Documentation of the AIRA Corpus

Caleb Rascon,[1, a)] Ivan V. Meza,[1] Aldo Millan-Gonzalez,[1] Ivette Velez,[1] Gibran Fuentes,[1] Dennis Mendoza,[1] and Oscar Ruiz-Espitia[1]

Universidad Nacional Autónoma de México, México

(Dated: 29 September 2020)

In this document, the Acoustic Interactions for Robot Audition (AIRA) corpus is fully documented. The aim of this corpus is to be used for research on sound source localization and separation, as well as for multi-user speech recognition, in circumstances where the sound source is outside of the microphone array. This provides great potential for Robot Audition applications, as well as for Auditory Scene Analysis in general, in the aspects of evaluation and model training. It employs two microphone array configurations: one is an equilateral triangle array, and another is a three-dimensional 16-microphone array set over a hollow plastic body. It was recorded in 6 real-life scenarios that varied in terms of noise presence and reverberation time: it ranged from an anechoic chamber to a considerably large and busy department store. It includes clean speech data for static sources and tracking information (both grounded and estimated) for mobile sources. Finally, it is freely available from https://aira.iimas.unam.mx/.

Pages: 1–12

I. INTRODUCTION

The area of Auditory Scene Analysis, and the specific area of research known as Robot Audition, aim to create a description of the auditory scene analysis, in terms of the audio sources location and classification, with an interest of speech sources for Robot Audition. Thus, speech recognition is also of great interest to be carried out, which implies a necessity to carry out sound source separation in cases of noisy environments or acoustic scenarios with a high amount of speech interferences (such as a restaurant or a department store).

To this effect, these types of algorithms are expected to be employed in real environments. This brings up two considerations that this paper acknowledges relevant to the topic of sound source localization and separation: evaluation and training.

In terms of evaluating these algorithms, it is important to measure their performance in real-life scenarios as to provide a grounded discernment of their capabilities. However, it is also important to carry out these evaluations in a repeatable manner when it is of interest to quantify improvements in performance. A corpus of real-life recordings can satisfy both of these conditions.

In terms of training, there has been a recently emerging interest in the use of data-based models for source localization and separation (Rascon and Meza, 2017), specifically those using techniques related to Deep Learning. In such cases, it is important to use a training data set that was acquired in real-life scenarios so that the model is able to extract information that may have been not included in a set of simulated data.

A few acoustic corpora have been specifically collected for localizing and tracking sound sources. One of the earliest such corpora is the RWCP Sound Scene Database in Real Acoustical Environments (RWCP-SSD) (Nakamura *et al.*, 2000) which contains audio recordings of static speech and non-speech sources as well as moving speech sources in different scenarios and conditions. RWCP-SSD was recorded using a lineal array of 14 microphones and semi-spherical array of 54 microphones, which remained static while recording. Another instance is the AV16.3 (Lathoud *et al.*, 2005), composed of video and audio recordings of meeting rooms with one or more speakers acquired using two-microphone arrays. The recordings include static and moving speakers but the two-microphone array remained static in all recordings. A corpus with a moving recording system is CAVA (Arnaud *et al.*, 2008). It consists of audio and video recordings acquired with a binaural microphone and a binocular camera mounted on a persons head in different complex and dynamic scenarios such as meetings with multiple moving speakers and noise sources.

Two corpora specifically collected for sound source localization and tracking in robotic platforms are AVASM (Deleforge *et al.*, 2014) and CAMIL (Deleforge and Horaud, 2011). Both of these corpora were recorded from a robotic head using a binaural array. In the AVASM corpus, the recordings include both static and moving sound sources and a static robotic head. On the other hand, the CAMIL corpus have recordings with and without robotic head movements of single static sources in an office environment.

In may 2018, the IEEE-AASP Challenge on Acoustic Source Localization and Tracking (LOCATA) took place with the aim of benchmarking sound source localization methods in realistic environments found in different ap-

[a)] caleb.rascon@iimas.unam.mx

plications. As part of the challenge, a corpus(Löllmann *et al.*, 2018) was released which consists of recordings in a real acoustic environment using four different microphone arrays in static and mobile platforms surrounded by mobile and static sources accompanied by their ground truth locations.

To the Robot Audition community, as well as to the Auditory Scene Analysis field, it is of great interest to have a corpus that incorporates the benefits of the aforementioned corpora, such as: a varying amount of microphones, with different array configurations, recorded in real scenarios, etc.

To this effect, this paper presents an extensive description of the Acoustic Interactions for Robot Audition (AIRA) corpus which we believe covers these benefits, since it has the following characteristics:

- It uses two array configurations: a triangular array and a 16-microphone three-dimensional array.
- It was recorded in 6 different real life scenarios, including an anechoic chamber as a reference point.
- There is a considerable amount of variations between the scenarios in terms of noise presence and reverberation time.
- Static speech sources were 'simulated' by high-end flat-response studio monitors reproducing the recordings from another cleanly recorded corpus in Mexican Spanish: the DIMEx100 corpus (Pineda *et al.*, 2010). All clean speech data from these static speech sources is provided along with the real-life recordings.
- Mobile speech sources were carried out by human volunteers, and their position through time are either provided by a laser-based tracking system or by an estimation from their start and end position (to simulate noisy localization results).

The corpus is composed by a set of audio recordings captured in six different environments, with two different hardware configurations. In some scenarios, up to 4 static sound sources were simulated via monitor speakers reproducing clean audio signals. In these scenarios, the corpus includes the clean audio signals as well as the direction-of-arrival of each sound source. In other scenarios, human volunteers acted as mobile speech sources, and their location through time is included. In all scenarios, a complete transcript of what each speech source enunciated is also included.

It is important to mention that a preliminary version of this corpus has already been presented as part of the evaluation of an algorithm that tracks multiple sound sources with a small amount of microphones (Rascon *et al.*, 2015). However, in this paper we aim to formally introduce the AIRA corpus, and describe it in full detail. It is also worth mentioning that even though the AIRA corpus was originally aimed at Robot Audition applications, we believe that its scope is much wider, such as the Auditory Scene Analysis field.

To simplify its use by the Robot Audition community, as well as the Auditory Scene Analysis area at large, the AIRA corpus is freely available at `https://aira.iimas.unam.mx/`.

