Rascon *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:11
DOI 10.1186/s13636-015-0055-8

EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Lightweight multi-DOA tracking of mobile speech sources

Caleb Rascon[1]*, Gibran Fuentes[1] and Ivan Meza[1]

Abstract

Estimating the directions of arrival (DOAs) of multiple simultaneous mobile sound sources is an important step for various audio signal processing applications. In this contribution, we present an approach that improves upon our previous work that is now able to estimate the DOAs of multiple mobile speech sources, while being light in resources, both hardware-wise (only using three microphones) and software-wise. This approach takes advantage of the fact that simultaneous speech sources do not completely overlap each other. To evaluate the performance of this approach, a multi-DOA estimation evaluation system was developed based on a corpus collected from different acoustic scenarios named Acoustic Interactions for Robot Audition (AIRA).

Keywords: Multi-DOA; Three-mic array; Mobile sound sources; Kalman filter

1 Introduction

Estimating the direction of arrival (DOA) of a sound source is a well written-about topic in signal processing and has found a considerable amount of areas of application where its usefulness ranges from a complementary source of information to an essential part of the application. Such applications vary from steering multiple feed parabolic dish antennas [1], to source enhancement in antenna arrays [2], to fault monitoring in aircrafts [3], to intricate robotic pets [4], to close-to-life insect emulation [5].

This usefulness has also been applied in areas where sources are mostly speech, and thus, speech enhancement is required, either for the benefit of communication clarity between the users or to benefit automatic speech recognition (ASR) between the user and a computer. Examples of these applications are the design of hearing aids [6], robot audition [7-9], automatic meeting processing [10], and generic human-computer interaction via ASR (such as with mobile phones or smart homes). To carry out speech enhancement, these applications usually have a pre-processing stage in their auditory scene analysis, which involves the automatic estimation of the DOA of the active speech sources in the environment. This is known as multi-DOA estimation, and when being carried out in real acoustic environments (with audio interferences and prevalent reverberation) with several mobile active speech sources, it has been shown to be a very challenging task, even if assuming that the sound sources are in the far-field region of the microphone array (which simplifies the signal model).

A current popular solution to the multi-DOA estimation problem was presented in [11], which is arguably the starting point of two important robot audition projects: ManyEars [12] and HARK [13]. This solution requires an eight-microphone hardware solution but is able to detect accurately four moving speakers and up to seven static speakers if given enough time. This high eight-microphone requirement has been worked around by minimizing the physical footprint of the microphone array, the result of which is the small but effective 8SoundsUSB audio interface [14] which is now being applied in Willow Garage's PR2 service robot with good preliminary results [15]. However, even with a small physical footprint, the amount of microphones makes it difficult to employ this solution in applications where space is limited, such as hearing aids or mobile phones. In addition, to carry out the multi-DOA estimation phase of the process, it required an off-site computer because of its processing and memory requirements [16], which, although it may

*Correspondence: caleb.rascon@iimas.unam.mx
Instituto de Investigaciones en Matematicas Aplicadas y en Sistemas, Universidad Nacional Autonoma de Mexico, Circuito Escolar S/N, 04510 Mexico, Mexico

be able to be countered with today's technology, can be considered high for some applications.

Other multi-DOA estimation approaches using fewer microphones (four) and less processing and memory requirements involve the use of Kalman filters to smooth the estimated trajectory of the sound source in a noisy environment [17,18], and with a bit more microphones (five), they are even able to do so in a 3D search space [19]. However, they are only able to track one mobile source. Important attempts have been successful in being able to estimate the location of multiple sources using very few microphones (two to three) [20,21], but the sources are assumed to be static. A more detailed literature review is provided in Section 2.

In this paper, we improve upon our earlier approach, presented in [22] and applied in [23]. We also evaluate its performance in a much more detailed, systematic manner, which involves the capture and use of an evaluation corpus called Acoustic Interactions for Robot Audition (AIRA) and F1-based evaluation metrics, and doing so in several acoustic scenarios.

The approach divides the problem into two phases. The first phase estimates one DOA, even in a multiple-source environment, by 1) applying a coherence threshold between the DOAs estimated by each microphone pair and by 2) taking advantage of the incomplete overlap of simultaneous speech. The second phase keeps track of the DOA of several sound sources by assigning the incoming DOA of the first phase to either a) an existing track or b) a new track, depending on its angular distance to existing tracks. A Kalman filter is applied to each track for the resulting multi-DOA estimation and for track smoothing.

This current version of the approach is now able to 1) track multiple mobile sound sources 2) while employing a light hardware setup (only three microphones in a triangular array) and with a small computational resource requirement. Both these features, as far as we know, are not present in conjunction in any of the current multi-DOA estimation approaches, which is discussed in detail in Section 2.1. Such combination, we believe, would be of interest to several areas of application, as well the audio processing community in general.

This work is organized as follows: Section 2 provides background on the technical details to carry out multi-DOA estimation, as well as discusses the issue of the amount of microphones versus the number of sources estimated. Section 3 presents some background on the nature of simultaneous speech and how the proposed approach takes advantage of it. Section 4 details the proposed system. Section 5 presents the evaluation method we employed to measure the proposed system's performance. Section 6 discusses the results, and Section 7 provides the conclusions and future work.

2 Background on multi-DOA estimation

One of the most widely used acoustic feature for DOA estimation is the inter-microphone time difference (ITD). It 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 ways to calculate the CCV is by calculating a Pearson-based correlation factor for each delay value in the CCV, described in Equation 1.

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

where x and y are the two discrete signals being compared, i is the time index, k is the point at which y is being linearly shifted (delayed) and the correlation is being calculated, and 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.

A good example of the use of this Pearson-based cross-correlation method for DOA estimation in a robot audition application is presented in [24], where it provided limited results. Unfortunately, issues have arisen when using this method in reverberative and noisy environments [25], as the CCV calculations insert bias errors in such circumstances, which result in incorrect ITD estimations. However, as it would be seen in Section 4, this can be compensated with a combination of a form of redundancy and calculating the CCV using a variation of the generalized cross-correlation with phase transform (GCC-PHAT) [26].

The GCC-PHAT methodology has been widely recognized as one of the primary techniques for ITD calculation because of its robustness against reverberation [27,28]. Because of this, it has been extensively applied for ITD calculation in a wide variety of microphone array scenarios. For example, in [29], for the authors to be able to carry out real-time spatial rendering of different acoustic scenarios with a wide variety of sound sources, their DOA was required to be estimated via first calculating the ITD using the GCC-PHAT methodology. In addition, variations on the GCC-PHAT have been employed, such as in [30] where the authors were able to estimate jointly the DOA and pitch of two moving sources using a linear array of six microphones in reverberative simulated scenarios. Another variation was presented in [31], where the authors build an acoustic map of a room using a 13-microphone linear array, based on the GCC-PHAT technique, to directly estimate the ITD of multiple users. A detailed description of the GCC-PHAT methodology is provided in Section 4.3.1.

Having calculated an ITD, the direction of the source (DOA) can be estimated using a variety of methods. One popular method assumes that the sound source is far enough from the microphone array (aka. in the far-field region) and that there is no diffracting objects inside the microphone array (aka. the free-field assumption) so that the sound wave can be approximated to a planar-type wave. This simplifies the ITD-to-DOA relation to an arcsin function (details of which will be provided in Section 4.3.2), but this introduces other issues in the ITD-DOA relation. In Figure 1, the DOA is plotted against the ITD calculated from a two-microphone array (assuming far-field and free-field situations). As it can be seen in the −50° to 50° range, the ITD-DOA relation seems close to being uniform. However, outside that range, the relation loses its near-uniformity and the angular resolution is reduced, which causes major errors when estimating angles that are located in the sides of the microphone array [9]. To counter this issue while still maintaining the simplicity of the far-field model, in the proposed system, this loss in angular resolution is dealt with via a proper selection of the microphone pair in a triangular array from which the reported DOA is calculated, while assuming a far-field source.

Another important issue to consider is the geometry of the microphone array employed. Some geometries suffer from what is known as *ambiguity* [32], where an ITD may belong to more than one DOA. As it can also be seen in Figure 1, a two-mic array, by only estimating DOAs in the −90° to 90° range, is not able to differentiate if an ITD will be used to calculate the DOA that is coming from the front or the back of the array. This can be surmounted by implementing 'artificial ears' and a hidden Markov model (HMM) monaural mechanism [33] to then be able to detect if the sound source is coming from either side of the array, but it has been deemed impractical, as any physical change to the ear (physically or its relative position to the microphone) or to the acoustic scenario requires re-training of the HMM. This can also be tackled by a two-phase strategy: a first pair of signals could be used to estimate an initial DOA, the audio acquisition system could then rotate briefly, and then another pair of signals could be acquired to estimate a second DOA. A comparison between the DOAs would result in an angle estimation in the −179° to 180° range (countering the front-or-back ambiguity, which, in this case, could be considered as a trivial ambiguity). Unfortunately, this approach has its own set of issues: it would require considerably more time than when using one DOA estimate, the required rotation would hinder mobility requirements in some applications, and the sound source could be moving as well, rendering the DOA comparison mute. Another possibility would be to enlarge the microphone array so that it surrounds the source, as in [29,31]. To do so, however, the microphone array needs to encompass a significant amount of the space used by the users, which may be impractical in some applications.

