A Framework for Infrastructure-Free Indoor Localization Based on Pervasive Sound Analysis

Ricardo Leonardo, Marilia Barandas, and Hugo Gamboa, *Member, IEEE*

***Abstract*— Even as modern indoor positioning systems become more precise and computationally lightweight, most rely on specific infrastructure to be installed, leading to increased setup and maintenance costs. As such, multiple infrastructure-free solutions were devised relying on signals such as magnetic field, ambient light, and movement. In this paper, we propose a framework for determining the user's location through the sound recorded by the user's device. With this goal, we present two algorithms: SoundSignature and SoundSimilarity. With SoundSignature, we extract acoustic fingerprints from the recorded audio and employ them in a support vector machine classifier. With SoundSimilarity, where we employ a novel audio similarity measure to detect if users are in the same location as other users or microphone equipped devices. Both of these algorithms require no infrastructure and are computationally lightweight, thus allowing their use either in conjunction with other infrastructure-free technologies or standalone. The training of these algorithms requires nothing more than a smartphone or a similar device under normal usage conditions, eliminating the need of any dedicated equipment.**

***Index Terms*— Indoor location, sound analysis, infrastructure-free, support vector machine, feature selection, SMOTE.**

I. Introduction

IN RECENT times there has been a continuous increase in the ubiquity, processing power and sensing capabilities of modern smartphones. This has made possible the emergence of new technologies that allows users to keep track of their health, activities and location.

In the context of location, the most well known and widespread technology is the Global Navigation Satellite System (GNSS). However, while this system's precision and satellite coverage are typically sufficient for outdoor applications, this is not the case when the user is inside a building. The presence of walls and ceilings between the user and the satellites greatly attenuates the latter's signal, and the reduced scale of typical paths in buildings compared to outdoor routes creates a demand for higher precisions.

Multiple alternative solutions have been proposed in the literature for indoor positioning systems. Many leverage signals transmitted between beacons and the device to be located, namely Radio Frequency signals, either Wi-Fi [1] or Bluetooth [2], infrared signals [3], ultrasound [4], visible light [5], among others. By estimating the distance of the device to each beacon through metrics such as received signal strength (RSS) [6] and time difference of arrival (TDoA) [7], trilateration may be used to locate the device. Other methods include using these metrics for fingerprinting techniques [8]. However, most current systems rely on infrastructure, leading to elevated setup and maintenance costs.

Some infrastructure-free solutions have also been proposed such as using pervasive signals such as ambient light [9] and perturbations in the Earth's magnetic field for fingerprinting techniques [10]. Other systems integrate many of the aforementioned developments with inertial tracking and map data to locate the user indoors [11]. An existing sound-based solution uses the power spectra of the audio signal as an acoustic fingerprint to differentiate between different rooms [12].

Furthermore, Jun-Wei Qiu and Yu-Chee Tseng [13] have shown that meetings between two or more users may be used to calibrate their respective potential locations. Sound-based proximity detection was achieved through use of the normalized cross-correlation coefficient between the spectra of the compared signals [14].

The aim of this paper is to provide a framework for indoor location based on pervasive sound, taking advantage of a sensor present in every smartphone and requiring no dedicated infrastructure. The training of these algorithms require nothing more than a smartphone or similar device under normal usage conditions, eliminating the need for dedicated and possibly expensive recording equipment and easing the acquisition process. With this goal we provide two distinct tools to locate the user through pervasive sound. The first one is *SoundSignature*, where we determinate in which location the user is based on previous recordings of said location. Relevant work on sound recognition includes algorithms for music recognition. Commercially available solutions [15] extract

Manuscript received December 22, 2017; revised February 8, 2018 and March 6, 2018; accepted March 8, 2018. Date of publication March 21, 2018; date of current version April 23, 2018. This work was supported in part by the North Portugal Region Operational Programme (NORTE 2020), Portugal 2020 and the European Regional Development Fund (ERDF) from the European Union through the project Symbiotic technology for societal efficiency gains: Deus ex Machina (DEM), NORTE-01-0145-FEDER-000026, and project Collective Transfer FhP, NORTE-01-0246-FEDER-000029. The associate editor coordinating the review of this paper and approving it for publication was Dr. Ashish Pandharipande. *(Corresponding author: Ricardo Leonardo.)*

R. Leonardo and M. Barandas are with the Fraunhofer Portugal Research Center for Assistive Information and Communication Solutions, 4200-135 Porto, Portugal (e-mail: ricardo.leonardo@fraunhofer.pt).

H. Gamboa is with the Fraunhofer Portugal Research Center for Assistive Information and Communication Solutions, 4200-135 Porto, Portugal, and also with the Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics, Faculdade de Ciłncias e Tecnologia, Universidade Nova de Lisboa, 2829-516 Lisbon, Portugal.

Digital Object Identifier 10.1109/JSEN.2018.2817887

relevant data from the recording's spectrogram and isolate local maxima from background noise to generate a fingerprint specific for that song in that time displacement.

The second tool is SoundSimilarity, with which we can determinate whether two users are in the same location. For that we compare the signals received from both users without requiring any prior information related to the location itself. This allows multiple users to calibrate their respective potential locations based on the comparisons between each other's received signals.