The paper is organized as follows: in Section II the microphone array configurations used during the capture process are described; in Section III, the hardware equipment is detailed; in Section IV, the software used to capture and reproduce audio is presented; in Section V, the capture protocol is presented; in Section VI, the environments in which AIRA was captured are described; in Section VII, the corpus structure is explained; and in Section VIII, our conclusions are provided.

II. MICROPHONE ARRAY CONFIGURATIONS

Each microphone array configuration was decided upon on the basis of a specific set of objectives. The AIRA corpus bares two different array configurations which are detailed in this section.

A. Triangular Array

This configuration employs an array of microphones set in an equilateral triangle. The objectives of this configuration are:

- For algorithms that only require a small amount of microphones.
- For circumstances where there are more sources than microphones.

In Figure 1, an schematic of the array is shown as well as the frame of reference used for measuring the position of the source.

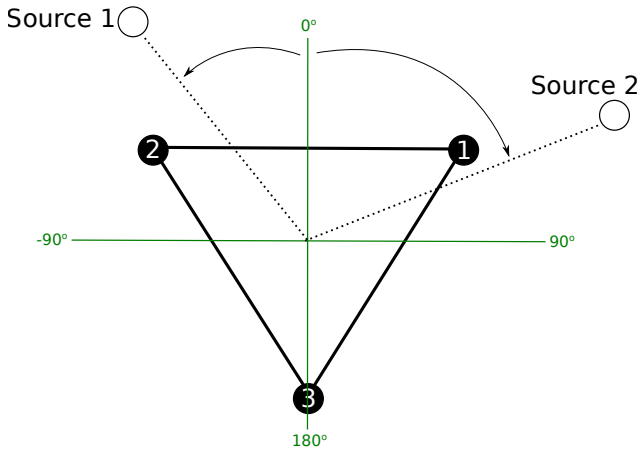

FIG. 1. Triangular Array Configuration.

The sources were positioned 1 m away from the center of the array, and their direction was registered in every recording. In terms of height, they were positioned in the same height as the imaginary plane formed by the

triangular array. Thus, no height information was registered.

The distance between microphones was either 0.18 m or 0.21 m. These distances were used to be close to the width of the human head, so that either pair of microphones can be used for free-field binaural algorithms. It is important to clarify that the distance between the microphones changed between recording environments (Section VI) because of the limited space in some of the environments. However, these changes were all registered as part of the corpus.

B. 3D Array

This configuration employs a three-dimensional array of 16 microphones, all set over a hollow plastic body. The objectives of this configuration are:

- For algorithms that require a high amount of microphones.
- For circumstances that break the inter-microphone free-field assumption.

In Figure 2, an schematic of the array is shown as well as the frame of reference used for measuring the position of the source.

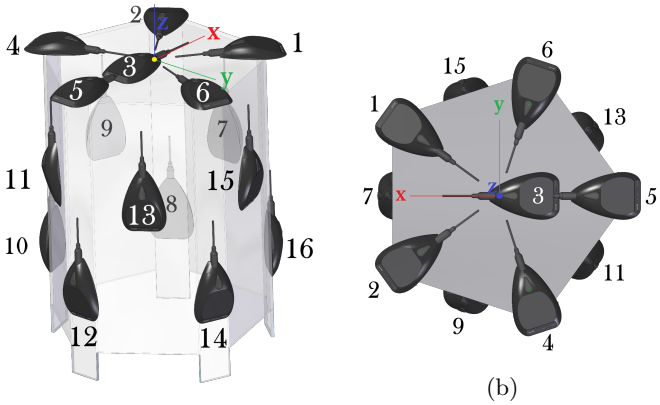

FIG. 2. 3D Array Configuration. (a) Side view, (b) top view.

As it can be seen, the center of the array is positioned in its top center part, so that the top 6 microphones can still be used as a free-field array. Additionally, microphone 3 is not positioned in the top center; it is positioned such that microphones 1, 2 and 3 form a close-to-equilateral triangle so that there can be comparison between array configurations. Additionally, the array as a whole can be used for algorithms that are robust against having a body between microphones, which breaks the free-field assumption. In fact, microphones pairs can be chosen (such as microphones 9 and 15) to be used with non-free-field binaural algorithms. The positions of the microphones are presented in Table I. These positions are using the frame of reference presented in Figure 2.

Plastic was used as the material for the body because of: ease of molding; ease of reproduction; ease of transportation; and enough sturdiness to hold the 16 microphones.

TABLE I. Microphone positions in the 3D array configuration.

Mic.	x	y	z
1	0.158	0.115	0.000
2	0.158	-0.115	0.000
3	-0.045	0.000	0.000
4	-0.050	-0.188	0.000
5	-0.195	0.000	0.000
6	-0.057	0.186	0.000
7	0.180	0.000	-0.168
8	0.158	-0.115	-0.335
9	0.056	-0.171	-0.168
10	-0.050	-0.188	-0.335
11	-0.128	-0.098	-0.168
12	-0.195	0.000	-0.335
13	-0.132	0.098	-0.168
14	-0.057	0.186	-0.335
15	0.056	0.171	-0.168
16	0.158	0.115	-0.335

III. HARDWARE

In this section, the hardware to capture the AIRA corpus is described. The equipment used is divided in:

- Audio interfaces
- Microphones
- Speakers

It is important to mention that different hardware was used between the array configurations, and this will be considered in the following descriptions. The reason for this change is purely logistical: the recordings for the triangular array were captured first with equipment that was available at the time and that was able to capture 3-channel audio data. Afterwards, the 3D array recordings were carried out with more up-to-date equipment that was able to capture 16-channel audio data. However, as it will be seen in the following descriptions, the differences between the microphones and speakers used for both sets of recordings were mostly minor.

A. Audio Interfaces

The audio interfaces used for each array configuration were as follows:

- **For Triangular Array:** *M-AUDIO Fast Track Ultra* (Avid Technology, Inc., 2007). It is able to handle up to 4 XLR inputs and 6 TRS outputs. Each XLR input has its own pre-amplification, the amount of which was changed according to the recording environment (detailed in Section VI). The TRS outputs, however, do not have any pre-amplification.

- **For 3D Array:** *Behringer X-32* (Behringer, 2016). It is a digital mixer that can act as an external audio interface, providing support for 32 XLR inputs and 32 XLR outputs. Each XLR input has its own pre-amplification, the amount of which was changed according to the recording environment (detailed in Section VI). The external audio interface module inside the digital mixer does not have any pre-amplification for the outputs.

