Evaluation of a multi-speaker system for socially assistive HRI in real scenarios

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

[1] Universidad de Jaén, Spain,
{amcolon, rviciana,jmperez}@ujaen.es,
[2] University of Southampton, United Kingdom, c.evers@soton.ac.uk,
[3] Imperial College London, p.naylor@imperial.ac.uk

Abstract. In the field of social human-robot interaction, and in particular for social assistive robotics, the capacity of recognizing the speaker's discourse in very diverse conditions and where more than one interlocutor may be present, plays an essential role. The use of a mic array that can be mounted in a robot supported by a voice enhancement module has been evaluated, with the goal of improving the performance of current automatic speech recognition (ASR) systems in multi-speaker conditions. An evaluation has been made of the improvement in terms of intelligibility scores that can be achieved in the operation of two off the self ASR solutions in situations that contemplate the typical scenarios where a robot of these characteristics can be found. The results have identified the conditions in which a low computational cost demand algorithm can be beneficial to improve intelligibility scores in real environments.

Keywords: beamforming, ASR, array, masking, intelligibility

1 Introduction

One of the main current concerns, both of the central and regional administrations, and of society in general, is to analyse the capacity to attend to and cover the growing needs of people in a situation of dependence, and in particular the elderly. As response to this challenge, in recent years, a great effort is being made to develop Socially Assistive Robots (SAR) as tools to support and leverage resources in different tasks in retirement homes [1]. 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 one more handicap to be considered. In this sense, Automatic Speech Recognition (ASR), together with voice synthesis, is a very powerful form of interaction to provide naturalness and social engagement to human robot interaction (HRI), as stated for other examples of computer based service system [2]. However, different studies within the SAR field have concluded about the preferences for other interaction modalities that a priori are not so straight forward or natural, mainly

due to the nowadays technological limitations of these systems [3]. 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 noise of very different kind.

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 corpus used for training, 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 Section 3, research is focused in mainly two approaches: dealing with the changing acoustic properties of the environments by employing de-reverberation algorithms to enhance the voice signal used as an input for these systems and using arrays of microphones (mic) to exploit their space filtering capabilities. There are also recent approaches based on combining these techniques by employing deep learning with complete training procedures as recently reviewed in [4]. However, there is an important trade-off to consider between responsiveness and intelligibility in order to keep interactivity and naturalness perception in HRI systems.

Regarding the use of beamforming techniques, this can be compared as an artificial way to provide attention to a specific source of acoustic information, in a similar way as humans attentional mechanism focuses in a specific conversation, but just exploiting the spatial filtering capabilities. One approach would be that reactive and deliberative agents of an attentional mechanism govern the signal enhancement module to improve the ASR input signal. This study analyses the extent to which beamforming algorithms may improve a commercial ASR engine from audio signals captured by a mic. array that can be mounted on SARs.

2 The use of ASR for socially assistive HRI

Voice synthesis and ASRs 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 in one-to-one interaction. The development of commercial service robots, such as Pepper (Softbank Robotics) and XR1 (CloudMinds), designed to provide services in public places has led to greater requirements in terms of voice recognition in multispeakers environments and with more challenging and dynamic conditions (reverberation, noise, interference, distances).

Thus, recent research projects such as EARS (FP7- 609465, Embodied Audition for RobotS) have focused in improving the auditory perception system of social robots and in the development of open-source robot audition [5] systems further 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 the MFCC features of an ASR embedded in SIG2 robotic head has been evaluated with a MIMO (multiple input multiple output) approach [6], and recent further research [7] is still being made in this line.

The scope of the robotic research into ASR systems is very broad, due to additional problems such as internal noises and the effect of the mic. integration in the robot structure. Nowadays, APIs such as those of Google, Microsoft, IBM, Nuance, etc. are being integrated in the robot system due to the improvement of multimedia distribution in networks and rapid access to SaaS in the Cloud. Thus, recent studies [8] have evaluated in real or "out of the lab" conditions, the performance of these ASR systems in terms of WER (Word Error Rate). In particular, Jankowski et al. have recently established that in Chinese most of them exhibit a degradation not bearable with SNR lower than 15 dB.

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 VR, the Cyborg's dilemma [9] 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. In [10], 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 NAO 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. Their results highlighted that Google outperformed the others with 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 [11], as 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 should be considered.