In the proposed approach, a triangular array is employed from which various DOAs are calculated; by using a redundancy measure, the aforementioned ambiguity issue is circumvented. This is detailed in Section 4.

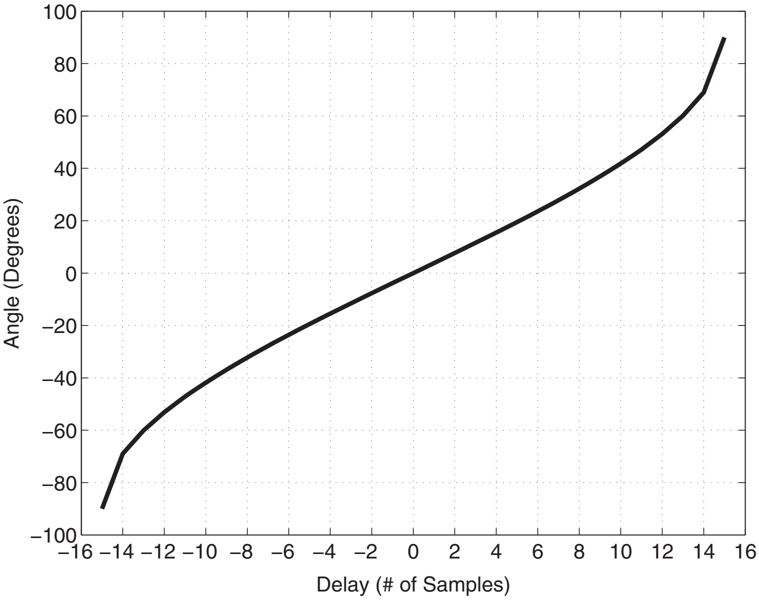

Figure 1 DOA vs. ITD. The graph shows DOA (or *Angle*) in degrees vs. ITD (or *Delay*) in number of samples and how it deviates severely from a uniform relation in the areas near the sides of the 1D array.

2.1 Amount of microphones employed vs number of sources estimated

It is important to note that the optimal amount of microphones to employ for multi-DOA estimation is a matter of discussion that is outside the scope of this work, but for ease of description, we are referring as a 'hardware-light' setup if it employs as much as four microphones, since there are portable commercial external audio interfaces that are able to handle that many microphones.

On the one hand, there is a large amount of approaches developed that carry out robust single-mobile-DOA estimation using a hardware-light setup [17-19]. On the other hand, there are important approaches in carrying out multi-mobile-DOA estimation which employ hardware-heavy solutions (which require specialized audio interfaces). It is important to mention that hardware-heavy approaches are practical for some applications; there are those that aim to conduct the audio acquisition once and process it offline, such as acoustical imaging [34] where it is not unusual to see 1,000-microphone arrays being used and no real-time localization is required.

The latter type of approaches (hardware-heavy) could be divided into two types of approaches: beamforming-based techniques and advanced statistical techniques.

An effective beamforming-based technique for multi-mobile-DOA estimation can be found in [35,36] and is somewhat representative of its type of approach. The idea is to create a noise map of the environment (or an acoustic map [31]) and then, by using metrics such as energy levels, propose possible sources of sound and their respective DOAs. It, technically, carries out a basic form of sound separation throughout all possible directions and then 'decides' which directions are valid sound sources and which are not. A good example of this decision is presented in [37] where, assuming there are only two sound sources in a meeting room, the beamforming technique 'decides' which sound source is assigned to which sound source class. These decisions are carried out at specific intervals, with which several methodologies can be employed for tracking purposes, such as Kalman filters [18] or particle filtering [38]. Another technique that is somewhat related to this type of approach, discussed and refined in [39], is known as the position-pitch (PoPi) plane, where, instead of mapping energy values onto the directional plane, a pitch spectrum is estimated per direction, providing possibilities to jointly estimate the DOA and pitch of sound sources, which, in turn, provides additional information to locate more than one source in the same direction.

As a whole, the beamforming type of approach has a pervasive issue: to increase precision and the quality of the validity metrics, it requires to obtain a high-resolution noise/acoustic map in both the amount of directions to search for and the quality of the separated sound from such directions. The sound quality requirement of the high-resolution map in turn requires a large quantity of microphones, since the quality of the separated sound is mainly defined by the signal-to-interference ratio (SIR), which is bound by the amount of microphones employed. This bound is summarized as 'the more microphones, the higher the quality'. Another important issue with this type of approach is that a balance needs to be struck between high-resolution maps and amount of computation resources required, since a high resolution results in a big search space where valid sound sources are to be found.

The techniques that approach the multi-mobile-DOA estimation problem by applying advanced statistics mostly rely on some variation of the popular technique known as Multiple Signal Classification (MUSIC) algorithm [1]. It carries out multi-static-DOA estimation by projecting the received signals in a DOA subspace, based on their eigenvectors, similar to principal component analysis (PCA). Although it has been widely reported that its performance decreases considerably in the presence of reverberation [25], it has been continuously enhanced in both resolution and computational costs [7], in handling mobile sources [40], as well as in handling an office-type amount of reverberation [10]. However, an important issue is that it is only able to estimate the DOA of as many sources as one less the amount of microphones (e.g., one source with two microphones, two sources with three microphones, etc.). This is because having more sources than microphones invokes the well-known 'more variables than observations' issue in PCA-based methods. In fact, it could be argued that if any sound interference may unexpectedly appear in the acoustic scene (which is not out of the ordinary in some scenarios) and, thus, increment the amount of sources in the acoustic scene over its upper limit, it could present instability issues in all of the provided estimations. Thus, to avoid this issue, the rule of thumb when using this type of techniques is 'just in case: the more microphones, the better'.

As it can be seen, there is an apparent tendency of employing hardware-heavy solutions to carry out multi-mobile-DOA estimation, while employing hardware-light solutions for single-mobile-DOA estimation. The proposed approach is a hardware-light solution that is able to carry out multi-mobile-DOA estimation, which we believe may be of interest to several areas of application and the audio processing community.

3 Background on simultaneous speech signals

As mentioned in Section 1, the proposed system assumes that the sound sources are from speech sound sources. This is taken advantage of, as detailed in the following section, for the purpose of multi-DOA estimation.

The scenario in which multiple speech sources are active simultaneously is bound to occur. Such is the case of crowded places (such as bars, restaurants, open-cubicle offices, etc.), where people are not participating in the same conversation but are near one another. However, it has been seen that users do not talk over each other with a 100% overlap. In fact, when analyzing speech recognition, 'spurts' of non-overlapping speech has been observed to last up to 500 ms [41]. For example, in Figure 2, two randomly chosen recordings from the DIMEx100 Corpus [42] are chosen. It can be seen that, even though these recordings are not from the same conversation (which is a very difficult scenario for multi-DOA estimation), when overlayed over each other, there are some portions with no overlap between them.

This means that a single-DOA estimation solution that is not only robust against multiple sources, but *is also sufficiently fast to 'catch' these single-source windows*, would be able to provide reliable results of single sources even in multiple-simultaneous-speech scenarios. However, because of the stochastic nature of the presence of single-user time blocks in the simultaneous audio timeline, such results would be provided in a sporadic fashion. Thus, these DOAs would be required to be 'grouped' afterwards and then be proposed as sound source directions, as it is carried out in the proposed system, detailed in the following section.

4 Proposed system

The proposed system uses a 'divide and conquer' strategy to solve the multi-DOA estimation challenge. It first provides a reverberation-robust DOA estimation of a single source, even in multi-source environments. It then takes advantage of the fact that even with simultaneous speech sources, there is not a 100% overlap between them. This means that the provided single-DOA estimations are being estimated from different sources but are 'given' in a stochastic manner. Thus, the objective of the next stage is to make sense of the incoming single-DOA estimations, by associating them to tracks or spawn new ones. The associated DOAs are then used to estimate a DOA for each track via Kalman filtering.

The proposed system is comprised by three modules:

1. *Audio acquisition and pre-processing.* Obtains audio data from the microphones and provides it to the single-DOA estimation module.
2. *Single-DOA estimation.* Estimates, from the audio data, a fast-but-reliable DOA estimation of a single sound source in the environment.
3. *Multi-DOA tracking.* Arranges the incoming single-DOA estimations from the single-DOA estimation phase into groups that are to be reported as sound sources with a Kalman-filtered DOA.

The data flow in the whole proposed system is summarized in Figure 3.