In all circumstances, the audio interfaces were configured to capture at 48 kHz with a 16-bit floating point accuracy.

B. Microphones

The microphones used for each array configuration were as follows:

- **For Triangular Array:** *Shure MX391/O* (Shure, 2007). It is omnidirectional with a flat frequency response across the vocal range. It has a Signal-to-Noise Ratio (SNR) of 79.5 dB with a reference signal of 94 dB SPL. It has a permanent connection to its cable.

- **For 3D Array:** *Shure MX393/O* (Shure, 2015). It is also omnidirectional with a flat frequency response across the vocal range. However, it has an SNR of 80 dB with a reference signal of 94 dB SPL, and can be unplugged from its cable, making it easier to install.

As it can be seen, these two microphones have a very similar SNR, thus the difference between the recordings using the triangular array and using the 3D array should not be that much different in terms of quality.

C. Speakers

The speakers used for each array configuration were as follows:

- **For Triangular Array:** *Behringer Truth B3030A* (Behringer, 2009). It is a top of the line studio monitor speaker, with a XLR/TRS input combo. It has a flat frequency response in the range of 50 Hz to 24 kHz. Its pre-amplification was changed according to the recording environment (detailed in Section VI). It has a maximum sound pressure level of 113 dB at 1 m.

- **For 3D Array:** *KRK VXT 4* (KRK, 2007). It is a more moderate type of studio monitor speaker, which also has a XLR/TRS input combo. It has a flat frequency response in the range of 66 Hz to 22 kHz. Its pre-amplification was also changed according to the recording environment (detailed in Section VI). It has a maximum sound pressure level of 107 dB at 1 m.

As it can be seen, the frequency range in which both speakers have a flat response is not that different, which should not impact in any significant manner the recordings in which they were used. However, it is important to note that there is a considerable difference between these two speakers in terms of their maximum sound pressure level. This is considered when choosing their pre-amplification level in each recording environment, as detailed in Section VI.

IV. SOFTWARE

The software modules used for reproduction and recording of audio were built in-house, based on the Jack Audio Connection Kit (Davis, 2002). This API presumes to provide a real-time, synchronous handling of multi-channel audio data, which is essential for its use with sound source localization and separation algorithms. In this section, both software modules (reproduction and recording) are detailed, as well as the integration script that binds them.

A. Reproduction: ReadCorpusMulti

This section details the software used for multi-channel speech reproduction through electronic speakers, referred to here as ReadCorpusMulti. As it will be described in Section VI, there were some recording environments that involved mobile speakers, in which cases human volunteers acted as the speech sources and this module was not used.

During the capture process, ReadCorpusMulti feeds audio data to the speakers. This audio is obtained from the DIMEx100 corpus (Pineda *et al.*, 2010) which is conformed by read speech recordings of 100 users in a recording studio with a very low amount of noise. In DIMEx100, each user was recorded saying 50 sentences that were balanced phonetically in terms of the Mexican Spanish language.

ReadCorpusMulti assigns to each electronic speaker a randomly-chosen user from DIMEx100, and only feeds audio data from that DIMEx100 user to simulate that the speaker is an specific person with their own vocal characteristics. This is important as it is consistent to the circumstances in which human volunteers acted as the speech sources. Once this assignment is carried out, ReadCorpusMulti randomly chooses 1 out of the 50 recorded sentences of the DIMEx100 user to reproduce in the assigned electronic speaker. It also inserts a silence between the DIMEx100 recordings, which can be configured as an argument in number of milliseconds (q).

ReadCorpusMulti was written such that it is able to do this with a pre-specified number of electronic speakers s, and for a pre-specified amount of time t.

B. Recording: WriteMicMulti

This section details the software used for multi-channel audio recording, referred to here as WriteMicMulti. It is a simple module that records the audio data captured by each microphone which it then stores in an audio file. Meaning that for every microphone in the array, WriteMicMulti creates one audio file. It uses the libsndfile (de Castro Lopo, 1999) library to write these audio files as standard, mono 16-bit WAV files. WriteMicMulti appropriately closes each WAV file once it receives an external terminate signal, and names each WAV file according to the number of the microphone from where it received audio data.

C. Integration

ReadCorpusMulti and WriteMicMulti are bound by an integration script that runs them both in parallel threads, and synchronizes the termination of WriteMicMulti with the termination of ReadCorpusMulti which is triggered when it finishes its reproduction process.

As mentioned before (and detailed in Section VI), there are some recording environments that employ mobile sources, in which human volunteers act as the speech sources. In these cases, instead of running ReadCorpusMulti, this script runs a simple program that reproduces a "bleep" sound through one speaker as a start signal for the human volunteers, waits a pre-specified amount of time (the same as the one fed to ReadCorpusMulti), and reproduces another "bleep" sound as a stop signal.

V. CAPTURE PROTOCOL

In this section, a summary of the protocol that was followed to capture the AIRA corpus is summarized. For ease of explanation, consider the following:

- N is the number of sources
- n is the id of the speech source, such that $0 < n \leq N$
- C_N is the set of configurations for a given number of sources N
- p_n^c is the position of source n in configuration c
- q_c is the number of milliseconds of silence between speaker reproductions for configuration c
- E is the number of 'evaluations' or repetitions (later clarified)
- e is the id of the 'evaluation', such that $0 < n \leq E$

To this effect, the following protocol was carried out at each recording environment:

1. For $N = range(1 : N_{max})$, For each c in C_N, For $n = range(1 : N)$
   (a) For each p_n^c in c
      i. Position source n at p_n^c
      ii. Register position p_n^c as part of the corpus
      iii. If sources are mobile, p_n^c describes a path that the human volunteers were asked to follow
   (b) For $e = 1 : E$
      i. Launch Integration script described in Section IV, with $t = 30s$, $s = N$ and a given $q = q_c$.
      ii. Reset source positions to initial configuration c

The value of N_{max} and all C_N's were decided for each recording environment, considering the time and space constraints.

It is important to clarify what we mean by 'evaluations', since it is used here in an unorthodox manner. It is basically a repetition of the same environment and source position configuration, but with different speech reproductions and/or sentences uttered by the humans. The reason why the term 'evaluation' is used lays on the fact that this corpus was heavily used by the authors to carry out evaluations for robot audition algorithms. Modifying this term in the whole of the corpus AIRA is a paramount task that we ask the reader to overlook and simply consider as 'repetition'.

