Multimedia Tools and Applications (2022) 81:3327–3350
https://doi.org/10.1007/s11042-021-11291-3

1188: ARTIFICIAL INTELLIGENCE FOR PHYSICAL AGENTS

An audio enhancement system to improve intelligibility for social-awareness in HRI

Antonio Martínez-Colón[1] · Raquel Viciana-Abad[1] · Jose Manuel Perez-Lorenzo[1] · Christine Evers[2] · Patrick A. Naylor[3]

Received: 6 February 2021 / Revised: 31 May 2021 / Accepted: 9 July 2021 /
Published online: 28 August 2021

Abstract

Improving the ability to interact through voice with a robot is still a challenge especially in real environments where multiple speakers coexist. This work has evaluated a proposal based on improving the intelligibility of the voice information that feeds an existing ASR service in the network and in conditions similar to those that could occur in a care centre for the elderly. The results indicate the feasibility and improvement of a proposal based on the use of an embedded microphone array and the use of a simple beamforming and masking technique. The system has been evaluated with 12 people and results obtained for time responsiveness indicate that the system would allow natural interaction with voice. It is shown to be necessary to incorporate a system to properly employ the masking algorithm, through the intelligent and stable estimation of the interfering signals. In addition, this approach allows to fix as sources of interest other speakers not located in the vicinity of the robot.

Keywords Beamforming · ASR · Array · Masking · Intelligibility

✉ Raquel Viciana-Abad
rviciana@ujaen.es

Antonio Martínez-Colón
amcolon@ujaen.es

Jose Manuel Perez-Lorenzo
jmperez@ujaen.es

Christine Evers
c.evers@soton.ac.uk

Patrick A. Naylor
p.naylor@imperial.ac.uk

[1] Universidad de Jaén, Jaén, Spain

[2] University of Southampton, Southampton, United Kingdom

[3] Imperial College London, London, England

1 Introduction

Social awareness can be defined as the ability to take the perspective of and empathize with others, to understand social and ethical norms for behavior, and to recognize community resources and supports. In robotics, this can be summarized as being always focusing the computational resources in achieving two objectives: the specific task (such as navigating to a goal), and the fulfilling of certain social rules (such as answering questions or looking towards interlocutors faces). Some of the requirements addressed to give a robot social awareness [14] are to integrate technical knowledge of hardware and software, psychological knowledge of interaction dynamics, and domain-specific knowledge of the target application. Regarding the interaction dynamics, multimodal human communication design is a key point where natural language interaction plays an important role that is still lacking in Human Robot Interaction (HRI) [16].

In recent years, a great effort has been made to develop domestic service robotics as tools to support and leverage resources in different tasks, in particular Socially Assistive Robots (SAR) for retirement homes [13]. If the development of an interaction mechanism adapted to the needs and tastes of the people is already complex, it is even more of a challenge when it is conceived for older people, in which the digital gap and the possible sensory limitations are additional factors to be considered.

In this sense, Automatic Speech Recognition (ASR), together with voice synthesis, is a very powerful form of interaction that makes naturalness and social engagement possible in HRI, as stated for other examples of computer based service system [1]. However, different studies within the SAR field have presented preferences for other interaction modalities that a priori are not so straightforward or natural, mainly due to the current technological limitations of these systems [17]. These are associated with ASR engines' difficulties to adapt well to dynamic environments with very diverse acoustic properties, together with the challenge of recognising one person's voice among others talking simultaneously, or the a.k.a. 'cocktail party' situation and the effect of noises of very different nature.

In fact, recently, RoboCupHome [16] has defined a strategy to gradually improve the conditions of the competitions that are held with the aim of allowing different teams to improve the capabilities of interaction through voice, using their proposed methods. In particular, two priorities to consider are: responding to an open discourse instead of just to direct commands, and to being able to respond in multi-speakers environments with inherent noise.

On the other hand, there are great advances in the field of commercial ASR integrated in VoIP value-added services and "voice bots", thanks to deep learning strategies and the extensive corpora used for training, that favour their increasing use and success rate. However, these systems still tend to be based on a one-to-one interaction where it is also understood more for the use of specific commands than for an open speech dialog. In this regard, as described in Sect. 2, research is being focused on enhancing the voice signal used as an input for these systems, mainly based in dereverberation algorithms to deal with the changing acoustic properties of the environments and spatial filtering by exploiting the capabilities of arrays of microphones (mic). There are new approaches based on combining these techniques by employing deep learning with complete training procedures as recently reviewed in [29]. It is also important to consider the trade-off between responsiveness and intelligibility in these systems, because the complexity of certain systems to improve intelligibly may reduce the interactivity and naturalness in HRI.

Regarding the use of beamforming techniques, this can be considered as an artificial way to provide attention to an specific source of acoustic information, in a similar way that human attentional mechanisms focus on a specific conversation, but exploiting specially the spatial filtering capabilities. Thus, the beamforming system can be understood as a tool to focus perception resources on a source of interest (SOI) or target highlighted by the robot attentional mechanism. In this regard, one approach would be that reactive and deliberative agents of this attention mechanism govern the signal enhancement module to improve the ASR input signal.

This study is focused on analysing the extent to which beamforming algorithms with mic. arrays that can be mounted on SARs may improve the performance of current commercial ASR engines when working in real scenarios and in real time. Thus, this work presents the integration of an acoustic signal conditioning system built upon a basic Delay-and-Sum beamformer (DSB) and a time-frequency masking technique with the aim of evaluating the improvement in intelligibility in a real scenario while maintaining the required interactivity, in terms of response times. The approach followed for the conditioning system has been previously contrasted with other techniques [19], with a proof of concept evaluation with just two speakers in both a simulated and a real scenario. However, in this study, the conditions set are closer to a real situation in which a SAR provided with social awareness capabilities would find itself and with environment conditions that reproduce the noise level associated with everyday tasks performed in a care-home living room. Thus the directions of arrivals (DoAs) of the SOI and IFP (Interfering Person) voice are not fixed, but dynamically calculated. In addition, the voice signal sent to the ASR service is conditioned in short segments in order to maintain interactivity, even with the consequent decrease in the possible rate of recognition.

2 The use of ASR for socially assistive HRI

Voice synthesis and ASR systems have been widely used in SARs as they are one of the most natural interaction modalities for HRI. However, these systems have been mainly considered as a one-to-one interaction system. The development of commercial service robots, such as Peeper (Softbank Robotics) and XR1 (CloudMinds), designed to provide services in public places, has led to greater requirements in terms of voice recognition in multi-speakers and dynamic environments (levels of reverberation, noise, interference, distances), that is to say in more challenging conditions. Thus, recent research projects such as EARS (FP7- 609465, Embodied Audition for RobotS) have focused on improving the auditory perception system of social robots. Open-source robot audition systems [23] have been integrated in different robots (Honda ASIMO, SIG2, Robovie-R2 and HRP-2). A similar approach to our proposal, but with post-filtering and masking integrated together with Mel Frequency Cepstral Coefficients features of an ASR embedded in SIG2 robotic head has been evaluated with a MIMO (multiple input multiple output) approach [31], and recent further research [4] is still being made in this direction. Thus, HARK[1] is an example of open-source robot audition software consisting of modules for sound source localization, sound source separation and automatic speech recognition that have been used with different robots in RoboCup@Home competitions since 2008.

[1] Honda Research Institute Japan Audion for Robots with Kyoto University, https://www.hark.jp/

The scope of the robotic research into ASR systems is very broad, due to additional problems such as internal noise and the effect of the mic. integration in the casing. Nowadays, APIs such as those of Google, Microsoft, IBM, Nuance, etc. are being integrated into the robot systems due to the improvement of multimedia data distribution in networks and rapid access to SaaS (Software as a Service) in the Cloud. Thus, recent studies have evaluated in real or "out of the lab" conditions, the performance of these ASR in terms of WER (Word Error Rate). In particular, Jankowski et al. [11] have recently established that in Chinese, most of them exhibit a degradation not supported with SNR (Signal to Noise Ratio) lower than 15 dB, which is a value typical for example in hospital rooms.

