is to use a suitably placed Microphone Array (MA). So, the purpose of this paper is to evaluate the MA front-end to our Spoken Dialog System controlling home appliances. We integrated base technologies - Automatic Speech Recognition (ASR), Tex-to-Speech (TTS) synthesis, Natural Language Processing (NLP), Virtual Face Animation (FACE) and Microphone Array Processing - to derive a Spoken Dialog System (SDS) [1].

In spite of ASR being a matured technology, recognition errors do occur and, to avoid executing wrong commands, a language model is used to correct or, at least, minimize the incidence of those errors. Also, the existence of multiple sound sources, such as more than one speaker in the room, music/sound (TV, hi-fi) devices, room reverberation and extraneous noises, certainly add up to the difficulty of the task. However, MAs can steer their directivity towards the sound source and, as such, minimize the influence of those adverse factors. Nevertheless, they have drawbacks, such as the difficulty to locate a moving "target" and, therefore, to adjust its directional characteristics without adversely impairing the signal spectra. In this paper, we describe the demonstration home automation system, with an emphasis on the MA and the algorithms that were implemented to locate the speaker in the room and to perform speech enhancement in order to send the resulting speech signal to the SDS.

This paper is organized as follows: section 2 is devoted to the description of the home automation system; in section 3 we describe the real-time implementation issues; in section 4 we present the experimental results and finally, in section 5, the conclusions are addressed.

2. SYSTEM DESCRIPTION

Our home automation demonstration system is based in a Virtual Butler (VB) that is always available to control some home devices. The user establishes a dialog with the VB in order to control some specific device. The butler begins by acknowledging the users request and, if more information is needed to disambiguate that specific request, automatically questions the user, engaging in a dialogue. The home automation system is divided in two main subsystems: the MA processing unit and the SDS. The MA, whose advantages are well known [2-4] to be repeated here, acquires the speech signal and outputs a multi-channel signal that is pre-processed in the Spatial Filtering Unit (SFU), for both Speech Enhancement and Direction of Arrival (DoA) estimation. The importance of the DoA is twofold: it enables spatial filtering and, also, its angular estimation is feed to the

SDS in order to steer the face of the VB towards the user. To increase the interaction of the SDS with the user(s), synthesized speech is generated to confirm the received command. In figure 1 we present a simplified block diagram of the VB that will be described in more detail in the following subsections.

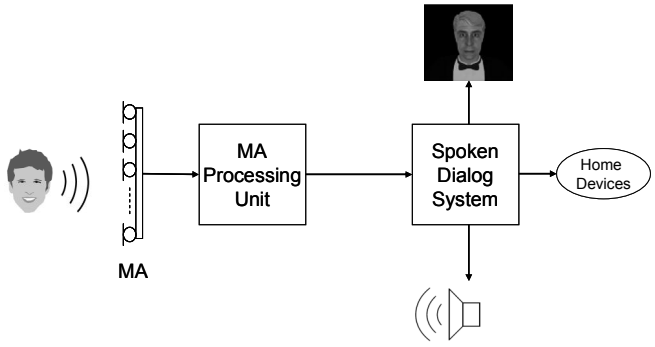

Figure 1: *Virtual Butler diagram.*

2.1. Microphone Array front-end

Figure 2 depicts the block diagram of the SFU that interfaces the MA with the SDS. The main objective of the SFU is to steer the directivity of the MA towards the sound source (the user) and, simultaneously, enhance the speech signal against environmental noise by spatial filtering (beamforming). Furthermore, the estimation of the DoA, sent to the FACE unit, allows us to build a better visual interface, since the VB can "turn its face" into the direction of the speaker. This behavior, added to the automatic generation of synthetic speech, is a step towards a more realistic human-machine interaction.