Once the protocol was finished for a recording environment, information complementary to the recordings was added. It is important to remember that electronic speakers reproducing speech recordings were used to simulate static speech sources, and in these cases the clean channel data for each source was available and stored as part of AIRA. However, in some recording environments, human volunteers were asked to participate as mobile speech sources and in these circumstances the clean channel data is unavailable. Thus, the following is the complete set of actions that were carried out after each capturing protocol:

1. Listen to every recording to check for artifacts from the capture process, such as clicks, pops, and/or system overruns. If so, remove the recordings containing them. Thus, there are some source position configurations of the corpus that do not have the same amount of evaluations as the rest.

2. For the remaining recordings, if the speech users were the electronic speakers reproducing the DIMEx100 speech recordings:
   (a) Add the clean channel recordings from each speech source by concatenating the appropriate DIMEx100 speech recordings. These recordings can be used as the ground truth for source separation purposes.

(b) Using the transcripts from DIMEx100, concatenate them to build the create the transcript for each clean channel. These can be used for multiple-source automatic speech recognition.

(c) Segment clean channel recordings between voice and silence, using the audioSegmentation module from the pyAudioAnalysis library (Giannakopoulos, 2015).

(d) Using the segmentation information, create the ground truth position file for each evaluation. The formatting of this file is later explained in Section V A.

3. Or if they were human volunteers:

(a) Copy over the text of the randomly-selected sentences from DIMEx100 as the transcript for each user.

(b) Listen to the recording and manually trim the sentences for each speech source where the recording stopped.

(c) Create a ground truth position file from the established path the user was asked to follow, assuming a constant speed from the start and end point. This implies that the path presented in these files is not technically a ground truth, but an estimation of their paths. The exception to this was the case of the recording environment of Office C, where this ground truth position file was created automatically via a laser-based user tracking system.

A. The Ground Truth Position File (MDOA)

The ground truth position files, as mentioned before, are where the positions of the speech sources are registered in AIRA. This position is presented as the direction of arrival of the speech source from the frames of references of the array configurations described in Section II.

These files are in a format referred here as MDOA (from Multiple Direction of Arrival), and presents the position of each source at each sample window. They are in fact text files that have the following format:

```
Sample:  00000
DOAS:
---
Sample:  00001
DOAS:
30.00
---
Sample:  00002
DOAS:
30.00
-45.00
---
Sample:  00003
DOAS:
---
```

In this example, no sources were active in sample window 0, a source at 30^o became active in sample window 1, another source became active at -45^o in sample window 2 along with the source at 30^o, and in sample window 3 both sources became inactive.

VI. RECORDING ENVIRONMENTS

In this section, the different environments in which AIRA was captured are detailed, as well as the hardware settings used for each. The noise level was measured in each recording environment by an SPL meter and are presented in dB SPL. The reverberation time (τ_{60}) was measured using a reverberation estimator that can provide measurements that are greater than 0.01 s.

Unless stated otherwise, all sources were placed 1 m. from the center of the microphone array. The azimuth angles are measured using the frame of reference presented in Figures 1 and 2. The heights are measured relative to the the center of the 3D array.

There were seven recording environments employed, here described.

A. Anechoic Chamber

This environment is located inside the full-anechoic chamber (Boullosa and Lopez, 1999) of the Instituto de Ciencias Aplicadas y Tecnología (ICAT, formerly known as the Laboratorio de Acústica y Vibraciones of the Centro de Ciencias Aplicadas y Desarrollo Tecnológico, CCADET) of the Universidad Nacional Autónoma de México (UNAM). It measures 5.3 m x 3.7 m x 2.8 m. A photo that is representative of the setup on-site is presented in Figure 3.

FIG. 3. Photograph of the setup in the Anechoic Chamber.

It has a very low noise level ($\approx$ 0.13 dB SPL) with a $\tau_{60} < 0.01s$. No other noise sources were present in this setting. Recordings have a SNR of $\approx$ 43 dB.

In this environment, the two array configurations were used, referred to as "Anechoic Chamber 3" and "Anechoic Chamber 16" in the corpus website for the triangular array and the 3D array respectively.

For the triangular array, the microphone pre-amplification was set such that it had a level of -20 dBFS when capturing a pure tone from one of the electronic speakers with a pre-amplification at 0 dB located at 0.5 m away from the microphone. The distance between microphones was set at 0.18 m. During the recordings, the speaker pre-amplification was set at -10 dB.

The source positions for 'Anechoic Chamber 3' are presented in Table II.

TABLE II. Source positions (DOA) for Anechoic Chamber 3.

1 Source	2 Sources		3 Sources			4 Sources			
S. 1	S. 1	S. 2	S. 1	S. 2	S. 3	S. 1	S. 2	S. 3	S. 4
45^o	-30^o	90^o	-30^o	90^o	-150^o	0^o	90^o	180^o	-90^o
	0^o	90^o	0^o	90^o	180^o				
	45^o	90^o							

For the 3D array, the microphone pre-amplification was set at 0 dB, since this was more or less which was necessary when using the triangular array (this is expected, since the noise was so low). During the recordings, the speaker pre-amplification was set at -30 dB to compensate for the maximum SPL difference between monitors. The microphones were located as specified in Table I.

The source positions for 'Anechoic Chamber 16' are presented in Table III.

B. Cafeteria

This environment is a cafeteria located inside the UNAM campus and was used during a 5 hour period of high customer presence. It has an approximate size of 20.7 m x 9.6 m x 3.63 m. A photo that is representative of the setup on-site is presented in Figure 4.

It has a high noise level (71 dB SPL) with an average $\tau_{60} = 0.27s$. Its walls and floor are made of concrete, and its walls are made of a mixture of concrete and glass. Noise sources around the array included: people talking, babies crying, tableware clanking, some furniture movement, and cooling fans of stoves. Recordings have a SNR of $\approx$ 16 dB. Its frequency-dependent reverberation time is shown in Figure 7.

In this environment, only the 3D array configuration was used, referred to as "Cafeteria 16" in the corpus website.

The microphone pre-amplification was set at 40 dB, as to obtain a level close to what a normal conversation

TABLE III. Source positions (DOA and height) for Anechoic Chamber 16.

