Multiple Direction-of-Arrival Estimation for a Mobile Robotic Platform with Small Hardware Setup

Caleb Rascon and Luis Pineda

Abstract Knowledge of how many users are there in the environment, and where they are located is essential for natural and efficient Human-Robot Interaction (HRI). However, carrying out the estimation of multiple Directions-of-Arrival (multi-DOA) on a mobile robotic platform involves a greater challenge as the mobility of the service robot needs to be considered when proposing a solution. This needs to strike a balance with the performance of the DOA estimation, specifically the amount of users the system can detect, which is usually limited by the amount of microphones used. In this contribution, an appropriately carriable small and lightweight hardware system (based on a 3-microphone triangular system) is used, and a fast multi-DOA estimator is proposed that is able to estimate more users than the number of microphones employed.

Key words: HRI, lightweight, microphone array, mobile, multiple direction of arrival, reverberation, service robot

1 Introduction

The problem known as Multiple Direction-of-Arrival (multi-DOA) Estimation provides a unique challenge when being carried out in a mobile hardware platform, such as service robots. However, it plays an essential part of a natural Human-Robot Interaction (HRI), as it is important to know from where the users are talking to the robot and how many are there in the environment.

Caleb Rascon
Universidad Nacional Autónoma de México, e-mail: caleb.rascon@iimas.unam.mx

Luis Pineda
Universidad Nacional Autónoma de México e-mail: lpineda@unam.mx

From the technical point of view, knowing the direction of the user in relation to the robot can benefit other system modules. For instance, once the direction of the user is known, voice recognition can be improved using directional noise cancellation [7] or by simply turning a directional microphone in the direction of the user.

In addition, it is well known that face detection and recognition provide rich information relevant to HRI: the identity of the user, the direction the user is looking at, his mood, etc. [4, 20]. However, such analysis is carried out by visual means, and the user cannot always be expected to be in the line of sight of the robot. When dealing with human beings, by sorts of providence, the mouth is expected to be in the vicinity of the face of the user, which means that knowing the direction of the user by speech alone also provides a good heuristic of the location of his/her face. Using this information to face the user tackles the visual-range-limitation issue straight on.

Moreover, the robot may be expected to be in a situation where several users are in the environment and actively speaking to the robot, such as taking a food order or while guiding a group of users in a tour. Knowledge of the amount of users and from where are they talking to the robot can be used to provide acoustic cues to separate several streams of audio data from the environment based on the Direction-of-Arrival (DOA) of the various sound sources and provide the single-source streams to the Automatic Speech Recognizer. This provides the functionality of being able to carry out ASR of multiple users that may be interrupting each other, an occurrence bound to happen in a multiple-user scenario.

From the perspective of the user, the action of the robot facing him/her when being talked to acts as a type of bodily feedback which the user will naturally interpret as if the robot is 'putting attention' to him/her. This interpretation is an important part of a successful HRI, as the robot reacts in a way expected by the user and, at the same time, provides important feedback that makes the user feel acknowledged at the very beginning of the interaction. Meaning that, with only this seemingly trivial act, a good preamble to HRI is put forward.

In addition, the location of the user is an important variable in HRI. During a human-robot conversation, the phrase "robot, come here" may be emitted by the human. In this situation, even if the phrase was recognized correctly, the robot may know that it needs to move, but, because the term 'here' lacks context, it will not know **where** to move. Knowing the direction of the user in regard to the robot is an essential variable in the estimation of the location of the user in the environment. In a 3-dimensional polar coordinate system, the horizontal angle (i.e. the direction of the user) is one of three values that define a location (the other two being: vertical angle and distance from origin). Using heuristics from the environment, the DOA of the user can be used to segregate the locations where the user is most probably at. This means that when using ASR and DOA estimation conjunctively, the aforementioned phrase can be contextualized and stripped of its vagueness. From the user's point of view, only a vague command is enunciated and the robot is able to carry it out, which is more 'natural' for the user than to position themselves in front of the robot.

Unfortunately, there are many challenges in the estimation of the DOA of the sound source. Reverberation is prevalent in the locations where a service robot is expected to be (supermarkets, restaurants, condominiums, etc.) and has been shown to hinder considerably the effectiveness of current DOA estimators [2]. Moreover, too many sound sources may drown the acoustic environment, complicating the estimation process. A sophisticated audio capturing system may be able to overcome these issues, such as the one proposed in [19] that used a 24-microphone 1-D array for precision. However, the application landscape of service robots provides a unique challenge for the multi-DOA estimation topic: a high amount of microphones may be impractical to carry by many of the currently-in-use service robots [22, 10], such as our in-house robot, Golem-II+, herein described.