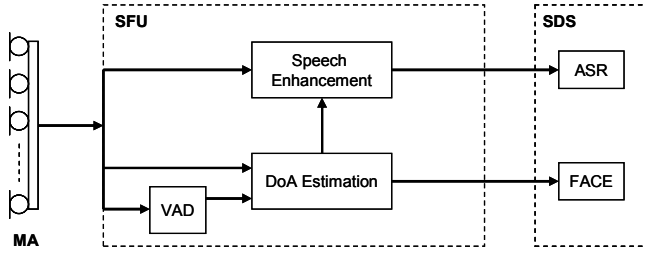

Figure 2: *SFU block diagram.*

A sixty-four linear and uniformly spaced MA, based on the NIST MarkIII MA [5], was built for both speech acquisition and DoA estimation. The distance between microphones was set to 2cm to allow for a 16 kHz sampling frequency without spatial aliasing. The audio signal is then 24-bit digitally converted with time-synchronized ADCs. The MA module connects to a remote computer by an Ethernet interface. The communication and data transfer are based on the standard UDP protocol, which provides this MA a generic interface to any computer.

Since the SDS input accepts a single channel input source, the multi-channel audio from the MA must be pre-processed. This task is done in real-time in the SFU, that also performs the DoA estimation. For speech enhancement, we apply the Delay-and-Sum Beamforming (DnSB) [6] algorithm that, when compared to the adaptive beamformers, has the advantage of providing less high-frequency spectral distortion to the desired speech signal and has a lower computational cost. For the DoA estimation, we apply the Generalized Cross Correlation with Phase Transform (GCC-PHAT) [7] algorithm. This estimation is activated whenever the speech signal is above the Voice Activation Detector (VAD) threshold. The underlying idea of this procedure is to assure that the animated face of the VB only steers to the users when they speak.

2.2. Spoken Dialog System

This system supports speech input through an ASR with NLP. The acknowledgements and/or questions from the SDS are converted into speech by the TTS module and synchronized with a 3D animated butler face. The SDS module can be divided in three blocks, as depicted in figure 3. The first one, the Input Output Manager (IOM) is where the interfaces of both the user and the butler are managed. The IOM comprises the following sub-blocks: the ASR, the TTS (to synthesize the speech of the butler) and the FACE to implement the 3D animated face of the VB. The second block of the SDS, the Dialog Manager (DM) module receives requests from the IOM in a XML format, determines the action(s) requested by the user, and directs them to the Service Manager (SM) for the execution of that action(s). This last module provides the DM with the necessary interface with a set of heterogeneous home devices grouped by domains, which users can control or interact.

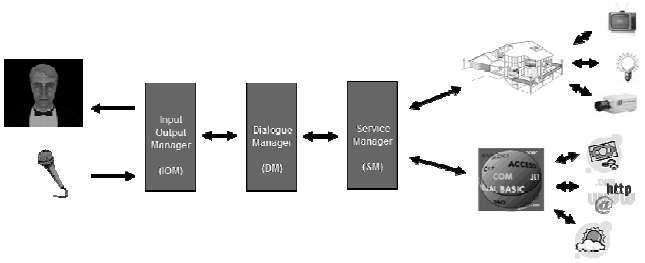

Figure 3: *SDS block diagram*.

The generic block approach enables our SDS to cope with different type of applications, and therefore be fully tailored to other applications that require speech (or dialog) interaction. The generic topology also allows the SDS to be independent from the input-output interface devices, and therefore the SDS can be accessed either locally or remotely from a wide range of devices, such as head-sets, PDAs, web browsers, mobile phones, just to mention a few.

3. IMPLEMENTATION

Our ubiquitous home automation prototype is currently deployed in our demonstration room and frequently tested with several users. The users are able to control the intensity of both the room and the table lights, can also control air-conditioning functions and room temperature or control the position of the window shades as well as their orientation, the hi-fi sound volume, the CD player and radio tuner. The home automation system can be extended to control other multimedia devices, such as TV and computers, and to retrieve web information like stock market data, personal banking services, weather reports and forecasts, flight and bus schedules, etc.