3 Implementation of an acoustic signal enhancement module for the source of interest in real time

In relation to the separation of sources of interest (SOI), when they are of different kind, and not only associated with speakers, there are many studies that propose to work with statistical models and deep learning strategies (neural networks, Gaussian mixture models, support vector machines) that contemplate the different nature of the signals present. As reviewed in [4], deep learning strategies, based on supervised training, have improved performance of data-driven approach in speech processing. However, these techniques require defining a proper training target and binaural or monoaural features as an input. Under the framework of Auditory Scene Analysis (ASA), the main benefits of the statistical techniques is that is not necessary to know "a priori" aspects such as the geometry of the room, number of sources, number of mics., which is referred to as Blind Source Separation (BSS) systems. In this sense, PCA (Principal Component Analysis), ICA (Independent component Analysis) and NMF (Non-negative Matrix Factorization) methods [12] are widely used, even though they are prepared for signals that vary mainly in amplitude and not in phase.

3.1 Beamforming techniques

Another approach under ASA highlights the role of auditory attention in stream segregation, based on knowledge about the target sound, such as previous knowledge of a speaker's presence (utterance including the name), source location or type of sound (alarm, crying, etc.). Considering location, beamforming techniques may be a gross separation technique previous to other attention-related parameters. These other parameters that allow a more refined separation are commonly based in applying algorithms in time-frequency (T-F) representations.

Different beamforming alternatives have been analyzed for separation purposes [12], and performance it is usually compared to that obtained from Delay and Sum Beamforming (DSB) by measuring the improvement due to attenuating interference from other directions, which is the more basic and less computational demand algorithm. DSB adds multiple mic. signals for target direction in phase, and its generic output for a beamformer focal point, $\mathbf{r}_p$, is given by:

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

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}:

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

However, DSB does not usually allow sufficient reduction of the interfering signal where the signal-to-interference ratio (SIR) is low, in the case of interfering signals energy greater than that of the signal of interest [12]. In particular, dynamic beamformers, such as MVDR (Minimum Variance Distortionless Response), LCMV (Linearly Constrained Minimum Variance) and GSC (Generalized Sidelobe Canceller) are an alternative in situations with moving speakers and unknown room conditions. In the case of MVDR, the objective is to calculate the direction vectors to minimize the output energy from not target directions, which requires the calculation of the co-variance matrix of the captured signals, processing it for each frequency bins and dealing with the complexity of non-inverted matrix computation. Therefore, it has a high computational cost and also requires knowing how much the SOI signal should be amplified to minimize the interfering energy, being difficult its implementation in real time environments. The possible use of LCMV (an evolution of MVDR) has been also analysed, because in addition to minimizing the interfering signal energy at the output, interfering signals can be cancelled if they come from known directions. This solution is interesting because the reactive module already implemented in [13] is able to localize an "a priori" unknown number of interference sources. However, previous studies [14] have indicated that despite the same demands in

terms of computational cost, the recalibration required for changes in the power of the SOI signal is not so limiting, but it is necessary to have a very precise computation of the directions of the interfering signals, which is not the case for a very dynamic scenario, as the once considered with the array mounted in a robot. A GSC implementation has been therefore considered, which is basically like LCMV but cancelling all directions that don't come from the SOI direction, thus simplifying the calculation of the conditioned part. GSC is based on the application of a fixed beamformer with the purpose of estimating the SOI output signal, having therefore part of the interfering signals eliminated, and on the other hand, making an estimation of the noise and interference through the use of a blocking matrix that eliminates the target source. The implementation of the interference modeling part can be done by calculating the adaptive filter weights based on the use of a least squares algorithm, in particular NLMS (Normalized Least Mean Square) [15]. Thus, achieving that in a dynamic way it is possible to model the present interference and directional noise sources. To avoid the computational cost of the null space calculation of the condition matrix, a time domain implementation of the classical adaptative beamformer Griffiths-Jim Beamformer (GJBF) [16] has been programmed. This approach consists in using a DSB in the fixed part and a blocking matrix associated with the subtraction of adjacent signals. The GSC output beamformer is given by:

$$y\left[n\right] = y_s\left[n\right] - \sum_{k=1}^{M-1} \boldsymbol{w}_k^T\left[n\right] \boldsymbol{z}_k\left[n\right] \tag{3}$$

where $y_s\left[n\right]$ represents a fixed beamformer computed with weights $(w_1, .., w_M)$ to each M mic., keeping its behaviour constant with time, $\boldsymbol{z}_k\left[n\right]$ is the kth Blocking Matrix output of O total samples, and $\boldsymbol{w}_k\left[n\right]$ is the kth column of the NLMS filter tap weight matrix $\mathbf{W}$ of length O. The adaptive filters are updated with the NLMS as follows:

$$\boldsymbol{w}_k\left[n+1\right] = \beta \boldsymbol{w}_k\left[n\right] + \mu y\left[n\right] \frac{\boldsymbol{z}_k\left[n\right]}{\left\|\boldsymbol{z}_k\left[n\right]\right\|^2} \tag{4}$$

