IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 21, 2020, accepted September 1, 2020, date of publication September 7, 2020,
date of current version September 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3022618

Acoustic Fingerprints for Access Management in Ad-Hoc Sensor Networks

PABLO PÉREZ ZARAZAGA[1], (Graduate Student Member, IEEE), TOM BÄCKSTRÖM[1], (Senior Member, IEEE), AND STEPHAN SIGG[2], (Member, IEEE)
[1]Department of Signal Processing and Acoustics, Aalto University, 00076 Espoo, Finland
[2]Department of Communications and Networking, Aalto University, 00076 Espoo, Finland
Corresponding author: Pablo Pérez Zarazaga (pablo.perezzarazaga@aalto.fi)

ABSTRACT Voice user interfaces can offer intuitive interaction with our devices, but the usability and audio quality could be further improved if multiple devices could collaborate to provide a distributed voice user interface. To ensure that users' voices are not shared with unauthorized devices, it is however necessary to design an access management system that adapts to the users' needs. Prior work has demonstrated that a combination of audio fingerprinting and fuzzy cryptography yields a robust pairing of devices without sharing the information that they record. However, the robustness of these systems is partially based on the extensive duration of the recordings that are required to obtain the fingerprint. This paper analyzes methods for robust generation of acoustic fingerprints in short periods of time to enable the responsive pairing of devices according to changes in the acoustic scenery and can be integrated into other typical speech processing tools.

INDEX TERMS Audio fingerprint, context based authentication, voice user interface.

I. INTRODUCTION

Multichannel array processing of acoustic signals has proven very useful in noise reduction applications and is a crucial part in many audio recording devices. However, these techniques rely on knowing the specific position of the microphones that capture the signal, which usually translates in expensive microphone arrays or multiple devices dedicated to a specific application. At the same time, we live surrounded by devices with microphones that can record and transmit audio signals, and, for that reason, wireless acoustic sensor networks (WASN) are becoming more popular. A WASN comprises of a group of independent low-resource devices connected to the same network, which record the audio in a room and process it in a distributed manner, behaving like a microphone array [1], [2]. This means that nearby mobile devices around us could collaborate, processing the audio signal that they record and providing improved services with respect to a single device. However, if multiple devices are recording, processing and sharing the audio signal in a room, we need to verify which devices are allowed to collaborate in this network. For example, if Alice is in a teleconference next to her coworker Bob, both their mobile phones could share their recorded signals and process them. However, when Alice goes into her office and closes the door, Bob's phone should not be allowed in the WASN anymore. For this reason, we need to define access management rules for WASNs that preserve the user's privacy in a distributed scenario.

Traditional multichannel processing techniques assume that the elements of the sensor array are located at known static positions. WASNs cannot make such an assumption, but analyze the acoustic environment to synchronize the multiple devices that are part of the network [3]. This enables the WASN to perform tasks such as beamforming or noise reduction in a distributed manner [4], [5].

An application where WASNs can provide a significant improvement is voice user interfaces (VUIs), where technologies are rapidly evolving to analyze the multiple aspects of the speech signal [6]–[8]. VUIs are growing in popularity because they allow us to interact with our devices using our voices, and they can be found in a wide range of applications such as voice over IP (VoIP) for communications or virtual assistants like Alexa or Siri [9]. However, to access a service, it is usually necessary to interact directly with the device that provides such a VUI. If all the devices around us collaborated in a WASN they could provide a distributed VUI that included all the services from each individual device. For example, if Alice wanted to make a phone call, she would only need

The associate editor coordinating the review of this manuscript and approving it for publication was Zheng Yan.

to say, ''call Bob'', her devices would then analyze her voice and send it to her phone to perform the call, regardless of the location of the phone.

Similar services can be found in applications such as Google's Nearby Share[1] and Apple's AirDrop,[2] which enable the sharing of files between nearby devices. These applications use the device's Bluetooth to analyze the proximity of the other device, such that devices in the range of the Bluetooth signal are considered close enough to share information. However, in a distributed VUI we do not only need to know if the devices are close to each other, it is necessary to identify the devices that would improve the usability of the VUI and remain within the user's trusted area.

When multiple devices collaborate and share our information, it is important to know how much information is shared and which devices are receiving it. If such an information flow is not handled properly, it can lead to violations of the users' privacy. Our voices contain a great amount of personal information, not only the content of a conversation, but also other private data such as our emotional state or health situation [10], [11]. If multiple devices record and share the user's voice, it is important to define which devices are allowed in the network so that the user's voice remains private. Therefore, if a device shares a speech signal with an unauthorized node outside the established VUI, that would entail a breach of privacy. This would also be the case if the shared personal information requested by an application was not essential for the proper operation of such an application. Thus, in order to preserve the user's privacy, it is necessary to define access management rules based on the information that the devices collect.

In human-to-human interaction, we have observed that people have awareness of who can hear them and adapt their speech accordingly [12]. We propose to apply a similar approach with devices, such that only devices which are within the reach of a user's voice are allowed to process it (see Fig. 1). Device authentication using ambient audio has been studied using a combination of acoustic fingerprints and fuzzy cryptography [13]–[16]. An audio fingerprint is generated from the ambient sound and used as a cryptographic key for the communication channel, such that, if the devices are in close proximity, they will generate the same key, allowing them to communicate with each other.