This paper starts with a description of the developed algorithms, in particular, the Sound Signature algorithm (Section II) and the Sound Similarity algorithm (Section III). After describing the algorithms, the evaluation process and the main results achieved are presented and discussed (Section IV). In Section V we analyse the time complexity of the developed algorithms. Section VI explains how the two developed algorithms complement and interact with each other. Finally, the main conclusions are drawn (Section VII) and some future work is identified (Section VIII).

II. SoundSignature: Indoor Location Through Location-Dependent Acoustic Fingerprints

Inspired by the aforementioned fingerprint-based solutions for indoor positioning systems, a system capable of locating people inside buildings by extracting location-dependent fingerprints from sound data is proposed.

Fingerprint-based solutions are typically divided into an offline stage and an online stage [16]. In the offline stage of the proposed solution, the recordings are split into windows of equal length from which features are extracted. The processing of this data is also done in this stage. During the online stage, data is collected in real time and used as the algorithm's input, returning an estimation of the user's location.

In the following subsections a detailed description of the developed algorithm is presented.

A. Acoustic Fingerprint Extraction

Every location has a distinct set of acoustic characteristics by which it can be identified. These characteristics are the result of constant background noises, such as the humming of computer fans and air conditioning systems. These noises are further modulated by the impulse response of the location itself, determined by its physical characteristics such as dimensions and materials used [17]. We propose that these characteristics can be considered sufficiently stable and unique for each considered location and as such they can be used in the construction of an acoustic fingerprint.

Given any audio signal, its spectrogram can be computed. This is done by dividing the segment into windows, computing each respective Fourier transform, discarding the redundant second half and multiplying by its complex conjugate. The result is a representation of the evolution of the frequency spectrum along time.

As shown in the spectrogram from figure 1, two distinct components can be identified: the background noise spectrum,

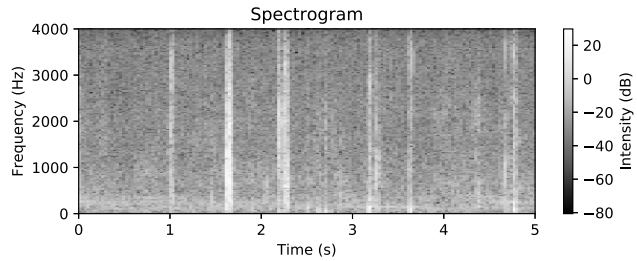

Fig. 1. In this spectrogram, we can differentiate the acoustic background spectrum from the transient sounds. The former remain constant throughout the spectrogram, while the latter are added to this and are represented in higher intensities (in white).

that relates to the aforementioned intrinsic acoustic characteristics of the location and remains consistent throughout the spectrogram, and short duration transient sounds, which are added to this spectrum.

To separate this background frequency from the remaining sounds, we could extract the minimum of each frequency along the spectrogram and merge them together to create a spectrum. However, this methodology may lead to erroneous results due to the inherent dynamic range compression of commonly available microphones, which leads to a decrease in gain proceeding loud sounds. Moreover, artefacts derived from the recording system's own signal processing and noise, be it from acoustic or electromagnetic nature, may also be a factor. As such, we choose instead the 5th-percentile of the power of each frequency as a value that is both an approximation to the minimum and more robust to these effects [12].

B. Feature Extraction

Achieved a frequency spectrum that can be used as an acoustic fingerprint, we proceed to extract features from said spectrum. Three subsets of features were considered:

1) Logarithm of Each Frequency: In signal analysis and processing it is common to use decibels (dB) [18] to better represent and compare the intensity of sound signals. The logarithmic nature of the decibel allows better differentiation between values spread out along different orders of magnitude. Given these resulting features will later be normalized by subtracting the mean value and dividing by the standard deviation, we choose to use just the logarithm of the intensity of each frequency.

2) Mel Frequency Cepstral Coefficients: The Mel Frequency Cepstral Coefficients (MFCC) are a group of features commonly used in speaker [19] and speech recognition [20]. These have also been used in recognition of human activities based in recorded sound [21].

The MFCC are based on the mel scale [22], a scale that translates frequencies in hertz to the way humans perceive pitch. This is achieved by applying a logarithm with base two, thus assuring that the difference between octaves remains constant. This scale is adjusted in such a way that it maps the values of 0 Hz and 1000 Hz to 0 mels and 1000 mels, respectively. The conversion from hertz to mels is done by the following equation (1):

$$m = 1000 \times log_2(1 + \frac{f}{1000}) \tag{1}$$

where m is the pitch in mels and f is the frequency in hertz.

To compute the MFCC from the obtained intensity spectrum, we start by applying a filter bank of triangular overlapping windows whose frequencies are equidistant in the mel scale. A discrete cosine transform is then applied to the logarithms of the dot products between each frequency band and the frequency spectrum. The number of filter banks used was 26, since this number is typically used in speech related applications.

Similarly to the previous subset, this subset was normalized by subtracting the mean value and dividing by the standard deviation computed from all MFCC values.

3) Additional Spectral Features: Other features were extracted from the background frequency spectrum. These features were *centroid*, *spread*, *skewness*, *kurtosis*, *slope*, *decrease* and *roll-off*. The formulations for these features are in the [23].

Each feature was individually normalized by subtracting the mean and dividing by the standard deviation.