Golem-II+ is a service robot built with a primary focus on HRI. It is integrated by a cognitive architecture focused on HRI, termed Interaction-Oriented Cognitive Architecture (IOCA) [13], which can take advantage of different types of information interpreted from the world, including the direction of the user. Because Golem-II+ is a conversational robot, it is of interest that it is able to detect and carry out conversations with several users at any point. This implies that the system that is to be estimating the multiple DOAs of the environment, needs to be sufficiently light on the hardware side for the robot to carry and not hinder its mobility, but robust enough in the software side to handle different types of noise and disturbances, as well as simultaneous speech from various sources. Moreover, such a system should be able to estimate the direction of the users in a -179°– 180° range, as no assumption can be made of the location of the users in the environment, and fast enough to do so with small utterances from the users. It is important to note, then, that the Multi-DOA Estimation problem is further complicated in a mobile robotic platform, and provides an interesting and unique challenge for current techniques.

This contribution is organized as follows: Section 2 is a brief review of current algorithms that aim to estimate the direction of one or more sound sources; Section 3 describes the proposed system; in Section 4, the results of the evaluation of the system on a service robot placed on a highly acoustically-complex scenario are provided; and in Section 5, conclusions and future work are discussed.

2 Background on Source Direction-of-Arrival Estimation

Estimating a Sound Source Direction of Arrival (DOA) is a well written-about topic in Signal Processing. It has been proven useful in applications ranging from fault monitoring in aircrafts [19], to intricate robotic pets [6], to close-to-life insect emulation [5]. In addition, the principles employed in DOA estimation have been applied in the design of hearing aids [7].

Having two audio sensors (i.e. microphones), the Inter-aural Time Difference (ITD) is the delay of a sound from one microphone to the other. Its calculation is usually based on the Cross-Correlation Vector (CCV) between the two captured signals. One of the simplest way to calculate the CCV is by applying Equation 1.

$$CCV[k] = \frac{\sum_i (x_i - m_x)(y_{i-k} - m_y)}{\sqrt{\sum_i (x_i - m_x)^2}\sqrt{\sum_i (y_{i-k} - m_y)^2}} \quad (1)$$

where x and y are the two discrete signals being compared; k is the point at which y is being linearly shifted and the correlation is being calculated; m_x and m_y are the mean values of x and y, respectively. The ITD is the k value of the highest correlation measure in the CCV. It is one the features most used for DOA estimation, particularly with two-microphone arrays, as in [9] where it provided limited results. The ITD yields a clear relation to the direction of the source, described in Equation (2).

$$\theta = arcsin\left(\frac{ITD \cdot V_{sound}}{F_{sample} \cdot d}\right) \quad (2)$$

where θ is the DOA angle; ITD is the Inter-aural Time Difference in number of samples; V_{sound} is the speed of sound ($\sim$ 343 m/s); F_{sample} is the sampling frequency; and d is the distance between microphones.

The Inter-aural Intensity Difference (IID) is the difference in magnitude between both microphones and can also be used for DOA estimation, although a training stage is usually necessary for it to be useful, as it was observed in [10].

In [2], the concept of Inter-aural Coherence (IC) is introduced, which is the highest correlation measure of the CCV. If a high IC is present, the signals are deemed coherent and, thus, an analysis using ITD and/or IID can proceed. This methodology was implemented in [6], and it was observed that it didn't improve DOA estimation when dealing with complex signals (e.g. more than one source, reverberation present, etc.).

A popular methodology for DOA estimation in robotic platforms is to use a microphone array with, usually, two microphones, as it is proposed in [11]. The reasoning behind using only two microphones in robotic platforms ranges from that of practicality (it is lightweight), to that of biological similarity [23, 3] where the robot is meant to be the most human-like possible. However, doing so comes with four main problems.

ITD-DOA Non-Linear Relation. In Figure 1, the DOA is plotted against the ITD, and it can be seen that in the -50°–50° range, the relation between both seem close-to-linear. However, in the outer ranges, the relation becomes exponential. This causes major errors when estimating angles that are located in the sides of the robot [11]. This issue can be overcome by only estimating DOAs in the linear range, but, as it will be described, the DOA range is already limited as it is.