We propose to encode the energy spectrum of the signal to directly compare two fingerprints and ensure that synchronization errors do not reduce the performance of the fingerprint. Previously proposed methods overcome this problem by using long non-overlapping analysis windows [13]. However, the long recordings needed to generate the fingerprints reduce the responsiveness of the system. A whole conversation can last less than 6 seconds, and if the pairing process takes longer than that, then the conversation could be finished before the network is established.

[1]See https://blog.google/products/android/nearby-share/
[2]See https://en.wikipedia.org/wiki/AirDrop

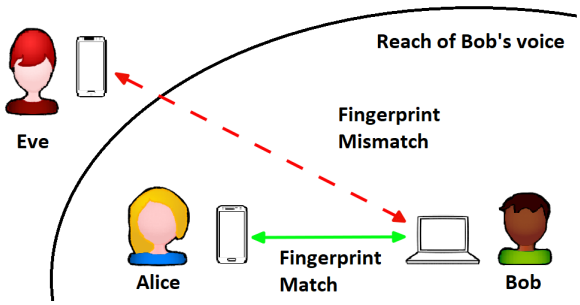

FIGURE 1. Ideal behavior of an access management system for distributed VUIs, where Alice and Bob share an acoustic space where their devices are allowed to collaborate. Eve, however, is outside the reach of Bob's voice and their devices are thus not permitted to collaborate.

Our objective is to authenticate devices in a wireless acoustic sensor network. The authentication process must 1) be robust to noise and distortions and 2) perform with low delay such that it is suitable for conversational applications. Our main contributions are four new acoustic fingerprinting methods based on 1) eigenvalue decomposition, 2) Wiener filtering, as well as 2D DCT with 3) entropy-based quantization and 4) mutual-information-based quantization of the spectrogram of the recorded signal. Our experiments demonstrate that the proposed methods improve the robustness of prior authentication methods, while reducing the length of the required audio recording. This enables the use of a distributed VUI in a conversational setting.

Section II analyzes device pairing methods and, more specifically, methods based on acoustic fingerprinting, describing their functioning and issues that need to be solved. The proposed acoustic fingerprinting methods operate over a log-energy spectrogram of the input signal. The procedure to extract such log-energy spectrogram is described in section III. Section IV depicts the different fingerprint extraction methods, which are combined with the quantization algorithms described in Section V. The generation process for the datasets used in the experiments, as well as the parameters of each fingerprint generator, are described in Section VI. Section VII presents an analysis of the results obtained from the multiple experiments. Finally, Section VIII summarizes the paper and discusses the obtained results.

II. RELATED WORK

Authentication between nearby devices using the environment's information has proven to be a useful feature in privacy preserving pairing of devices. Using the sensors of a device to obtain information of the environment allows us to recognize if multiple devices are located close to each other and, therefore, they could work together for a specific application.

Previous methods have used the signals recorded by the device's sensors to generate cryptographic keys that will enable secure communication between nearby devices. When the devices are close to each other, they would record similar features that will result in the same cryptographic key.

Sensors like the accelerometer [14], [15] or the microphone [13], [16] provide features that have proven useful for device pairing. More specifically, in audio applications, the features of the recorded audio are encoded into an acoustic fingerprint that can be easily compared to decide if two devices are hearing the same audio signal.

Acoustic fingerprinting is a popular technique used to compress the information of a segment of audio, which allows to efficiently compare multiple audio signals. This makes audio fingerprints useful in multiple fields, such as speaker recognition [17] or music retrieval [18], [19]. These fingerprints extract features that are useful for the specific application they will be used for. However, these approaches do not easily adapt to device authentication applications.

The method proposed in [13] uses the extracted acoustic fingerprint to derive the cryptographic key that can be used for secure communication. The 6.375 s long audio signal is divided into 17 non-overlapping time frames with a length of 0.375 s, scaled using a Hann window. They are transformed into the frequency domain and the resulting frequency components are then grouped into 33 energy bands. The fingerprint is then calculated as

$$FP(k,t) = \operatorname{sign}([E(k+1,t+1) - E(k,t+1)] - [E(k+1,t) - E(k,t)]), \quad (1)$$

where $E(k,t)$ represents the energy value on the k-th energy band and t-th time frame. The sign of each component is then quantized as a binary value, resulting in a 512 bit fingerprint. In order to allow mismatching bits caused by noise in the acoustic fingerprints, a fuzzy vault scheme is applied to calculate the cryptographic key that will be compared in the pairing process [20].

While this method allows secure pairing of devices using audio signals, the extensive length of the required recording makes it unsuitable for a conversational setting. The long non-overlapping windows are used to avoid synchronization errors between the communicating devices and reduce the correlation between time frames, which improves the statistical properties of the fingerprint. To reduce the length of the recordings, a combination of shorter overlapping windows and decorrelating transforms has been presented [21], providing higher robustness than previous methods.

III. PREPROCESSING

Our goal is to analyze the performance of different acoustic fingerprint methods in device authentication. The resulting fingerprints should maintain or surpass the performance of previous authentication methods [13], [21], including better adaptation to the authentication process.