C. Feature Selection

Given the large amount of features, a feature selection algorithm was employed to improve the accuracy and computational performance of the algorithm. The chosen algorithm was Sequential Forward Feature Selection [24]:

1) Start with an empty feature set $Y_0 = \{\}$, an accuracy $a_0 = 0$, an objective function J and $k = 0$;
2) Select the feature x^+ that maximizes $J(Y_k + x)$;
3) If $J(Y_k + x^+) > a_k$, update $Y_{k+1} = Y_k + x^+$, $a_{k+1} = J(Y_k + x^+)$ and $k = k + 1$ and go back to *2)*, otherwise continue;
4) Keep only the feature set Y_k and discard the rest.

D. Classification Algorithm

The classification algorithm employed was Support Vector Machines [25] using the *one-versus-rest* approach for multi-class classification, resulting in k individual binary classifiers where k is the number of classes [26]. To allow non-linear separations, the *kernel trick* [27] was employed using the radial basis function (*rbf*). Synthetic samples were generated with the SMOTE algorithm [28] to compensate for any possible class unbalance.

In the offline stage, acoustic fingerprints are extracted from labelled data, from which the selected features are extracted. These features and associated labels are then used to train the SVM classifier.

In the online stage, the user's location in t_0 is estimated by computing the acoustic fingerprint from the audio recorded in the interval $[t_0 - 5, t_0]$. The selected features are extracted from this fingerprint and used as the classifier's input, where the associated output will be the user's estimated position.

Furthermore, the probability of an input belonging to each class can be obtained with the method proposed by Wu *et al.* [29],

E. Validation

The validation method of the resulting classifier was a stratified 10-fold cross validation. This method tries to maintain the ratio between classes in each fold, thus minimizing the possible class unbalance caused by randomly selecting data for each fold. This method was used to validate the classifier and as the objective function J in feature selection.

III. SoundSimilarity: Detecting Co-Location Based on Audio Similarity

In order to identify whether two microphones are in the same location, we must employ some measure of similarity between the sounds recorded. Such measurement must be able to identify similarities in the shape of the waveforms along time, but impervious to phase differences derived from possible small time misalignments.

A. Cross-Correlation

Cross-correlation can be defined as measure of similarity between two series as a function of the displacement between them. For two real valued discrete signals x_1 and x_2, it can be formulated as:

$$(x_1 \star x_2)[n] = \sum_{m=-\infty}^{\infty} x_1[m]x_2[m+n] \tag{2}$$

Given the finite length of the data, common practice is to extend the series with leading and trailing zeros. However, this methodology is prone to errors due to the resulting tendency to give more weight to central values. Therefore, we instead employ circular cross-correlation, where the input series are extended with periodic summations. Given the discrete-time Fourier transform already employs this extension, we can employ the cross-correlation theorem and formulate the discrete circular cross-correlation as:

$$(x_1 \star x_2)[n] = \mathcal{F}^{-1}\{\overline{\mathcal{F}\{x_2\}} \cdot \mathcal{F}\{x_2\}\}[n] \tag{3}$$

This formulation avoids the aforementioned artefact and greatly improves computational performance.

B. Measuring the Similarity of Audio Segments

Given the circular cross-correlation between two finite series, the series will be similar if a pronounced peak is present. As such, we propose a novel measurement for audio similarity (MAS) correlated to the presence of this peak.

Let x_1 and x_2 be two series corresponding to the recordings of two microphones in a window of t_{win} with a time misalignment $t_{delay} < t_{win}/d$, where the parameter d determines the proportion between the length of the cross-correlation and the width of the centre region that contains the correlation peak, as determined in figure 2. Given a circular cross-correlation C_{x_1,x_2} between these two series, we start by taking the absolute value of each value in the correlation:

$$C_{abs} = |C_{x_1,x_2}| \tag{4}$$

Then we define a region R centred around $t_{win}/2$ and with width t_{win}/d:

$$R = [\frac{t_{win}}{2} - \frac{t_{win}}{2d}, \frac{t_{win}}{2} + \frac{t_{win}}{2d}] \tag{5}$$

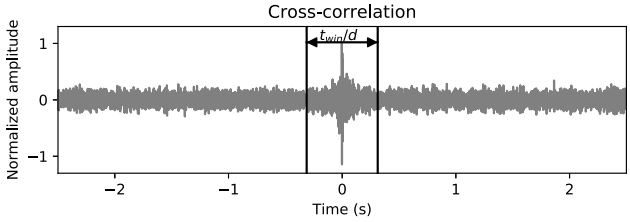

Fig. 2. Cross-correlation between two audio signals with a length of 5 seconds. The presence of a pronounced peak in the central region indicates that the signals are similar and therefore were recorded in the same room at the same time. The parameter d determines the width of the region where we expect this peak to appear.

Finally, we define the measurement for audio similarity (MAS) as:

$$MAS_{f,g} = 1 - \min\left(\frac{\max_{t\in\overline{R}} C_{abs}}{\max C_{abs}}, 1\right) \tag{6}$$

If a pronounced correlation peak is present, it will be contained in the region R. Therefore, the overall maximum of the correlation will be larger than the maximum in $\overline{R}$, and the greater this difference the closer to zero the ratio $\frac{\max_{t\in\overline{R}} C_{abs}}{\max C_{abs}}$ will be.

On the other hand, if no correlation peak is present, we can suppose that the correlation has the properties to those of noise. In this case, two situations can occur: if the overall maximum is contained in $\overline{R}$, we can obviously say that it is equal to the maximum in $\overline{R}$, and therefore the ratio between them will be one; if the overall maximum is contained in R, it will be approximately equal to the maximum in $\overline{R}$, and therefore the ratio between them will be approximately one.