1 Source	2 Sources		3 Sources		
S. 1	S. 1	S. 2	S. 1	S. 2	S. 3
-45^o	-30^o	0^o	-30^o	90^o	-150^o
0.08 m	0.08 m	0.08 m	0.08 m	0.08 m	0.08 m
-90^o	-30^o	0^o	0^o	90^o	180^o
0.08 m	-0.17 m	0.23 m	0.08 m	0.08 m	0.08 m
0^o	-30^o	90^o	0^o	90^o	180^o
0.08 m	0.08 m	0.08 m	0.23 m	0.08 m	-0.17 m
0^o	-30^o	90^o	60^o	90^o	120^o
0.23 m	-0.17 m	0.23 m	0.08 m	0.08 m	0.08 m
0^o	0^o	90^o	60^o	90^o	120^o
-0.17 m	0.08 m	0.08 m	0.23 m	0.08 m	-0.17 m

4 Sources			
S. 1	S. 2	S. 3	S. 4
0^o	90^o	180^o	-90^o
0.08 m	0.08 m	0.08 m	0.08 m
0^o	90^o	180^o	-90^o
0.08 m	0.23 m	0.08 m	-0.17 m
45^o	90^o	135^o	0^o
0.08 m	0.08 m	0.08 m	0.08 m
45^o	90^o	135^o	0^o
0.08 m	0.23 m	0.08 m	-0.17 m
60^o	90^o	120^o	30^o
0.08 m	0.08 m	0.08 m	0.08 m
60^o	90^o	120^o	30^o
0.08 m	0.23 m	0.08 m	-0.17 m

FIG. 4. Photograph of the setup in the Cafeteria.

is recorded at. During the recordings, the speaker pre-amplification was set at -25 dB to compensate for the maximum SPL difference between monitors. The microphones were located as specified in Table I.

The source positions for 'Cafeteria 16' are presented in Table IV.

C. Department Store

This environment is a sizable department store (comparable to a Walmart or Tesco) known as "Tienda UNAM" located inside the UNAM campus. It has an approximate size of 91 m x 62 m x 6 m. It was used during a 5 hour period of moderate customer presence

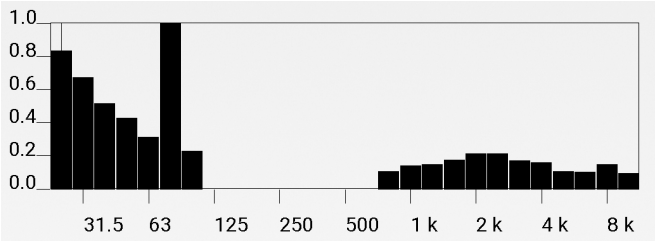

FIG. 5. Frequency dependent reverberation time of the Cafeteria setting.

TABLE IV. Source positions (DOA and height) for Cafeteria 16, Department Store 16 and Office A 16.

1 Source	2 Sources		3 Sources		
S. 1	S. 1	S. 2	S. 1	S. 2	S. 3
-45^o	30^o	0^o	30^o	-90^o	150^o
-0.02 m	-0.02 m	-0.02 m	-0.02 m	-0.02 m	-0.02 m
-90^o	30^o	0^o	0^o	-90^o	180^o
-0.02 m	-0.17 m	0.22 m	-0.02 m	-0.02 m	-0.02 m
0^o	30^o	-90^o	0^o	-90^o	180^o
-0.02 m	-0.02 m	-0.02 m	0.22 m	-0.02 m	-0.17 m
0^o	30^o	-90^o	-60^o	-90^o	-120^o
0.22 m	-0.17 m	0.22 m	-0.02 m	-0.02 m	-0.02 m
0^o	0^o	-90^o	-60^o	-90^o	-120^o
-0.17 m	-0.02 m	-0.02 m	0.22 m	-0.02 m	-0.17 m

4 Sources			
S. 1	S. 2	S. 3	S. 4
0^o	-90^o	180^o	90^o
-0.02 m	-0.02	-0.02 m	-0.02 m
0^o	-90^o	180^o	90^o
-0.02 m	0.22 m	-0.02 m	-0.17 m
-45^o	-90^o	-135^o	0^o
-0.02 m	-0.02 m	-0.02 m	-0.02 m
-45^o	-90^o	-135^o	0^o
-0.17 m	0.22 m	-0.17 m	0.22 m
-60^o	-90^o	-120^o	-30^o
-0.02 m	-0.02 m	-0.02 m	-0.02 m
-60^o	-90^o	-120^o	-30^o
-0.17 m	0.22 m	-0.17 m	0.22 m

(as only allowed by the store administration). A photo that is representative of the setup on-site is presented in Figure 6.

It has a high noise level (63 dB SPL), and it has an average $\tau_{60} = 0.16s$. It has a high ceiling made of plastic and aluminum, its walls are made of brick, and its floor made of concrete. As it can be seen, this environment is quieter and less reverberant than the Cafeteria environment. Noise sources around the array included: people talking, people walking by with trolleys, and general announcements. Recordings have a SNR of ≈ 17 dB. Its frequency-dependent reverberation time is shown in Figure 7.

FIG. 6. Photograph of the setup in the Department Store.

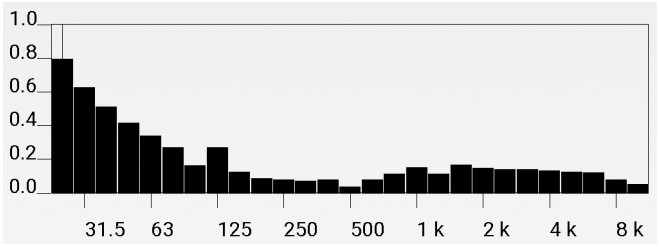

FIG. 7. Frequency dependent reverberation time of the Department Store setting.

In this environment, only the 3D array configuration was used, referred to as "Department Store 16" in the corpus website.

The microphone pre-amplification was set at 40 dB, as to obtain a level close to what a normal conversation is recorded at. During the recordings, the speaker pre-amplification was set at -25 dB to compensate for the maximum SPL difference between monitors. The microphones were located as specified in Table I.

The source positions for 'Department Store 16' are the same as 'Cafeteria 16', which are presented in Table IV.

D. Hall

This environment is the hallway of the Computer Science Department of the Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS) of the UNAM. It has an average width of approximately 1.5 m, a length of approximately 28 m, and a height of 2.1 m. A diagram of the setting is presented in Figure 8.