β is the forgetting factor ($0 < \beta < 1$), the $\left\|\right\|^2$ is the calculation of the Euclidean norm, and μ is the step size parameter ($\mu > 0$) determining how much the filter tap changes with each iteration. Large values of μ result in fast convergence toward a steady-state signal with large misadjustment, and small values result on the contrary situation. β modifies the influence of previous calculated tap weights on the future weights. Therefore, both parameters affect the stability of the NLMS filters and they have been set to $\mu = 0.1$ and $\beta = 0.9$ as in [17].

3.2 Masking techniques

Masking techniques are used together with beamformers to treat speech separation as a supervised learning problem based on computing two-dimensional masks in T-F. Recent studies have evaluated a basic implementation of a DSB

and a binary mask, featured by its low computational demands, for distributed array of mics. in simulated and real scenarios. In particular, this study [17] showed the benefits of a masking implementation in terms of the intelligibility with a distributed mic. array. In this experiment, instead of a distributed approach, we have proposed the use of a small circular MEMS array and the evaluation has considered environments with multiple stationary sound sources, such as human speakers, air-conditioner (AC) machine, TV noise, etc.

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:

$$Y\left[k,i,\mathbf{r}_p\right]=\sum_{q=1}^{Q}G_{pq}\left[k\right]\cdot X\left[k,i,\mathbf{r}_q\right] \tag{5}$$

where i is the index of a particular window, k is the frequency index (frequency bin), $X\left[\cdot\right]$ is a time frequency representation of an audio source signal located at position $\mathbf{r}_q$, and $G_{pq}\left[k\right]$ 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:

$$S_{pq}\left[k,i\right]=\frac{\left|Y\left[k,i,\mathbf{r}_p\right]\right|^2}{\left|Y\left[k,i,\mathbf{r}_q\right]\right|^2} \tag{6}$$

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:

$$T_{pq}=\begin{cases}1, & if\ S_{pq}\left[k,i\right]\geq 1\\ 0, & if\ S_{pq}\left[k,i\right]<1\end{cases} \tag{7}$$

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:

$$T_p\left[k,i\right]=\prod_{q=1,q\neq p}^{Q}T_p\left[k,i\right] \tag{8}$$

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

$$Y^{'}\left[k,i,\mathbf{r}_p\right]=T_p\left[k,i\right]\cdot Y\left[k,i,\mathbf{r}_p\right] \tag{9}$$

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

3.3 Simulated evaluations of intelligibility

STOI (short-time objective intelligibility)[18], which measures the correlation between the short-time temporal envelops of a reference utterance and a separated utterance, is the most commonly used metric in recent years [4] in studies with prerecorded signals and simulated scenarios. The outcomes of these studies have been later contrasted with another intelligibility measurements associated with the use of ASR and scenarios where real noise sources are considered. Studies [19] have evaluated word intelligibility scores with respect to SNR from different kinds of interference (multiple voices, broadband noise and complex tone). Results have proved a speech reception threshold (SRT) of around -2dB for two speakers and a SRT of 2dB for broadband noise, but when the interference is due to voices a higher degradation in terms of word intelligibility score in terms of SNR(dB) has been found considering a threshold of 50% of word intelligibility.

4 Overview of the general system and implementation

The algorithms evaluated have been implemented as part of a system conceived to endorse a robot with capabilities of exhibiting a socially-accepted behaviour. For that purpose, three main modules have been designed, a perception system supported by sensors (visual, audio, odometer, laser, etc) that can be included within a robot housing; a module that implement 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 commanded by two planners: a reactive planner in charge of allowing fast reactions in the robot as a response to the detection of new events (i.e. an alarm, or the a arrival of a new person) and a deliberative one that focuses the robots attention in specific points commanded by rules that consider the task to perform and the environment state. One of the tasks included in the sensorial processing system is the capability of detecting, tracking and identifying the existing people in the surroundings. As part of this module it is vital to recognise the content of a conversation with the robot in typical 'cocktail party' situations, where multiple speakers must be talking simultaneously. In this situation, the attentional mechanism based in the information obtained from the sensorial processing module and the environment model must indicate who is the SOI based in their identification, relative positions, speech content and the task being performed.