WINDOWING

While applications like music retrieval are able to scan through the reference audio to find the optimum synchronization, fingerprints for device authentication must assume that the two recording devices are already synchronized. It is likely, however, that two devices will not start recording at the exact same time, and if this delay becomes too long, the audio information will be displaced to neighboring time frames, the generated fingerprints will not match and the communication between the devices will be restricted. To compensate for possible errors in synchronization, prior fingerprint generators have used long analysis windows to segment the audio signal. Long windows ensure that the features that are represented in the fingerprint are located in the same time frame. However, to obtain enough features for a robust fingerprint the length of the recorded audio may become too long for the application.

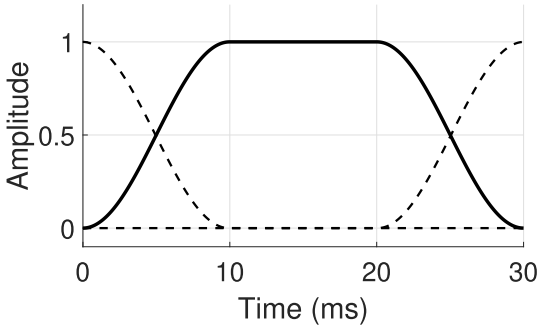

FIGURE 2. Windowing function used to segment the audio signal in the proposed methods.

We present methods for device authentication in voice user interfaces, where speech information is transmitted between devices. Therefore, we choose to adapt the parameters of the fingerprint extraction to the processing applied by a speech codec. This way, some of the processes carried out by the fingerprint extraction could be shared with the speech codec that would transmit the signal.

We have therefore chosen to use 30 ms overlapping windows with an overlap of one third of a window (10 ms). This will allow us to adapt the number of time frames to the requirements of the fingerprint without significantly compromising the length of the recorded audio. The windowing function is defined in three steps of 10 ms each. Both sides of the window are defined by slopes following a Hann window function and a flat middle section with an amplitude 1. As we can observe in Fig. 2, the windowing function allows the perfect reconstruction of the signal using the overlap-add method.

ENERGY GROUPING

The recorded signal is analyzed in the frequency domain, therefore, a short-time Fourier transform is applied to the windowed signal. The raw audio data in the frequency domain is, however, vulnerable to any noises present in realistic recordings. For that reason, we group the spectrum of the signal into energy bands. This will make the fingerprint extraction more robust against noise in the recorded signal.

From the spectrum of each frame, we extract the logarithmic energy of uniformly distributed spectral bands. In other words, for spectral coefficients x_f, the logarithmic energy of

band k is $b_k = \log\left[\sum_{f=kB}^{(k+1)B-1} |x_f|^2\right]$ where B is the width of the bands. We will also discuss averaging energy bands over a number of time frames to adapt the length of the data segment to the requirements of the whitening transformations that are applied later. This time averaging will also improve the robustness of the fingerprint against synchronization delays.

TIME FRAME AVERAGING

A central objective of this work is to reduce the length of time over which fingerprints are generated to make the authentication process more responsive. The main challenge in reducing the length of the recording is that such reduction tends to make fingerprints less robust to errors in synchronization. To improve the robustness, we therefore apply smoothing over time with an averaging filter.

The averaging of multiple time frames allows us to control the length of the recording without modifying the signal extraction methods. For example, a recording with a duration of 2.57s fits 128 windows. Averaging every 4 time frames, we can obtain 32 averaged frames for the corresponding audio length. Analogously, if the recording duration is 1.29s, which contains 64 windows, averaging every 2 time frames would also produce a sample with 32 averaged frames. This way, the input to the fingerprint extraction method will always have the same size.

IV. FINGERPRINT EXTRACTION

Our objective is to obtain a fingerprint that 1) contains characteristic information of the acoustic environment and 2) is robust against noise.

To that end, we apply decorrelation methods over the time and frequency axes of the energy band spectrum of the recorded speech signal. The resulting components are then quantized, thus obtaining the fingerprint as a binary array.

We consider two approaches to decorrelation, where signals are decorrelated in their time-frequency neighborhood or over the whole length of both axes. After decorrelation, the signals are quantized based on their respective entropies, as described in Section V.

SPECTRAL CONTEXT

One approach to decorrelation would be to diagonalize the covariance matrix between the energy bands over time and frequency. Since the decorrelation results in 32 energy bands and time frames, our data vector contains 1024 samples and calculations with a 1024×1024 covariance matrix would lead to an unbearable complexity.

To process the data more efficiently, we assume that the maximum correlation between energy bands is found in neighboring frequencies and time frames, and we will consider components that are far enough apart in frequency and time to be sufficiently decorrelated. Therefore, we define a spectral context, as shown in Fig. 3, over which we will calculate a whitening transformation matrix. In Fig. 3, x_0 stands for the energy of the target time-frequency bin and components x_1 to x_8 represent the neighborhood that we will analyze.

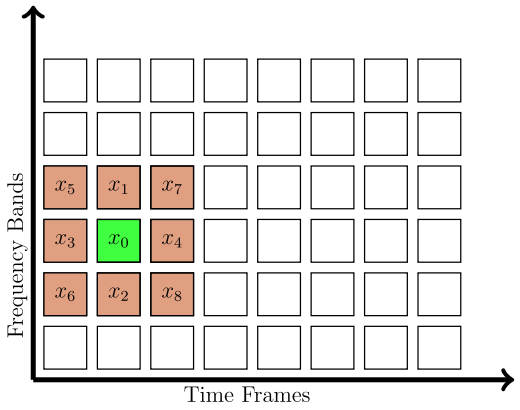

FIGURE 3. Example of the spectral context used to calculate a whitening transformation.