The line in blue represents an approximate path of the robot's movement which carried the microphone array.

It has a low noise level (48 dB SPL), and it has an average $\tau_{60} = 0.21s$. The ceiling is made of plaster, and the walls and the floor are made of concrete. Noise sources around the array included: inter-office chatter and robotic motor ego-noise. Recordings have a SNR of ≈ 10 dB. Its frequency-dependent reverberation time is shown in Figure 7.

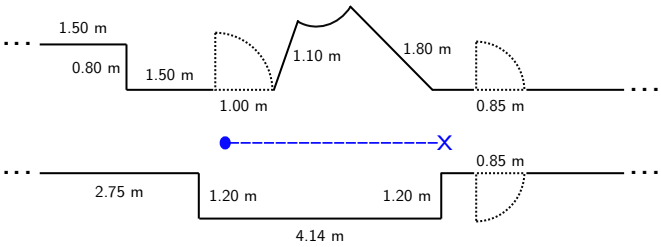

FIG. 8. Map of the Hall setting.

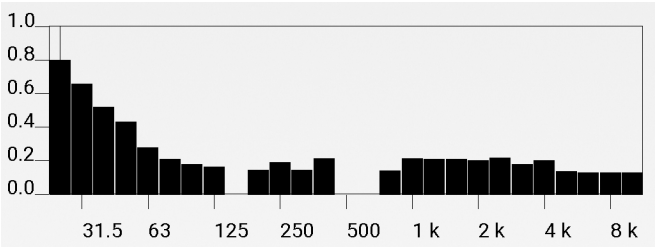

FIG. 9. Frequency dependent reverberation time of the Hall setting.

In this environment, only the triangular array configuration was used, referred to as "Hall 3" in the corpus website.

The microphone pre-amplification was set such that it had a captured level close to -40 dBFS at a background level, as to avoid clipping from the audio interface. The distance between microphones was set at 0.21 m. During the recordings, human volunteers acted as speech sources and were asked to read pre-specified randomly chosen sentences from the DIMEx100 Corpus while they were moving in a pre-specified path. This path only had a starting and stop position, thus the positions in the ground truth position files are estimations which assumed they moved in a uniform fashion.

In addition, the microphone array was also moving, being placed over a service robot called Golem-II (Pineda *et al.*, 2015), that was lent from the Golem Group. The robot was programmed to move in a straight line, through the hallway, at a speed of 0.13 m/s. The MDOA files were created according to the center of the microphone array.

The source paths for 'Hall 3' are presented in Table V. If a path is shown as a one number (instead of path, such as "x to y"), it implies that the user was walking along with the robot.

It bares clarifying that in this setting the distance of the sources to the center of the microphone array is dynamic.

E. Office A

This environment is one of the computer labs of the the Computer Science Department of the IIMAS of the UNAM. It has an approximate size of 5.7 m x 6.6 m x 2.1 m. A diagram of the setting is presented in Figure 10.

TABLE V. Source paths (DOA) for Hall 3.

1 Source	2 Sources		3 Sources		
S. 1	S. 1	S. 2	S. 1	S. 2	S. 3
31^o to 149^o	31^o to 149^o	-90^o	31^o to 149^o	-90^o	90^o

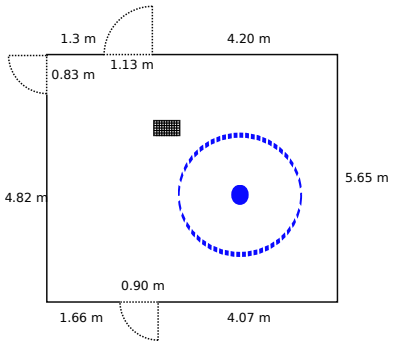

FIG. 10. Map of the Office A setting.

The blue dot represents the center of the array, and the circle surrounding represents the imaginary line where the sources were positioned. A photo that is representative of the setup on-site is presented in Figure 11.

FIG. 11. Photograph of the setup in Office A.

It has a somewhat low noise level (52 dB SPL) with an average $\tau_{60} = 0.20s$. The ceiling is made of plaster, the walls a mix between concrete and glass, and the floor of concrete. Noise sources around the array included: inter-cubicle chatter and computer cooling fans. Recordings have a SNR of $\approx$ 21 dB. Its frequency-dependent reverberation time is shown in Figure 12.

In this environment, the two array configurations were used, referred to as "Office A 3" and "Office A 16" in the corpus website for the triangular array and the 3D array respectively.

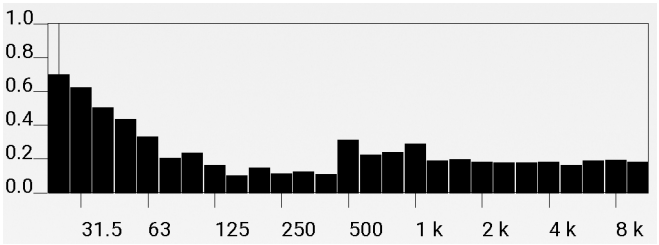

FIG. 12. Frequency dependent reverberation time of the Office A setting.

For the triangular array, the microphone pre-amplification was set such that it had a captured level close to -40 dBFS at a background level, as to avoid clipping from the audio interface. The distance between microphones was set at 0.21 m. During the recordings, the speaker pre-amplification was set at -6 dB. The source positions for 'Office A 3' are presented in Table VI.

TABLE VI. Source positions (DOA) for Office A 3.

1 Source	2 Sources		3 Sources			4 Sources			
	S. 1	S. 2	S. 1	S. 2	S. 3	S. 1	S. 2	S. 3	S. 4
S. 1	-30^o	90^o	-30^o	90^o	-150^o	0^o	90^o	180^o	-90^o
0^o	0^o	90^o	0^o	90^o	180^o				
	45^o	90^o							

For the 3D array, the microphone pre-amplification was set at 40 dB, as to obtain a level close to what a normal conversation is recorded at. During the recordings, the speaker pre-amplification was set at -30 dB to compensate for the maximum SPL difference between monitors. The microphones were located as specified in Table I.

The source positions for 'Office A 16' are the same as 'Cafeteria 16', which are presented in Table IV.

F. Office B

This environment is another of the computer labs of the Computer Science Department of the IIMAS of the UNAM. It has an approximate size of 10.5 m x 4.9 m x 2.1 m, and is divided into three 3.5-m-wide spaces that are acoustically connected. A diagram of the setting is presented in Figure 13.