Limited DOA Range. As it can also be seen in Figure 1, a 2-mic array only estimates DOAs in the -90°– 90° range. This can be surmounted by implementing 'artificial ears' that can detect if the sound source is coming from the front or back of the robot, but it has been proven impractical [16]. This can also be tackled by a two-phase strategy: a first pair of signals can be used to estimate an initial DOA, the robot can then rotate briefly, and then another pair of signals can be acquired

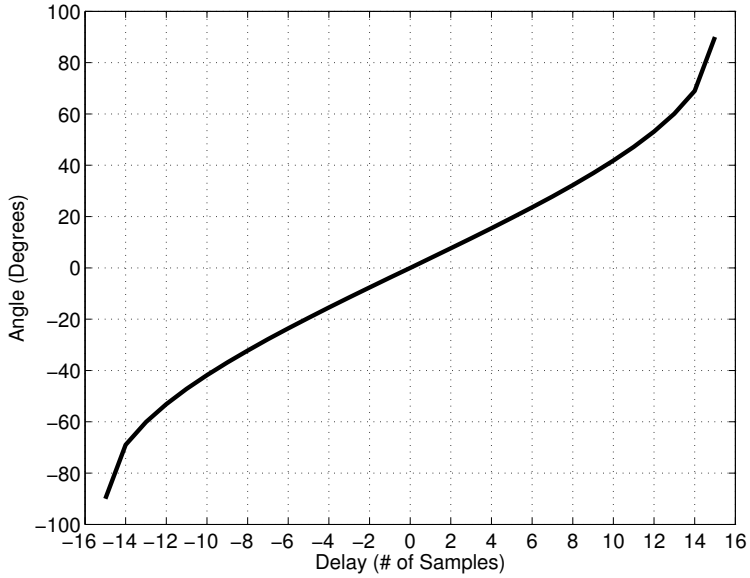

Fig. 1 DOA (or *Angle*) in degrees vs. ITD (or *Delay*) in # of samples.

to estimate a second DOA. A comparison between the DOAs results in an angle estimation in the -179°– 180° range, but has its own set of issues: it requires considerably more time than when using one DOA estimate, the required rotation may hinder navigational requirements, and the user may be moving as well, rendering the DOA comparison mute.

Reverberation Sensitivity. The estimation of the ITD, based on the calculation of a CCV, can be very sensitive to reverberations and other noise sources [23](pp. 213-215). This may result in significant errors in the DOA estimation without any form of redundancy.

Number of Microphones. A 2-microphone array has rarely been used for multi-DOA estimation, as it provides sparse information from the environment. Adding more microphones generalizes the strategy, as a 2-microphone array is an instantiation of classic reverse beamforming techniques [19], which create a noise map of the environment, and then, by using metrics such as energy levels, propose possible sources of sound and their respective DOAs. However, to obtain a high resolution noise map, and, thus, a precise DOA estimation, beamforming techniques require a large quantity of microphones, which is impractical for mobile robotic platforms. In addition, the more popular 1-dimensional (1-D) beamforming methodologies are also bounded by the first three problems described earlier, and 2-D arrays can be cumbersome to the mobility of the robot.

The topic of how many microphones to use in a service robot is intrinsic to the nature of the application, as it is important for the audio capture system to be mobile.

A many-microphone solution may provide good results, such as the one proposed in [22] where the sources were separated from each other, in order to enhance speech recognition, and as a preamble for DOA estimation. However, it required an array of 8 microphones positioned in a cube-like manner to work, doubling the height the robot occupied without it.

The other side of the argument is to use one microphone, such as the work described in [16], where the DOA of a source was able to be estimated by implementing an 'artificial ear'. Unfortunately, the sound was required to be known *a priori* and any modification to the ear (even its location in relation to the microphone) required re-training.

A popular technique is the Multiple Signal Classification algorithm (MUSIC) [17], which is able to detect the Direction of Arrival (DOA) of as many sources as one less the available microphones (e.g. 1 source with 2 microphones, 2 sources with 3 microphones, etc). It does this by projecting the received signals in a DOA subspace, based on their eigenvectors, similar to Principal Component Analysis. It was applied in [8] with good results, although it has been observed that its performance decreases considerably in the presence of reverberation [23] (pp. 169).

In this contribution, a technique is proposed where a small hardware system (based on only 3 microphones) is able to estimate multiple DOAs, as much as 4 simultaneous sources.

3 Proposed System

The work carried out in [15], which, in turn, is based on the proposal presented in [14] is the basis of the proposed system in this contribution. It is comprised by three modules that are described extensively in the rest of this section:

1. *Audio Acquisition*. Obtains audio data from the microphones and provides it to the Initial DOA Estimation module.
2. *Initial DOA Estimation*. Estimates, from the audio data, an initial, fast, but reliable DOA estimation of a single sound source in the environment.
3. *Multi-DOA Tracking*. Carries out dynamic clustering of the incoming DOA estimations, and proposes clusters of DOAs as sound sources.