Given these properties, the result of this measurement is a value between 0 and 1 where the greater the value, the more similar are the audio signals used as input. Furthermore, by defining a range where the correlation peak can be instead of a fixed point, we make the algorithm robust to small misalignments in time.

C. Co-Location Detection

To apply this measurement in real time, for each second a window of the last t_{win} seconds is extracted from each signal source and compared using the above described algorithm. To avoid short duration artefacts, the results are then smoothed by applying an IIR low-pass filter.

In order to identify whether two devices are in the same location, a threshold for this measurement must be set. With that goal, for a range of potential thresholds we plot the resulting $sensitivity$ against $1 - specificity$, generating a Receiver Operating Characteristic (ROC) curve. From this curve a threshold is obtained by choosing the one that maximizes Youden's J statistic [30]:

$$J = sensitivity + specificity - 1 \tag{7}$$

Measurements above this threshold indicate the devices are in the same location, while measurements below this value indicate the devices are in different locations.

TABLE I
EFFECT OF SMARTPHONE POSITION ON THE RESULTS

Position for validation	Accuracy
Calling	86.78%
Hand	50.22%
Pocket	34.58%
Texting	79.91%

TABLE II
EFFECT OF DIFFERENT SMARTPHONES ON THE RESULTS

Smartphone for validation	Accuracy
Nexus 5 A	91.21%
Nexus 5 B	84.51%
Samsung S5	82.23%

IV. RESULTS

A. SoundSignature

To train and test the proposed algorithm, a series of experiments for data collection were conducted. Several routes were designed and labeled according to the location they were in. The user would then go through these routes with a smartphone in a predetermined position. The audio recording was collected while users simply walk through each planned route from a starting to an ending point.

Preliminarily, a small dataset for distinction between four rooms was acquired to test the effect of different smartphones and positions. The positions used were in calling position, in hand as if the user was texting, swinging in the right hand and in the right pocket of jeans. The smartphones used were two Nexus 5 and a Galaxy S5.

Table I shows the accuracies achieved with by leaving each position for validation while using the remaining for training. Our results show that while the Calling and Texting positions yield acceptable results, the Hand and Pocket positions are not fit for this application.

To test the effect of different smartphones, a similar approach was used. The data recorded with each smartphone was left out for testing while using the remaining for training. The "Hand" and "Pocket" positions were excluded. Table II shows that it is possible to identify the location of the user when comparing to data from other smartphones.

To validate the algorithm with a larger amount of locations and under different conditions, two distinct datasets were recorded. The first one was recorded in 8/8/2017 with the air conditioning system turned off, while the second one was recorded in 13/11/2017 with the air conditioning system turned on. While the locations used as classes were the same, the routes were not. The datasets were recorded by two different persons.

These datasets used a total of 16 locations, the disposition of which can be seen in figure 3. Each recording was split into non overlapping windows with a length of 5 seconds each. This results in a total of 495 labeled samples. For each of these an acoustic fingerprint was extracted using the aforementioned algorithm. The number of points used in the FFT for the spectrograms was 512, generating spectra with

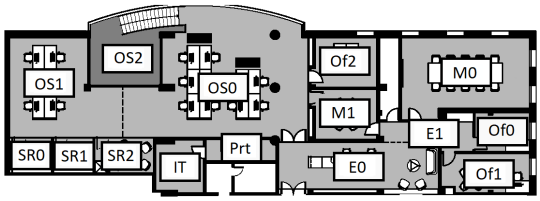

Fig. 3. Disposition of the locations used to validate the SoundSignature algorithm. There is one location that could not depicted: a small enclosed room in the bottom floor we named "Lab." There is no physical barrier between "OS0" and "OS1," as well as between "E0" and "E1."

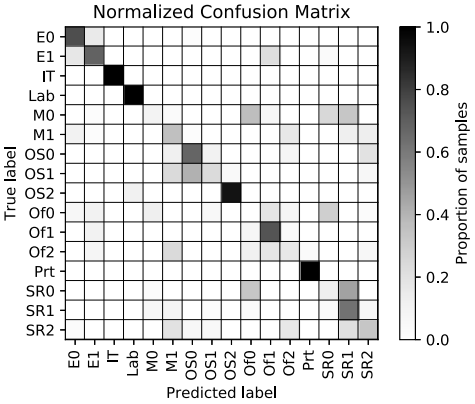

Fig. 4. Normalized confusion matrix obtained by using data recorded in different days for training and testing.

257 frequency bins each. The full set of features was extracted for each of these fingerprints.

Using the features extracted from the data in the first dataset, the above described feature selection algorithm was applied, reducing the number of features from 290 to 13. All features correspond to logarithms of powers of frequencies from the acoustic fingerprint, and all from frequencies bellow 2000 Hz. Validation through 10-fold stratified cross-validation yields an accuracy of 90.28%.