The blue dot represents the center of the array, and the circle surrounding represents the imaginary line where the sources were positioned.

It has a low noise level (42 dB SPL) with an average $\tau_{60} = 0.42s$. The ceiling is made of plaster, the walls a mix between concrete and glass, and the floor of concrete. As it can be seen, it is a bit quieter but more reverberant than Office A. Noise sources around

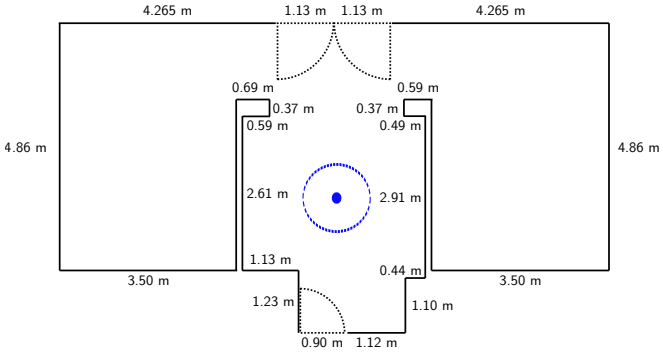

FIG. 13. Map of the Office B setting.

the array included: inter-cubicle chatter and computer cooling fans. Its frequency-dependent reverberation time is shown in Figure 14.

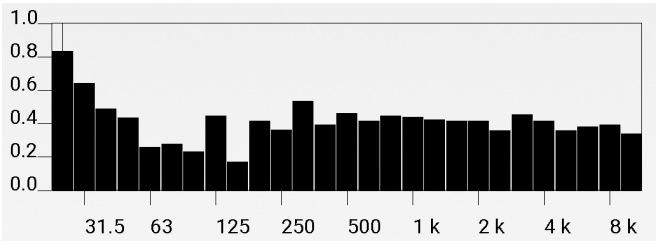

FIG. 14. Frequency dependent reverberation time of the Office B setting.

In this environment, only the triangular array configuration was used, referred to as "Office B 3" in the corpus website.

For the triangular array, the microphone pre-amplification was set such that it had a captured level close to -40 dBFS at a background level, as to avoid clipping from the audio interface. The distance between microphones was set at 0.21 m. During the recordings, human volunteers acted as speech sources and were asked to read pre-specified randomly chosen sentences from the DIMEx100 Corpus while they were moving in a pre-specified path. This path only had a starting and stop position, thus the positions in the ground truth position files are estimations which assumed they moved in a uniform fashion.

The source paths for 'Office B 3' are presented in Table VII.

G. Office C

This environment was the same as Office A. However, in this case the sources were mobile and were automatically tracked. The tracking system used the laser on the bottom part of the Golem-II robot, and the human volunteers were asked to use an accessory on their legs that facilitated their tracking. Thus, the ground truth

TABLE VII. Source paths (DOA) for Office B 3.

1 Source		
S. 1		
90^o to 0^o to -90^o to -180^o to 100^o		
90^o to 0^o to -80^o		
2 Sources		
S. 1	S. 2	
0^o to 90^o to 170^o	180^o to 90^o to 10^o	
90^o to 0^o to -80^o	-90^o to 180^o to 100^o	

position file was created automatically. Recordings have a SNR of $\approx$ 14 dB.

In this environment, only the triangular array configuration was used, referred to as "Office C 3" in the corpus website.

For the triangular array, the microphone preamplification was set such that it had a captured level close to -40 dBFS at a background level, as to avoid clipping from the audio interface. The distance between microphones was set at 0.21 m.

The source paths for 'Office C 3' are presented in Table VIII.

TABLE VIII. Source paths (DOA) for Office C 3.

1 Source	2 Sources		3 Sources		
S. 1	S. 1	S. 2	S. 1	S. 2	S. 3
90^o to 0^o	90^o to 0^o	0^o to -90^o	90^o to 0^o	0^o	0^o to -90^o

VII. CORPUS STRUCTURE

In this section, a brief summary is provided of the directory structure of the AIRA corpus for reference and ease of use. For each recording environment, the following directory structure was followed:

- *File* **HardwareSetupforEval.py :** it is a Python script that describes the hardware setup used for the recording environment, specifically the microphone positions and distance between them.
- *File* **Audio Setup.txt :** It is a text file describing in a general manner the configuration of the microphones and speakers in terms of their preamplification and positioning.
- *Directory* **X Source :** It is a directory that holds all the recordings carried using an X number of sources. It follows the following structure:
  - *Directory* **Source Configuration :** it is a directory that holds all the recordings for a given source position configuration. Its name provides a summary of the source DOAs, their heights (if appropriate), and the amount of silence between speaker reproductions. It follows the following structure:
    - *File* **GetDOACorpusConfig.py :** it is a Python script describing all the variables set for this configuration, such as source locations, number of 'evaluation' (aka. repetitions), length of recordings, etc.
    - *Directory* **Evaluation E:** it is a directory that holds the recorded data for the Eth repetition (referred to as 'evaluation' for the reasons explained in Section V) of the source configuration. It follows the following structure:
      - *File* **goldstandard.mdoa :** it is the ground truth position file of the sources. If the sources were mobile and an estimation of their path is given, this file is named **estimation.doa**. An exception to this is the Office C set of recordings, since the source were tracked by a laser-based system, and their true positions are given through time.
      - *File* **pristine_channelY.wav :** it is the clean audio data of source Y. These files are not provided when the sources were human volunteers.
      - *File* **speech_channelY.txt :** it is the transcript of the what source Y said during the recording.
      - *File* **wav_micZ.wav :** it is the recorded data from microphone Z.
      - *File* **readcorpus__config.txt :** it is a text file that contains an overview of the data reproduced via the ReadCorpusMulti program, such as the amount of time that was reproduced, the amount of silence between speech reproductions, what specific DIMEx100 recording was used for each channel, etc. This file is omitted when the sources were human volunteers.
- *File* **Chirp :** Directory with the same structure as the Source Configuration folder, but it contains the recording of a sine-sweep signal from one speaker with a starting frequency of 50 Hz and ending frequency of 4 kHz.