3.1 Audio Acquisition

As it will be described in the following section, the hardware is comprised by three omnidirectional microphones, and, because the DOA estimation is based on an ITD measure, it requires that the audio from the three microphones be acquired simultaneously as well as in real-time. For this purpose, the JACK Audio Connection Toolkit [1] was employed, as it can sample at rates of 44.1 kHz and 48 kHz, provid-

ing a good resolution for ITD calculations, and it does so without slowing down the other robotic software modules.

3.2 Initial DOA Estimation

The Initial DOA Estimation is carried out by the technique described in [14]. It avoids the problems that arise when estimating a DOA using 1-D microphones arrays (described in Section 2), and maintains a relatively small hardware setup: an equilaterial-triangular-array, as it is shown in Figure 2. To this effect, the system obtains a set of 3 simultaneous sample windows.

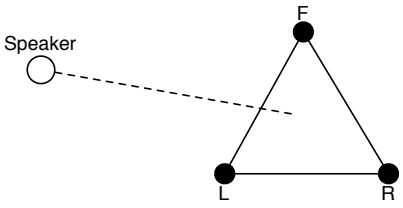

Fig. 2 Hardware setup of the proposed system.

The audio data is passed through various serialized sub-modules: a band-pass filter, a Voice Activity Detection stage, multi-ITD estimation, a redundancy check, and, finally, a final DOA estimation. The flow of data is summarized in Figure 3.

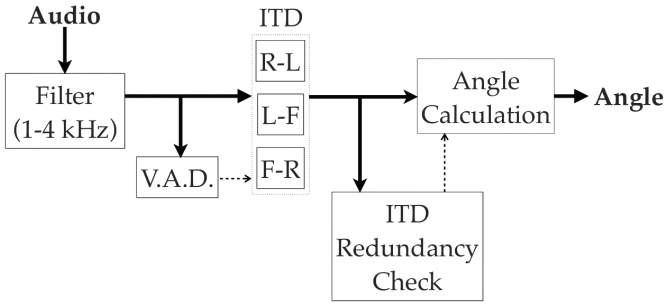

Fig. 3 Initial DOA Estimation flow of data.

3.2.1 Band-Pass Filter

A general infinite impulse response band-pass filter is used at the beginning of the process, to remove general ambient noise that is outside the human speech frequency bands. The filter model is described in Equation (3):

$$\begin{aligned} y_n = {} & 0.0348 \cdot x_n - 0.0696 \cdot x_{n-2} + 0.0348 \cdot x_{n-4} \\ & + 3.2680 \cdot y_{n-1} - 4.1247 \cdot y_{n-2} + 2.3984 \cdot y_{n-3} - 0.5466 \cdot y_{n-4} \end{aligned} \quad (3)$$

where y_i is the output of the filter, x_i is the input, and n is the number of the current sample.

It was observed that this filter made the system less sensitive towards unwanted noises that should always be ignored, such as high-pitch sounds, microphone hiss, etc. Concurrently, it did not degrate the sensitivity of the system towards human speech.

3.2.2 Voice Activity Detection

To trigger the ITD estimation described in the next section, Voice Activity Detection (VAD) needs to be carried out. Because the robotic platform may be changing position and environments, the VAD system is required to adjust to such changes automatically. To this effect, a simple VAD algorithm is proposed that is based on adjusting the baseline of the environmental noise to any sound that is emitted with a pre-specified delay.

Two history buffers of sample window energy values are kept in memory and shifted based on the specified delay (2 seconds provided good results). One is always being refreshed by new sample window energy values (*avg_buffer*), and is used to calculate the current average energy value (*avg_value*). The other buffer (*min_buffer*) is used to calculate the current average minimum value (*min_value*), and is refreshed with a new energy value if it is less than the current *min_value* or if the difference between it and *avg_value* is less than the difference between *avg_value* and *min_value* (which would mean that its value is close to the values of *min_buffer*).

The VAD is triggered if the energy value of the current sample window is greater than the average between *avg_value* and *min_value* by a multiplicative threshold (1.5 provided good results).

3.2.3 Multi-ITD Estimation, Redundancy Check, & Angle Calculation

Once the VAD is triggered, the Multi-ITD Estimation follows. Three possible ITDs can be calculated using cross-correlation between sample window R and L (I_{RL}), L and F (I_{LF}), and F and R (I_{FR}). 2 DOAs are calculated from each ITD: one using

Equation (2), and another shifting the first DOA to its possible counterpart on the 'backside' of the microphone pair.