To test the algorithm with data from different days for training and validation, a different method was employed. Using the same selected features, the first dataset was used for training while the second was used for validation. Posteriorly, the second dataset was used for training while the first was used for validation. The achieved accuracy with this method was 48.08%, and the resulting confusion matrix can be observed in figure 4. By analysing this matrix, we can observe that while the algorithm failed for some locations, others were successfully identified. We believe this is due to noise sources that define the intrinsic acoustic properties of these locations:

- "OS0" is where the computers are located, emitting a characteristic noise from their fans;
- "IT" is where the servers are located, also emitting a characteristic sound from their fans;
- In "Prt" we can hear the sound of "IT" from behind a door, which acts as a natural low-pass filter;
- "OS2" and "Lab," being underground, have dedicated ventilation systems that produce noise;
- "E1" has a small fridge that emits a characteristic sound;
- We could not find an explanation for "E0" or "SR1."

Finally, by validating both datasets together with 10-fold stratified cross-validation, the accuracy achieved is 77.89%. The resulting confusion matrix and its analysis is in figure 5.

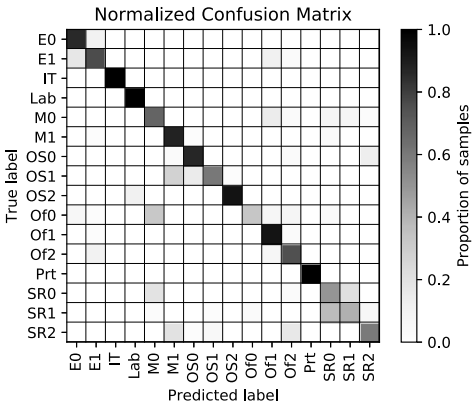

Fig. 5. Normalized confusion matrix obtained by applying the SoundSignature algorithm to the datasets recorded in 8/8/2017 and 13/11/2017, validated with 10-fold stratified cross-correlation. We can observe that there is some misclassification between "OS0" and "OS1," between which there are no physical barriers. Similarly, there is some confusion between "E0" and "E1." There is also some notable confusion between "SR0" and "SR1," locations that are physically nearly identical.

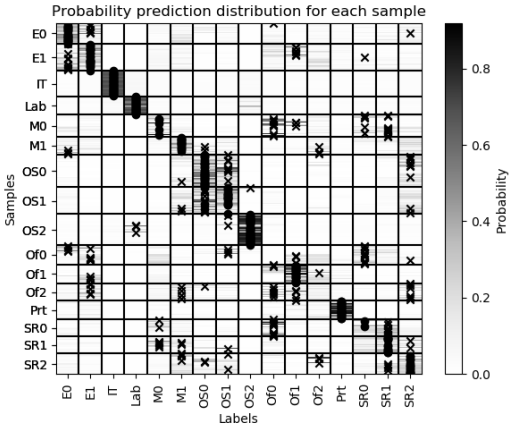

Fig. 6. Class membership probability estimates for each sample for each recorded sample when validating the classifier under the conditions of figure 4. Each line represents a sample, the circles represent correct classifications and the crosses represent wrong classifications.

Figure 6 shows the class membership probability estimates for each sample when validating the classifier under the conditions of figure 4. This figure shows that even when a sample is misclassified, the probabilistic response shows that it does so with a low degree of certainty and the correct class often has a higher than average probability.

All data was acquired using a recorder application with tap counter functionality that eased the process of annotating samples. The datasets were collected using a LG Nexus 5 Android smartphone. Data processing was executed in Python 3.5 using algorithms from the library *scikit-learn* [31].

B. SoundSimilarity

For data acquisition, a series of experiments were designed. These involved two microphones: one in a fixed location and another carried by the subject. While both microphones are recording the subject walks through a predetermined path. These paths start in the same location as the fixed microphone, leave to another location and then return to the starting point. The devices used included an HP laptop, an iPad Air tablet,

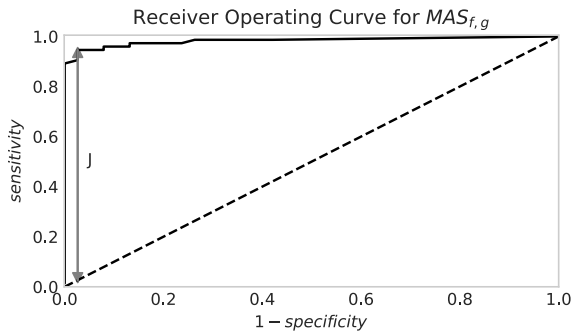

Fig. 7. Receiver Operating Curve (ROC) for the MAS. The area under the curve is 0.9829. Youden's J statistic for a certain threshold is equal to the distance from the random guess line (plotted in dashes) to the ROC curve. The ideal threshold is the one that maximizes this statistic, in this case 0.261.

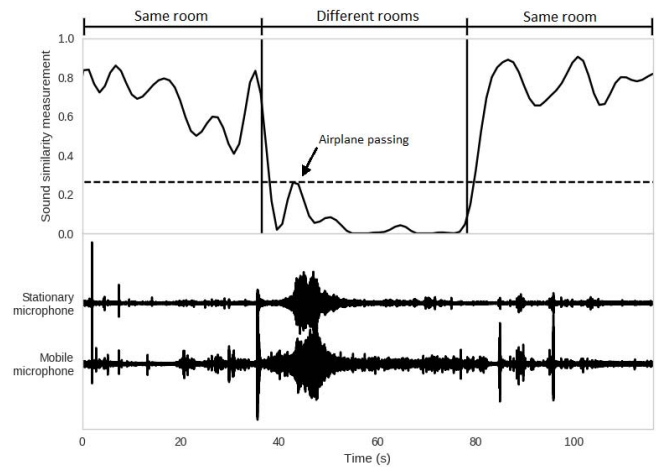

Fig. 8. Graphical representation of the MAS over time between two separate audio signals: one from a stationary microphone and another from a smartphone carried by a subject walking along a predefined route. These signal are represented in the bottom half. In the top half, along with the MAS are shown the ground truth and the found threshold. Annotated is the passing of an airplane while the two microphones are in different rooms, showing that the algorithm is resilient and adapts to this event.

a LG Nexus 5 smartphone and an iPhone 6 smartphone, and no pair of recordings used two of the same device.