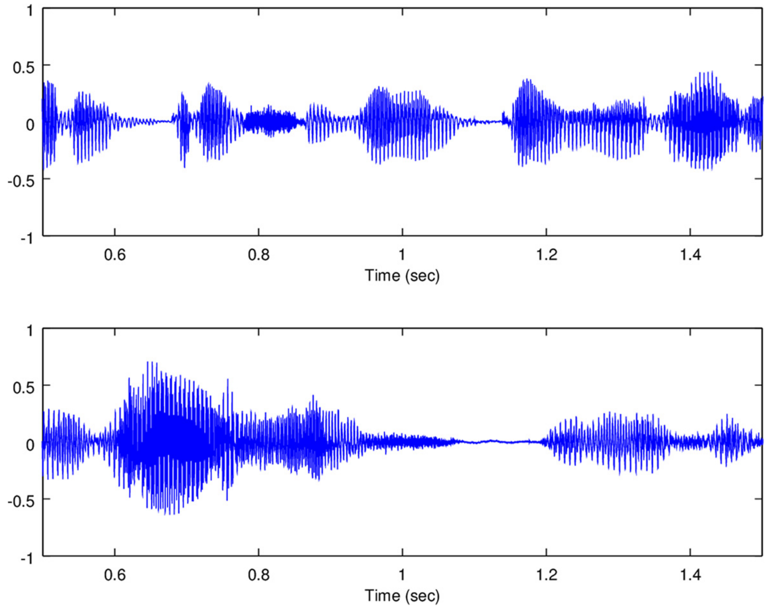

Figure 2 Non-overlapping simultaneous speech. An illustrative example of how even in simultaneous speech there is not a 100% overlap, a fact that can be taken advantage for multi-DOA.

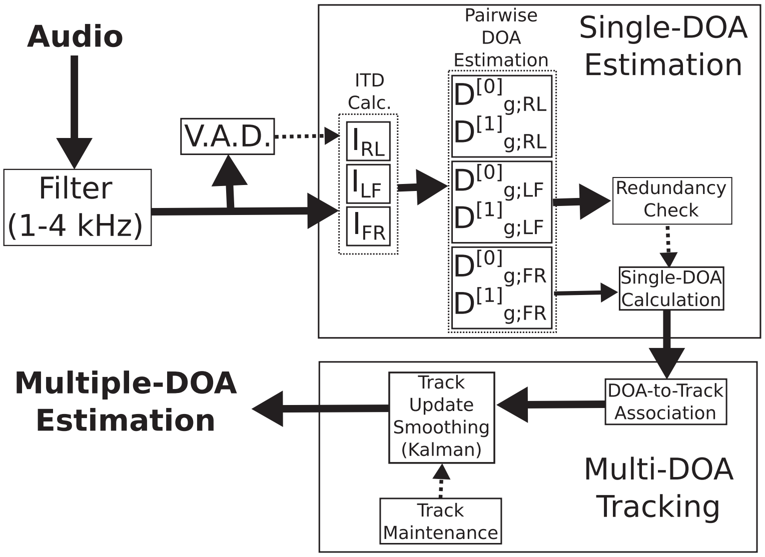

Figure 3 The flow of data throughout the proposed system. The audio data flows through a band-pass filter, voice activity detection (VAD) stage, the single direction-of-arrival (DOA) estimation (which includes multiple inter-microphone time-difference (Multi-ITD) calculation, a redundancy step, and single-DOA calculation), and the multiple direction-of-arrival (DOA) tracking (which includes DOA-to-track association, track updating, and track maintenance).

4.1 Hardware settings

To avoid the problems that arise when estimating a DOA using 1D microphone arrays (described in Section 2) and, at the same time, to maintain a hardware-light configuration, an equilateral triangular array is used, as shown in Figure 4. As it can be seen, the array maintains the free-field assumption, i.e., there are no objects inside the array that may diffract the sound waves coming into the the three microphones. Figure 5 presents one of the scenarios in which the system was tested.

4.2 Audio acquisition and pre-processing

As described in the last section, an equilaterial triangular array is used. Thus, the audio acquisition in the proposed system requires that the audio from three microphones be acquired simultaneously, in real time. For this purpose, the

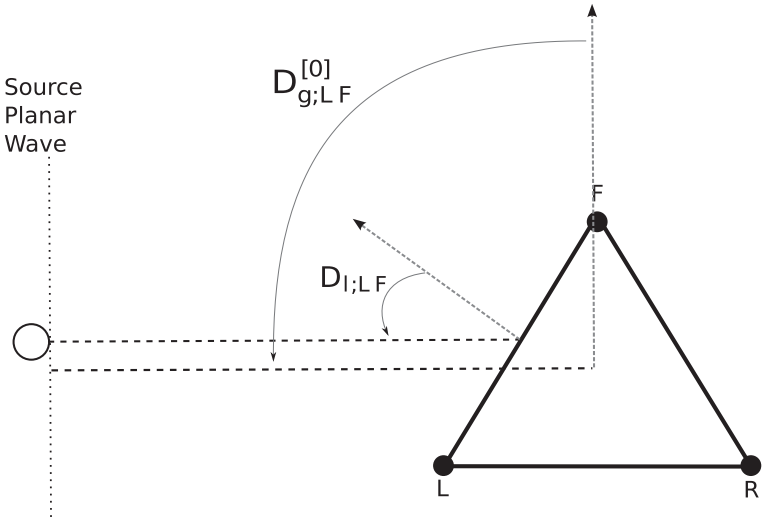

Figure 4 Microphone array setup of the proposed system and its coordinate system. The setup uses an array with three microphones, in a 2D array positioned horizontally. There are two coordinate systems being used: the local coordinate system per microphone array $D_{l;xy}$ that uses the front of the pair as its reference, and the global coordinate system $D^{[0]}_{l;xy}$ that uses the front of the whole array as reference.

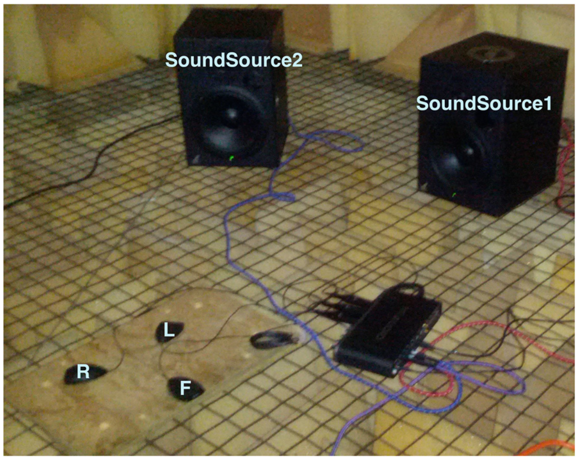

Figure 5 The anechoic chamber testing scenario of the proposed system. This scenario set in the anechoic chamber was one in which the system was tested. It shows that there are no diffracting objects inside the microphone array.

JACK Audio Connection Toolkit [43] was employed for audio capture. It can sample at rates of 44.1 and 48 kHz, providing a good resolution for ITD calculations, without slowing down other software modules running in the same computer.

4.2.1 Band-pass filter

Since the focus of the multi-DOA estimation system is to estimate the direction of speech sources, 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 (between 1 and 4 kHz). Thus, sound sources that are outside of this frequency range are rejected. This filter was designed using a single Butterworth-based second-order solution.

4.2.2 Voice activity detection

To trigger the single-DOA estimation described in the next section, voice activity detection (VAD) needs to be carried out. A state-based VAD is proposed which can be in one of two possible states: SILENCE and NON-SILENCE, whose switch is triggered when the current energy values (obtained by averaging the values of the three microphones of a whole sample window) cross a pre-specified threshold (t_{change}) above the noise energy value (e_{noise}; obtained by averaging the most recent energy values before the state change, e_{hist}). When in NON-SILENCE, a sub-state ACTIVE can be achieved by surpassing an additional threshold (t_{vad}). The reasoning behind the ACTIVE sub-state is that in this manner, a 'precedent' of the noise energy value is set and such sub-state is only activated when the audio energy is over that value during the whole time the system is in the NON-SILENCE state. e_{noise} is 'reset' every time there is a state change, providing robustness in dynamic scenarios. Some considerations are now presented and discussed (all energy values are in the linear scale of 0 dBFS):

$\boldsymbol{e}_{\textbf{hist}}$: The amount of energy values stored in e_{hist} will determine the VAD flexibility to changes in the noise level while also determining how much information is used to estimate the environmental noise. In [44], it was shown that moderately long time segments (750 ms) were preferable for estimating real types of noise. If using windows of 4,800 samples, and sample rates of up to 48 kHz, it is advisable to use more than eight energy values to store in e_{hist} for realistic scenarios.

$\boldsymbol{t}_{\textbf{change}}$: Since t_{change} defines if a sound should be detected or not, its value is strongly linked to the signal-to-noise ratio (SNR) of the signals that are desired to be detected. A small value will detect more signals, but more potential noise will go through; a high value will be more strict but potentially will not let actual source signals through. It is advised to use an SNR value over the amplitude ratio of 1 (1 dB) but lower than 2 (6 dB), to be adequate in noisy environments (since high SNRs cannot be expected in such circumstances). Thus, having measured the average noise level in the environment and having chosen a desired SNR value, a value of t_{change} (in the linear scale of 0 dBFS) can be calculated using:

$$t_{\text{change}} = 10^{(\text{SNR}+N)/20} - 10^{N/20} \tag{2}$$

where N is the noise level in dBFS and SNR is the desired signal-to-noise ratio in dB. In addition, to be certain that the activated sample windows are of a signal of considerable energy, the ACTIVE sub-state should be triggered with a t_{vad} such that $t_{\text{vad}} \gtrsim t_{\text{change}}$.