In particular, the module of the acoustic signal enhancement is fed with the position information (cylindrical coordinates) of the target speaker and in the case of the masking system the position of the main disrupting speaker is also provided. The implementation is therefore made through three components: acquisition and recording of audio captured; the real time enhancement process (beamforming and masking) and the ASR agent implemented by calls to the REST API of an external ASR engine. Moreover, although still not connected to the enhancement agent, two algorithms for the source localization have been

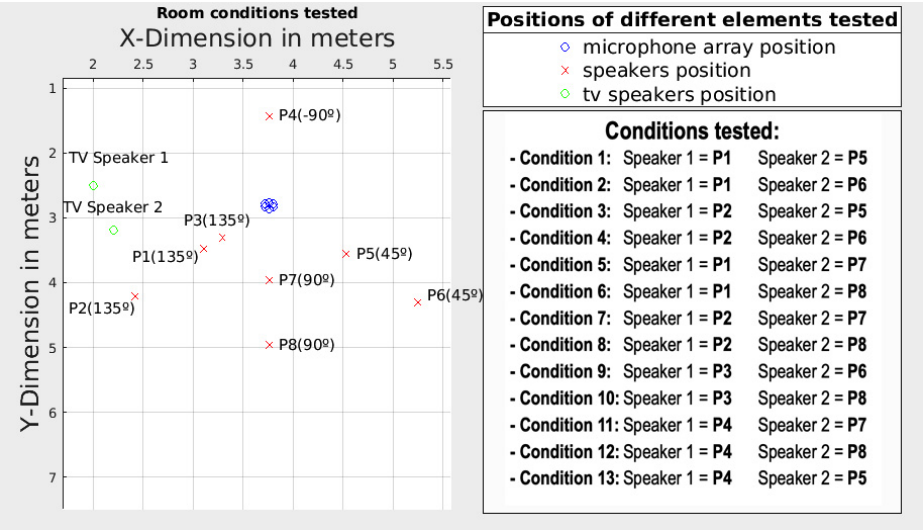

Fig. 1. Conditions tested in a research lab that model some of the most challenging situations in terms of relative positions between the target speaker and the interferer, that could take place on a real environment in an AAL.

already implemented in C++ and evaluated: KDS-EM (Kurtosis-driven split-Expectation Maximization) binaural algorithm [13] and a SRP-PHAT algorithm with multichannel audio captures [21] obtained from a circular array, in particular a XMOS array based on 6 Micro Electrical-Mechanical System (MEMS) mics. These localization modules have been implemented to allow their execution in real time in C++ and through the use of Robocomp, a robotic middleware [3] that among other aspects allows the connection of different agents and components as well as the data exchange through ICE communication middleware[4].

The implementation of DSB, GSC and DSB+Masking algorithms has been carried out considering a capturing time of 15 sec., with a processing window of 80 ms and an overlap of 50%.

The experiment considers the performance analysis in terms of recognition accuracy of two ASR engines, Google Speech API [5] and Watson Speech to Text API from IBM [6], that can be accessed in the ASR component with externals calls through their APIs. In particular, for this study both engines are used through the corresponding web browser utility. In the case of Google Speech API, the model with higher fit in each case has been chosen (Default and Command).

[4] https://zeroc.com/products/ice
[5] https://cloud.google.com/speech-to-text?hl=es-419
[6] https://speech-to-text-demo.ng.bluemix.net/

Table 1. Average and standard deviation of STOI measurements obtained simulating the 13 different situations showed in Fig. 1 in an anechoic room using Matlab

.

Intelligibility	raw	DSB	GSC	Masking
STOI	0.69 (0.08)	0.73 (0.07)	0.71 (0.05)	0.84 (0.04)

5 Evaluation

5.1 Off-line analysis of intelligibility

The three algorithms (DSB, GSC and DSB+Masking) have been evaluated in Matlab by simulating the lab conditions without any noise source and the cocktail party effect with anechoic recordings of two speakers (one male and one female) by using the "Image method for efficiently simulating small-room acoustics"[7], for the specific case of a 192.4 m^3 rectangular room. The speakers have been placed in the 13 conditions showed in Fig. 1. Outcomes of the enhancement process have been evaluated in terms of STOI in average for all the conditions and results are shown in Table 1. These results indicate that in these challenging conditions (close interferer and use of small array of mic) the improvement in terms of STOI of the binary masking applied is high (DSB+Masking 21,7%) but as the expense of knowing the interferers' position. In contrast, the improvement due to the beamformers is low (DSB: 5,8%; GSC: 2,9%).

