# A Spoken Dialog System Speech Interface based on a Microphone Array

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
R. Alves Redol, 9
1000-029 LISBOA, Portugal
www.l2f.inesc-id.pt
{gustavo.coelho, antonio.serralheiro, joao.neto} @ l2f.inesc-id.pt

**Abstract.** In this paper we present a Spoken Dialog System (SDS) with a Microphone Array (MA). Our goal is to create a hands-free home automation system with a speech interface to control home devices. The MA interface enables to create ubiquitous speech acquisition for the SDS. The implemented system allows any user – in any position in a room – to establish a dialog with a virtual butler that is able to control a wide range of home appliances (room lights, air-conditioner, windows shades and hi-fi features). This virtual butler has a 3D animated face that is, while the dialog is engaged, able to steer to the user's position and respond to his/hers commands with synthesized speech. The presented results show that the MA, as distant talk interface, performs quite well and is a step towards a more realistic human-machine interaction.

**Keywords:** Home Automation, Microphone Arrays, Automatic Speech Recognition.

## 1. Introduction

Considering that 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 speech as well as the usual non-speech forms. To pursue this goal, adequately speech acquisition is imperative to provide the best recognition performances. Close-talking microphones (e.g. head-set, lapel) have the advantage of high Signal-to-Noise Ratio (SNR). However, they are intrusive and if the speaker needs to moves inside a large room, or to an adjacent one, other ways of communication with the computers are mandatory. Another approach is to use a single far-field microphone in a fixed place. However, preliminary tests show degradation on the recognition performances, whenever a user utters at increasing distances from that fixed microphone. For instance, in a quiet room, the Word Error Rate (WER) goes from circa 14% to 24% when the distance from the microphone is increased from 1 to 3.5 meters. If the acoustic environment now includes some noise

sources (even at moderate levels, typical in real acoustic environments) the WER increases to 95% at 1m distance. Briefly, a single far-field microphone is definitely not adequate for practical usage.

Seeking to create ubiquitous speech interfaces and to avoid the nuisance of wearing close-captioning microphones we used a suitably placed a Microphone Array (MA), as our speech acquisition front-end. MAs offer a principled approach to recovering a particular person's speech from a mixture of distant microphones signals. A MA is composed of a multiple omni-directional microphones arranged in purposeful geometries in a room. MAs filter the received signals according to the spatial configuration of speech sources and noise sources, enabling thus to focus on a sound originating from a particular location. Contrary to the single close-talk microphones, MAs are also capable of locating sound sources in reverberant enclosures, separation of the sources and enhancement of speech signals from desired sources.

One of the main problems with MA (in terms of speech recognition) is the robust acquisitions of the speech signal given the adverse conditions in most real acoustic environments. Real environments are often reverberant and they suffer from significant background noise. Close talking microphones alleviate many of these problems and give the highest accuracy from speech recognition system. However, MA processing techniques offers an increasingly viable alternative with overcomes many advantages of close-talk microphones. MA speech enhancement generally involves *Beamforming*, which consists of filtering and combining the individual microphone outputs in such way as to enhance signals coming from a specific location, while attenuating signals from other locations.

Projects like CHIL [1], AMI [2] and the recent DICIT [3], addressing the development of advanced technologies for speech/acoustic processing and interpretation based on MA devices, are examples of the wide spreading of this technology.

In this paper we evaluate the viability of a MA as the speech acquisition front-end of a Spoken Dialogue System (SDS) whose purpose is to control a set of home appliances. The SDS [4] comprises the following base technologies: Automatic Speech Recognition (ASR), Tex-to-Speech (TTS) synthesis, Dialog Management (DM), Virtual Face Animation (FACE) and Microphone Array Processing. The main advantage of a SDS is the capability of interaction with the users to overcome recognition errors that can impair the execution of some uttered command.

This paper is organized as follows: in section 2 the description and implementation of the home automation system is presented; in section 3 experimental results with speech data are presented to evaluate our system and finally, in section 4, the conclusions are addressed.

## 2. The Virtual Butler System

The implemented SDS is currently tailored to work with Portuguese language[1], including both ASR and TTS systems. Our home automation demonstration system is based in a Virtual Butler (VB) that is always available to control the home devices (figure 1). Users can control a specific device with different speech commands - e.g. it is possible to turn on the ceiling light with either "*liga a luz*" (turn on the light), or "*acende a luz*" (switch on the light), or "*ligar luz da sala* " (turn on the room light), or even "*liga-me a luz*" (switch me the light).