The three DOA pairs are used to check if the three ITDs are from a sound source located in the same angle sector. To do this, the average of the differences between the DOA pairs is calculated using Equation (4).

$$C_{pqr} = \frac{|D^p_{RL} - D^q_{LF}| + |D^q_{LF} - D^r_{FR}| + |D^r_{FR} - D^p_{RL}|}{3} \tag{4}$$

where a D^i_{xy} is the ith DOA of the DOA pair from I_{xy}. A set of 8 C_{pqr} are calculated, where p, q, and r can be either 0 or 1, depending on which DOA of the DOA pair is being compared. Of the 8, the minimum is considered as the *incoherence* of the sample window set. As it can be seen in Figure 4a there is no combination of p, q, and r DOAs that provide low incoherence, while in Figure 4b, the combination $p = q = r = 1$ provides good coherence, all three pointing towards the source.

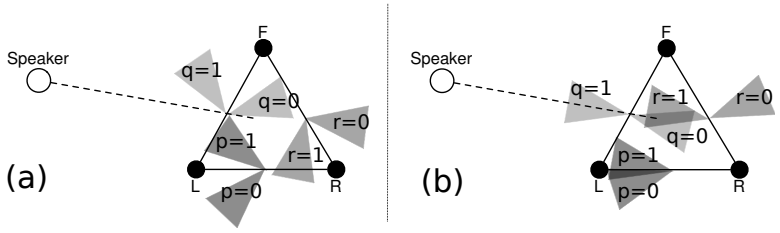

Fig. 4 a) A highly incoherent ITD set. b) A coherent ITD set ($p = q = r = 1$).

A pre-specified *incoherence threshold* (measured in degrees of separation between the DOAs; a value between 30° and 40° provided good results) is used to reject sample window sets. A high incoherence implies that the sample window set either has too much reverberation to be trustworthy for further processing or that it contains **more than one sound source**. This rejection step serves as a type of redundancy check *per sampling window set*.

If all of the DOAs are coherent/redundant with each other, a preliminary DOA value (θ_m) can be calculated using Equation (5),

$$\theta_m = arcsin\left(\frac{I_{min} \cdot V_{sound}}{F_{sample} \cdot d}\right) \tag{5}$$

where I_{min} is the ITD with the lowest absolute value of the three (I_{RL}, I_{LF}, I_{FR}). θ_m is then shifted to the appropriate angle sector in relation to the orientation of the robot, resulting in the final DOA value (θ).

Using I_{min} ensures that θ_m is calculated from the microphone pair that is the most perpendicular to the source. This means that the resulting θ is always estimated using a θ_m inside the -30°– 30° range (well within the close-to-linear -50°– 50° range), because of the equilateral nature of the triangular array. Meaning that all through the -179°– 180° range, there is always a close-to-linear ITD-DOA relation.

Because of both the redundancy check and the close-to-linear relation, the maximum error of this system can be known beforehand using Equation (6).

$$|error^{\circ}_{max}| = arcsin\left(\frac{I_{>30^{\circ}} \cdot V_{sound}}{F_{sample} \cdot d}\right) - arcsin\left(\frac{I_{<30^{\circ}} \cdot V_{sound}}{F_{sample} \cdot d}\right) \quad (6)$$

where $I_{>30^{\circ}}$ and $I_{<30^{\circ}}$ are the ITDs that provide the closest ceil and floor measurements, respectively, to 30°. For example, sampling at 44.1 kHz and with the microphones spaced at 18 cm, a maximum error of $\pm 2.8747^{\circ}$ can be expected. In the same set of circumstances, when using a 2-Mic 1-D array, the maximum expected error, which occurs when the sound source is close to either side of the robot, is of $\pm 15.0548^{\circ}$.

3.3 Multi-DOA Tracking

The DOA estimator described in the previous section only provides results when there is considerable confidence of only one sound source being detected in a small sample window (up to 100 ms). Although, it has been shown that people tend to not talk over each other while in conversation [21], even in simultaneous-speech, it has been seen that users are not expected to talk with a 100% overlap over each other. In fact, when analyzing speech recognition, 'spurts' of non-overlapping speech has been considered to the order of 500 ms [18]. For example, in Figure 5, it can be seen how two randomly chosen tracks from the DIMEX corpus [12], when overlayed over each other, still have some portions with no overlap between them.

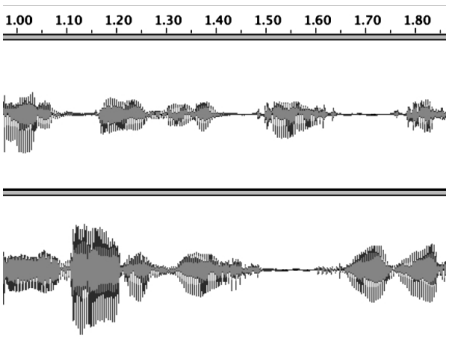

Fig. 5 Non-overlapping simultaneous speech.