5.2 Experimental procedure and variables

The experiments consider an emulation of a common room of a retirement home, where a social robot may be interacting with residents. In these rooms, it is normal the presence of caregivers talking among them and typical noise sources associated with TV and AC systems. In particular, the emulated scenario considers two simultaneous speakers where the goal of the ASR is recognising the open commands of the resident speaker, being more difficult than to work with fix or template provided grammars.

Both speakers talk in Spanish, to properly measure ASR performance. One of them emulates a resident asking the robot for help regarding five basic daily life matters (taking food or medicines, luminosity and temperature conditions, TV volume, visits). They were chosen of different length and including in certain cases the name of the robot (Felipe) as part of the sentence. The other speaker was emulating typical sentences of a worker or caregiver, that may also have common key words with the one associated to the sentences addressed to the robot. 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.

[7] http://web.engr.uky.edu/~donohue/audio/Arrays/MAToolbox.htm

The objective of evaluating two ASRs is to analyse whether the enhancement module behaves in the same way with ASRs with different performances under the optimal conditions of a one-to-one interaction. Google ASR has been chosen because it has already been labelled as one of the best options [10], and IBM because it has lower performance. The experiment has been done first with just one speaker (ideal one-to-one interaction) and after that with multiple speakers. Testing in both cases ASR engines in two conditions: with and without the presence of noise. Regarding the positions of the speakers, they were chosen with the goal of analysing the worst possible conditions in terms of speakers closeness (positions and angles) to the array for both, the target and the interfering speaker. Thus, the 13 conditions showed in Fig. 1 have been evaluated. For each condition, 4 tests were carried out: two in which the woman acts as the target speaker saying two phrases from the list of target phrases while the male says long disruptive sentences, and two tests in which they change their roles. This procedure gives rise to 52 tests carried out under conditions without background noise and another 52 with noise. The noise condition includes the AC system of the laboratory working and a TV program (news with reporters talking).

The preliminary one-speaker test was conducted by placing the less average power voice speaker (woman) in the four positions around the array considered in the experimental set-up. In each position, the woman has said five different sentences and the sentences analysed with the two ASR systems were analysed for the raw data and after applying the three enhancement techniques.

ASR systems usually under-perform with voices associated with young users or women in the presence of male voices, most of the time due to the lower power voice and pitch differences, but also as stated in [11] due to the existence of less corpora for children and/or women used to generate models. For this reason, one of the speaker was a man (speaker 2) and the other a woman (speaker 1) in order to evaluate the behaviour of the ASR engines after applying the enhancement algorithms in this complex situation.

In order to evaluate the ASR intelligibility with the goal of analysing the extent to which results correspond with the STOI measurements, two measurements have been employed:

- **WER** (Word Error Rate): Ratio between the number of wrong recognised words and the total number of words in the target speaker's sentence.
- **RAR** (Relaxed Accuracy Rate): Ratio between the number of key words (adjectives, nouns, verbs and adverbs) recognised between the total amount of key words in target speaker's sentence.

5.3 Setup for the real environment conditions

The real scenario is a research lab with similar dimensions to the common rooms typically used in some residences. Its dimensions are: $5.83\ m \times 10\ m \times 3.34\ m$. The XMOS mic. array has been placed on a table in coordinates ($x = 3.76$, $y = 2.81$, $z = 1.21$) expressed in metres. Three walls are smooth and covered with

Table 2. WER ratio [% RAR] measurements obtained with Google and IBM ASR engines for only one speaker (the woman as she is the speaker with lower voice power) in conditions with and without environment noise.

	Non noisy environment			Noisy environment-SNR (dB)[-0.8, 3.7]		
ASR	raw	DSB	GSC	raw	DSB	GSC
Google	**0.02 [97]**	**0 [100]**	**0 [100]**	**0 [99]**	**0 [100]**	**0 [100]**
IBM	**0.37 [59]**	**0.18 [82]**	**0.29 [69]**	0.64 [32]	**0.42 [56]**	0.65 [34]

Table 3. WER ratio [%RAR] measurements with two simultaneous speakers obtained with Google and IBM ASR for the 13 conditions tested with and without environment noise.