4.3 Single-DOA estimation

Once the VAD reaches an ACTIVE state, the single-DOA estimation is triggered. This phase is divided into an ITD calculation, a pairwise DOA estimation stage, redundancy check, and single-DOA calculation described herein.

4.3.1 ITD calculation

In this stage, ITD calculation is carried out between signals captured in pairs of microphones, based on a variation of the reverberation-robust generalized cross-correlation with phase transform [26].

Equation 1 is a time-based version of the Pearson-based cross-correlation method. Its frequency-based equivalent is presented in Equation 3.

$$\text{CCV}_F[f] = X[f]\, Y[f]^* \tag{3}$$

where X and Y are the Fourier transforms of the hamming-windowed x and y signals of Equation 1, respectively; f is the frequency bin; and the * operator stands for the complex conjugate operation.

The resulting cross-spectrum CCV_F relates to the cross-correlation vector CCV via the Fourier transform, such that $\mathrm{CCV}_F = \mathcal{F}(\mathrm{CCV})$. Thus, the size of the resulting CCV will be dependent on the size of the sample window.

The generalized cross-correlation (GCC) applies a weighting function ($\psi[f]$) to Equation 3 to improve CCV calculation. In the proposed approach, the weighting function used is shown in Equation 4.

$$\psi[f] = \begin{cases} \frac{1}{|X[f]Y[f]^*|}\ , & \text{if } f_{\min} < f < f_{\max} \\ 0\ , \text{otherwise} \end{cases} \tag{4}$$

where, if the first condition is reached (in our case, $f_{\min}$ is 1 kHz and $f_{\max}$ is 4 kHz), $\psi[f]$ applies the phase transform [27,28] (canceling out the magnitudes of both signals to estimate delta functions in the CCV domain). If not, the information from unused frequency bands is filtered to reduce estimation errors when carrying out $\mathrm{CCV} = \mathcal{F}^{-1}(\mathrm{CCV}_F)$. It is important to note that this rectangular cross-spectrum filtering introduces in the cross-correlation a leak in the CCV space, which means that some CCV bins may suffer some distortion. However, this type of distortion does not increase the value of the distorted bins to the point of being higher than the highest values of the CCV, and since the k index with the highest correlation value in the CCV ($k_{\max}$) is provided as the ITD of the two signals (I_{xy}), this is not an issue. As it can be gathered, the weighting is uniform inside the $[f_{\min}, f_{\max}]$ range, which can result in providing the same weights to a noisy frequency bin as to one that belongs to a speech source. Although automatic weighting in this step could be possible, it would require online noise estimation which is not carried out in this version of the proposed system. It is definitely being considered for future versions, though.

In addition, to avoid providing ITDs of reverb-bounces or noise sources, this part of the system only provides an ITD if its correlation value is higher than the CCV mean ($\overline{\mathrm{CCV}}$) plus a pre-specified threshold (C_{thresh}), thus $I_{xy} = k_{\max}$. If $\mathrm{CCV}[k_{\max}]$ is not higher than that, it is considered a 'noisy' ITD, and no ITD value is provided, thus $I_{xy} = \mathrm{NULL}$. The selection of C_{thresh} should consider three points: 1) Although the theoretical lower bound for C_{thresh} is 0, choosing this value would result in erroneous ITD calculations in silent sample windows that are considered active after the sound source has become inactive, as only ambient noise would be present. 2) Although its theoretical upper bound is $1 - \overline{\mathrm{CCV}}$, correlation values near 1 in real circumstances are rare; thus, a more practical generally applicable upper bound is 1. 3) A worst case scenario would be to have a large amount of values very close to $\mathrm{CCV}[k_{\max}]$ (expected with noise/high reverberation), which in turn would result in $\overline{\mathrm{CCV}}$ having a value very close to $\mathrm{CCV}[k_{\max}]$. Thus, to avoid filtering out correct ITD estimations in such scenarios, C_{thresh} should be very small compared to its upper bound and very close to its lower bound, meaning $0 \lesssim C_{\text{thresh}} \ll 1$.

It is important to note that, even though the IIR filter that precedes this step is rejecting information from outside the 1- to 4-kHz range, the resolution of the CCV (aka. the spatial resolution) is not hindered, since, in this case, it is highly dependent on the sample window size. Thus, a large window size (such as 4,800) is preferable for the purpose of having a high spatial resolution, in conjunction with a high sample rate (such as 48 kHz) to not hinder the overall responsiveness of the system.

4.3.2 Pairwise DOA estimation

Using the ITD calculation procedure, three ITDs are calculated between sample windows obtained from microphones R and L (I_{RL}), L and F (I_{LF}), and F and R (I_{FR}). If any ITD is NULL, the window set is discarded. Otherwise, a local DOA ($D_{l;xy}$) is calculated from each ITD (I_{xy}) using Equation 5.

$$D_{l;xy} = \arcsin\left(\frac{I_{xy}\ V_{\text{sound}}}{F_{\text{sample}}\ d}\right) \tag{5}$$

where x and y are identifiers of the signals received from the microphones R, L, or F; V_{sound} is the speed of sound (343 m/s); F_{sample} is the sampling rate (in Hz); and d is the distance between microphones (in m). All DOAs (referred to with the base notation D) hereafter are expressed in degrees.

It is important to note that this model assumes a planar-type sound wave, which in turn assumes that the sound source is located in the far-field region of the microphone array. If r_{source} is the distance of the sound source to the center of the array and r_{mic} is the distance of any microphone to the center of an equilateral array, then the far-field requirement is $r_{\text{source}} \gg r_{\text{mic}}$ [45]. This means that small array sizes ($d = 2r_{\text{mic}}\cos(30°) < 0.25$ m) are preferable to satisfy the far-field assumption in appropriate conversation scenarios ($r_{\text{source}} \approx 0.70$ m). In the case of the proposed system, because of the use of large sample windows, the CCV maintains a high resolution even with small array sizes. However, since the following calculations only require the resulting $D_{l;xy}$ to carry on, other models (such as near-field techniques) could be used in place of Equation 5 without requiring any further modifications to the proposed system and is definitely cause for future work.

Having calculated $D_{l;xy}$, a pair of global DOAs ($D^{[0]}_{g;xy}$ and $D^{[1]}_{g;xy}$) are then calculated using Equations 6, 7, and 8.

As described in Figure 4, while the local DOA $D_{l;xy}$ provides an angle from the perspective of the microphone pair alone, the global DOAs provide an angle from the perspective of the whole array. Specifically, $D_{g;xy}^{[0]}$ is the global DOA calculated as if $D_{l;xy}$ was coming from the front of the xy microphone pair, and $D_{g;xy}^{[1]}$ as if it was coming from the back.

$$D_{g;\mathrm{RL}}^{[0]} = \begin{cases} 0^\circ & \text{, if } D_{l;\mathrm{RL}} = 0^\circ \\ -D_{l;\mathrm{RL}} & \text{, otherwise} \end{cases}$$
$$D_{g;\mathrm{RL}}^{[1]} = \begin{cases} -180^\circ + D_{l;\mathrm{RL}} & \text{, if } D_{l;\mathrm{RL}} > 0^\circ \\ 180^\circ + D_{l;\mathrm{RL}} & \text{, otherwise} \end{cases} \qquad (6)$$

$$D_{g;\mathrm{LF}}^{[0]} = \begin{cases} D_{l;\mathrm{LF}} - 240^\circ & \text{, if } (120^\circ - D_{l;\mathrm{LF}}) > 180^\circ \\ 120^\circ - D_{l;\mathrm{LF}} & \text{, otherwise} \end{cases}$$
$$D_{g;\mathrm{LF}}^{[1]} = D_{l;\mathrm{LF}} - 60^\circ \qquad (7)$$

$$D_{g;\mathrm{FR}}^{[0]} = \begin{cases} D_{l;\mathrm{FR}} + 240^\circ & \text{, if } (-120^\circ - D_{l;\mathrm{FR}}) > 180^\circ \\ -120^\circ - D_{l;\mathrm{FR}} & \text{, otherwise} \end{cases}$$
$$D_{g;\mathrm{FR}}^{[1]} = D_{l;\mathrm{FR}} + 60^\circ \qquad (8)$$

4.3.3 Redundancy check

The three global DOA pairs $\left(\left[D_{g;\mathrm{RL}}^{[0]}, D_{g;\mathrm{RL}}^{[1]}\right], \left[D_{g;\mathrm{LF}}^{[0]}, D_{g;\mathrm{LF}}^{[1]}\right], \left[D_{g;\mathrm{FR}}^{[0]}, D_{g;\mathrm{FR}}^{[1]}\right]\right)$ are used to check if the three ITDs are from a sound source located in the same angle sector as a type of rejection step. To do this, the average of the differences between the DOA pairs (E_{pqr}) is calculated using Equation 9.

$$E_{pqr} = \frac{|D_{g;\mathrm{RL}}^{[p]} - D_{g;\mathrm{LF}}^{[q]}| + |D_{g;\mathrm{LF}}^{[q]} - D_{g;\mathrm{FR}}^{[r]}| + |D_{g;\mathrm{FR}}^{[r]} - D_{g;\mathrm{RL}}^{[p]}|}{3} \qquad (9)$$

where p, q, and r, each, can be either 0 or 1. This provides eight possible E_{pqr}, the lowest of which is considered the *incoherence* of the sample window set. The global DOA set that is represented in the minimum E_{pqr}, meaning $\left[D_{g;\mathrm{RL}}^{[p]}, D_{g;\mathrm{LF}}^{[q]}, D_{g;\mathrm{FR}}^{[r]}\right]$, is proposed as the DOA set for this window sample set.