The correlation between neighboring frequency bands has already been used to predict the value of specific features in post-filtering techniques for speech enhancement [22]. Therefore, our aim is to decorrelate the signal by reducing the correlation of each bin with its neighboring ones.

EIGENVALUE EXTRACTION OVER SPECTRAL CONTEXT

The first proposed method consists of extracting the first eigenvalue [23] of the context components around each time-frequency bin of the energy spectrogram. This method aims to concentrate the relationship between the context components into one value which will then be quantized. The context matrix is transformed into a one dimensional vector $x_n = [x_0, x_1, \ldots, x_8]$ and, following the definition of eigenvalue decomposition, we calculate its eigenvalues and eigenvectors as

$$C_{xx} = V^T \Lambda V. \tag{2}$$

where C_{xx} represents the covariance matrix of the context arrays x_n. The covariance matrix is defined as $C_{xx} = E[X^T X]$, where X represents the set of context arrays x_n for a specific component, concatenated over the whole training set. The square diagonal matrix Λ contains the eigenvalues of C_{xx} and V contains their corresponding eigenvectors as columns. Since the covariance matrix is Hermitian, it will not be necessary to distinguish between the left and right eigenvectors as one will be the transposed version of the other. The calculated eigenvectors will then define the transformation that we will apply to the context data as

$$y_n(k,t) = V^T x_n(k,t). \tag{3}$$

where the resulting vector $y_n(k,t)$ represents the eigenvalues of the corresponding spectral context component $x_n(k,t)$ in the k-th frequency band and t-th time frame. Finally, we only keep the largest eigenvalue, which will then be quantized.

As the process of estimating the covariance matrix and calculating the corresponding eigenvalues every time we analyze a signal would pose a significant computational load, we decided to train the transformation matrix over a large set of speech data. This allows us to reduce the complexity of the transformation to a multiplication with a predefined matrix. Using a fixed transformation matrix might not adapt well to different noise environments that are not present in the training dataset. However, we do not consider that this will entail a significant issue, as the objective of this transformation is to simplify the comparison of two signals after the quantization of the fingerprint.

Since only the first eigenvalue is necessary, the complexity of the method can be further reduced by turning the matrix multiplication into a vector product. Each row of V^T contains an eigenvector of the transformation, therefore, to obtain the first eigenvalue, we only need to multiply our input data by the first row of the transformation matrix.

WIENER FILTER-BASED TRANSFORMATION

In our second approach, we consider the differences between signals in the same environment as additive uncorrelated noise. We propose to model the context of each frequency component using a Wiener filter [6] whose transfer function is

$$H_{xy}(z) = \frac{S_{yy}(z)}{S_{xx}(z)}, \tag{4}$$

where $S_{yy}(z)$ represents the energy spectrum of the clean signal and $S_{xx}(z)$ corresponds to the noisy one, which can be defined as $x_n(k,t) = y_n(k,t) + n_n(k,t)$. Following Eq. (4), we can then define our transformation:

$$y_n(k,t) = \mathbf{C}_{yy}\mathbf{C}_{xx}^{-1}x_n(k,t). \tag{5}$$

We assume the noise component is not correlated to the signal, therefore, to estimate $\mathbf{C}_{xx}$ in Eq. (5) we add a diagonal matrix to the covariance of our data such that $\mathbf{C}_{xx} = \mathbf{C}_{yy} + \alpha\mathbf{I}$. The term α represents a scale factor that we will set to 1 for the following experiments. In a similar way as the eigenvalue decomposition, in this case only one value of the final result will be used in the quantization. This means that we only need to apply one row from the transformation matrix, considerably reducing the complexity of the fingerprint calculation.

2-D TIME-FREQUENCY DCT

Previously proposed methods [21] present how the DCT can be used as a decorrelating transform to generate a robust acoustic fingerprint. Applying the transform only over the frequency bands, however, does not decorrelate the time structure of the signal. The time structure of the audio signal contains important information about the audio scene and, by decorrelating it, we expect to improve the quality of the generated fingerprints.

To decorrelate the time structure, we apply the DCT over both the time and frequency dimensions. Such 2D-DCTs are popular in image compression [24], which take advantage of

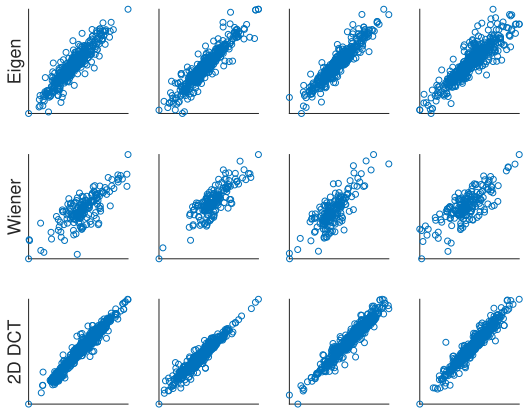

FIGURE 4. Scatter-plot for matching components of the proposed fingerprint extraction methods. (First row) Eigenvalue over the spectral context, (second row) Wiener filter over the spectral context, (third row) 2D-DCT transformation.