	Non noisy environment				Noisy environment [(dB)[-0.8, 3.7]]			
ASR	raw	DSB	GSC	Masking	raw	DSB	GSC	Masking
Google	**0.44 [55]**	**0.40 [60]**	**0.47 [53]**	**0.12 [88]**	0.50 [49]	**0.48 [51]**	0.60 [41]	**0.14 [86]**
IBM	0.8 [22]	0.62 [38]	0.78 [23]	0.52 [49]	0.95 [5]	0.85 [16]	0.94 [6]	0.72 [28]

plaster and one of them is formed by large windows. The room is featured with a reverberation of 1 s (RT_{60}).

The noise condition considers real noise sources (AC system and TV) placed in the positions represented in Fig. 1. As there is utterance by utterance variability in power, the SNR in dB has been computed with the average values of the one-speaker audio recordings in the different positions, obtaining SNR values within a range of [-0.8 dB, 3.7 dB]. In these conditions,[19] has specified that the word recognition accuracy for a person with broadband noise is around 50%. The SIR (Signal to Interference Ratio) has been measured considering the relative speakers' voice power in the 4 positions where they have been placed, by recording 5 different sentences in each position. It has been obtained an average SIR considering all the conditions within a range of [-11 dB, 3 dB] when the target source is a woman and within a range of [-3 dB, 11 dB] when the target source is a man. Miler's study [19] has also established that for these ranges, considering one voice interference, the word intelligibility would be between 50%-80% when the woman is the target speaker, and between 70%-90% when the target speaker is a man.

6 Results

The performance of the two ASR engines in terms of WER and RAR, in the ideal situation with one-to-one interaction (only a woman), is shown in Table 2. As can be seen, results indicate that, Google's ASR always reaches the highest intelligibly scores with and without noise, while IBM performance fails mainly due to conditions where the speaker is at 2 m from the array. In the case of IBM's ASR engine the WER reaches 0.37 (59% for RAR) in the condition without noise,

Table 4. Results with two simultaneous speakers in terms of WER ratio and [% RAR] using Google ASR for the most challenging situations (A: both speakers close; B: distant speakers; C:inteferer closer than the target speaker), with and without noise.

			Non noisy environment			Noisy environment		
Cat.	Angle	SOI	DSB	GSC	Masking	DSB	GSC	Masking
A	45°	female	1 [0]	1 [0]	**0.14 [83]**	1 [0]	1 [0]	0.67 [33]
		male	**0.11 [83]**	**0 [100]**	**0 [100]**	**0.11 [83]**	**0.11 [83]**	**0.11 [83]**
	90°	female	1 [0]	1 [0]	**0.25 [75]**	1 [0]	1 [0]	**0.11 [88]**
		male	**0.07 [90]**	**0 [100]**	**0 [100]**	**0 [90]**	0.59 [35]	**0 [100]**
	180°	female	**0.31 [83]**	**0.27 [69]**	**0 [100]**	0.61 [38]	0.75 [27]	**0 [100]**
		male	**0 [100]**	**0 [100]**	**0 [100]**	**0 [100]**	**0.07 [90]**	**0 [100]**
	225°	female	**0.5 [50]**	**0.43 [62]**	**0 [100]**	1 [0]	1 [0]	**0 [100]**
		male	**0.1 [87]**	**0.05 [100]**	**0 [100]**	**0 [100]**	**0 [100]**	**0 [100]**
B	45°	female	0.84 [19]	1 [0]	**0.11 [92]**	1 [0]	1 [0]	0.66 [42]
		male	**0 [100]**	**0.18 [87]**	**0 [100]**	**0.12 [87]**	**0.5 [50]**	**0.18 [75]**
	90°	female	0.75 [25]	1 [0]	**0 [100]**	1 [0]	1 [0]	**0.15 [85]**
		male	**0.14 [80]**	**0.57 [50]**	**0.15 [80]**	0.57 [40]	0.88 [20]	**0.03 [100]**
C	45°	female	1 [0]	1 [0]	0.69 [30]	1 [0]	1 [0]	1 [0]
		male	**0.47 [51]**	0.62 [37]	**0.4 [62]**	0.75 [25]	1 [0]	**0.25 [75]**
	90°	female	1 [0]	1 [0]	0.57 [37]	1 [0]	1 [0]	0.57 [37]
		male	1 [0]	1 [0]	**0.07 [91]**	1 [0]	1 [0]	**0.07 [91]**
	180°	male	**0.35 [66]**	**0.5 [50]**	**0.07 [91]**	**0.5 [50]**	**0.21 [74]**	**0 [100]**

and the degradation not detected in Google's ASR due to the environment noise, in this case has been high (WER score of 0.64). The effect of beamformers is positive but most evident for DSB, with a reduction in the WER scores of: 52% without noise and 35% with noise. Thus, as expected, in these noise conditions (TV, AC-system), that is to say broadband noises or background interference, DSB outperforms GSC in terms of ASR scores improvement. The same tests have been carried out for the other speaker (a man) detecting the same performance.