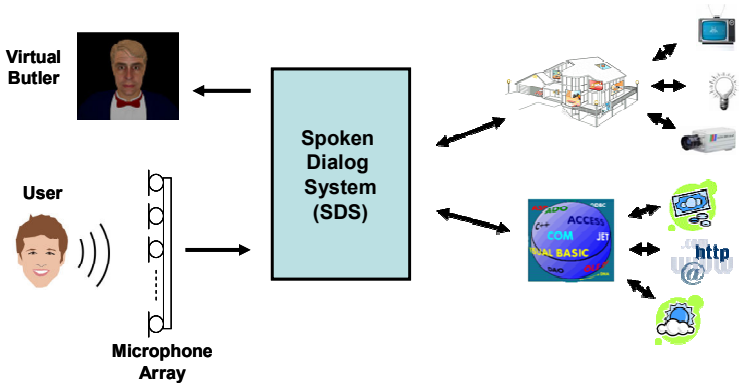


**Figure 1 -** Block diagram of the Home Automation system with a Virtual Butler.

The user gets the VB "attention" by uttering its name "*Ambrósio*", followed by a command to control a specific device. The butler acknowledges the users request and, if more information is needed to disambiguate that order, automatically questions the user, engaging a dialogue. This ambiguity can arise, for instance, directly from the previous request example, since it is possible to control both table and ceiling lights in the room. Therefore, the VB needs to complete the command "*liga a luz*" (turn on the light) knowing which light will be switched on. So, the VB questions the user with the synthesized sentence "*qual a luz que pretende ligar?*" (which light do you want to switch on?); then, the user must answer "*da sala*" (room light) or "*da mesa*" (table light), to complete the command. Other cause of ambiguity can be erroneous recognition of uttered commands. The VB acknowledgements and/or questions are converted into speech by the TTS module and synchronized with a 3D animated butler face (including face expressions and movements of the lips).

The home automation system is divided in two main subsystems, the SDS and the MA processing unit, described is the following sub sections. The SDS provides the interface between the user's speech and the VB, briefly mentioned above. The MA front-end acquires the user's speech and performs the enhancement of the signal before delivering it to the SDS input. The MA processing unit also estimates the

[1] The usage with other languages involves the modification of, at least, the acoustic models and the language models, not to mention the TTS.

user's direction and signals the SDS with that information to steer the VB face towards the user.

### 2.1. Spoken Dialog System

The SDS module is divided in three main blocks, as depicted in figure 2.

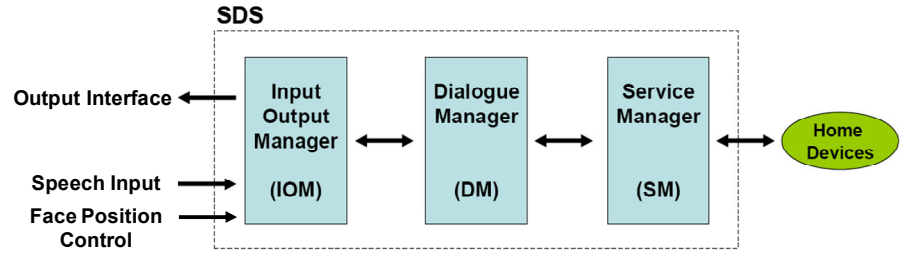


**Figure 2 -** SDS block diagram.

The first block, 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 (to recognize the user's speech commands), 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. This generic block approach enables our SDS to cope with different types of applications and, therefore, be fully tailored to other applications that require speech (or dialog) interaction. As an example, the SDS is currently applied to create a virtual personal assistant enabling automatic scheduling for meeting and other events, telephone answering and redirection, etc; and also a virtual home banking system, where users can access their banking information and services by telephone.

As mention earlier, one of the drawbacks of MA applied to ASR systems is the poor speech recognition results, namely when compared to close talk microphones, since speech data varies greatly with the acoustic environment, and therefore causes further degradation in the recognition performance. However, home automation systems are limited-domain ASR applications; we mitigate this drawback by tailoring 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 "*Ambrósio*". 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 us to create 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 covering all current possible home automation speech commands.

The ASR is based on the Audimus [5], 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). The ASR is used to recognize the enhanced speech processed by the MA.

The TTS module (DIXI +) [6] is a concatenative-based synthesizer, based on the Festival framework. This framework supports several voices and two different types of unit – fixed length units (such as diphones), and variable length units. This latter data-driven approach can be fine tuned to limited domain applications, by an adequate design of the corpus. The TTS is used to synthesize the VB speech output.

The FACE module [7] is a Java 3D implementation of a synthetic talking face with a set of visemes for the Portuguese phonemes and a set of emotions and head movements. The VB face representation is accomplished with this module.