This means that the DOA estimator described in the last section is able to provide reliable results of single sources even in multi-user scenarios. However, because of the stochastic nature of the presence of single user sample windows in the simultaneous audio timeline, such results would be provided in a sporadic fashion. To this effect, a simple tracking system is proposed that dynamically clusters similar DOAs into candidate sound sources.

The tracker maintains in memory the last DOAs provided by the initial DOA estimator in a specific time frame. When a new DOA is estimated, the tracker carries out the following:

1. If the new DOA is not 'close enough' to the average DOA of any current cluster (good results were obtained when using 5° for clusters with one DOA, 10° for clusters with more than one DOA, as thresholds for closeness), or there are no clusters in the environment: create a new cluster with the new DOA.
2. If it is close enough to a current cluster, add the new DOA to it, and re-calculate its new average DOA.

If a DOA is too old (an age of 10 seconds provided good results), it is 'forgotten' by removing it from its respective cluster and re-calculating its average DOA.

Every cluster is considered a *candidate sound source*, until it has a pre-specified number of DOAs attributed to it (2 DOAs provided good and fast results), when it becomes a 'sound source' and its average DOA becomes its main estimated DOA.

4 Trials & Results

The test scenario was as follows: three microphones, 20 cm. apart from each other, were installed in the upper base of the Golem-II+ robot. In turn, it was placed in a large room with a high sonic complexity (considerably reflective materials, with a low ceiling, hard floor, cement columns in the middle, and moderate reverberation). Two electronic speakers emitting, simultaneously, random recordings from the DIMEX corpus [12] for 20 seconds, were placed at 1.5 meters from the robot, one at 0°, another at -45°.

The Audio Acquisition module was sampling at 48 kHz, and providing sample windows of 4800 samples (100 ms). The buffers in the VAD were 10 energy values long, and considering a 2 second delay for adjustment to the environment noise. The DOA estimator had a 40° incoherence threshold (any sample window set with a higher incoherence was rejected). The multi-DOA tracker considered a new DOA as part of a cluster with more than one DOA if it was 10° or closer to its average DOA; if the cluster only had one DOA, 5° or closer was considered as part of the cluster. The results of the test are shown in Figure 6.

As it can be seen, the tracking system performed well with 2 sound sources (in the Figure referred to as 'Users').

The system was then tested with an additional simultaneous source: a human emitting continuously the phrase "golem i am over here (pause)" placed at 35°. The results of this scenario are shown in Figure 7.

As it can be seen, the system tracked the human and one of the electronic speakers (placed at -45°) well. The other of the two electronic speakers (placed at 0°) was 'missed' for a moderate amount of time, however, in any other moment, the tracking system was able to track it considerably well.

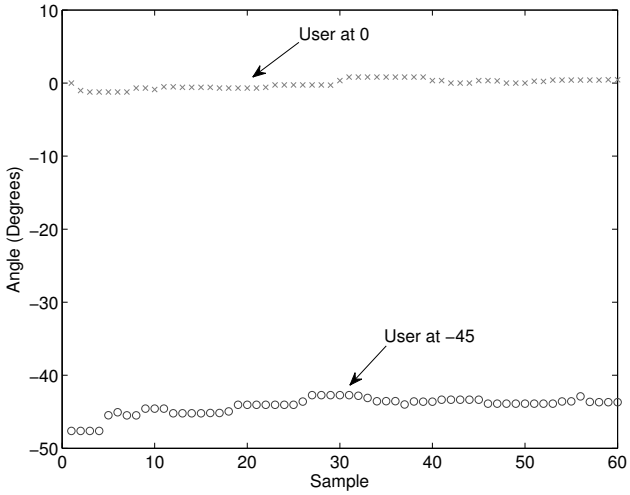

Fig. 6 Tracking 2 simultaneous sources (2 electronic speakers).

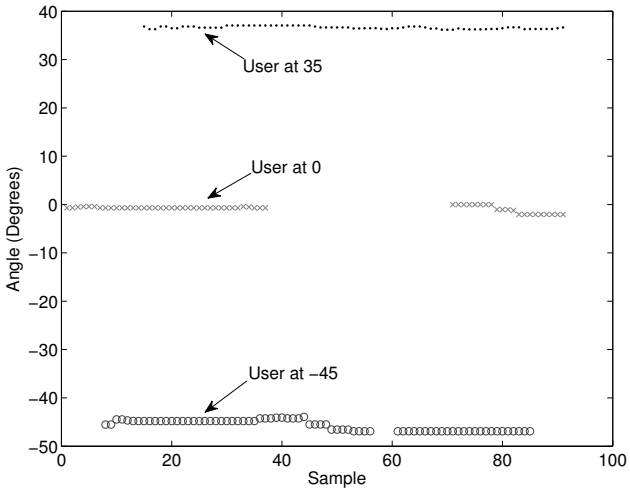

Fig. 7 Tracking 3 simultaneous sources (2 electronic speakers, 1 human).