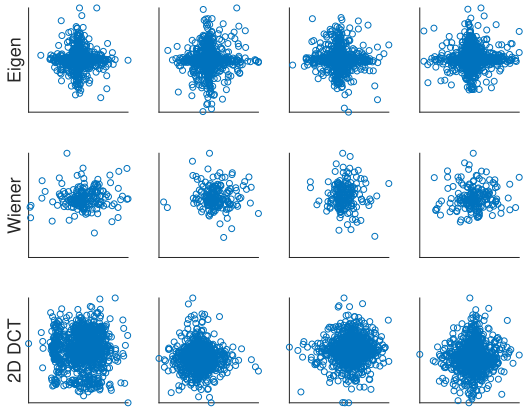

FIGURE 5. Scatter plot for non-matching components of the proposed methods. (First row) Eigenvalue over the spectral context, (second row) Wiener filter over the spectral context, (third row) 2D-DCT transformation.

frequency components over vertical and horizontal dimensions. An important reason for using the 2D-DCT for compression is because the information is grouped to the lowest frequency components, which results in a smaller number of components holding the majority of the information.

Figures 4 and 5 illustrate the relationship between the values in matching and non-matching components of the transformed log-energy spectrograms. We can observe that, while the matching cases are in a thin strip and thus highly correlated, the components of non-matching pairs do not show such a structure and can be assumed to be uncorrelated. Therefore, we expect that quantizing the transformed log-energy spectrograms will provide robust fingerprints to recognize matching scenarios.

V. BIT ALLOCATION

Features generated by these decorrelation methods are represented as real-valued scalars, which we will need to quantize to obtain the final binary fingerprints. Considering that the decorrelation methods that were applied result in features with different statistical properties, the final distribution of

bits in the fingerprint will change depending on the decorrelation.

Previously proposed methods apply one-bit quantization to each of the components of the transformed energy spectrogram [13], [21]. These methods assume that all components have the same statistical properties, however, this is not usually the case for different frequency components of the spectrum and it is more noticeable after applying the proposed time-frequency decorrelation.

To take into account the varying statistics of components, we propose two methods of entropy-based quantization to perform a more efficient distribution of the available bits of the fingerprint on the transformed spectrograms. In the first approach, we estimate the entropy of each component using their individual variance values. In the second approach, we estimate the entropy of the signal according to mutual information [25] between matching and non-matching pairs.

Additionally, we observed that the context-based methods, i.e. eigenvalue and Wiener, present a uniform distribution of bits over all the components. For that reason, we decided to apply one-bit quantization to both context-based methods and test the entropy and mutual information based approaches only on the 2D-DCT transformed spectrograms.

ENTROPY-BASED QUANTIZATION

Informal observations of the 2D-DCT components showed that the distribution of their values resembled a Gaussian distribution. In order to maintain the simplicity of the model and generalize to possible noise affecting the measurements, we assume that each of the components of the 2D-DCT spectrogram follows a normal distribution. Therefore, the number of bits assigned to each component is proportional to its entropy, defined as $\frac{1}{2}\log_2\left(\sigma_x^2\right)$, where σ_x^2 is the variance of each of the 2D-DCT components. We define

$$\text{bits}_i = \frac{1}{2}\max\left(\log_2\left(\sigma_i^2\right) + \mathbf{C}, 0\right) \tag{6}$$

such that the total number of bits is

$$\text{bits}_{total} = \frac{1}{2}\sum_i \max\left(\log_2\left(\sigma_i^2\right) + \mathbf{C}, 0\right). \tag{7}$$

The expression for the number of bits assigned to each component is represented in Eq. (6). The term $\mathbf{C}$ stands for a bias element to control the total number of bits. $\mathbf{C}$ is then determined by selecting the target number of bits of the fingerprint and solving Eq. (7).

Fig. 6 shows the estimated bit distribution for the components based on the individual entropy of each feature. As we can observe, the majority of the bits are concentrated in the lowest frequencies, which have the highest variances. However, we remove the lowest component corresponding to the overall energy of the signal, as it is more dependent on microphone configuration than the desired source signal.

The bit distribution estimated using mutual information method is illustrated in Figure 7. We can observe a similar behavior as in the entropy-based method.

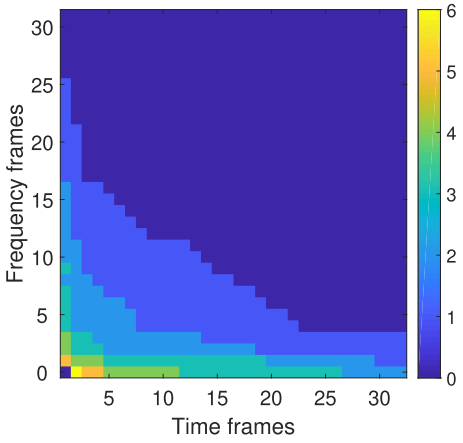

FIGURE 6. Bit distribution of 2D-DCT components quantized using individual entropy method.

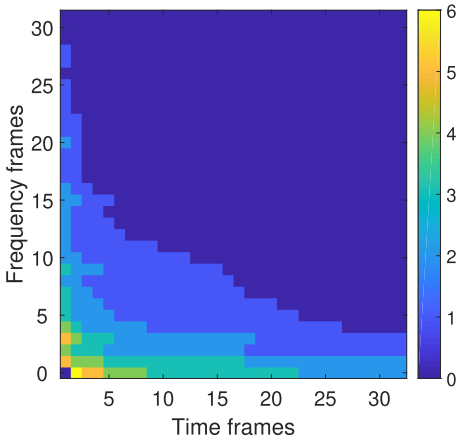

FIGURE 7. Bit distribution of 2D-DCT components quantized using the mutual information method.