This 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.

### 2.2. Microphone Array front-end

The MA, whose advantages were already mentioned [8-10], 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. Figure 3 depicts the block diagram of the SFU that interfaces the MA with the SDS. The main objective of the SFU is to virtually steer the directivity of the MA towards the sound source (the user) and, simultaneously, enhance the speech signal against environmental noise by means of 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.

A sixty four linear and uniformly spaced MA, based on the NIST MarkIII [11] MA, was built for both speech acquisition and DoA estimation [12]. The distance between microphones was set to 2cm to allow for a 16 kHz sampling frequency without spatial aliasing. The audio signal from all microphones is then 24-bit digitally converted with time-synchronized ADCs (simultaneous in-phase sampling). 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.

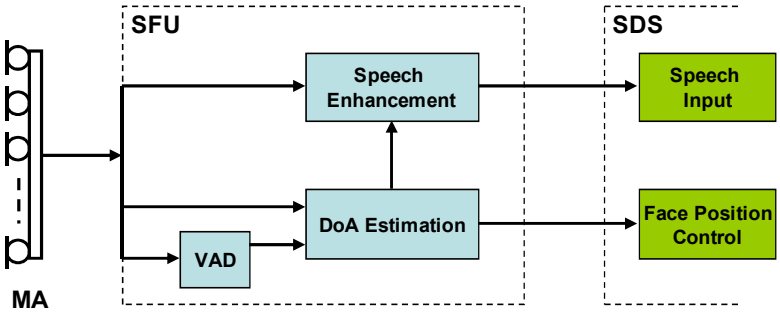


**Figure 3 -** SFU block diagram.

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. For speech enhancement, we apply the Delay-and-Sum Beamforming (DnSB) [13] 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. The virtual steering process mention earlier is implemented by means of software, with the DnSB algorithm, maintaining the MA physically fixed in a pre determined location. The resulting enhanced signal from the DnSB output is then sent to the SDS input. For the DoA estimation, we apply the Generalized Cross Correlation with Phase Transform (GCC-PHAT) [14] algorithm. This estimation process 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, avoiding the VB to steer towards the noise sources, and to avoid *noise beam-steering* (aiming the MA virtual beam towards noise sources).

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 incoming 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 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 also through an Ethernet connection.

The MA virtual beam steering direction is done according to the DoA estimations. The DnSB receives all 64 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 also through Ethernet to the SDS, for ASR processing.

As an example of the spatial capabilities of the implemented speech enhancement algorithm, in figure 4, is observed the DnSB spatial response when the current MA is electronic steered (or virtually steered) towards the endfire steering direction (DoA = 180º). The simulated spatial filtering response show a frequency variant attenuation of the signals acquired with the MA, due the large bandwidth of speech signals. However, this simulation shows that signals arriving from the desired direction (180º) are passed (0dB), while signals in other directions are attenuated (<0dB) in a wide spectral region. Because the inter-microphone spacing determines the spatial sampling frequency, for frequency above the Nyquist frequency the resulting beam response will exhibit a spatial aliasing phenomena. As a result, grating lobes (0 dB) will appear out of the steering direction for frequencies > 8 kHz, creating spatial ambiguities, as depict in figure 4. As mentioned earlier, the working sampling frequency is 16 kHz and, therefore, the spatial aliasing does not constitute a problem to the overall system.

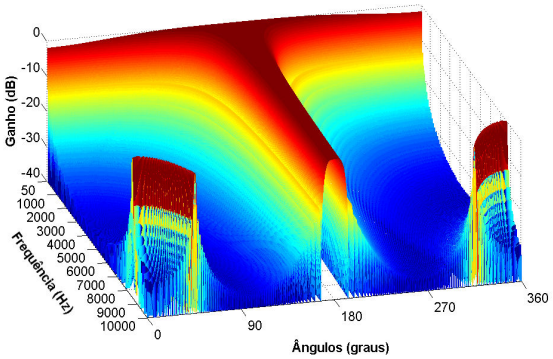


**Figure 4:** Spatial filtering response of the implemented MA: DnSB aiming towards the endfire steering direction.

## 3. Experimental Evaluation

In order to assess the recognition performance of our SDS with a MA interface 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 different spoken Portuguese sentences (234 words), corresponding to the home automation task, e.g. "*diminuir a temperatura da sala*" (lower the room temperature). All 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%, and will be our base line for the ASR evaluation. Then, the recorded speech data was played with loudspeakers in 3 different locations, as depicted in figure 5. 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, validating thus the MA purpose for the far-field speech acquisition. 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.