To assess if the 'missed' tracking issue was with the electronic speaker itself, and, in addition, to observe if the tracker is able to better identify humans than electronic speakers, an additional simultaneous source was added to the environment: another human emitting continuously the phrase "one two three (pause)" placed at -100°. The results of this final scenario are shown in Figure 8.

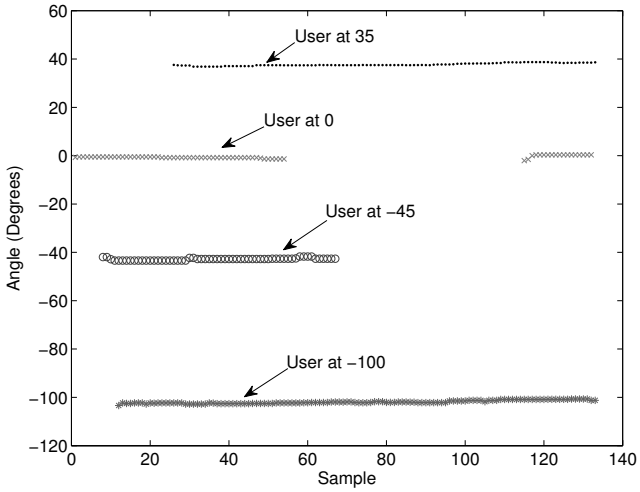

Fig. 8 Tracking 4 simultaneous sources (2 electronic speakers, 2 humans).

And, as it can be seen, the electronic speaker placed at 0° was again 'missed' in a similar fashion than in the 3-user scenario, which suggests a failure with the specific characteristics of the electronic speaker (positioning, volume, frequency enveloping of speech, etc.). However, both humans were tracked very well, and the electronic speaker placed at -45° was tracked relatively well. These results imply that the proposed system is well suited for tracking simultaneous human speech.

In addition, and more significantly, for a moderate amount of time, the 4 simultaneous sources were being tracked well. Considering that the system only employs 3 microphones, it showed that it was able to monitor more sources than the number of microphones present, a feat that the popular approach known as MUSIC is unable to accomplish [17]. In fact, the number of sources that can be simultaneously tracked by the proposed system may not have a theoretical boundary, as speech overlap is the primary limiter, and, as previously stated, people do not tend to interrupt each other [21]. However, further testing is required to explore this topic.

The authors would like to remind the reader that the setting of the test scenario were considerably harsh: the sonic complexity of the room was high, there was moderate reverberation, the human user placement can be expected to be inconsistent, and no reverb adequation was carried out. When considering all of this, the proposed system has shown it is an adequate solution to the multi-user DOA estimation problem in a robotic mobile platform.

5 Conclusion & Future Work

Human-Robot Interaction benefits from a rich perception of the world. Having the robot orient itself towards the user during a conversation enhances HRI from the point of view of both the user and the robot: the 'naturality' of the conversation is improved, and the acquisition of more information from the user (face recognition, voice context, etc.) is simplified. To do this, however, the direction of the user is required. Because a conversation is carried out via voice, it is appropriate that the direction of the user be estimated by sound analysis.

In addition, multi-user scenarios are common in the day-to-day dynamics of a service robot. However, the multi-DOA estimation problem is further complicated when applied in mobile robotics, as it presents a unique challenge: the mobility of the robot should not be compromised, thus the hardware should be small and lightweight (limited amount of microphones), but it should be robust and flexible enough to be able to carry out DOA estimation in acoustically complex settings.

In this contribution, a 3-microphone system was proposed, built upon earlier work published by the authors. It provides a reliable Multiple Direction-of-Arrival estimation, and it was shown that it was able to track more users than the amount of microphones used. Moreover, it did so while being light enough to be carried by a service robot. It also provided a robust estimation in the presence of moderate reverberation and high sonic complexity.

However, during the evaluation, were human speech and electronic-speakers were emitting simultaneously, it was observed that the human speech overcame the electronic speakers. Although this might be attributed to specific characteristics of the hardware, it was observed that human speech was consistently tracked well, which is something desirable as it will be employed with real-life human speech.

This system is planned to be a preamble for a consequent module that will perform online source separation based on the DOA of the source, which will then provide the Automatic Speech Recognizer with speech data. This will result in a multiple-simultaenous-speech recognition, with a small hardware setup and redundant estimation.