Average ASR results for two simultaneous speakers in the 13 conditions shown in Fig. 1, with and without environment noise, are shown in Table 3. As expected, results indicate a clear degradation of performance without any signal enhancement (nearly 40% increase of WER for raw data), for both engines, as they employ training models considering one person speech recognition. In conditions, where the target and interferer are both relatively close between them and the array, just the T-F masking algorithm has improved the scores, which agrees with the simulations performed in terms of STOI, but indicating that the intelligibly measured with STOI underestimated the improvement of the masking in real conditions. Regarding the differences between engines in the effect of the target voice enhancement, for Google the masking has increased the performance in more than 70% for conditions with and without noise. In the case of the IBM ASR engine, this improvement could bring performance to a bearable

value (0.5 WER and 50%RAR) only for conditions without environment noise, with an increase of 35% (WER reduction from 0.8 to 0.52).

From the 52 tests performed, the results obtained in the most challenging situations have been analysed in detail, classifying them according to: the difference in metres between SOI and interferer, the azimuth angle separation, whether the target speaker was a man or a woman, resulting 3 categories associated with the experimental conditions of Fig. 1, as follows:

A: Both speakers close: 1 m. (C5- 45º, C1-90º, C11-180º, C13-225º)
B: Both speakers far away: 2 m. (C8-45º, C4-90º)
C: The interferer closer (1 m) than the target speaker (2 m) (C7 female-45º, C6 male-45º, C10 male -45º, C3 female-90º, C2 male-90º, C12 male-180º).

When analysing the results by category, it becomes clear that in the case of a good ASR engine like Google's (See results in Table 4), the obvious need for masking starts to be evident for a SRI below 0dB. In our experiment, in the case of a man being the target speaker (-3dB, +11dB) this is observed in category C and in the case of a woman (-11dB, +3dB) in all the categories analyzed in detail. Even in category C the masking no longer allows the WER to be below 0.5 or the RAR to exceed 50%. In the case of an engine like IBM's, performance was worse due to the higher effect of a low SNR and SIR. The beamformers effect has not been able to improve sufficiently the signal in these challenging cases and only the masking has been able to improve performance when the SIR is higher than zero. In our experiment this happens only in category A and only when the target speaker is a man.

7 Conclusions and future research

The study presented has made it possible to evaluate the extent to which a simple signal conditioning algorithm with low computational cost makes it possible to improve the degree of use in terms of intelligibility of a low-cost array that can be easily integrated into a robot. Despite the existence of numerous studies that evaluate different proposals for a module that allows speech signal enhancement of a target speaker in a multi-speaker scenario with simulated environments and/or standardized test-benches, the experiment performed in challenging conditions for a robot equipped with an ASR in conditions closer to real ones has highlighted the feasibility of improving efficiency of this natural interaction modality by using a basic approach DSB+Masking already tested with distributed array in a room, which is less flexible for a solution based on a mobile robot.

The time domain implementation of a dynamic GSC beamformer has exhibited very low results being outperformed by a fix DSB beamformer mainly due to the problem of the SOI signal cancellation of the blocking matrix, which is in accordance with previous studies that have proposed as a solution implementations based on adaptive filter coefficients in linear arrays [20] or other type of null-steering algorithms [22]. Thus, these approaches that avoid the need of

detecting the interferer position as in the masking approach should be tested in dynamic situations and with small arrays to analyse their feasibility.

On the other hand, assuming that the localization module has a certain dynamic precision determination of at least one speaker and one interferer, the basic masking proposal evaluated has been addressed as beneficial in terms of intelligibility as previously obtained with a distributed array. The 'cocktail party' situation with more than two participants has been emulated with a realistic noise condition and intelligibility performance has remained with feasible scores. However, further research is needed once the integration among the localization, signal enhancement, and ASR modules is successfully completed, in order to address the behaviour not only in terms of intelligibility but keeping under certain threshold recognition delays that may interfer with interactivity perception.