MUTUAL INFORMATION-BASED QUANTIZATION

The goal of the generated fingerprints is to remain as similar as possible in matching cases and differ from each other when the audio samples do not match. Therefore, we can include this criterion into the choice of bit distribution, assigning more bits to the components that are more correlated in the matching cases and reducing them for higher correlations in non-matching cases.

In this case, instead of the variance of individual components, we calculate the convariance matrix of a set of matching component pairs. The covariance matrix of a reference set of components with a matching pair can be seen as

$$\Sigma\left(x_R, x_M\right) = \begin{bmatrix} \sigma_R^2 & \sigma_R\sigma_M \\ \sigma_M\sigma_R & \sigma_M^2 \end{bmatrix}, \tag{8}$$

where x_R stands for the reference data and x_M represents the components of the corresponding matching segments of audio. The corresponding non-matching covariance has zeros on the off-diagonal.

By assuming that the components follow a Gaussian distribution, the mutual information data will follow a bivariate normal distribution. Its entropy can be defined as $\frac{1}{2}\log_2(\det(\Sigma(x_R, x_M)))$, the bit assignment can be thus defined as

$$\text{bits}_i = \frac{1}{2}\max\left(\log_2(\det(\Sigma(x_R, x_M))) + \mathbf{C}, 0\right) \quad (9)$$

and correspondingly,

$$\text{bits}_{total} = \frac{1}{2}\sum_i \max\left(\log_2(\det(\Sigma(x_R, x_M))) + \mathbf{C}, 0\right). \quad (10)$$

VI. EXPERIMENTAL SETUP

As we explained in Section I, the goal of this project is to define robust audio fingerprint algorithms that can be used for the authentication of devices in the same acoustic space. To evaluate the quality of these fingerprints, we will analyze their robustness and statistical properties. A robust fingerprint will maintain a high similarity with its matching scenarios even when it is affected by noise and inaccurate synchronization between the devices. For the statistical analysis, we assume that the fingerprints will be used as described in [13], using fuzzy cryptography to transform the fingerprints into cryptographic keys. Therefore, we will analyze the performance of the generated acoustic fingerprints to be used as cryptographic keys in the communication.

DATA GENERATION

The methods described in Sections IV and V are trained using a database that aims to resemble multi-channel recordings of noisy speech. This database was generated combining clean speech samples from the TIMIT database [26] and noise samples from the QUT corpus [27]. The QUT database contains noise samples in different locations such as a café and different areas of a house, which resemble realistic noises that would be present in a VUI scenario.

Figure 8 shows the simulated scenario of the multi-channel recording. The clean speech samples simulate a dialogue between two people located on different sides of the sensor array. Three microphones are located in the scene as depicted. Two of the microphones are placed near the speakers and the third one is placed in a position farther away from the conversation. Environment noise is then simulated using multiple noise sources randomly located around the whole scene to resemble diffuse noise.

To add the effect of the acoustics of each room to the TIMIT recordings, we used the Pyroomacoustics library [28]. To resemble the locations where the different environmental noises are located, we used the parameters described in Table 1.

We generated a training dataset using the training section of the TIMIT database, resulting in 12 hours of noisy speech equally distributed for SNR values in the range of $[-12, 12]$ dB in intervals of 3 dB. Similarly, the test section of the TIMIT database was used to generate a noisy speech test set that contained 3 hours of noisy speech with SNR values between -6 and 6 dB. The audio files in both training and test sets were around 10 seconds long with a sampling rate of 16 kHz, containing two different speech segments simulated from each of the described sound sources.

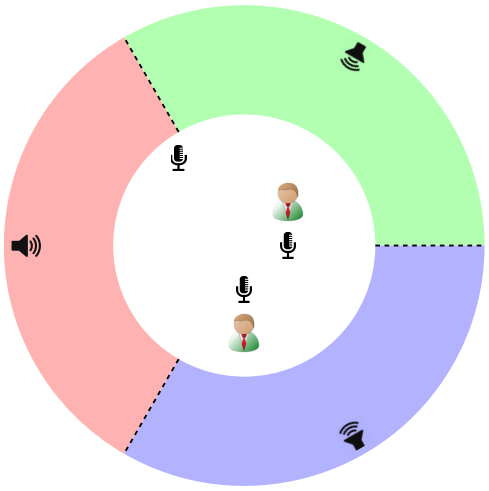

FIGURE 8. Distribution of the acoustic sensors, speech and noise sources in the generation of the noisy database.

TABLE 1. Different room scenarios and acoustic properties.

Room Type	Dimensions (m)	Avg. Absorption Coefficient
Café	$15 \times 10 \times 3$	0.3
Kitchen	$4 \times 4 \times 3$	0.2
Living Room	$5 \times 5 \times 3$	0.3
Car	$3 \times 2 \times 2$	0.3
Street	$30 \times 30 \times 30$	0.99
Pool (High reverb)	$15 \times 10 \times 3$	0.1

The training and test sets are built using different speech samples. However, while the noise samples are extracted from different segments of the noise database, both training and test sets contain the same scenarios. Considering that the proposed methods require to be trained, it is possible that they show a different performance if they evaluate scenarios that have not been observed during the training process. However, the goal of this paper is to provide an overview of the proposed methods in scenarios that could contain a VUI, such that we can focus our efforts on the ones that show better results. Every method can then be further optimized to adapt to different types of scenarios.