VIII. CONCLUSION

In this paper, the AIRA corpus is formally introduced and detailed. It can be used to model or to evaluate techniques for sound source localization and separation, as well as multi-user speech recognition, in the aspects of evaluation and model training. It uses two microphone array configurations, was recorded in 6 very varied acoustic scenarios, it includes clean speech data for static sources and tracking information (both grounded and estimated) for mobile sources, and it is freely available from https://aira.iimas.unam.mx/.

This version of AIRA was captured over a span of seven years. However, there are plans to continue complementing it with recording in different recording environments. Additionally, it is of interest to provide different Signal-to-Noise ratio per recording environment by employing several speaker pre-amplification amplitudes. It is also of interest to use different bodies with which to set the microphone arrays, as to explore different materials as well as different array topologies. Finally, it is also of interest to expand the AIRA corpus into other application scenarios, such as autonomous drones. In fact, we have made an important push in this regard (Ruiz-Espitia *et al.*, 2018) that the official AIRA website links to, and that we encourage the readers to follow.

ACKNOWLEDGMENTS

The authors thank the support of CONACYT through the projects 81965, 178673 and 251319, PAPIIT-UNAM through the project IN107513 and ICYTDF through the project PICCO12-024. In addition, the authors would like to thank the Golem Group and their lead researcher, Dr. Luis Pineda, who provided the DIMEx100 corpus on which the AIRA recordings were based on and loaned the Golem-II+ robot for some of the recordings. Furthermore, we would like to specifically thank the support of Oscar Aguilar, Rodolfo Petrearce, Varinia Estrada, and Alfonso Vega for their help during the capture process of the corpus. Finally, a special recognition is given to Dr. Santiago Jesus Perez Ruíz, of the Laboratorio de Acústica y Vibraciones of the Instituto de Ciencias Aplicadas y Tecnología (formerly known as the Laboratorio de Acústica y Vibraciones of the Centro de Ciencias Aplicadas y Desarrollo Tecnológico) of the Universidad Nacional Autónoma de México, for his invaluable support in the recollection of the AIRA Corpus.

Arnaud, E., Christensen, H., Lu, Y.-C., Barker, J., Khalidov, V., Hansard, M., Holveck, B., Mathieu, H., Narasimha, R., Taillant, E., Forbes, F., and Horaud, R. (**2008**). "The CAVA corpus: Synchronised stereoscopic and binaural datasets with head movements," in *Proceedings of the 10th International Conference on Multimodal Interfaces*, pp. 109–116.

Avid Technology, Inc. (**2007**). *Fast Track Ultra User Guide*, https://c3.zzounds.com/media/071206_FTUltra_UG_EN01-a4d756ff6b3b638a95d99cecca31fa70.pdf.

Behringer (**2009**). *Behringer Truth B3030A User Guide*, https://img.musicworld.bg/pdf/i/3/1/8/25813/B3030A_oper-instr.pdf.

Behringer (**2016**). *Behringer X32 User Guide*, https://media.music-group.com/media/PLM/data/docs/P0ASF/X32_M_EN.pdf.

Boullosa, R. R., and Lopez, A. P. (**1999**). "Some acoustical properties of the anechoic chamber at the centro de instrumentos, universidad nacional autonoma de mexico," Applied Acoustics **56**(3), 199–207.

Davis, P. (**2002**). *JACK Audio Connection Kit*, http://jackaudio.org.

de Castro Lopo, E. (**1999**). *libsndfile*, http://www.mega-nerd.com/libsndfile/.

Deleforge, A., Drouard, V., Girin, L., and Horaud, R. (**2014**). "Mapping sounds onto images using binaural spectrograms," in *European Signal Processing Conference*, pp. 2470–2474.

Deleforge, A., and Horaud, R. (**2011**). "Learning the direction of a sound source using head motions and spectral features," Technical Report .

Giannakopoulos, T. (**2015**). "pyaudioanalysis: An open-source python library for audio signal analysis," PloS one **10**(12).

KRK (**2007**). *KRK VXT 4 User Guide*, https://s3.amazonaws.com/gibson-pro-audio/krk/manuals/vxt_manual.pdf.

Lathoud, G., Odobez, J.-M., and Gatica-Perez, D. (**2005**). "AV16.3: An audio-visual corpus for speaker localization and tracking," in *Machine Learning for Multimodal Interaction*, edited by S. Bengio and H. Bourlard, pp. 182–195.

Löllmann, H. W., Evers, C., Schmidt, A., Mellmann, H., Barfuss, H., Naylor, P. A., and Kellermann, W. (**2018**). "The LOCATA challenge data corpus for acoustic source localization and tracking," in *IEEE Sensor Array and Multichannel Signal Processing Workshop*.

Nakamura, S., Hiyane, K., Asano, F., and Nishiura, T. (**2000**). "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proceedings of the International Conference on Language Resources and Evaluation*, pp. 965–968.

Pineda, L. A., Castellanos, H., Cuétara, J., Galescu, L., Juárez, J., Llisterri, J., Pérez, P., and Villaseñor, L. (**2010**). "The Corpus DIMEx100: Transcription and Evaluation," Language Resources and Evaluation **44**, 347–370.

Pineda, L. A., Rodríguez, A., Fuentes, G., Rascon, C., and Meza, I. V. (**2015**). "Concept and Functional Structure of a Service Robot," International Journal of Advanced Robotic Systems **12**(2), 6, doi: 10.5772/60026.

Rascon, C., Fuentes, G., and Meza, I. (**2015**). "Lightweight multi-DOA tracking of mobile speech sources," EURASIP Journal on Audio, Speech, and Music Processing **2015**(11).

Rascon, C., and Meza, I. (**2017**). "Localization of sound sources in robotics: A review," Robotics and Autonomous Systems **96**, 184210, http://www.sciencedirect.com/science/article/pii/S0921889016304742, doi: 10.1016/j.robot.2017.07.011.

Ruiz-Espitia, O., Martinez-Carranza, J., and Rascon, C. (**2018**). "Aira-uas: an evaluation corpus for audio processing in unmanned aerial system," in *2018 International Conference on Unmanned Aircraft Systems (ICUAS)*, pp. 836–845.

Shure (**2007**). *Shure MX391 User Guide*, http://www.shure.pl/dms/shure/products/microflex/documents/mx391/mx391-user-guide_MULTI_816kb/mx391-user-guide_MULTI_816kb.pdf.

Shure (**2015**). *Shure MX393 User Guide*, https://pubs.shure.com/view/guide/MX39x/en-US.pdf.