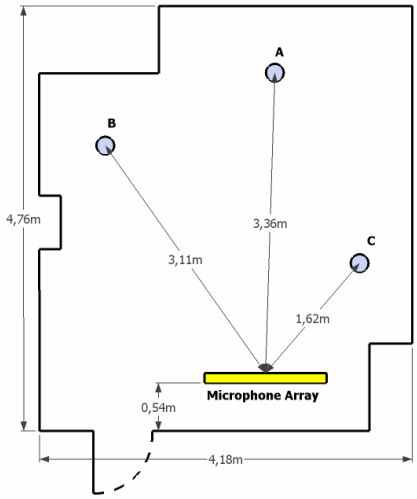


**Figure 5 -** Experimental setup with 3 different positions. The DoA is 92º for location A and 55º and 131º for B and C, respectively.

**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 |

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 compare comfortably with the theoretical limit SNR gain for the DnSB of $10\log(N) \approx$ 18dB, where $N$ is the number of microphones of the MA. In practice, the DnSB is only able to attenuate spatial uncorrelated noise. Therefore, it was expected to observe a SNR gain less then 18dB.

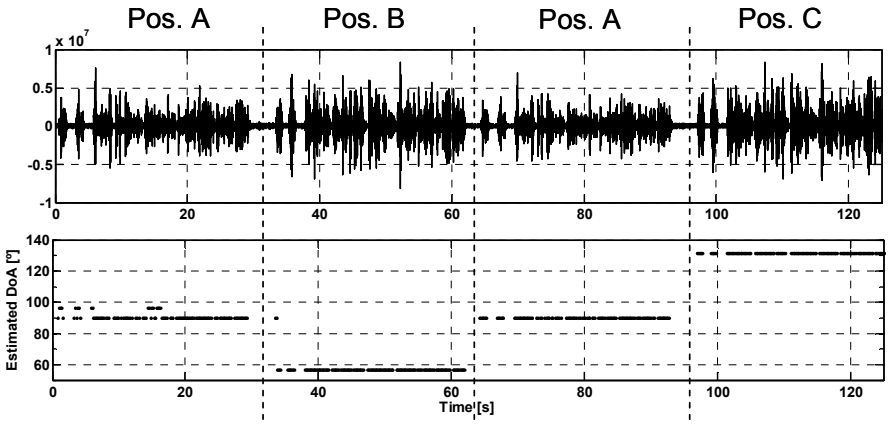


**Figure 6 -** DoA estimation results with GCC-PHAT: (above) audio from #29 microphone; (below) DoA results for the acquired speech in different positions.

Finally, we present DoA estimation results (figure 6) 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. At 3.5m distance from the MA, this error corresponds to a 0.15 m location mismatch. Since the width of the loudspeaker (used to play the recorded speech data) is ≈0.2 m, the resulting error is within the physical size of the sound source. Considering that this error can occur due the user's face movements, this error is less than the normal length of the human face and, therefore, acceptable.

As mentioned, the VAD disables the GCC-PHAT estimation during silence periods, thus preventing erroneous beam-steering. As depicted in figure 6, the estimated DoA values are present only in speech intervals. During non-speech intervals, no estimation is done and the DnSB maintains beam steering to the previously estimated DoA.

## 4. 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.

## 5. Acknowledgments

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

## 6. References

[1] "CHIL - Computers In the Human Interaction Loop," http://chil.server.de/.

[2] "AMI - Augmented Multi-party Interaction," http://www.amiproject.org/.

[3] "DICIT - Distant-talking Interfaces for Control of Interactive TV," http://dicit.fbk.eu/.

[4] J. P. Neto, R. Cassaca, M. Viveiros, and M. Mourão, "Design of a Multimodal Input Interface for a Dialog System," in *PROPOR 2006 - Computational Processing of the Portuguese Language*, Brazil, 2006, pp. 170-179.

[5] 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 - 6th International Workshop on Computational Processing of the Portuguese Language*, Portugal, 2003.

[6] S. Paulo and L. C. Oliveira, "Reducing the Corpus-based TTS Signal Degradation Due to Speaker's Word Pronunciations," in *Interspeech, ISCA*, Portugal, 2005, pp. 1089-1092.

[7] M. Viveiros, "Cara Falante - Uma interface visual para um sistema de diálogo falado," Graduation thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, 2004.

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

[9] 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.

[10] 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.

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

[12] G. E. Coelho, A. J. Serralheiro, and J. Neto, "Microphone Array front-end interface for Home Automation," in *Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Trento, Italy, 2008, pp. 184 - 187.

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

[14] 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.