A. FINGERPRINT PARAMETERS

The goal of this project is to generate robust fingerprints while reducing the length of the required speech signal. For this reason, we will aim to keep a constant length in the audio recording and generate the same number of bits from each method. The lengths of the required audio recording and the generated fingerprint depend on the parameters that we choose for each proposed method. In order to allow a

TABLE 2. Defined parameters for each of the tested methods to generate 512 bit fingerprints.

Method	Audio length (s)	Frame length (s)	STFT step (s)	# of windows	Ener. bands (Freq)	Avg. time frames
Δ time-frequency [13]	6.375	0.375	0.375	17	33	No
DCT + Δ time [21]	2.45	0.21	0.14	17	32	No
Context based methods	2.17	0.03	0.02	108	34	6
2D-DCT based methods	2.57	0.03	0.02	128	32	4

proper comparison between the different methods, we need to carefully choose the different parameters of the fingerprint generation.

Table 2 shows the configuration of the different fingerprint extraction methods. They have been designed such that the length of the audio signal is as similar as possible, except for the Δ-time-frequency method, which is defined as it was presented [13]. As it is explained in Section III, the methods that we propose in this paper increase the length of their time frames by averaging a few of them after the energy band grouping. The context methods group 6 time frames to obtain 18, which will result in a 32 × 16 matrix after the transformation. The 2D-DCT based methods are converted into a 32 × 32 matrix, from which many components will be discarded in the quantization, as shown in Fig 6.

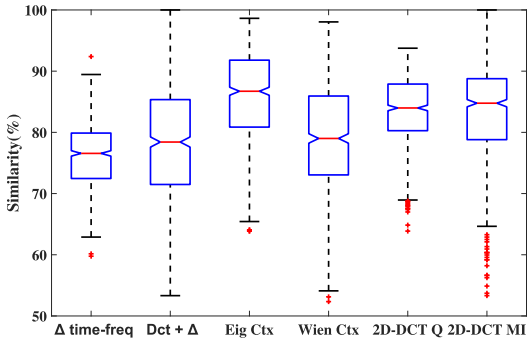

FIGURE 9. Boxplot of the similarity distribution for fingerprint pairs in matching scenarios, expressed as a percentage of the total number of bits of the fingerprint. 1) Δ-time-frequency (Δ time-freq), 2) frequency dct + Δ (DCT +Δ), 3) the first eigenvalue over the context (Eig Ctx) 4) Wiener filter over the context (Wie Ctx), 5) 2D-DCT with entropy based quantization (2D-DCT Q), 6) 2D-DCT with mutual information based quantization (2D-DCT MI).

VII. RESULTS

The acoustic fingerprints were designed such that they contained the information of the conversation and the acoustic environment. If two devices are located in the same acoustic space, the signals that they record will be similar and so will the generated fingerprints. Therefore, to authenticate two devices, pairs of fingerprints are compared and, depending on their similarity, we can determine whether they are in the same space or not.

To analyze the quality of the proposed methods, we evaluate three different aspects of the generated fingerprints that are important in an authentication process. The proposed fingerprints are compared with the performance of previously proposed methods, which we denote as Δ time-freq [13] and Dct+Δ [21]. The robustness of fingerprints is analyzed based on the similarity between fingerprints in matching scenarios. In order to preserve the privacy of the user, it is important that the fingerprint used for authentication is hard to predict by external devices. Therefore, we analyze the statistical properties of the generated fingerprints. Finally, we analyze the computational complexity of the fingerprint generation methods. If the fingerprints are going to be generated by mobile devices, efficiency is a significant factor for their implementation.

ROBUSTNESS

A good fingerprint should match across devices in the same acoustic space even in noisy conditions, and diverge when in different spaces. Therefore, as a measure of similarity of the fingerprints, we determined the number of matching bits between two fingerprints. Ideally, the similarity should be near 100% for matching scenarios and around 50% when the signals do not match, resembling the comparison between two random sequences.

We consider that two recordings are a matching scenario when they are extracted from two microphones in the same environment recording simultaneously. The recordings will be considered a mismatch if they are recorded in two different environments or they are not synchronized in time, i.e. both recordings contain the same speech sample but one is recorded 5 seconds after the other.

Figure 9 depicts the distribution of the similarity values over matching pairs in the test dataset. While all the methods present a substantial variance in the similarity results, the methods proposed in this article provide an overall improvement of 5% to 10% with respect to previous approaches. The best performing method is the eigenvalue extraction, providing a median similarity of 85%. Concurrently, the eigenvalue method uses the shortest recordings (see Table 2). If we assume that a new fingerprint begins to be calculated right after another, a short recording length means that the authentication system will be more responsive to changes in the acoustic scene.