To simulate usage of the algorithm in real time, the first step was to align the signals from both microphones. Since that at the start both the subject and the fixed microphone will be at the same location, we can assume that a peak will be present in the correlation between the first seconds. As such, for two signals f and g, a correlation between the first $t'_{win} = 10s$ of each signal is computed. The misalignment t_{delay} is given by the following expression:

$$t_{delay} = \frac{t'_{win}}{2} - t[argmax(f \star g)] \tag{8}$$

Once the signals are aligned, for each second a window of the last t_{win} of each signal was extracted. The similarity between each pair of windows was then computed through use of the previously described algorithm, returning the similarity of signals over time. This output was then smoothed with a low-pass IIR filter of order 5 with a cut-off frequency of 0.3Hz. The free parameters were set to $t_{win} = 5s$ and $d = 16$, chosen empirically.

After applying this process to every pair of recordings, a ROC curve is computed (figure 7). The area under this curve (AUROC score) is 0.9829, and maximizing Youden's statistic in this curve yields an ideal threshold of 0.248.

Graphs of audio similarity over time such as figure 8 were generated for each of the 11 pairs of simultaneous recordings. These graphs show that most errors occur in transition between moments where the devices are in the same location and moments where they are not.

V. Time Complexity Analysis

A. SoundSignature

The extraction of the fingerprint begins with the computation of a spectrogram without window overlap, which for n sample points and w *fft* points performs $\frac{n}{w}$ *fft* operations with a time complexity of $O(w \log w)$. This results in a total time complexity of $O(n \log w)$ and $\lfloor \frac{w}{2} \rfloor + 1$ frequencies with $\frac{n}{w}$ values each.

The chosen sorting algorithm for the extraction of the 5th percentile of each frequency was heapsort, resulting in a time complexity of $O(\frac{n}{w} \log \frac{n}{w})$ for each frequency and $O(n \log \frac{n}{w})$ for this whole step. Since both MFCC's and the spectral features taken from [23] were consistently discarded during the feature selection algorithm, their time complexity won't be taken into account.

Using the *libsvm* library [32], the decision function for classification is as follows:

$$y_{pred} = sgn(\sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b) \tag{9}$$

where $y_i \alpha_i$ and b are parameters generated in the offline phase, l the number of support vectors selected during training, x_i the i-th support vector, x the evaluated feature vector, K the kernel function (in this case *rbf*) and y_{pred} the predicted class. The time complexity of this method therefore is $O(lm)$, where m is the number of selected features during feature selection. As such, when using the *one-versus-rest* approach for multi-class classification, the time complexity for prediction using k binary classifiers is $O(klm)$.

The total time complexity of the SoundSignature algorithm is therefore $O(n \log w + n \log \frac{w}{n} + klm)$. Given that w, k, l and m are typically much smaller than n, the computational cost is usually observed to be proportional to n.

By applying this algorithm to synthetic data 10000 times using an Intel Core i7-5500u, the average running time of $6.64 \times 10^{-4} \pm 7.48 \times 10^{-4}$ seconds per iteration.

B. SoundSimilarity

For two signals with n samples each, the correlation is implemented using the correlation theorem which consists of two *fft* operation and a sum of the resulting arrays, resulting in a time complexity of $O(n \log n)$. The remaining operations are done in either linear or constant time, resulting in a time complexity of $O(n \log n)$ for the SoundSimilarity algorithm.

VI. Integration of Both Algorithms

By using both algorithms simultaneously, it is possible to calibrate the predicted locations of users if proximity between them is detected.

In a practical example, on the one hand using SoundSignature Alice's smartphone reports "M0" as her predicted location, with a probability of 83%. On the other hand, Bob's smartphone reports a predicted location of "SR1" with a confidence score of 62%. If the SoundSimilarity algorithm reports that both subjects are in the same location, Bob's location will be to "M0," as the probability associated to his prediction is lower than Alice's.

VII. Discussion and Conclusions

In this paper, we presented two different algorithms for indoor localization based on sound. These algorithms allow to locate users and calibrate their locations by comparing the signals by them perceived, relying exclusively on pervasive sound and requiring no infrastructure.

A. SoundSignature

The first algorithm, SoundSignature, identifies the location the user is in. This algorithm can be divided into two stages: an offline stage and an online stage. During the offline stage, acoustic fingerprints are extracted from the training data by filtering out transient sounds from the background noise spectrum. From these fingerprints a large group of features is extracted and oversampled through use of the SMOTE algorithm. From these, the best features are selected through the use of a feature selection algorithm. Finally, these selected features are used to train an SVM classifier. During the online stage, the previously selected features are extracted from real-time audio to predict the location of the user.