Until recently, the use of systems that hinge on effective verbal communication in scenarios where the dialogue requires more than a command, has caused the discouragement of the users or even their adaptation to the robot system. This is also typical in HCI for Virtual Reality, the Cyborg's dilemma [2] that refers to the paradoxical situation in which the development of increasingly "natural" and embodied interfaces leads to "unnatural" adaptations or changes in the user. To avoid this, there are already studies focused on evaluating the ability of ASR in social robotics to adapt to the speech of elder or children, where existing models used by the ASR engines may not work well with pronunciation, pitch differences, or speech difficulties, that may be typical in these users. In [12], commercial ASR engines (Google, Bing, Sphinx, Nuance) have been evaluated using fixed, all spontaneous, and clean spontaneous kids speech utterances recorded during their interaction with a robot, in typical conditions of social HRI scenarios. Their performance metrics not only consider WER, but also matches with "relaxed accuracy", as they can be enough for an ASR system to recognise sentences, such as the human ability of recognising incomplete utterances. In particular, the experimental evaluation has been developed by recording the interaction of children with a NAO robot array (two mics.) and two other types of mics. They have created a database of utterances including spontaneous speech, closed single-word sentences and multi-word sentences and evaluated intelligibility with cloud-based ASR engines, including the Nuance ASR engine used by NAO. They have considered location and angles, type of errors, background noise, therefore presenting a detailed analysis of the limitations of a widely used robot with child in terms of ASR performance. Their results highlighted that Google's ASR outperformed the others with the three types of utterances.

The differences in the gender representation in broadcast corpora used to train ASR engines is being also an aspect being considered [8]. In the case of children and women, it may have a negative impact in HRI while being used with general population and, together with voice differences, this should be considered.

3 Implementation of a system to improve the intelligibility of the source of interest

In order to promote a natural interaction through voice, one of the most important parts is the enhancement of the SOI's speech. Most of the speech enhancement techniques based on multi microphone processing rely on one fundamental cue that is mostly unknown, the positions of the human speakers around the robot. Once the speakers have been detected, it is normal to "pay more attention to the speaker of interest", either by making an effort to pay attention to that person's speech or by getting closer to that person. Thus, some of the strategies of DoA estimation, and in particular Steered Response Power-Phase Transform (SRP-PHAT) algorithm, which is the method implemented, are described in Sect. 3.1. Once the SOI and possible interfering sources have been located, it is possible to apply spatial filtering techniques in order to reinforce the data coming from the SOI, as described in Sect. 3.2.

3.1 DoA estimation of multiple sources

The problem of acoustic source localization has a high impact in domains such as speech enhancement and indoors tracking of acoustics sources. Recently, an approach based on applying learning based methods, such as DNN (Deep Neural Networks) [3, 28] is being used within this field. Indeed, with the higher availability of advance hardware architectures, the increase of computational resources, there are promising studies (different architectures, target, experimental setups) that are being tested with extensive datasets. However, the training required for this approach is even more complex within a HRI scenario where the array is not placed in a fix position, but mounted on a mobile platform, and where different rooms and conditions may be found during interaction in a real scenario. Moreover, if the estimation is not only of speakers but is understood as an acoustic sensor that allows the localisation and identification of sources of different nature, model-based methods are normally used, rather than black-box training, which allow the association of auditory information with its position based on mathematical and physical models. In this sense, there are proposals based mainly on the use of beamforming or high-resolution spectral estimation functions.

A recent review about localization of sound sources in robotics [26] has highlighted studies where beamforming techniques are used to detect multiple DoAs by first meshing the search spaces of possible candidates DoAs, sequentially beamforming each targeting candidate DoA, measuring the beamformer response in each direction (usually the output energy), building a function with the responses in each direction, and taking, the local maxima of this function to correspond to the source locations.

Some examples of the use of beamforming techniques applied to the localization of acoustic sources in robotics can be found in [15, 30]. The most commonly used beamforming algorithm for localization purposes is the SRP-PHAT, which exhibits less problems when facing reverberant environments. This algorithm computes the likelihood of each potential source position on the basis of the generalized cross-correlation estimations between pairs of microphones. SRP-PHAT combines the robustness of the steered beamforming methods with the insensitivity to signal conditions afforded by the PHAT filtering, which is used to weight the incoming signals. The advantage of using PHAT is its resistance to reverberation and room conditions. Moreover, to increase the estimation confidence, an histogram from different consecutive windows can be constructed from the steering response of each potential source position or in long inter-microphone distances arrays the use of weighting functions to emphasize the contributions coming from specific pairs. Another advantage of building an statistical model upon the steering response obtained from different angles is the feasibility of determining the likelihood of more than one active source by searching for more local maxima, which can be associated with possible sources. However, one of the main drawbacks of this algorithm is the computational demand that increases with the number of pairs (more cross-correlation estimations) and the search grid. Indeed in the case of just one source, there are implementations based on using coarse detection with gross search areas (low cut-off frequency). Once selected the strongest speech energy this area can be divided into equi-spaced thin areas. Following this approach, recent studies [24, 32] have implemented single source versions by using Density Peak Clustering or multi sound source localization with a circular array reducing the computational demand to 50% real time.

In [18], an algorithm based on a statistical decision built upon the output of SRP-PHAT applied on multi-channel audio captures obtained from a 4 mics. circular array has been evaluated for an audio-visual localization system. In particular, the approach followed is represented in Fig. 1, where the SOI's position is selected as the point of the grid exhibiting the maximum value of the steering response, expressed as:

$$P_n(\vec{x}) = \sum_{n\in Z}\sum_{l=1}^{M} \left| w_l m_l(n - \tau_l(\vec{x}))\right|^2 \tag{1}$$

being $\vec{x}$ the spatial point where the SRP is computed, n the time frame of window length (T), M the number of microphones, $m_l(t)$ denotes the signal output for a given microphone l, w_l is a weight, and $\tau_l(\vec{x})$ is the propagation time of the direct path from the point $\vec{x}$ to the microphone l.

Removing some terms of fixed energy, the part of $P_n(\vec{x})$ that is variable with $\vec{x}$ can be expressed as in [5]:

$$P'_n(\vec{x}) = \sum_{v=1}^{M}\sum_{l=v+1}^{M} R_{m_v m_l}(\tau_{vl}(\vec{x})) \tag{2}$$

being $R_{m_v m_l}$ the GCC for the pair of microphones (v, l), and $\tau_{vl}(\vec{x})$ the Inter-Microphone Time-Delay Function (IMTDF), which can be expressed as:

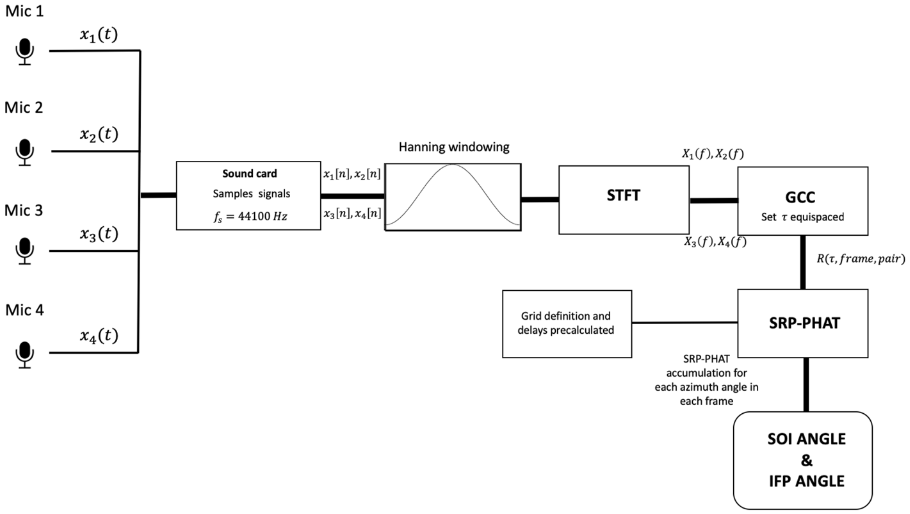

Fig. 1 SRP-PHAT applied to frames of 1.4 *s* length with windows of 90 *ms*. DoAs of SOI and IFP are computed as the two local maximun values of an histogram built accumulating the steered power obtained for positions placed at 36 possible azimuth angles

$$\tau_{vl}(\vec{x}) = \frac{||\vec{x} - x_v|| - ||\vec{x} - x_l||}{c} \tag{3}$$

with c being the speed of sound, and x_v, x_l the points where microphones k and l are respectively located.