A pre-specified *incoherence threshold* (E_{thresh}; measured in degrees of separation between the DOAs) is used to reject sample window sets. A high incoherence implies that there is an absence of 'consensus' in the sample window set for a DOA estimation, usually caused by having too much noise on the CCVs because of interferences in the acoustic scene or more than one source being active. This rejection step serves as a type of redundancy check *per sampling window set*. The selection of the value of E_{thresh} depends on the minimum ITD-DOA resolution of each microphone pair, since the global DOAs may provide an angle calculated in different regions of Equation 5, which in turn depends on F_{sample} and d. This minimum resolution is calculated by obtaining the difference between the DOA calculated with the highest possible ITD and the DOA obtained with the second highest possible ITD. In addition, ITD errors can occur in the range of ±1 bins. Thus, the worst case scenario is calculating a DOA in the minimum resolution range while having an ITD error. To this effect, a value for E_{thresh} that is not too small to reject window sets that may be sufficiently coherent, while rejecting those that are, is found around twice the minimum resolution.

As an example, Figure 6 shows two examples of two different data sets. The example shown in Figure 6a shows the scenario of two sources being active, which produces a highly incoherent set with no $\left[D_{g;\mathrm{RL}}^{[p]}, D_{g;\mathrm{LF}}^{[q]}, D_{g;\mathrm{FR}}^{[r]}\right]$ combination 'pointing' to a clear direction. On the other hand, Figure 6b shows the scenario of having only one source active (for a brief moment, the other source became inactive), which results in having a data set configuration with very low incoherence, which is $\left[D_{g;\mathrm{RL}}^{[1]}, D_{g;\mathrm{LF}}^{[0]}, D_{g;\mathrm{FR}}^{[1]}\right]$, with the combination $p = 1, q = 0, r = 1$.

4.3.4 Single-DOA calculation

If the sample window set is considered coherent, its reported DOA value (θ) is chosen from the member of

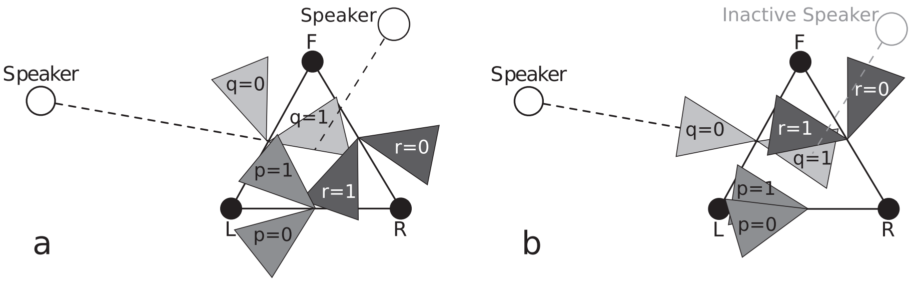

Figure 6 Incoherence example. **(a)** An example of a highly incoherent data set is presented, due to having two active sources. **(b)** A data set with very low incoherence is shown, with configuration $p = 1, q = 0, r = 1$, due to one of the sources being inactive at that brief moment.

its DOA set $\left[D^{[p]}_{g;\mathrm{RL}}, D^{[q]}_{g;\mathrm{LF}}, D^{[r]}_{g;\mathrm{FR}}\right]$ whose ITD from which it was calculated has the lowest absolute value ($I_{\mathrm{RL}}, I_{\mathrm{LF}}, I_{\mathrm{FR}}$).

This decision ensures that θ is based upon the microphone pair that is the most perpendicular to the source, and because of the equilateral nature of the triangular array, this implies that it is always estimated using a local DOA ($D_{l;xy}$) with a value inside the −30° to 30° range, well within a close-to-uniform area of Equation 5. Thus, the proposed system will always have a nearly uniform resolution between ITD and DOAs, which will result in low error rates throughout the −179° to 180° range, overcoming the issue of loss of resolution when the sound sources are located in certain directions relative to the array (detailed in Section 2).

4.4 Multi-DOA tracking

The single-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).

As described in Section 3, simultaneous speech has a non-overlapping nature, which means that the single-DOA estimator would be able to provide reliable results of single sources even in multi-user scenarios. However, these results would be provided in a sporadic fashion, as the presence of single-user sample windows in the simultaneous audio timeline is stochastic.

The objective of the final phase of the proposed system is to gather these sporadic DOAs, associate them into existing tracks or new ones, and use such information to estimate their current location. This approach is based on radar tracking techniques [46] that are useful when trying to report data from different targets using consecutive radar observations.

For each sample window, 1) if deemed coherent, the single-DOA estimator will provide at most one DOA estimation, and 2) the multi-DOA tracker will update. Each update of the multi-DOA tracker goes through four steps:

Single-DOA estimation to track association. This is solved via an acceptance gate approach (similar to nearest neighbor), in which the incoming single-DOA estimation is compared to the estimated DOA of all pre-existing tracks (the manner in which these are estimated is explained in the 'Track update' step). The single-DOA estimation is associated to its nearest track, only if the angular distance between them is below a pre-specified threshold (the same threshold used for redundancy check in the ITD estimation). If no track is within that threshold, a new track is spawned with only the single-DOA estimation associated to it.

Track update/smoothing. All tracks are updated, regardless of having new associations or not. This is carried out by estimating the current DOA by solely using the DOAs that are associated with it, and, thus, reporting multiple simultaneous estimations per sample window. To model the movement of the speech sources, we consider linear dynamic and measurement models, with the state-space representation presented in Equation 10:

$$\begin{aligned} \mathbf{s}_{t+1} &= \mathbf{A}s_t + \mathbf{w} \\ \mathbf{m}_t &= \mathbf{H}s_t + \mathbf{v} \end{aligned}, \tag{10}$$

where $\mathbf{s}_t$ is the dynamic state vector, $\mathbf{m}_t$ is the measurement vector, $\mathbf{A}$ is the transition matrix, $\mathbf{H}$ is the measurement matrix, and $\mathbf{w} \sim \mathcal{N}(0, Q)$ and $\mathbf{v} \sim \mathcal{N}(0, R)$ are zero-mean process and measurement noises with covariances Q and R, respectively. The state vector $\mathbf{s}_t = [s_t, m_t, \dot{s}_t, \dot{m}_t]^T$ contains the Cartesian coordinate and velocity of the DOA, and the transition and measurement matrices are defined in Equation 11:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & S_w & 0 \\ 0 & 1 & 0 & S_w \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{11}$$

where S_w is the size of the sample window in seconds. The estimation of the system state is carried out via the Kalman filter technique [47]. It was chosen since it provides an optimal and efficient solution for linear Gaussian models, such as the one being used here (see [46] for details). In addition, it reduces the influence of 'noisy' DOAs associated with the track, which is helpful in the case of close sources.

Track initiation. If the single-DOA estimation was not associated to any pre-existing track, a new track is spawned with this single-DOA estimation associated with it. However, the track is labeled as *tentative* until it has enough DOAs associated with it. What defines the final responsiveness of the system is the amount of DOAs associated with a new track such that it can be labeled *confirmed* (n_{doa}). To avoid providing 'noisy' tracks, more than 1 DOA is recommended. Since the single DOAs are provided stochastically, an upper bound for this value is difficult to set; however, values close to 1 will provide greater responsiveness. Thus, $n_{\mathrm{doa}} \gtrsim 1$ is a good rule of thumb.

Track maintenance. If a speech source goes inactive, it is important for the system to stop reporting it, to avoid false positives. However, it is also important to be robust against situations in which the speech source went inactive temporarily. Thus, a track is labeled again as *tentative* if a number of sample windows have gone by without any new association (n_{missed}). As described in Section 3, 'spurts' of non-overlapping speech has been observed to last up to 500 ms [41]. Thus, a value of sample windows that represents such an amount of time provides a good balance between responsiveness and low false positives.

5 Evaluation methodology

Four evaluation strains were carried out, each with its own objective, using an evaluation corpus based on the DIMEx100 Corpus [42], called AIRA.

The following parameter values were used during all evaluations:

- Sampling rate: 48 kHz
- S_w: 0.1 s (4,800 samples)
- e_{hist}: calculated using ten energy values
- C_{thresh}: 0.03645 (175 in an unnormalized FFT vector of 4,800 samples).
- E_{thresh}: 30°
- n_{doa}: 2
- n_{missed}: 5

The results of these evaluations are discussed in Section 6.

5.1 Anechoic chamber with static sources

To evaluate if the algorithm is able to detect the number and DOA of simultaneous speech sources, several experiments were carried in a 5.3 m × 3.7 m × 2.8 m full-anechoic chamber of the Laboratorio de Acústica y Vibraciones of the Centro de Ciencias Aplicadas y Desarrollo Tecnológico (CCADET) of the Universidad Nacional Autónoma de México (UNAM), a detailed description of which can be found in [48].