Preliminary tests show that the algorithm can use data recorded from other smartphones without significant effects on the results, but positions of the device that introduce noise do affect them. For further validation, two datasets were recorded in different days under different conditions. Applying the algorithm to the first dataset and validating it through 10-fold cross-validation, we achieve an accuracy of 90.28%. Applying this same validation method to both datasets together yields an accuracy of 77.89%, while validating data from one day with data from another day returns an accuracy of 48.08%. Our results show that while the generated acoustic fingerprints are stable enough to classify with good accuracy data collected under the same conditions, this only holds true for some locations when recording under different conditions. We also show that training the classifier with data collected under different conditions significantly improves the results.

The features selected by the feature selection algorithm are all logarithms of the power of frequencies bellow 2000 Hz, indicating that MFCC and the other additional spectral features may be inadequate for this application. The fact that all chosen frequencies were bellow 2000 Hz may be an indicator that the sampling frequency may be reduced from 8000 Hz to 4000 Hz, further increasing the algorithm's computational performance.

While sound-based indoor location has been achieved with good results by Tarzia *et al.* [12], we consider our algorithm shows significant improvements over the current state of the art regarding the subject:

- Our recording process was significantly different: instead of using dedicated recording equipment with a static placement, we used a smartphone placed on the user while they walked through a predetermined route. This decision was led by the fact that we consider these the conditions under which an indoor positioning system would be used;
- The length of the window used was 5 seconds, allowing a swift feedback to the user and response to location changes, a requirement for real-time indoor navigation systems;
- When validating data with a classifier trained in different conditions, we achieved an accuracy of 48.08%, as opposed to the 17.9% achieved in an experiment similarly designed by Tarzia *et al.* [12];
- The classifier used allows not only classification of inputs but also obtaining class membership probability estimates.

The class membership probability estimates allow easy integration of the proposed algorithm with other probability-based methods, such as similar approaches to geomagnetic or radio signals and systems based on particle filters.

B. SoundSimilarity

The second algorithm, SoundSimilarity, detects if a user is in the same location as another user or a device. This algorithm relies on a novel measure of sound similarity (Measurement for Audio Similarity or MAS) that is based on the presence a correlation peak. This measurement's associated area under the ROC curve of 0.98 validates its use for the proposed function. By maximizing Youden's statistic in this curve we obtain an optimum threshold for the measurement. Observation of the graphs of sound similarity over time indicate that the algorithm only fails in moments of transition, showing a delay of approximately 1.5s. We believe this delay is due to the use of the t_{win} second window and to the low pass filtering.

Other similarity metrics were considered, namely covariance, correlation coefficient and mutual information score, all in both time and frequency domains. The only metric showing acceptable results was the correlation coefficient in the frequency domain, showing an AUROC score of 0.8803. The other considered metrics showed AUROC scores between 0.45 and 0.55, being therefore unusable for this application.

Most notably, we compared the obtained results to the results obtained by Satoh *et al.* [14]. While the algorithms performed similarly in static conditions, our method showed imperviousness to perturbations such as the sound of footsteps and the passing of airplanes, while the compared method behaved poorly under these conditions. The AUROC score obtained by applying the method proposed by Satoh *et al.* to our data is 0.8736.

C. Concluding Remarks

The SoundSignature algorithms allows devices to predict their location through the sound perceived by their microphones. The SoundSimilarity algorithm allows corrections to

these predictions, should proximity be detected but different predictions obtained.

While these were designed to work together, they are capable of functioning independently of each other. Furthermore, they can be used in conjunction with other indoor positioning systems already proposed in the literature as a new layer of information.

VIII. Future Work

Although the developed work shows promising results, further experimenting is needed in how different recording conditions affect the results of the SoundSignature algorithm. Moreover, we will also study the effects of imposing limitations on transitions between locations, thus excluding physically those that are impossible. We will also implement a rejection class by thresholding the degree of confidence returned by the classifier to try and prevent unexpected behaviours in unknown locations, as well as assuring that predictions are made with more certainty.

Different experiments will be executed to test the integration between both developed algorithms, allowing the SoundSimilarity algorithm to correct possible misclassifications made by the SoundSignature algorithm.

Finally, a mobile application will be developed to streamline the processes of data acquisition and training and validation of the developed algorithms, by allowing these to be tested in real-time.

References

[1] B.-G. Lee and W.-Y. Chung, "Multitarget three-dimensional indoor navigation on a PDA in a wireless sensor network," *IEEE Sensors J.*, vol. 11, no. 3, pp. 799–807, Mar. 2011.

[2] R. Faragher and R. Harle, "Location fingerprinting with Bluetooth low energy beacons," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 11, pp. 2418–2428, Nov. 2015.

[3] S. Hijikata, K. Terabayashi, and K. Umeda, "A simple indoor self-localization system using infrared LEDs," in *Proc. 6th Int. Conf. Netw. Sens. Syst. (INSS)*, Jun. 2009, pp. 1–7.

[4] J. Bordoy *et al.*, "Bank of Kalman filters in closed-loop for robust localization using unsynchronized beacons," *IEEE Sensors J.*, vol. 16, no. 19, pp. 7142–7149, Oct. 2016.

[5] S. Pergoloni, Z. Mohamadi, A. M. Vegni, Z. Ghassemlooy, and M. Biagi, "Metameric indoor localization schemes using visible lights," *J. Lightw. Technol.*, vol. 35, no. 14, pp. 2933–2942, Jul. 15, 2017.