Subspace search is usually based on the study of the vector sub-spaces formed by the spatial correlation matrix of the array outputs formed by M microphones to which D signals arrive from different locations, with $D < M$. One of the most popular of these methods is MUSIC (MUltiple SIgnal Classification), which is based on the computation of the eigenvalues (and their respective associated eigenvectors) of the correlation matrix. The D largest eigenvalues correspond to the set of eigenvectors that generate the so-called signal subspace. This subspace, if the necessary assumptions are satisfied, allows a perfect decomposition of any of the D vectors representing a given DoA. The remaining eigenvectors associated with the smaller eigenvalues form the so-called noise subspace, which in the ideal case would be orthogonal to a vector representing a source at a given DoA. These two subspaces once computed allow the determination of the sources' DoAs by computing the maxima of the spatial spectrum. In a recent approach, MUSIC outputs have been used for Acoustic Simultaneous Localization and Mapping (aSLAM) [7] to simultaneously map the 3D positions of multiple sound sources while passively localizing the observer within the scene map. This approach makes it feasible for a robot to estimate the positions of different SOIs while navigating within a room just by exploiting its acoustic inputs. Indeed, the Locata Challenge[2] and SPARK source code repository provide access to implementations of localization techniques based on MUSIC. However, it requires assuming that the signals are coherent, a condition that is difficult to meet in changing scenarios with possible conditions of high reverberation and with the presence of multiple audio sources, such the ones where a mobile indoors robot would interact.

Most of the speech enhancement techniques based on multi microphone processing rely on the source position. Thus, the need for a reliable target position estimation in the beamforming applications is one of the reasons for the increasing interest in the acoustic source localization and tracking topic, which is even more critical when simultaneous sources are present as in a HRI scenario. Beamforming techniques and methods based on subspace search estimate the source location for a certain time window defined by a past number of samples. However, temporal clustering techniques estimate a single DoA in each time window and group different DoAs in clusters representing a sound source. For example, the Kalman filter is used as part of a multi-DoA localization system in [25]. In [6] a Gaussian mixture model (GMM) is used for temporal clustering. A modification of the K-means++ algorithm, that does not assume the number of sources, is used in [10], and a similar binaural approach has been implemented with a Kurtosis-driven split-Expectation Maximization (KDS-EM) [27] and tested as part of a HRI attentional mechanism.

3.2 Techniques for acoustic enhancement

The basic idea of beamforming is to generate a directive beam that targets the desired direction based on the combination of the signals arriving to an array of microphones, and it has been broadly investigated as a pre-processing stage in order to enhance the

[2] https://www.locata.lms.tf.fau.de/challenge-description/

recorded signal that might be used for any speech application. As reviewed in [19], most of beamforming techniques are widely used to filter the SOI's information. The beamformer is a spatial filtering system, that enhances the signals coming from a predetermined direction and reduces those coming from unwanted directions. Basically, a beamformer combines the signal collected from multiple microphones so that signals coming from the desired direction are constructively added, while signals coming from other directions are diffusely or destructively combined. The simplest beamforming technique is known as DSB. Another well-known technique is GSC (Generalized Sidelobe Canceller). This beamformer is an efficient implementation of the LCMV (Linear Constraint Minimum Variance) beamformer that seeks to minimize the output power of an array, while preserving the power in one or more specified directions.

Beamforming techniques may be understood as a gross separation technique based in location being applied previous to consider another attention-related parameters, that may allow a more refined separation, commonly based in applying algorithms in T-F representations. Different beamforming alternatives have been analyzed for separation purposes [22], and are often compared to the DSB, by measuring the improvement due to attenuating interference from other directions, since the DSB is the more basic and less computationally demanding algorithm. Using time alignment, DSB adds multiple mic. signals for a target direction in phase, and its generic output for a beamformer focal point, $\mathbf{r}_p$, is given by:

$$y\left(\mathbf{r}_p\right)[n] = \frac{1}{M}\sum_{m=1}^{M} x_m\left[n - \tau_{pm}\cdot f_s\right] \tag{4}$$

where x_m is the mth mic. response for sample n, f_s is the sample frequency, $1/M$ (M the number of mics.) is the traditionally selected weights magnitude, and τ_{pm} are delays due to sound signals propagation. This delay is due to the distance between $\mathbf{r}_p$ and the mics. array at the sound speed c, thus being τ_{pm}:

$$\tau_{pm} = \frac{d_{pm}}{c} = \frac{\sqrt{\left(x_p - x_{mic}\right)^2 + \left(y_p - y_{mic}\right)^2 + \left(z_p - z_{mic}\right)^2}}{c} \tag{5}$$

However, the DSB does not usually provide sufficient reduction of the interfering signal where the signal-to-interference ratio (SIR) is low and in the case of interfering signals with energy greater than that of the signal of interest [22]. In particular, dynamic beamformers, such as GSC, are an alternative in situations with moving speakers, imprecise computation of the directions of the interfering signals and unknown room conditions. A GSC implementation [9] has been therefore considered, which basically is like LCMV but cancelling not only the known directions but everything that does not come from the direction of interest, thus simplifying the calculation of the conditioned part. An evaluation of these two approaches has been recently performed in terms of intelligibility with simulated conditions through STOI (Short-Time Objective Intelligibility) [29] measurements and ASR performance metrics in [19] with a very controlled study. Although the STOI analysis has indicated an improvement in intelligibility, ASR performance measured in terms of WER has not addressed a significant improvement compared to just using a DSB.

A different approach to beamformers or used together with beamforming strategies treats speech separation as a supervised learning problem based on computing two-dimensional masks in T-F (Time-Frequency Masking). Recent studies [20] have evaluated a basic implementation of a DSB and a binary mask, characterized by its low computational demands, for distributed array of mics. in simulated and real scenarios. In particular, that study showed the benefits of a T-F Masking implementation in terms of the intelligibility with a distributed mic. array. The main drawback is that reliable speaker's position estimations (SOI and Interferers) are required to correctly apply the masks.

Given an environment with Q sound sources distributed through the room, the T-F masking algorithm consists of a short-time windowing (20-50 ms) and the spectrum computation of the signal after being steered by a beamformer to each source position. Thus, the discrete function Y represents the T-F of a beamformed signal as follows:

$$Y\left[k, i, \mathbf{r}_p\right] = \sum_{q=1}^{Q} G_{pq}[k] \cdot X\left[k, i, \mathbf{r}_q\right] \tag{6}$$

where i is the index of a particular window, k is the frequency index (frequency bin), $X[\cdot]$ is a time frequency representation of an audio source signal located at position $\mathbf{r}_q$, and $G_{pq}[k]$ is the discrete beamformer transfer function for the sound source located at position $\mathbf{r}_q$ with the beamformer pointing to $\mathbf{r}_p$.

Even though the beamformer has its highest gain at the focal point, a T-F window can be dominated by an interferer in particular moments when the SOI doesn't speak or when the interferer speaks louder than the SOI. A spectral power ratio is used to determine T-F windows where the SOI is the dominant source and those in which the interferer is the dominant source:

$$S_{pq}[k, i] = \frac{\left|Y\left[k, i, \mathbf{r}_p\right]\right|^2}{\left|Y\left[k, i, \mathbf{r}_q\right]\right|^2} \tag{7}$$

being $\mathbf{r}_p$ the position where the SOI is located and $\mathbf{r}_q$ the position of an interferer. A binary mask is chosen as:

$$T_{pq} = \begin{cases} 1, & S_{pq}[k, i] \geq 1 \\ 0, & S_{pq}[k, i] < 1 \end{cases} \tag{8}$$

If several interference sources are present in the environment, the mask is chosen as a multiplication (or binary "AND" operation) of each mask corresponding to individual sources:

$$T_p[k, i] = \prod_{q=1, q \neq p}^{Q} T_p[k, i] \tag{9}$$

Thus, the output of a signal spectrum for a specific T-F window is given as:

$$Y'\left[k, i, \mathbf{r}_p\right] = T_p[k, i] \cdot Y\left[k, i, \mathbf{r}_p\right] \tag{10}$$

Finally, the time domain signal can be reconstructed processing its inverse FFT. Once T-F areas where predominate interfering sound sources are masked, the intelligibility of the SOI improves.