3.1. MA processing

The MA works originally with a sampling frequency of 22.05 kHz, sending all 64 digital audio channels through an Ethernet connection to a remote SFU. The SFU is programmed in Java and splits the incoming audio channel to the DnSB, GCC-PHAT and VAD, respectively, since these algorithms concurrently process the audio data. All audio data is windowed in 4096 samples (≈190 ms) with no overlap. The GCC-PHAT implements the DoA estimation using only 2 of the 64 available microphones. This pair of microphones is chosen according to prior correlation and precision analysis, weighting two contradictory factors: microphones should simultaneously be close enough to assure that correlation coefficients are acceptable and, conversely, the pair must be separate enough to ensure precision in the DoA estimations. The GCC-PHAT is controlled by a VAD, in order to ensure that DoA is estimated only when speech is present. The VAD is implemented by calculating the energy over the windowed audio data from a single microphone in the MA, and sets a threshold to define the speech/non-speech decision. The estimated DoA is then sent from the SDS to the FACE unit through Ethernet, to steer the butler animated face towards the user direction.

The speech enhancement is implemented by the DnSB, steering the MA virtual beam according to the DoA estimations. This DnSB receives all audio channels from the MA e returns a single audio channel with the enhanced speech data. The resulting single audio channel from the DnSB is down sampled to 16 kHz, since this the working sampling frequency of our ASR. This audio is sent, through Ethernet to the SDS for ASR processing.

3.2. SDS processing

Usually, one of the drawbacks of MA applied to ASR systems is the poor speech recognition results, namely when compared to close talk microphones. It is evident that the speech data acquired with MA varies greatly with the acoustic environment, and therefore causes further degradation in the recognition performance. Since home automation systems are limited-domain ASR applications, we mitigate the poor speech recognition drawback, limiting the recognition vocabulary to the specific domain needs. Consequently, our speaker-independent (SI) home automation system with the MA interface is able to perform home automation tasks with no specific adaptation of the acoustic models. Nevertheless, it is possible to personalize the SDS system, tagging the butler commands with an activation word, namely the butler's name. With this feature, the VB is able to respond only to the specific user's speech, while speech commands are processed in a SI basis.

To accomplish home automation tasks, a specific grammar is loaded into the SDS. This grammar was written according to SRGS specification format and contains a hierarchical structure defining all possible home automation commands rules. The SRGS specification format allows creating a flexible speech commands, enabling the user to order a specific command in many different ways. The vocabulary and lexicon of the SDS is automatically generated from the previous loaded SRGS grammar. The present vocabulary can be easily extended or modified and comprises 65 words, generating a total of 530 different sentences.

The ASR is based in Audimus [8], a hybrid speech recognizer that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of multilayer perceptrons (MLPs).

4. EXPERIMENTAL EVALUATION

In order to assess the recognition performance of the MA we include, as a reference, results obtained with a close-talk (headset) microphone. Furthermore, we also present recognition results using one single microphone (#32 from the MA) in a far-talk setup. To begin with, all speech data was recorded in a clean acoustical environment using a headset. Our test corpora is composed of 73 spoken Portuguese sentences (234 words), corresponding to the home automation task, e.g. "*diminuir a temperatura da sala*" (lower the room temperature). All the experiments were obtained with off-line processing, using the previous described recordings. The recognition Word Error Rate (WER) for the close-talk microphone was 2.14%. Then, the recorded speech data was played with loudspeakers in 3 different locations, as depicted in figure 4. To assess the speech enhancement performance, the recorded speech audio was contaminated with a Gaussian white noise source, located in the same 3 positions. The objective of this experiment is to show that the DnSB is able to enhance the speech from a specific direction while attenuating the noise source in other directions. As a result, the DnSB should

increase the WER, when compared with the clean speech recorded by the headset, and decrease when compared with the single far-talk microphone. The experimental results with a single microphone in far-field conditions were carried out in mild noise and reverberant conditions and the WER ranged from over 94% to 98%! These results do show how inappropriate a single far-field microphone is.