[6] W. Xue, W. Qiu, X. Hua, and K. Yu, "Improved Wi-Fi RSSI measurement for indoor localization," *IEEE Sensors J.*, vol. 17, no. 7, pp. 2224–2230, Apr. 2017.

[7] A. Makki, A. Siddig, M. Saad, J. R. Cavallaro, and C. J. Bleakley, "Indoor localization using 802.11 time differences of arrival," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 3, pp. 614–623, Mar. 2016.

[8] L. Chen, K. Yang, and X. Wang, "Robust cooperative Wi-Fi fingerprint-based indoor localization," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 1406–1417, Dec. 2016.

[9] J. Liu *et al.*, "The uses of ambient light for ubiquitous positioning," in *Proc. IEEE/ION Position, Location Navigat. Symp. (PLANS)*, May 2014, pp. 102–108.

[10] Y. Ma, Z. Dou, Q. Jiang, and Z. Hou, "Basmag: An optimized HMM-based localization system using backward sequences matching algorithm exploiting geomagnetic information," *IEEE Sensors J.*, vol. 16, no. 20, pp. 7472–7482, Oct. 2016.

[11] V. Guimarães *et al.*, "A motion tracking solution for indoor localization using smartphones," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat. (IPIN)*, Oct. 2016, pp. 1–8.

[12] S. P. Tarzia, P. A. Dinda, R. P. Dick, and G. Memik, "Indoor localization without infrastructure using the acoustic background spectrum," in *Proc. 9th Int. Conf. Mobile Syst., Appl., Services (MobiSys)*. New York, NY, USA, Jun. 2011, pp. 155–168.

[13] J.-W. Qiu and Y.-C. Tseng, "M2M encountering: Collaborative localization via instant inter-particle filter data fusion," *IEEE Sensors J.*, vol. 16, no. 14, pp. 5715–5724, Jul. 2016.

[14] H. Satoh, M. Suzuki, Y. Tahiro, and H. Morikawa, "Ambient sound-based proximity detection with smartphones," in *Proc. 11th ACM Conf. Embedded Netw. Sensor Syst.* New York, NY, USA, Nov. 2013, p. 58.

[15] A. L.-C. Wang, "An industrial-strength audio search algorithm," in *Proc. 4th Int. Conf. Music Inf. Retr.*, Oct. 2003, pp. 7–13.

[16] L.-H. Chen, E. H.-K. Wu, M.-H. Jin, and G.-H. Chen, "Intelligent fusion of Wi-Fi and inertial sensor-based positioning systems for indoor pedestrian navigation," *IEEE Sensors J.*, vol. 14, no. 11, pp. 4034–4042, Nov. 2014.

[17] H. Okubo, M. Otani, R. Ikezawa, S. Komiyama, and K. Nakabayashi, "A system for measuring the directional room acoustical parameters," *Appl. Acoust.*, vol. 62, no. 2, pp. 203–215, Feb. 2001.

[18] W. H. Martin, "Decibel—The name for the transmission unit," *Bell Syst. Tech. J.*, vol. 8, no. 1, pp. 1–2, Jan. 1929.

[19] T. Kinnunen *et al.*, "Low-variance multitaper MFCC features: A case study in robust speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 1990–2001, Sep. 2012.

[20] S. Umesh and R. Sinha, "A study of filter bank smoothing in MFCC features for recognition of children's speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 8, pp. 2418–2430, Nov. 2007.

[21] E. Garcia-Ceja, C. E. Galván-Tejada, and R. Brena, "Multi-view stacking for activity recognition with sound and accelerometer data," *Inf. Fusion*, vol. 40, pp. 45–56, Mar. 2018.

[22] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, Jan. 1937.

[23] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," Dept. Anal./Synth., IRCAM, Paris, France, Tech. Rep., 2004.

[24] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153–158, Feb. 1997.

[25] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[26] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[27] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Dept. Comput. Sci., Nat. Taiwan Univ., Taipei, Taiwan, Tech. Rep., Jul. 2003.

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[29] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, Dec. 2004.

[30] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

[31] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[32] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

Ricardo Leonardo received the B.S. degree in biomedical engineering from the Faculty of Sciences and Technology, NOVA University of Lisbon, Portugal, where he is currently pursuing the M.Sc. degree in biomedical engineering. In 2017, he was invited to elaborate his master's thesis at the Fraunhofer Portugal AICOS, where he is currently a Research Assistant. His current research focuses on smartphone-based solutions for indoor locations.

Marilia Barandas received the M.Sc. degree in biomedical engineering from the Faculty of Sciences and Technology, NOVA University of Lisbon, Portugal. After completing her master's thesis, she was invited to lecture the laboratory classes of Medical Information Systems, FCT-UNL, and to join the CA3 Research Group, Centre of Technology Systems. Since 2015, she has been a Scientist at the Fraunhofer Portugal AICOS, focusing on indoor locations solutions based on smartphones' built-in inertial sensors.

Hugo Gamboa received the Ph.D. degree in electrical engineering from the Instituto Superior Técnico, Universidade de Lisboa. He is currently an Assistant Professor with the Physics Department, Faculty of Sciences and Technology, NOVA University of Lisbon, Portugal.

He is also a Senior Scientist at the Fraunhofer Portugal AICOS. He has authored more than 100 papers in conferences and journals. His research activities focus on biomedical instrumentation and biosignals processing and classification.