Results obtained with a similar approach to [20], but with experimental results obtained with an XMOS array of 6 integrated mics, have also highlighted an improvement in off-the-shelf ASRs performance, in particular, Google Speech API, compared with two beamformers implementations (DSB and GSC) [19]. In that previous study, the DoA angles associated with SOI and IPF were set manually, knowing the experimental positions, and the result of the signal conditioning was done on frames of 10 s. In that situation, ASR recognition rates reached nearly a 40% improvement with the voice signal processed with the DSB+Masking algorithm compared to DSB. However, the voice recognition has been done on the basis of 10 s processed signals with the online ASR web service and with only two people. It is necessary to address the dependencies with reliable dynamic computation of target and interferers' positions (DoAs of both SOI and IFP) and analysing the feasibility of improving ASR performance in conditions that allow interactivity, that is to say, with low signal duration.

4 Overview of the general system and implementation

The algorithms evaluated have been implemented as part of a system conceived to provide a robot with capabilities to exhibit a socially awareness behaviour. For that purpose, as shown in Fig. 2, three main modules have been conceived: a perception system supported by different type of sensors (visual, audio, deep cameras, odometer, laser, RFID readers, etc) that can be included within a robot housing; a module that implements the agents of an attentional mechanism in charge of focusing the robot attention in different points of interests depending on the task to develop and the world's state (environment model); and a sensorial processing system.

The attentional mechanism is controlled by two planners: a reactive planer in charge of allowing fast reactions in the robot as a response to the detection of new events (i.e. an alarm, or the arrival of a new person) and a deliberative one that focuses the robots attention in specific points controlled by rules that consider the task to perform and the environment state. That is to say, in the case of a conversation, the focus should be on the person.

One of the tasks included in the sensorial processing system is the capability of detecting, tracking and identifying the existing people in the surroundings, in such a way that

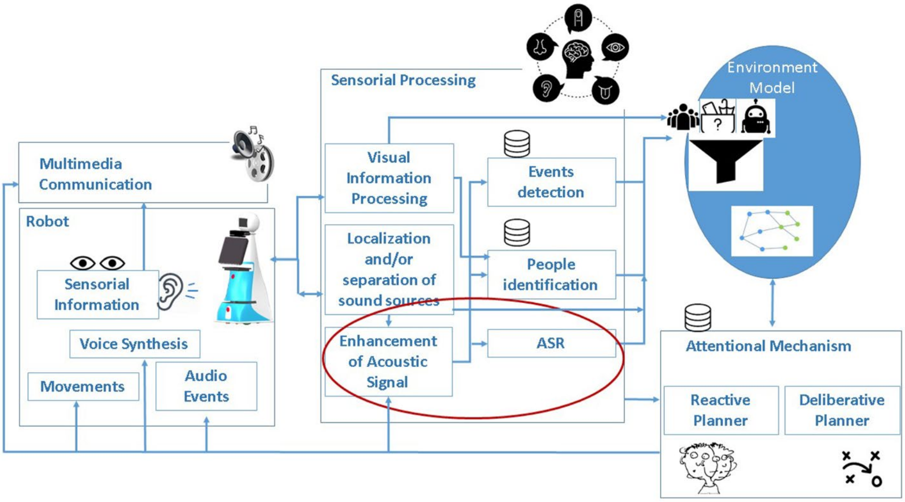

Fig. 2 System implemented that includes as part of the Sensorial Processing module the conditioning system implemented

the robot can adapt its behaviour to the possible interlocutors. As part of this module, it is vital to be able to recognise the content of a conversation taken with the robot in typical ’cocktail party’ situations, where multiple speakers may be talking simultaneously. In this situation, the attentional mechanism based on the information obtained from the sensorial processing module and the environment model must indicate who is the SOI or the person of interest based in their identification, relative positions, speech content and the task being performed.

Prior to the development of this high-level deliberative module, this study aims to assess to what extent the signal conditioning system evaluated in [19] improves the ASR performance, in a situation where another processing module based on a SRP-PHAT is dynamically computing the DoA of two speakers present in the environment. Moreover the temporal limitations of the system when it is operating in real-time should be addressed in terms of processing times and the size of the voice frames necessary for the ASR system to maintain an adequate level of recognition.

In particular, the four modules implemented in order to improve the performance of an ASR system working in a multi-speaker environment are:

- A localization system that computes the DoA for the SOI and the main IFP. This system has been programmed to obtain the two local maximum of the accumulative steering response algorithm applied to the audio frame, as described in Fig. 1.
- Two acoustic signal enhancement systems fed with the previous DoAs. DSB and GSC are implemented just for comparison purposes, and the DSB+Masking algorithm is implemented as described in Sect. 3.
- An acquisition module that prepares the multi-channel audio signals obtained from a 6 mics. circular array with a diameter of 8.6 cm. during the configured frames of duration 1.4 s.
- An ASR agent is implemented by calls to the REST API of the Google ASR engine.

All these module together with recording modules, in charge of recording audio (pre and post processed) and ASR outputs used for evaluation purposes, are implemented as Robocomp[3] components. Robocomp is an open-source robotic framework, that among other aspects, allows running agents and components in different PCs connected in a network, as well as the data exchange through ICE[4] communication middleware.

Indeed, as explained in detail in Sect. 5.3, complex conditions have been considered by analysing frames where two speakers are simultaneously talking nearly all the time and in close positions. However, an operating scenario has been assumed as a simplification, in which the robot would already be at the same distance or closer to the SOI than to others and that the SOI is always placed in a static position and facing the robot, simulating the effect that the interfering people in the room are talking to other people. This simplification means that, in frames where both people are talking simultaneously, in a very high percentage of cases it can be assumed that the speaker of interest is located at the first local maximum and the interfering person at the second maximum in the histogram constructed from the steered power response accumulated during the windows processed in a 1.4 s duration frame.

[3] https://robolab.unex.es/index.php/robocomp/

[4] https://zeroc.com/products/ice

4.1 Processing time and configurations

Motivated mainly by the computational cost of the SRP-PHAT localization module, only 4 of the 6 mics XMOS array have been used. The acquisition module has been programmed to send through Robocomp's socket based publishing system a data structure associated with the acquisition of 1.4 s of multi-channel signal. This frame-size allows the localisation module to perform steered power accumulation over 30 windows of around 90 ms (50% overlap). With a computer equipped with an Intel Core i7-4790 CPU 3.60 GHz x 8. SRP-PHAT takes on average 900 ms, and the conditioning module (DSB+masking) takes 70 ms (DSB needs 30 ms). Although a GSC component has also been implemented for comparative purposes (See Sect. 6.3), it has not been used due to its high computational demands (1.2 s). The module that communicates with the Google ASR requires 32 ms. In total, this means that the content of a 1.4 s signal could be recognized every 1 s. However, after initial tests, it was found that the good performance obtained with 10 s frames was greatly degraded when sending signal frames of only 1.4 s. This is logical if the algorithm performs recognition based on the semantics of what is being sent and also because the possible word fragmentation. Tests were made, and with 5 s frames, good recognition levels were obtained. Thus, the system has been configured to request to the ASR every 3 frames of 1.4 s, which means that a response is obtained for each audio of 4.2 s with a delay of 1 s. As an example, the average time to say in Spanish the target speaker's sentences included in Annex I is 3.28 s. Sentence 11 is the shortest with 2.2 s and sentence 15 is the longest with 4.2 s, which make reasonable the configuration chosen.

5 Method

5.1 Participants

Twelve participants (6 men, 6 women) were recruited from the Polytechnic School of Linares at the University of Jaén. Participants were aged between 24 and 60, and included researchers, as well as administrative and services staff. No compensation or reward was offered for their participation. Due to COVID19 restrictions, all participants wore masks.