In this evaluation setting, the number of sources was varied from 1 to 4 in several positions in the whole horizontal degree range. For each number of sources, ten sessions were carried in which randomly chosen voice recordings from the DIMEx100 Corpus were reproduced for 30 s at 0 dB gain through studio-grade monitors acting as speech sources placed at 1 m away from the center of the triangular array. The microphones were set at 0.18 m apart, with 0 dB gain in the audio interface. The average sound level received at each microphone during voice activity was around −22 dBFS. Since the amount of noise inside the anechoic chamber was very low, the values for t_{change} and t_{vad} required to satisfy $t_{\text{change}} \gtrsim 0$, and $t_{\text{vad}} \gtrsim t_{\text{change}}$ (0.0005 was chosen for both). Figure 5 shows a photograph of the testing environment.

Two evaluation metrics were calculated:

Number and location of sources For each session, a sound source was considered as 'estimated' if it was detected for at least 25% of the duration of the session (30 s). If a source is estimated within a ± 15° range of the expected direction of an actual sound source, it is considered a true positive. If a sound source is estimated outside that range from an actual sound source, it is considered a false positive. If an actual sound source is not estimated during the experiment, it is considered a false negative. Using these metrics, the precision, recall, and F1 scores [49] (Chapter 8) of the proposed system's ability to detect sound sources are calculated.

Average error Once estimated, an average absolute error is calculated for every sound source that is deemed true positive, from the direction it is actually located.

In Table 1, these scores are provided per set of experiments.

5.2 Real acoustic setting with static sources

In the past evaluation, the evaluation was carried out in a controlled acoustic setting. However, it is important to measure the performance of the proposed approach in a real acoustic setting. An open-cubicle office was chosen for this test: 5.9 m × 7.9 m × 2.1 m (photo in Figure 7), with a typical indoor reverberation ($\text{RT}_{60} = 0.47$ s) and some presence of noise (−45 dBFS with 0 dB gain in microphones). A desired SNR for the VAD phase was set at 2.75 dB (for high sensitivity); thus, a t_{change} of 0.002 (in the linear scale of 0 dBFS) was used to achieve that, and a t_{vad} of 0.004 was used, satisfying $t_{\text{vad}} \gtrsim t_{\text{change}}$.

Table 1 Evaluation results reported in F1-type measure and average errors in the anechoic chamber

# of sources	Recall (%)	Precision (%)	F1 (%)	Average errors (in degrees)	Actual DOAs (in degrees)
1	100	100	100	1.22	45
2	100	100	100	7.36	−30
				3.49	90
3	90.00	100	94.74	7.15	−30
				3.50	90
				2.80	−150
4	72.22	100	83.87	0.04	0
				1.25	90
				0.61	180
				0.53	−90

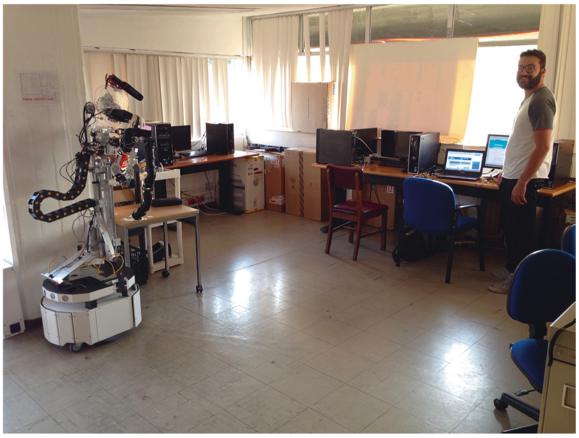

Figure 7 Office A photograph. A photograph of the office A test scenario. The microphone array was set in the middle open space.

The same type of evaluation as in the 'Anechoic chamber with static sources' evaluation was carried out in this acoustic setting, while having the microphones set at 0.21 m apart and varying the number of simultaneous sources from 1 to 4 that were placed 1 m away from the center of the array. The average SNR captured during these experiments was of 21 dB. The results are shown in Table 2.

5.3 Real acoustic setting with mobile sources

To evaluate the performance of the proposed approach with mobile sound sources, pre-specified randomly chosen sentences from the DIMEx100 Corpus read by expert volunteers were recorded while they were moving in a pre-specified manner around the recording station in another open-cubicle office of size 10.5 m × 5.9 m × 2.1 m, which is divided into three 3.5-m-wide spaces that are acoustically connected. However, in acoustic terms, it is very similar to the acoustic scenario of the 'Real acoustic setting' evaluation. Their main difference is that there are walls much closer to the microphone array (photo in Figure 8), amplifying moderately the effects of reverberation ($RT_{60} = 0.51$ s), and it had slightly more noise (−41 dBFS with 0 dB gain in microphones). The SNR captured during these experiments was 17 dB. The same t_{change} and t_{vad} values as in the 'Real acoustic setting' evaluation were used. The microphones were also set 0.21 m apart.

An approximation of the movement of each source was calculated. Then an estimated acoustic scene description file was created that describes an approximation of a linear trajectory around the audio acquisition base, with start and stop times and DOAs. Because of this limitation, the movement of the sound source was limited to simple 'go from left to right' (or vice versa) trajectories so their approximations can be considered representative of their real trajectories.

Figures 9 and 10 present a representative plot of the tracking carried out by the proposed approach with one and two mobile speech sources respectively presented over the expected behavior (plotted as dashed lines).

An evaluation similar to the 'Anechoic chamber with static sources' evaluation was carried out, only that in this case the average errors, true positives, false positives, and false negatives were reported from the reference of the expected trajectory previously described, instead of a static value. These scores are provided in Table 3.

5.4 Resource requirements

The implementation of the proposed approach occupies a 14.6 MB memory footprint, which includes a graphical interface for data visualization. In addition, the proposed system occupies up to 7% of the CPU resources when active. To put this in perspective, the ManyEars application [12], running in the same machine, occupies

Table 2 Evaluation results reported in F1-type measure and average errors in an office-type acoustic setting with static sources

# of sources	Recall (%)	Precision (%)	F1 (%)	Average errors (in degrees)	Actual DOAs (in degrees)
1	100	100	100	2.77	0
2	65.00	100	78.79	8.25	−30
				0.07	90
3	56.67	100	72.34	2.53	−30
				0.32	90
				0.67	−150
4	45.00	94.74	61.02	1.89	0
				1.27	−90
				0.50	180
				3.97	90

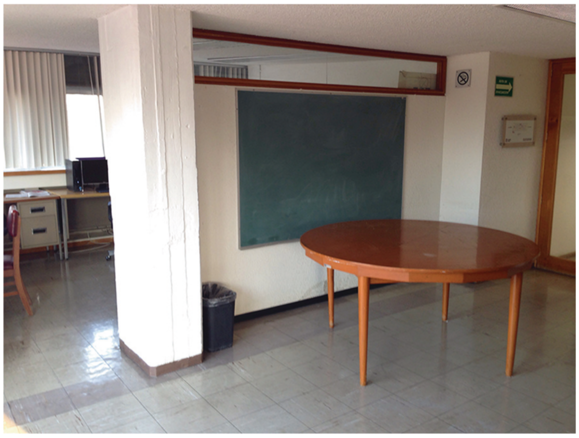

Figure 8 Office B photograph. A photograph of the office B test scenario. The microphone array was set in the table in the middle open space.

a 38.6 MB memory footprint and up to 61% of the CPU resources. This implies that the proposed approach is carrying multi-DOA estimation, with high F1 scores and low error rates, using a very small resource footprint.

6 Results discussion

As it can be seen in Tables 1, 2, and 3, in most circumstances, the average error is very low ($< \pm 5°$). This is the result of the fact that the resulting θ from the single-DOA estimator is calculated from the ITD of the microphone pair most perpendicular to the sound direction. Considering the triangular geometry of the microphone array, this results in having this ITD well within the close-to-uniform range area of Figure 1, providing a uniform resolution for the resulting θ.

Another observation is that the performance of the proposed system decreases as more sources are in the environment, which is to be expected. However, the decrease was not as pronounced when in the anechoic chamber. This implies that reverberation/noise is an important influence in the proposed system. When observing the precision and recall metrics, it can be observed that the addition of more sources affects the recall metric much more then it does the precision metric. An explanation for this is that the chance of single-user sample windows occurring will tend to decrease when the number of sources there are in the environment increases, resulting in false negatives. However, varying the number of sources does not affect the precision scores as much (if at all), which implies that once a sound source direction is reported, it is highly likely that it is an actual sound source. This is evidence of the effectiveness of 1) the coherence-based redundancy measures and 2) the application of the variations to the phase transform in the GCC calculation. Both measures filter out noisy sources and reverberation bounces and, thus, provide a relatively high amount of true positives.