Acknowledgements The authors would like to thank the support of CONACYT through the project 81965, and PAPIIT-UNAM with the project IN115710-3.

References

1. Davis, P.: Jack, connecting a world of audio. http://jackaudio.org/ (2001)
2. Faller, C., Merimaa, J.: Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. The Journal of the Acoustical Society of America **116**(5), 3075–3089 (2004)
3. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. Robotics and Autonomous Systems **42**(3-4), 143–166 (2003)

4. Hjelmås, E., Low, B.K.: Face detection: A survey. Computer Vision and Image Understanding **83**(3), 236–274 (2001)
5. Horchler, A.D., Reeve, R.E., Webb, B., Quinn, R.D.: Robot phonotaxis in the wild: a biologically inspired approach to outdoor sound localization. In: Sound Localization," 11 th International Conference on Advanced Robotics, (ICAR '03), pp. 1749–1756 (2003)
6. Liu, R., Wang, Y.: Azimuthal source localization using interaural coherence in a robotic dog: modeling and application. Robotica **First View**, 1–8 (2010)
7. Lockwood, M.E., Jones, D.L., Bilger, R.C., Lansing, C.R., Jr., W.D.O., Wheeler, B.C., Feng, A.S.: Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms. Journal of the Acoustical Society of America **115**(1), 379–391 (2004)
8. Mohan, S., Lockwood, M.E., Kramer, M.L., Jones, D.L.: Localization of multiple acoustic sources with small arrays using a coherence test. Journal of the Acoustical Society of America **123**(4), 2136–2147 (2008)
9. Murray, J.C., Erwin, H., Wermter, S.: Robotics sound-source localization and tracking using interaural time difference and cross-correlation. In: AI Workshop on NeuroBotics (2004)
10. Murray, J.C., Erwin, H.R., Wermter, S.: Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks **22**(2), 173–189 (2009). What it Means to Communicate
11. Nakadai, K., Okuno, H.G., Kitano, H.: Real-time sound source localization and separation for robot audition. In: in Proceedings IEEE International Conference on Spoken Language Processing, 2002, pp. 193–196 (2002)
12. Pineda, L., Castellanos, H., Cuetara, J., Galescu, L., Juarez, J., Listerri, J., Perez, P., Villaseñor, L.: The corpus dimex100: Transcription and evaluation. Language Resources and Evaluation **44**(4), 347–370 (2010)
13. Pineda, L., Meza, I., Aviles, H., Gershenson, C., Rascon, C., Alvarado-Gonzalez, M., Salinas, L.: Ioca: Interaction-oriented cognitive architecture. Research in Computer Science **54**, 273–284 (2011)
14. Rascon, C., Aviles, H., Pineda, L.A.: Robotic orientation towards speaker for human-robot interaction. Advances in Artificial Intelligence - IBERAMIA 2010 **6433**, 10–19 (2010)
15. Rascon, C., Pineda, L.: Lightweight multi-doa estimation on a mobile robotic platform. In: Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2012, WCECS 2012, 24-26 October, 2012, San Francisco, USA, pp. 665–670 (2012)
16. Saxena, A., Ng, A.Y.: Learning sound location from a single microphone. In: ICRA'09: Proceedings of the 2009 IEEE international conference on Robotics and Automation, pp. 4310–4315. IEEE Press, Piscataway, NJ, USA (2009)
17. Schmidt, R.: Multiple emitter location and signal parameter estimation. Antennas and Propagation, IEEE Transactions on **34**(3), 276–280 (1986)
18. Shriberg, E., Stolcke, A., Baron, D.: Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In: in Proceedings of Eurospeech 2001, pp. 1359–1362 (2001)
19. Smith, M., Kim, K., Thompson, D.: Noise source identification using microphone arrays. Proceedings of the Institute of Acoustics **29**(5) (2007)
20. Stiefelhagen, R., Ekenel, H.K., Fugen, C., Gieselmann, P., Holzapfel, H., Kraft, F., Nickel, K., Voit, M., Waibel, A.: Enabling multimodal human-robot interaction for the karlsruhe humanoid robot. IEEE Transactions on Robotics **23**(5), 840–851 (2007)
21. Stivers, T., Enfield, N.J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J.P., Yoon, K.E., Levinson, S.C.: Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences **106**(26), 10,587–10,592 (2009)
22. Valin, J., Rouat, J., Michaud, F.: Enhanced robot audition based on microphone array source separation with post-filter. In: in Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems, pp. 2123–2128 (2004)
23. Wang, D., Brown, G.J. (eds.): Computational auditory scene analysis: Principles, Algorithms, and Applications. IEEE Press/Wiley-Interscience (2006). URL http://www.casabook.org/