5.2 Apparatus

The experimental set-up reproduced a real scenario of a common room of a nursing home, where a social robot may be interacting with residents. In these rooms it is also expected that carers talk among each others and to patients and noise from TVs and AC systems. In particular, the emulated scenario considers two simultaneous human talkers where the goal of the ASR is recognising open commands said by the resident person, as it is more difficult for them to talk with fix or template provided grammars. As shown in Annex I, 20 sentences in Spanish are used as the target talker's phrases (one of the 12 participants) and 5 sentences are said in a loop by the IFPs (another participant), with similar topics. The target sentences are shuffled with two possible orders to avoid the possible effect of coincident pairs. The topic of the sentences is related to five basic daily life matters (taking food or medicines, luminosity and temperature conditions, TV volume or noise, visits). The English translations of the 25 sentences are detailed in Annex I. The sentences were selected to reflect different lengths and including in certain cases the name of the robot

(Felipe) as part of the sentence. The sentences said by the participants playing the role of a interferer simulate possible conversations among workers or carers, that may also have common key-words with the target talker's sentences. They are longer sentences as the goal is analysing the system behaviour in the worst case, that is to say, when most of the time both speakers are talking simultaneously. The target participants were also instructed to adapt their voices to levels that they would use in a real-life situation where they had to make themselves understood, in the different conditions they would encounter during the experiment.

The scenario is a research lab with similar dimensions to the common rooms typically used in some residences to allow the interaction among residents. The lab dimensions are: 5.83 m x 10 m x 3.34 m. The XMOS mic. array has been placed on the tripod (simulating the robot position) shown in Fig. 4, in coordinates ($x = 3.0, y = 5.0, z = 0.9$) expressed in meters, as can be seen in Fig 3. Three of the walls are smooth and covered with plaster and one of them is formed by large windows. The reverberation value (RT_{60}) of the room is 1 s.

5.3 Procedure and variables

Upon arrival all the participants read 15 sentences in the position labelled SOI in Fig. 3, with and without the presence of TV noise. This initial tests were made to let the participants to familiarize with the sentences and also to measure the ASR performance just using a DSB beamformer with the utterances of each participant. Thus, this test has allowed to address possible sentences where the ASR could have problems recognising the sentence due to inherent pronunciation biases.

For the performance evaluation in a cocktail party, two participants were talking simultaneously and the noise condition was simulated in the same way (TV on). The

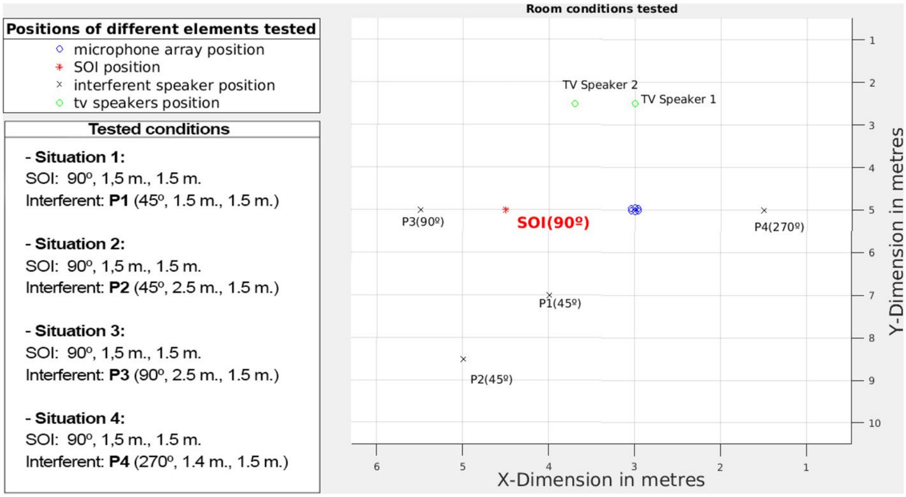

Fig. 3 Conditions tested in a research lab that model different situations in terms of relative positions between the SOI and IFP that could take place on a real environment in a nursing home. The target speaker is always placed at * mark and the IFP is placed at 4 possible locations

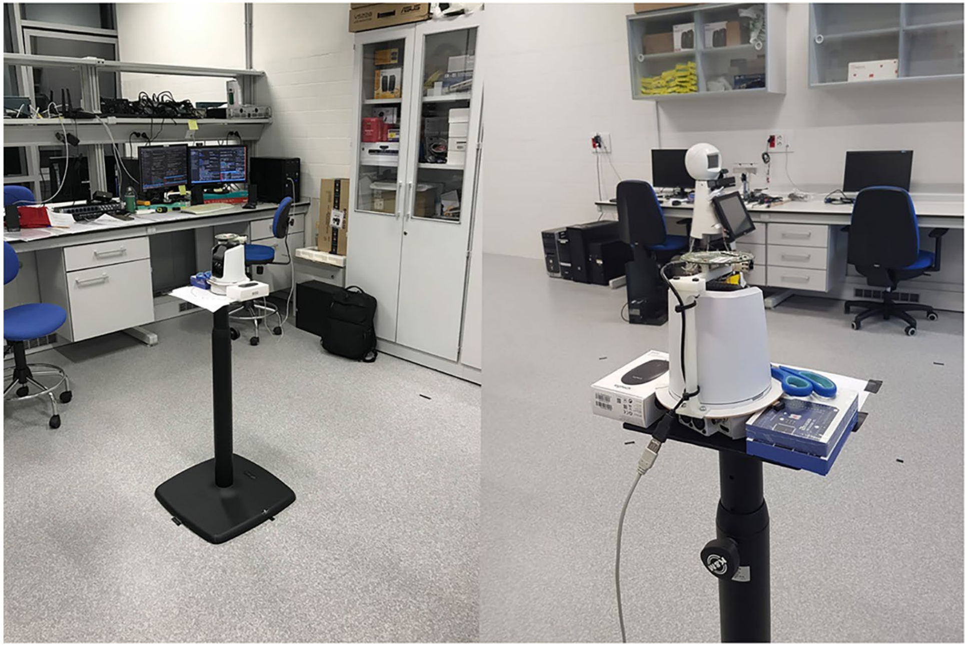

Fig. 4 Real scenario of the tests carried out with the microphone array

positions of the talkers can can be seen in Fig. 3, the SOI was always placed in front of the array (DoA of 90°) at a distance of 1.5 m from the array, simulating a situation of a possible dialog taking place with a robot. The positions of the IFP's were chosen to test the impact of disrupting conversations taking place around the robot and SOI, in four different conditions (see P1, P2, P3, P4 in Fig. 3).

Each participant repeated the test in four conditions, by considering two independent variables, the environment noise and the gender of the IFP. In each of the 4 positions considered, the target talker was instructed to say 4 sentences while the disrupting person was already reciting the disruptive phrases.

The AC system of the laboratory with two outlets in the ceiling was always on, and the noise condition includes always the same TV program (news with reporters talking) being played at the same volume.

The system performance measure has been associated with the ASR's ability to recognise sentences using the Word Success Rate (WSR) measure, which is computed as the Ratio between the number of successful recognised words and the total number of words in the target talker's sentence.

5.4 Setup for the real environment conditions

The evaluated condition with noise (AC system and TV) corresponds with real noise and not simulated pink or white noise. The specific positions are represented in Fig. 3. As there is utterance by utterance variability in power, the SNR in dB has been computed with the average values of the audio recordings with just one speaker talking without noise, obtaining SNR values within a range of [-4.3 dB, 9.6 dB]. In these conditions, [21] has specified that the word intelligibility score for a person with broadband noise is around 50%. Moreover they have evaluated the speech reception threshold as the required SNR level for a 50% intelligibility score for two or more simultaneous talkers. The SIR has been measured considering the relative talkers' voice power obtained while they were talking alone in the different positions considered. It has been obtained an average SIR considering all the conditions within a range of [-5,2 dB, 3 dB] when the target source is a woman and within a range of [-1 dB, 11 dB] when the target source is a man. Miller's study [21] has also established that for these ranges considering one voice interference the word intelligibility would be between 60%-80% when the woman is the target speaker, and between 70%-90%, when the target speaker is a man, due to these SNR differences.