It can also be seen that there was a near 20% decrease in the F1 metric when moving from an anechoic environment (Table 1) to a realistic one (Table 2), per test. For example, the F1 score for two sources went from 100%

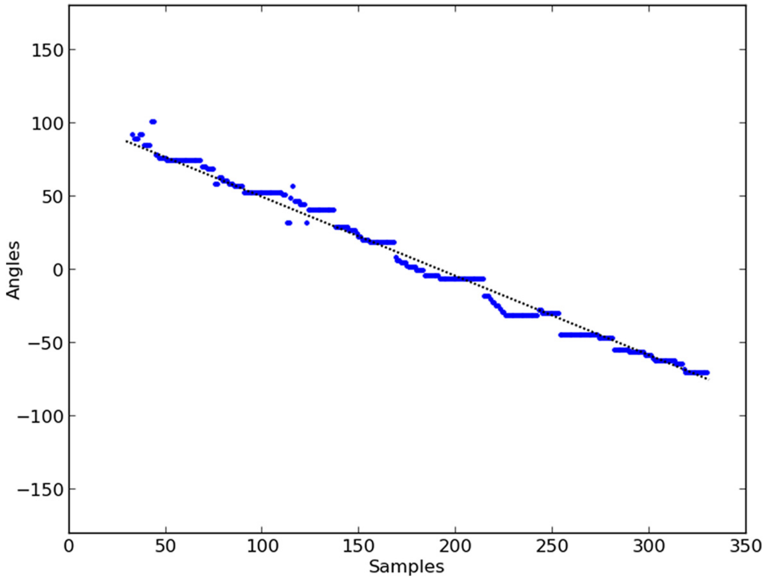

Figure 9 Office B with one source. A representative plot of the tracking carried out by the proposed approach of the acoustic scene office B with a source moving from 90°, passing through 0° and ending at −80°.

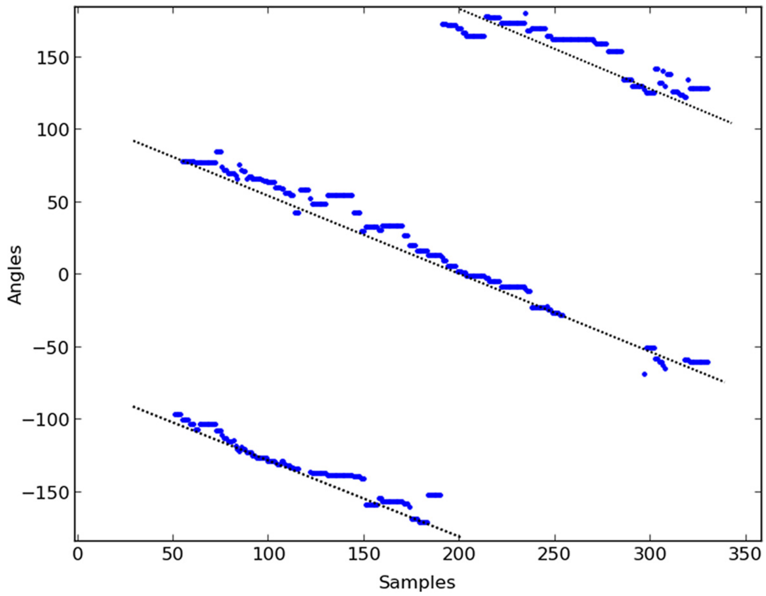

Figure 10 Office B with two sources. A representative plot of the tracking carried out by the proposed approach of the acoustic scene office B with a source moving from 90°, passing through 0° and ending at −80°, and another source moving −90°, passing through 180° and ending at 100°.

in an anechoic chamber to 78.79% in the office environment. This 20% decrease is consistent throughout most of the multi-DOA experiments, which, although significant, can be considered reasonable since the difference between both scenarios is quite notable. Moreover, it is worth mentioning that the system performed admirably in all the scenarios with one source, which is the baseline of a good multi-DOA estimation solution. Likewise, it is important to note that the system was able to carry out multi-DOA estimations with more static sources than microphones in a realistic environment.

In addition, when going from a realistic environment with static sources (Table 2) to one with mobile sources (Table 3), the decrease in the F1 metric is not as significant, which implies that the movement of the sources does not affect the performance of the proposed system as much as the noise/reverberation does (if at all). Furthermore, as evidenced in Figures 9 and 10 and Table 3, although the tracks are considerably 'noisy', it can be seen that the proposed approach is able to track the movement of the sound sources in a realistic acoustic environment and, more importantly, it is doing so with more than one mobile source, using a few-microphone solution, and with a small resource footprint requirement.

It is important to note that the recall score definitely calls for improvement. However, we believe that the proposed system, in the overall sense, has struck a balance that is ideal for real acoustic conditions and speech sources, since noise and reverberation are consistently 'tuned out' by the system while keeping track of the sound sources that are coherent and active. In fact, even if at the moment it may call for improvement, the level of performance the system is showing can currently benefit applications such as complementing the ASR of a service robot serving as a waiter in a restaurant [23], or in the analysis of the acoustic scene of a cocktail party, or in the design of hearing aids.

7 Conclusions

Multiple direction-of-arrival estimation can benefit a large array of audio applications. Current approaches tend to either be able to track multiple mobile sound sources

Table 3 Evaluation results reported in F1-type measure and average errors in an office-type acoustic setting with mobile sources

# of sources	Recall (%)	Precision (%)	F1 (%)	Average errors (in degrees)	Estimated trajectory
1	100	100	100	6.56	90° → 0° → −80°
2	60.00	92.31	72.73	7.10	90° → 0° → −80°
				8.73	−90° → 180° → 100°

using hardware-heavy solutions or use lightweight techniques to track one mobile sound source.

The proposed approach is based upon our earlier work and is now able to track, in reverberant scenarios, multiple moving sound sources while being hardware-light. We believe this is of interest in several areas of applications as well the audio processing community.

The technique carries out multi-DOA estimation by taking advantage of the fact that simultaneous speech sources do not completely overlap over each other and, thus, employs a 'divide and conquer' strategy. It first performs a quick-but-robust single-DOA estimation with single-source sample windows and then proposes sound source tracks from incoming single-DOA estimations. For each track, a Kalman filter is created to estimate the movement of the sound source the track represents.

To evaluate the performance of the proposed approach in a real-life acoustic scene, but in a consistently repeatable manner, the AIRA corpus was created in conjunction with an evaluation methodology based on the F1 score. It presented very good results in terms of the precision metric, thanks to its redundancy measures, the variations of the phase transform while carrying out GCC, and that it is always estimating a direction in the close-to-uniform range of the ITD-to-DOA function. These results confirmed that the proposed approach is able to track more than one source while using a small amount of microphones, in several acoustic scenarios.

For future work, we plan to improve upon the results presented here in terms of the recall metric, as well as in the manner of reducing its noisiness and responsiveness. One way to do this is by carrying an automatic weighting to remove noisy influences during the calculation of the GCC-PHAT. Another is to investigate other types of ITD-to-DOA models, specifically near-field techniques so that the proposed system is usable in more types of applications. In addition, it is of interest to develop a much more sophisticated DOA-to-track association technique, since it is very probable that such associations will break when having sound sources cross each other; however it is also important to maintain the low footprint requirement in such associations, which, unfortunately, is not satisfied by current techniques (such as the ones based on particle filters). We also plan to implement a stricter evaluation methodology for moving sources, providing a more representative evaluation metric for their tracking. Although keeping the number of microphones low is an important topic of this work, we plan to explore the addition of microphones to this solution for the DOA estimation in a 3D space to broaden the range of fields of where this approach can be applied. Moreover, we will explore its implementation in some of the aforementioned applications, specifically in the area of hearing aid design, where the small amount of microphones has been a limiting requirement. Another applicable area is that of Bioacoustics, by carrying census of animal species that interact with each other by song (birds, marine mammals, etc.) in acoustic scenarios that are far more dynamic and complex than the ones presented in this work and do not assume the presence of speech sources.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors thank the support of CONACYT through the project CB-178673 and PAPIIT-UNAM with the project IN107513. In addition, the authors would like to thank the Golem Group and their lead researcher, Dr. Luis Pineda, who provided many insights and resources to carry these findings out. Furthermore, we would like to specifically thank the support of Varinia Estrada, a member of the Golem Group, as well as Dr. Santiago Jesus Perez Ruiz, of the Laboratorio de Acústica y Vibraciones of the Centro de Ciencias Aplicadas y Desarrollo Tecnológico of the Universidad Nacional Autónoma de México, for their invaluable support in the recollection of the AIRA Corpus.