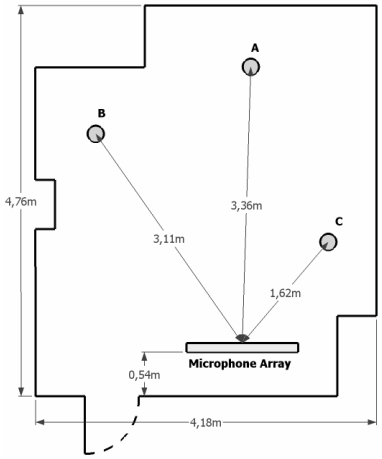

Figure 4: *Experimental setup with 3 different positions. The DoA is 92º for location A and 55º and 131º for B and C, respectively.*

Table 1 depicts the WER results for both clean speech and noise source in different positions. It can be observed that position C achieves the lower WER, since it is the nearest to the MA. Conversely, the higher WER is achieved when the noise source is closest to the MA. The SNR gain, calculated from the #32 microphone signal and the DnSB output, is presented in column 4 of table 1. These results comfortably compare with the theoretic limit of $10\log(N) \approx 18$dB for the noise attenuation, where N is the number of microphones. In practice, the DnSB is only able to attenuate spatial uncorrelated noise. Therefore, it is expected to observe a SNR gain lower than 18dB.

Table 1 : *DnSB experimental results.*

Speaker	Noise Source	DnSB DoA, º	SNR gain, dB	WER, %
A	B	92	10.6	12.8
B	A	55	11.0	18.0
B	C	55	12,6	24.8
C	B	131	12.9	6.4

Finally, we present DoA estimation results, figure 5, using microphones #29 and #36. It can be observed that DoA estimation provides an accurate direction of the speech sources with a maximum error smaller than ±2.5 degrees. As mentioned, the VAD disables the GCC-PHAT estimation during silence periods, thus preventing erroneous beam-steering.

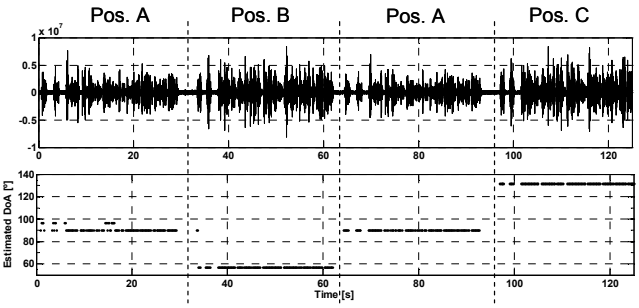

Figure 5**:** *DoA estimation results with GCC-PHAT:(above) audio from #29 microphone;(below) DoA results for the acquired speech in different positions.*

5. CONCLUSIONS

In this paper we presented a Spoken Dialog System with a Microphone Array as the speech acquisition interface, being a step forward to a ubiquitous Home Automation system, where users can control some home devices establishing a dialog with the virtual butler. The presented home automation prototype has been deployed in our demonstration room and has been successfully tested with several users.

As expected, close-talk microphones achieve better results in terms if ASR performance but, obviously, they are not a practical solution. However, the presented results show that MAs, besides providing speech enhancement, achieve sufficiently small WER to enable home automation tasks.

6. ACKNOWLEDGMENTS

This work was funded by PRIME National Project TECNOVOZ number 03/165.

7. REFERENCES

[1] J. P. Neto, R. Cassaca, M. Viveiros, and M. Mourão, "Design of a Multimodal Input Interface for a Dialog System," in *PROPOR 2006*, Brasil, 2006, pp. 170-179.

[2] M. Brandstein and D. Ward, *Microphone Arrays*: Springer, 2001.

[3] W. Kellermann, H. Buchner, W. Herbordt, and R. Aichner, "Multichannel Acoustic Signal Processing for Human/Machine Interfaces - Fundamental Problems and Recent Advances," in *Proc. Int. Conf. on Acoustics (ICA)*, Kyoto, Japan, 2004.

[4] H. Buchner, J. Benesty, and W. Kellermann, "Generalized Multichannel Frequency-Domain Adaptive Filtering: Efficient Realization and Application to Hands-Free Speech Communication," *Signal Processing,* vol. 85, pp. 549-570, 2005.

[5] "The Nist Mark-III Microphone Array," *http://www.nist.gov/smartspace/cmaiii.html*.

[6] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*: Prentice Hall, 1993.

[7] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech Signal Processing,* vol. 24, pp. 320 - 327, 1976.

[8] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, "AUDIMUS.media: a Broadcast News speech recognition system for the European Portuguese language," in *PROPOR 2003*, Portugal, 2003.