6 Results and discussion

6.1 ASR performance with just one talker

The ASR performance measured through the WSR is on average for all the tested conditions and participants within the range of 70-95%[5]. Therefore, none of the sentences have been removed. When analysing the data by gender, there was a decrease of almost 10% and 20% in ASR performance in the case of a woman being the SOI. These results seem to corroborate what already seemed to be evident in the first tests carried out in [19], and

[5] It should be noted that all tests were conducted with participants wearing a mask

Table 1 Average of WSR results (%, [Standard Deviation]) obtained with Google ASR engine in conditions where the interfering person is placed at 4 different positions (P1, P2, P3, P4), with and without environment noise. The total values refer to the average value considering all the positions

Position	DSB		DSB+Masking	
	Non-noisy	Noisy	Non-noisy	Noisy
P1	44 [26]	44 [26]	54 [21]	49 [24]
P2	74 [13]	67 [16]	76 [10]	67 [12]
P3	76 [15]	71 [18]	71 [17]	67 [20]
P4	57 [21]	57 [24]	67 [17]	61 [20]
TOTAL	63 [23]	60 [23]	67 [17]	61 [21]

they are in line with what has been pointed out by several studies about a possible worse performance of these systems in the case of women and children [8], which may be due to inherent characteristics of the voice such as pitch or power, or to the training of the ASR engines themselves.

One of the subjects had to be discarded, because despite being fluent in Spanish, he had a strong accent (French and Arabic are his mother tongues) which caused the recognition rate to be around 40% and 60% for the conditions with and without noise.

6.2 Speech recognition results in multi-speaker's situations

The average WSR obtained from the raw data, 65%, already indicates a reduction of around 20% of the ASR performance obtained with only one speaker, as expected for ASR technologies prepare for one-to-one interaction.

Results in average of the effect of the two conditioning systems tested in real time are shown in Table 1. No significant difference between algorithms can be found if we consider results obtained in total. The benefits of using the DSB+Masking can be seen with an increase of 10% (without noise) and 5% (with noise) just in conditions where SOI and the IFP are placed at a similar distance from the array (P1 and P4). These results do not corroborate the high improvement detected by the use of DSB+Masking, in the evaluation performed in [19]. Although conditions were not the same (positions, 6 mics. and no mask being used), the performance improvement reported has been of around 20% in conditions similar to these ones (Category A).

In order to isolate the possible problem, an off-line analysis of masking results have been made as in [19], that is to say by processing the raw data split by conditions (P1 to P4) and 4.2 s duration windows with fix DoA angles associated to SOI and IFP in the known position, as can be seen in Table 2. It should be noted that although the time windows do not exactly match those sent to the ASR in real time, in general terms this analysis of the information captured without processing and with the DSB+Masking off-line processing can help to understand the differences in behaviour. If we compare results of Tables 1 and 2, Raw data results indicate that real time behaviour of the two algorithms implemented are not really improving what could be obtained just by feeding the ASR with the data captured. Thus, the real benefit can be seen in results obtained with DSB+Masking computed by fixing DOA values. In particular, for P3 (same DoA for SOI and IFP) the IFPs DoA used was the extracted by the SRP-PHAT in most of these conditions (125°), as it cannot be used the same value.

The high difference in DSB+Making results between on-line and off-line processing indicates the need to include a deliberative module that, by means of tracking and

Table 2 Average of WSR results (%, [Standard Deviation]) obtained with Google ASR engine with raw audio captured (without processing) and DSB+Masking executed offline with fixed DoAs, in conditions where the interfering person is placed at 4 different positions (P1, P2, P3, P4), with and without environment noise. The total values refer to the average value considering all the positions

Position	Offline Raw Data		Offline Masking	
	Non-noisy	Noisy	Non-noisy	Noisy
P1	52 [22]	49 [27]	72 [22]	69 [23]
P2	73 [15]	65 [19]	88 [7]	86 [10]
P3	76 [13]	75 [16]	85 [9]	86 [7]
P4	64 [18]	62 [24]	88 [9]	88 [10]
TOTAL				
	66 [20]	63 [23]	83 [14]	80 [15]

employing an intelligent decision system, keeps the detection of the IFP's DoA more accurate and stable, as it is where more difference have been found among the SRP-PHAT results computed by 1.4 s. Indeed, this negative effect has been more evident in the conditions with noise, due to the constant alternations between the DoAs associated to the TV speakers and the interfering person. Although this reactive behaviour is interesting in order to constantly infer the presence of new sound sources, a deliberative module should filter the angles used in the conditioning module. This means that in Fig. 2, the module "localization and/or separation of sound sources" should be connected to the "Enhancement Acoustic System" through a module that filters these angles depending of previous information gathered from inputs coming from the localization and attentional mechanism modules.

The experimental results were also analysed separately in terms of gender to identify if recognition results are in line with those obtained with just one speaker. The off-line analysis of raw data indicated similar results than those shown in Tables 3 and 4 for DSB. For women a WSR of 59% (without noise) and 51% (with noise) and for men 70% and 71%, respectively. These results are inline with Miller's study [21] due to the differences in SIR ratios obtained. As can be seen in Tables 3 and 4, the negative effect of both, the

Table 3 Average of WSR results (percentage, [Standard Deviation]) obtained with Google ASR engine in conditions where the target speaker is a woman and the interfering person is placed at 4 different positions (P1, P2, P3, P4), with and without environment noise. The total values refer to the average value considering all the positions

General women results (12)						
Positions	DSB		Masking		Offline-Masking	
	non-noisy	noisy	non-noisy	noisy	non-noisy	noisy
P1	43 [24]	33 [25]	51 [19]	38 [24]	68 [22]	61 [26]
P2	70 [14]	62 [18]	74 [10]	61 [12]	89 [7]	84 [12]
P3	75 [16]	64 [20]	68 [20]	62 [14]	85 [6]	86 [6]
P4	49 [24]	46 [27]	61 [19]	59 [22]	88 [10]	90 [9]
TOTAL						
	59 [23]	51 [25]	64 [19]	55 [20]	81 [16]	80 [13]

Table 4 Average of WSR results (percentage, [Standard Deviation]) obtained with Google ASR engine in conditions where the target speaker is a man and the interfering person is placed at 4 different positions (P1, P2, P3, P4), with and without environment noise. The total values refer to the average value considering all the positions

General men results (12)

Positions	DSB		Masking		Offline- Masking	
	non-noisy	noisy	non-noisy	noisy	non-noisy	noisy
P1	46 [30]	54 [22]	57 [22]	60 [17]	75 [23]	77 [16]
P2	78 [11]	73 [13]	78 [11]	73 [11]	88 [8]	89 [7]
P3	78 [14]	78 [14]	73 [13]	71 [15]	84 [11]	86 [7]
P4	66 [16]	67 [15]	72 [13]	64 [18]	88 [8]	86 [11]
TOTAL						
	67 [22]	68 [17]	70 [16]	67 [15]	84 [14]	83 [11]

TV noise and the presence of the IFP is higher (10%) on ASR performance for women, which is in line with results already obtained with only one speaker. Thus, it seems that it may be inherent to the ASR performance when working with lower SNR or SIR. In fact, in the case of men the presence of additional noise does not seem to have any effect and the ASR shows a good performance even without any processing. In the case of women, the degradation due to the presence of noise is evident by almost 10% and, in particular, when the interfering speaker is closer to the SOI (P1). The off-line analysis made with the DSB+Masking algorithm shows that its application helps to reduce these differences in the ASR behaviour.

The effect of the IFP's gender has been analysed separately in the case of the SOI being a man or a woman. In the case of the SOI being a man there are no significant differences in the ASR performance. However, as can be seen in Table 5, in the case of a SOI being a woman, there is a WSR degradation of around 15% when the IFP is a man for DSB (IFP Woman: 62% IFP Man: 48%) and of around 7% for DSB+Masking (IFP Woman: 63% IFP Man: 56%). Thus, it seems that in the case of implementing a more stable and accurate estimation of IFP's DoA, the masking algorithm could minimize to a certain extent gender differences in terms of recognizing performance.

6.3 Impact evaluation of the chosen set-up

In order to prioritise the fast response of the system and therefore the interactivity needed in an HRI scenario, the following decisions were taken:

- the use of only 4 mics. instead of the 6 mics of the array.