Received: 11 December 2014 Accepted: 10 April 2015
Published online: 07 May 2015

References

1. R Schmidt, Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propagation. **34**(3), 276–280 (1986)
2. L Griffiths, C Jim, An alternative approach to linearly constrained adaptive beamforming. IEEE Transac. Antennas Propagation. **30**(1), 27–34 (1982)
3. MG Smith, KB Kim, DJ Thompson, Noise source identification using microphone arrays. Pro. Inst. Acoust. **29**(5) (2007)
4. R Liu, Y Wang, Azimuthal source localization using interaural coherence in a robotic dog: modeling and application. Robotica. **28**(7), 1013–1020 (2010)
5. AD Horchler, RE Reeve, BH Webb, RD Quinn, in *Sound Localization, 11th International Conference on Advanced Robotics, (ICAR '03)*. Robot phonotaxis in the wild: a biologically inspired approach to outdoor sound localization (Coimbra, Portugal, 30 June 2003), pp. 1749–1756
6. ME Lockwood, DL Jones, RC Bilger, CR Lansing, WD O'Brien Jr, BC Wheeler, AS Feng, Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms. J. Acoust. Soc. Am. **115**(1), 379–391 (2004)
7. K Nakamura, K Nakadai, HG Okuno, A real-time super-resolution robot audition system that improves the robustness of simultaneous speech recognition. Adv. Robot. **27**(12), 933–945 (2013)
8. J-M Valin, J Rouat, F Michaud, in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*. Enhanced robot audition based on microphone array source separation with post-filter (Sendai, Japan, 28 September 2004), pp. 2123–2128
9. K Nakadai, HG Okuno, H Kitano, in *Proceedings IEEE International Conference on Spoken Language Processing, 2002*. Real-time sound source localization and separation for robot audition, (2002), pp. 193–196
10. F Asano, K Yamamoto, J Ogata, M Yamada, M Nakamura, Detection and separation of speech events in meeting recordings using a microphone array. EURASIP J. Audio Speech Music Process. **2007**(027616) (2007)
11. J-M Valin, S Yamamoto, J Rouat, F Michaud, K Nakadai, HG Okuno, Robust recognition of simultaneous speech by a mobile robot. IEEE Trans. Robot. **23**(4), 742–752 (2007)
12. F Grondin, D Létourneau, F Ferland, V Rousseau, F Michaud, The ManyEars open framework. Autonomous Robots. **34**(3), 217–232 (2013)
13. K Nakadai, T Takahashi, HG Okuno, H Nakajima, Y Hasegawa, H Tsujino, Design and implementation of robot audition system HARK - open source software for listening to three simultaneous speakers. Adv. Robot. **24**(5–6), 739–761 (2010)
14. D Abran-Cote, M Bandou, A Beland, G Cayer, S Choquette, F Gosselin, F Robitaille, DT Kizito, F Grondin, D Letourneau, USB synchronous multichannel audio acquisition system. http://sourceforge.net/projects/eightsoundsusb/files/Technical%20Paper/USB%20Synchronous%20Multichannel%20Audio%20Acquisition%20System.pdf. Accessed 26 June 2014

15. A Canoso, Giving ears to PR2 with 8Sounds and ManyEars. http://www.willowgarage.com/blog/2013/04/12/giving-ears-pr2-8sounds-and-manyears. Accessed 12 April 2013
16. J-M Valin, F Michaud, B Hadjou, J Rouat, in *Proceedings of IEEE International Conference on Robotics and Automation*. Localization of simultaneous moving sound sources for mobile robot using a frequency- domain steered beamformer approach, vol. 1 (New Orleans, LA, USA, 26 April 2004), pp. 1033–1038
17. Z Liang, X Ma, X Dai, Robust tracking of moving sound source using multiple model Kalman filter. Appl. Acoust. **69**(12), 1350–1355 (2008)
18. H Sun, P Yang, H Sun, L Zu, in *Control, Automation and Systems Engineering (CASE), 2011 International Conference On*. Outliers based Double-Kalman filter for sound source localization (Singapore, 30 July 2011), pp. 1–4
19. D Bechler, MS Schlosser, K Kroschel, in *Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference On*. System for robust 3D speaker tracking using microphone array measurements, vol. 3 (Sendai, Japan, 28 September 2004), pp. 2117–21223
20. S Mohan, ME Lockwood, ML Kramer, DL Jones, Localization of multiple acoustic sources with small arrays using a coherence test. J. Acoust. Soc. Am. **123**(4), 2136–2147 (2008)
21. NTN Tho, S Zhao, DL Jones, in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference On*. Robust DOA estimation of multiple speech sources (Florence, Italy, 4 May 2014), pp. 2287–2291
22. C Rascon, L Pineda, in *IAENG Transactions on Engineering Technologies, Special Issue of the World Congress on Engineering and Computer Science 2012. Lecture Notes in Electrical Engineering*. Multiple direction-of-arrival estimation for a mobile robotic platform with small hardware setup, vol. 247 (Springer Netherlands, 2014)
23. C Rascon, I Meza, G Fuentes, L Salinas, LA Pineda, Integration of the multi-DOA estimation functionality to human-robot interaction. Int. J. Adv. Robot. Syst. **12**(8) (2015)
24. JC Murray, H Erwin, S Wermter, in *AI Workshop on NeuroBotics*. Robotics sound-source localization and tracking using interaural time difference and cross-correlation (Ulm, Germany, 20 September 2004)
25. (D Wang, GJ Brown, eds.), *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. (IEEE Press/Wiley, Interscience, 2006). http://www.casabook.org
26. MS Brandstein, HF Silverman, in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference On*. A robust method for speech signal time-delay estimation in reverberant rooms, vol. 1 (Munich, Bavaria, Germany, 21 April 1997), pp. 375–3781
27. H Teutsch, *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition. Lecture Notes in Control and Information Sciences*, vol. 348. (Springer, Berlin Heidelberg, 2007)
28. J Benesty, J Chen, Y Huang, *Microphone Array Signal Processing. Springer Topics in Signal Processing*, vol. 1. (Springer, Berlin Heidelberg, 2008)
29. E Gallo, N Tsingos, G Lemaitre, 3D-audio matting, postediting, and rerendering from field recordings. EURASIP J. Adv. Signal Process. **2007**(47970) (2007)
30. S Gerlach, J Bitzer, S Goetze, S Doclo, Joint estimation of pitch and direction of arrival: improving robustness and accuracy for multi-speaker scenarios. EURASIP J. Audio Speech Music Process. **2014**(1) (2014)
31. A Brutti, M Omologo, P Svaizer, Multiple source localization based on acoustic map de-emphasis. EURASIP J. Audio Speech Music Process. **2010**(147495) (2010)
32. A Manikas, C Proukakis, Modeling and estimation of ambiguities in linear arrays. IEEE Trans. Signal Process. **46**(8), 2166–2179 (1998)
33. A Saxena, AY Ng, in *ICRA'09: Proceedings of the 2009 IEEE International Conference on Robotics and Automation*. Learning sound location from a single microphone (IEEE Press Piscataway, 2009), pp. 4310–4315
34. WS Gan, *Acoustical Imaging: Techniques and Applications for Engineers*. (John Wiley & Sons, Ltd, 2012)
35. J-M Valin, F Michaud, J Rouat, D Letourneau, in *Intelligent Robots and Systems, 2003. (IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference On*. Robust sound source localization using a microphone array on a mobile robot, vol. 2 (Las Vegas, USA, 27 October 2003), pp. 1228–12332
36. J-M Valin, F Michaud, J Rouat, Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. Robot. Autonomous Syst. **55**(3), 216–228 (2007)
37. R Chakraborty, C Nadeu, T Butko, Source ambiguity resolution of overlapped sounds in a multi-microphone room environment. EURASIP J. Audio Speech Music Process. **2014**(18) (2014)
38. P Pertila, T Korhonen, A Visa, Measurement combination for acoustic source localization in a room environment. EURASIP J. Audio Speech Music Process. **2008**(278185) (2008)
39. M Wohlmayr, M Kepesi, in *Proceedings of the 8th Annual Conference of the International Speech Communication Association*, ed. by ISCA. Joint position-pitch extraction from multichannel audio (Antwerp, Belgium, 27 August 2007), pp. 1629–1632
40. T Otsuka, K Nakadai, T Ogata, HG Okuno, in *INTERSPEECH'11*. Bayesian extension of MUSIC for sound source localization and tracking (Florence, Italy, 15 August 2011)
41. E Shriberg, A Stolcke, D Baron, in *Proceedings of Eurospeech 2001*. Observations on overlap: findings and implications for automatic processing of multi-party conversation (Scandinavia, Aalborg, Denmark, 3 September 2001), pp. 1359–1362
42. LA Pineda, H Castellanos, J Cuétara, L Galescu, J Juárez, J Llisterri, P Pérez, L Villaseñor, The Corpus DIMEx100: transcription and evaluation. Lang. Res. Eval. **44**, 347–370 (2010)
43. P Davis, JACK connecting a world of audio. http://jackaudio.org. Accessed 27 April 2015
44. C Ris, S Dupont, Assessing local noise level estimation methods: application to noise robust {ASR}. Speech Commun. **34**(1–2), 141–158 (2001)
45. J Huang, T Supaongprapa, I Terakura, F Wang, N Ohnishi, N Sugie, A model-based sound localization system and its application to robot navigation. Robot. Autonomous Syst. **27**(4), 199–209 (1999)
46. KV Ramachandra, *Kalman Filtering Techniques for Radar Tracking*. (Marcel Dekker, Florida, USA, 2000)
47. RE Kalman, A new approach to linear filtering and prediction problems. Trans. ASME–J. Basic Eng. **82**(Series D), 35–45 (1960)
48. R Ruiz-Boullosa, A Perez-Lopez, Some acoustical properties of the anechoic chamber at the Centro de Instrumentos, Universidad Nacional Autonoma de Mexico. Appl. Acoust. **56**(3), 199–207 (1999)
49. CD Manning, P Raghavan, H Schutze, *Introduction to Information Retrieval*. (Cambridge University Press, US, 2008)