Considering the final objective of including a robot as part of an AAL that serves to as tools that facilitate the development of certain monotonous but important tasks, such as actively monitoring an older person's degree of autonomy or assisting them in everyday tasks, it is necessary to evaluate the capacity of the system overcoming age-impairment interaction capabilities and gender limitations, to draw conclusions about the feasibility of assuring ASR performance with limited dialogues in real conditions.

8 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'.

References

1. Kriegel, J. et al.: Socially Assistive Robots (SAR) in In-Patient Care for the Elderly. Stud Health Technol Inform. 260, 178–185 (2019).
2. Beckert, E. et al.: Event-based experiments in an assistive environment using wireless sensor networks and voice recognition. In: 2nd International Conference on PErvasive Technologies Related to Assistive Environments (PETRA09), pp. 1–8. ACM, Corfu (2009).
3. Martínez, J. et al.: Towards a robust robotic assistant for Comprehensive Geriatric Assessment procedures: updating the CLARC system. In: 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 820–825. IEEE Press, Nanjing (2018).
4. Wang, D., Chen, J.: Supervised Speech Separation Based on Deep Learning: An Overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 26, 1702–1726 (2018).

5. Okuno, H.G., Nakadai, K., Kim, H.: Robot Audition: Missing Feature Theory Approach and Active Audition. Springer Tracts in Advanced Robotics (14th Conference Robotics Research), 70, 227–244 (2009)
6. Valin, J., et al.: Robust Recognition of Simultaneous Speech by a Mobile Robot. IEEE Transactions on Robotics 23, 742–752 (2007).
7. Chang, X., et al.: MIMO-Speech: End-to-End Multi-Channel Multi-Speaker Speech Recognition. In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 237–244 (2020).
8. Jankowski, C., Mruthyunjaya, V., Lin, R.: Improved Robust ASR for Social Robots in Public Spaces (2020).
9. Biocca, F.: The cyborg's dilemma: embodiment in virtual environments. In: Second International Conference on Cognitive Technology Humanizing the Information Age, pp. 12–26. Japan (1997).
10. Kennedy, J. et al.: Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations. In: 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 82?90. IEEE/ACM, Vienna (2017).
11. Garnerin, M., Rossato, S., Laurent, B.: 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), pp. 3–9. ACM, New York (2019).
12. Nikunen, J., Diment, A., Virtanen, T.: Separation of Moving Sound Sources Using Multichannel NMF and Acoustic Trackings. IEEE/ACM Transactions on Audio, Speech, and Language Processing 26, 281–295 (2018).
13. Reche, P. J., et al.: Binaural lateral localization of multiple sources in real environments using a kurtosis-driven split-EM algorithm . Eng. Appl. Artif. Intell. 69, 137–146 (2018).
14. Souden, M., Benesty, J., Affes, S.: A Study of the LCMV and MVDR Noise Reduction Filters. IEEE Transactions on Signal Processing. 58, 4925?4935 (2010).
15. Yu, Z. L., Er M. J.: An extended generalized sidelobe canceller in time and frequency domain. In: 2004 IEEE International Symposium on Circuits and Systems, pp. 629?633. IEEE, Vancouver (2004).
16. Griffiths, L., Jim, C.: An alternative approach to linearly constrained adaptive beamforming. IEEE Transactions on Antennas and Propagation. 30, 27–34 (1982).
17. Morgan, J.P.: Time-Frequency Masking Performance for Improved Intelligibility with Microphone Arrays. Master Thesis in the College of Engineering at the University of Kentucky. (2017)
18. Taal, C. H., et al.: A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: IEEE Transactions on Antennas and Propagation, pp. 4214–4217. IEEE. Dallas (2010).
19. Miller, G. A. The masking of speech. Psychological Bulletin. 44, 105–129 (1947).
20. Ni, F., Zhou, Y., Liu, H.: A Robust GSC Beamforming Method for Speech Enhancement using Linear Microphone Array. In: IEEE 21st International Workshop on Multimedia Signal Processing (MMSP). pp.1–5. IEEE. Kuala Lumpur (2019).
21. Martinez-Colon, A. et al. Attentional Mechanism Based on a Microphone Array for Embedded Devices and a Single Camera. In: Workshop of Physical Agents. pp. 165–178. Springer. Madrid (2018).
22. Li, C., Benesty, J., Chen, J. Beamforming based on null-steering with small spacing linear microphone arrays. J Acoust Soc Am. 143, 2651–2664. (2018)