Table 5 Gender effect in terms of average WSR results (%, [Standard Deviation]) obtained with Google ASR engine in conditions where the target talker is a woman

Interferer's Gender	DSB		DSB+masking	
	non-noisy	noisy	non-noisy	noisy
Woman	65 [23]	60 [26]	67 [19]	59 [21]
Man	53 [23]	43 [24]	60 [19]	51 [20]

Table 6 Average percentage of adequate SOI position identification by SRP-PHAT algorithm

Percentage SOI position well detected		
Configuration	Environment	
	non-noisy	noisy
C1: 4 mics. and 2048 window samples	59	59
C2: 4 mics. and 4096 window samples	68	60
C3: 6 mics. and 2048 window samples	63	61
C4: 6 mics. and 4096 window samples	**73**	**69**

- the speech fragments size has been selected of 1.4 s, in order to keep more decisions with a static speaker but once the system is used with possible interferers walking (5km/h), frames of 500 ms should be considered for the position estimation.
- the processing of SRP-PHAT decisions in 90 ms windows to keep a more accurate representation of the signal spectrum and to improve the result of the cross-correlation between mic. pairs in environments with high reverberation (typical of very empty and large spaces such as hospital or residential rooms). However, a window of 2048 samples (45 ms) should allow detecting delays associated with distances among speakers of 4 m.
- The implementation of a simple and fast masking algorithm (200 ms) that has given good results in a proof of concept study, but whose main limitation is its dependence not only on the SOI DoA, but also of the possible interfering sources (IFP's DoA).

In order to determine the adequacy of the choice of these parameters, a smaller study has been carried out to analyse the effect of their variation with 3 participants (SOI-IFP genders: famale-male, male-female, male-male).

Firstly, we have analysed the possible impact of the choice of the SRP-PHAT processing window, which was initially set at 4096 samples (90 ms), in order to reduce the possible effect of room reverberation. A reduction has been made to a window of 2048 samples in the conditions with and without TV noise and with audio segments of 1 s. The computational cost is the same but, as can be seen in the first two rows of Table 6, the SOI angle is estimated properly with a slightly higher percentage for 90 ms windows. This study has also identified, as mentioned in Sect. 6.2, that the main limitations for the DSB+Masking algorithm is that in a high percentage of cases the DOAs associated to the SOI and IFP are switched, making even more complex the ASR behaviour with the audio coming from the SOI.

The same evaluation has been made with 6 mics. In this case, the computational cost of the SRP-PHAT component has increased to 1.3 s compared to the 0.6 s necessary with 4 mics. As can be seen in Table 6, slightly better results are obtained that with 4 mics. and also with a higher window (4096 samples).

Secondly, in order to address the improvement of the masking technique compared to another widely used approach as reviewed in Sect. 3, its performance has been compared with GSC dynamic beamformer, according to the implementation described in [19]. In particular the test has been made with a configuration of 6 mics. and 90 ms processing windows. In these conditions, the processing times have been in average of 200 ms and 1600 ms, for DSB+Masking and the time domain implementation of the GSC respectively.

Table 7 Average of WSR results (percentage, [Standard Deviation]) obtained with Google ASR the interfering person is placed at 4 different positions (P1, P2, P3, P4) . The total values refer to the average value considering all the positions

Positions	raw data	DSB+Masking	GSC
P1	47 [25]	74 [10]	69 [9]
P2	63 [13]	79 [7]	90 [1]
P3	74 [13]	78 [13]	79 [9]
P4	66 [24]	70 [11]	67 [13]
TOTAL	62 [11]	75 [4]	76 [10]

Table 7 shows the results in terms of WSR. Although with a high standard deviation and only 3 participants, the results show that the improvement of the GSC is similar to that obtained by DSB+Masking, but with a much higher computational cost. Furthermore, in the P3 condition, the improvement over the response obtained without any ASR processing (raw conditions) is almost the same, since the two speakers are at the same angle and the only difference is that the speaker of interest is closer to the array and facing in that direction. It should be noted that in the case in which the positions obtained from the localisation component are dynamically adjusted without introducing any additional tracking or deliberative decision system, the negative effect that the swapping between the DOAs of the IFP and SOI can have is much greater for the DSB+Masking than for the GSC. In this sense, it is again evident that DSB+Masking is a good option to consider in a system with a small array as the one used, but whose performance is more influenced by errors in the localisation of the speakers.

7 Conclusions and future research

This work has allowed us to verify the benefits and feasibility of integrating simple acoustic signal conditioning techniques (DSB and masking) using low-cost microphone arrays in real-world environments. The benefits of the proposed techniques was evaluated based on the improvement in WSR of an ASR. In particular, the system has been able to recognise open sentences of around 3 s length with 1 s delay due to the use of concurrent modules programmed with an open robotic development platform. The evaluation highlighted the feasibility of taking decisions upon the recognition of open sentences or after requesting prior confirmation of these commands, without introducing excessive delays and therefore keeping certain responsiveness.

The deployment of components that can be run on different machines requires further analysis about the initial parallelization among capturing and processing modules. As well as the analysis of the trade-off between the positive impact of reducing the audio segments length to overcome limitations of real time data delivery in the network among components, which increase with the number of microphones, and its negative impact for the semantic behaviour of the ASR.

The experiment involved 12 participants and have helped to identify inherent fault designs regarding the need of a deliberative system that mediate between the localization and conditioning modules. This system should include a long-term auditory memory or tracking system to allow higher level decisions to be made based on the indications of the attentional mechanism (task to be developed, focus of interest regardless of whether it is closer or not, etc.) or on other information obtained from the environment (identification and visual localization of people, static positions of constant noise sources of noise, etc). However, it should also keep the strongly reactive capacity inherent to auditory processing that allows the rapid localisation of auditory events in a very wide area (alarms, falling objects, etc.), and which can therefore help to provide a robot with the capacity to better understand its surroundings and to emit reactions that allow its behaviour to be perceived as more natural.

On the other hand, in line with current roadmap proposals associated with the use of voice as a mechanism of interaction in the field of HRI [16], additional problems associated with the robot's own noise, navigation, etc. need to be evaluated together with the difficulty already inherent in the problem of the operation of an ASR in multi-speaker environments.

Despite the small sample size and the negative effect of excessive variability in the detection of interfering sources, the study of the differences in the algorithmic performance due to the gender of the speakers indicates that the masking algorithm can improve to some extent possible difference in performance of the ASR, due mainly to differences in SNR or SIR. On the other hand, although it has not been possible to test with elderly people, the strong dependence that the ASR engine exhibits with possible accents of a language, seems to indicate that it will be an added complexity to work adequately with people who may have some kind of diction problem derived from diseases associated with ageing.

All tests have been done with the speakers wearing a mask, an aspect that should be eliminated in future tests. While this may have a negative impact on the performance of the system, both in terms of speaker's localization and ASR performance, it may also be closer to real-life conditions where the person, due to physical limitations, may not constantly looks at the robot or speak clearly, or even due to their health condition, may have to wear a mask.

8 Annex. sentences

8.1 Sentences of the target speaker or SOI

1. Felipe, I can't hear the television.
2. Felipe, can you show me on your screen what they are saying over the loudspeaker? I can't hear it well.
3. I can't see well, can you tell them to turn on the light?
4. I'm hungry, how much is left to eat?
5. I prefer the first menu you showed me.
6. Felipe, I dropped something on the floor, can someone come and pick it up?
7. I'm cold, Felipe, can you turn up the heating?
8. I'm hot, Felipe, can you open the window?
9. Could I go for a walk now?
10. What time is my family coming to visit me?
11. Felipe, what's the weather like today?
12. Felipe, I need to go to the bathroom, can you call someone or come with me?
13. Can you close the window? There seems to be a draught..

14. I really like that program, can you turn it up?
15. Felipe, do you know if I've taken my pill yet?
16. Felipe, can you help me find Juan?
17. Felipe, do you know what time the next activity starts?
18. Felipe, can you tell them to help me change position?
19. It hurts here, can you call the nurse?
20. I hear strange noises, can you tell me what's wrong?

8.2 Sentences of the interfering speaker

1. It seems that the doctor is not coming today because he has a problem, but the appointments will be rescheduled for tomorrow.
2. We should clean up the TV room because it was a mess yesterday.
3. Maybe we should change the sheets in Luis' room today while he is at his doctor's appointment.
4. Maria's medication is going to be changed, we should change the administration schedule on the computer.
5. Tell the maintenance man to come by tomorrow, the heating system doesn't seem to be working properly.

Acknowledgements This work has been funded by the National Research Project TEST-RTI2018-099522-A-C44': "Test-beds for the Evaluation of Social Awareness in Assistance Robotics" and thanks to the collaboration with CSP group at Imperial College London, funded by the Spanish Ministry of Science, Innovation and University through the lectures mobility program (Jose Castillejo's 2018 grant). Most of the information about the typical life in a retirement house and Felipe's robot name have been gathered from the experiences during the work developed in Vitalia Teatinos and supported by the Regional Project AT17-5509-UMA 'ROSI'.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

1. Becker E, Le Z, Park K, Lin Y, Makedon F (2009) Event-based experiments in an assistive environment using wireless sensor networks and voice recognition. In Proceedings of the 2nd International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '09). Association for Computing Machinery, New York, NY, USA, Article 17, 1-8. https://doi.org/10.1145/1579114.1579131
2. Biocca F (1997) The cyborg's dilemma: embodiment in virtual environments. In Proceedings of Second International Conference on Cognitive Technology Humanizing the Information Age, Japan, pp 12-26. https://doi.org/10.1109/CT.1997.617676

3. Chakrabarty S, Habets EAP (2019) Multi-Speaker DOA estimation using deep convolutional networks trained with noise signals. In IEEE J Sel Top Sign Proces vol. 13, no. 1, 8-21. https://doi.org/10.1109/JSTSP.2019.2901664
4. Chang X, Zhang W, Qian Y, Roux JL, Watanabe S (2020) MIMO-Speech: end-to-end multi-channel multi-speaker speech recognition. In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 237-244. https://doi.org/10.1109/ASRU46091.2019.9003986
5. DiBiase JH, Silverman HF, Brandstein MS (2001) Microphone arrays: signal processing techniques and applications. M. S. Brandstein and D. Ward, Eds. Springer-Verlag
6. Evers C, Moore AH, Naylor PA, Sheaffer J, Rafaely B (2015) Bearing-only acoustic tracking of moving speakers for robot audition. In Proceedings of 2015 IEEE International Conference on Digital Signal Processing (DSP), Singapore (Singapur)
7. Evers C, Naylor PA (2018) Acoustic SLAM. IEEE/ACM Trans Audio, Speech and Lang Proc 26, 9, 1484-1498. https://doi.org/10.1109/TASLP.2018.2828321
8. Garnerin M, Rossato S, Laurent B (2019) Gender representation in French broadcast corpora and its impact on ASR performance. In: 1st International Workshop on AI for Smart TV Content Production, Access and Delivery (AI4TV 19), ACM, New York, pp 3?9. https://doi.org/10.1145/3347449.3357480
9. Griffiths L, Jim C (1982) An alternative approach to linearly constrained adaptive beamforming. IEEE Trans Antennas Propag 30, 27-34. https://doi.org/10.1109/TSP.2010.2051803
10. Hu J, Yang C, Wang C (2009) Estimation of sound source number and directions under a multisource environment. In Proceedings of 2009 IEEE/RSJ Int Conf Intell Robots Syst (IROS 2009). St, Louis, MO, USA
11. Jankowski C, Mruthyunjaya V, Lin R (2020) Improved robust ASR for social robots in public spaces. https://arxiv.org/abs/2001.04619
12. Kennedy J, Lemaignan S, Montassier C, Lavalade P, Irfan B, Papadopoulos F, Senft E, Belpaeme T (2017) Child speech recognition in human-robot interaction: evaluations and recommendations. In: 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE/ACM, Vienna, pp 82?90. https://doi.org/10.1145/2909824.3020229
13. Kriegel J, Grabner V, Tuttle-Weidinger L, Ehrenmuller I (2019) Socially Assistive Robots (SAR) in in-patient care for the elderly. Stud Health Technol Inform 260: 178-185. https://doi.org/10.3233/978-1-61499-971-3-178
14. Lazzeri N, Mazzei D, Cominelli L, Cisternino A, De Rossi D (2018) Designing the mind of a social robot. Appl Sci 8, 302. https://doi.org/10.3390/app8020302
15. Lim H, Yoo I, Cho Y, Yook D (2015) Speaker localization in noisy environments using steered response voice power. IEEE Trans Consum Electron 61(1):112–118
16. Matamoros M, Harbusch K, Paulus D (2018) From commands to goal-based dialogs: A roadmap to achieve natural language interaction in RoboCup@Home. In: Holz D., Genter K., Saad M., von Stryk O. (eds) RoboCup 2018: Robot World Cup XXII. RoboCup 2018. Lect Notes Comput Sci vol 11374. Springer, Cham. https://doi.org/10.1007/978-3-030-27544-0_18
17. Martinez J et al (2018) Towards a robust robotic assistant for Comprehensive Geriatric Assessment procedures: updating the CLARC system. In Proceedings of 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), IEEE Press, Nanjing, pp. 820-25. https://doi.org/10.1109/ROMAN.2018.8525818
18. Martinez-Colon A, Perez-Lorenzo JM, Rivas F, Viciana-Abad R, Reche-Lopez P (2018) Attentional mechanism based on a microphone array for embedded devices and a single camera. In Proceedings of the 19th International Workshop of Physical Agents (WAF 2018), November 22-23, Madrid, Spain. https://doi.org/10.1007/978-3-319-99885-5_12
19. Martinez-Colon A, Viciana-Abad R, Perez-Lorenzo JM, Evers C, Naylor PA (2021) Evaluation of a multi-speaker system for socially assistive HRI in real scenarios. Bergasa, Luis M., Ocana, Manuel, Barea, Rafael, Lopez-Guillen, Elena and Revenga, Pedro (eds.) In Advances in Physical Agents II, WAF 2020 vol. 1285, Springer, pp 151-166. https://doi.org/10.1007/978-3-030-62579-5_11
20. Morgan JP (2017) Time-frequency masking performance for improved intelligibility with microphone arrays. Master Thesis in the College of Engineering at the University of Kentucky
21. Miller GA (1947) The masking of speech. Psychol Bull 44:105–129. https://doi.org/10.1037/h0055960
22. Nikunen J, Diment A, Virtanen T (2018) Separation of moving sound sources using multichannel NMF and acoustic trackings. IEEE/ACM Trans Audio Speech Lang Process 26, 281-295. https://doi.org/10.1109/TASLP.2017.2774925
23. Okuno HG, Nakadai K, Kim H (2009) Robot audition: missing feature theory approach and active audition. Springer Tracts in Advanced Robotics (14th Conference Robotics Research), 70: 227-244. https://doi.org/10.1007/978-3-642-19457-3_14

24. Pavlidi D, Puigt M, Griffin A, Mouchtaris A (2012) Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2625-2628. https://doi.org/10.1109/ICASSP.2012.6288455
25. Rascon C, Fuentes G, Meza I (2015) Lightweight multi-DOA tracking of mobile speech sources. EURASIP J on Audio, Speech, and Music Processing 1:1–16
26. Rascon C, Meza I (2017) Localization of sound sources in robotics: A review. Robot Auton Syst 96:184–210
27. Reche PJ et al (2018) Binaural lateral localization of multiple sources in real environments using a kurtosis-driven split-EM algorithm. Eng Appl Artif Intell 69, 137-146. https://doi.org/10.1016/j.engappai.2017.12.013
28. Takeda R, Komatani K (2016) Discriminative multiple sound source localization based on deep neural networks using independent location model, In: 2016 IEEE Spoken Language Technology Workshop (SLT), pp. 603-609. https://doi.org/10.1109/SLT.2016.7846325
29. Wang D, Chen J (2018) Supervised Speech Separation Based on Deep Learning: An Overview. IEEE/ACM Trans Audio Speech Lang Process 26: 1702-1726. https://doi.org/10.3233/978-1-61499-971-3-178
30. Valin J, Michaud F, Hadjou B, Rouat J (2004) Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach. In Proceedings of the IEEE International Conference on Robotics and Automation, 2004 ICRA ’04, New Orleans, USA
31. Valin J, Yamamoto S, Rouat J, Michaud F, Nakadai K, Okuno HG (2007) Robust recognition of simultaneous speech by a mobile robot. IEEE Trans Robot 23: 742-752. https://doi.org/10.1109/TRO.2007.900612
32. Zhuo DB, Cao H (2021) Fast sound source localization based on SRP-PHAT using density peaks clustering. Appl Sci 11, 445. https://doi.org/10.3390/app11010445