ENTROPY ANALYSIS

As mentioned before, the authentication between devices needs to be robust enough to ensure that unauthorized devices cannot collaborate in the distributed VUI. Simultaneously, an eavesdropper that managed to predict the corresponding fingerprint of an authentication process would also have access to our private information. This means that the

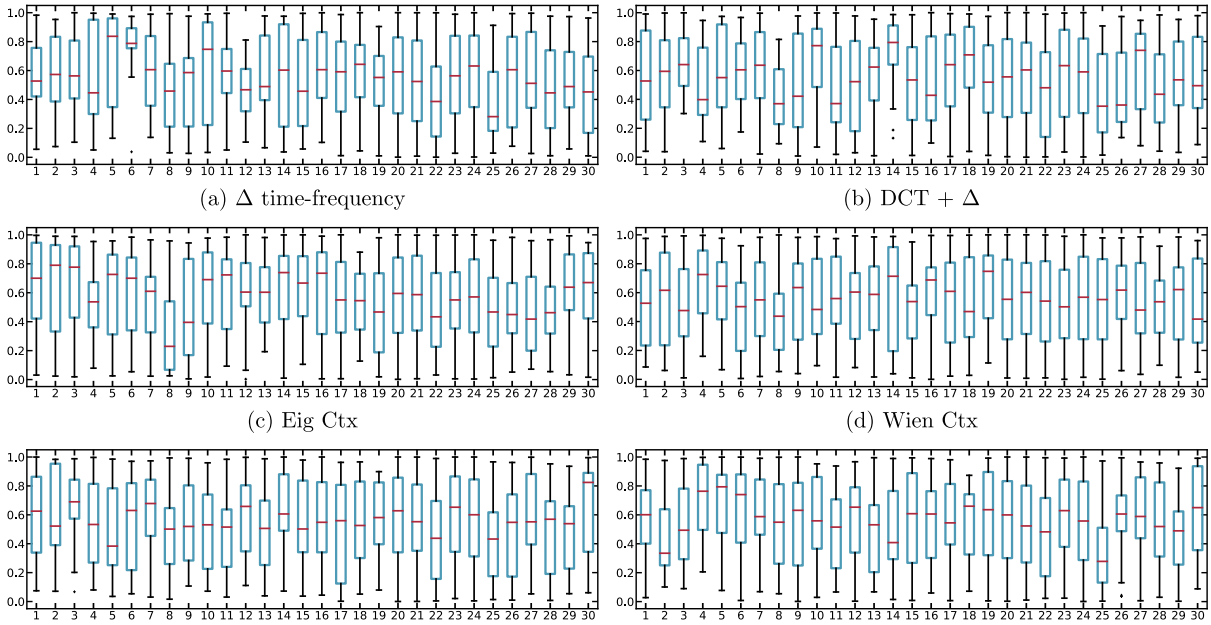

FIGURE 10. Distribution of p-values achieved for keys after 20 runs of the DieHarder set of statistical tests. Tests are: 1) birthdays 2) operm5 3) rank32 × 32 4) rank6 × 8 5) bitstream 6) opso 7) oqso 8) dna 9) count-1s-str 10) count-1s-byt 11) parking 12) 2D circle 13) 3D sphere 14) squeeze 15) runs 16) craps 17) marsaglia 18) sts monobit 19) sts runs 20) sts serial 21) rgb bitdistr. 22) rgb min dist. 23) rgb perm. 24) rgb lagged sum 25) rgb kstest 26) dab bytedistr. 27) dab dct 28) dab filltree 29) dab filltree 2 30) dab monobit 2.

TABLE 3. Results of ENT pseudorandom test for a set of fingerprints generated with the different proposed methods. The measured parameters are 1) entropy, 2) arithmetic mean (A. M.), 3) Monte Carlo value (M. C.), 4) serial correlation coefficient (S. C. C.).

Method	Entropy	A. M.	M. C.	S. C. C.
Worst case	0	0 or 1	–	±1
Best case	1	0.5	π	0
Δ time-freq	0.999995	0.499	3.39	−0.256
DCT + Δ	0.999917	0.494	3.35	−0.150
Eig Ctx	0.999918	0.4947	3.06	0.079
Wien Ctx	0.9988	0.4797	3.05	0.130
2D-DCT Q	0.997966	0.503	3.12	−0.020
2D-DCT MI	0.999991	0.501	3.05	−0.011

fingerprints also need to be secure from a cryptographic perspective.

To achieve this, we aim to generate fingerprints that resemble random sequences of bits without noticeable patterns. The generated fingerprints were evaluated using the ENT pseudo-random test [29] and the DieHarder set of statistical tests [30]. The ENT pseudo-random test analyzes a sequence of symbols and returns the information density of the sequence in terms of its entropy, arithmetic mean and serial correlation. DieHarder performs multiple tests to evaluate whether the proposed methods are robust against bias in the generated random sequences.

Table 3 shows the results of the ENT pseudo-random test for each of the algorithms, as well as the best and worst possible values of each parameter. The fingerprints were generated from the same audio files, randomly sampled from the test database described in Section VI. We observe that the results provided by the proposed methods barely deviate from the optimal values.

Each test in the DieHarder set computes a p-value between 0 and 1, which we can interpret as the probability that a random number generator would produce the analyzed sequence. After enough runs of the DieHarder tests, the computed p-values should be uniformly distributed between 0 and 1. Conversely, if the results computed by one of the tests would be concentrated towards an individual value, then the sequence would have failed the test.

Figure 10 represents the distribution of p-values for each proposed method over 20 runs of the DieHarder test. We can observe that the computed p-values are mostly uniformly distributed and centered around 0.5. Note that some of the methods present a tendency towards a specific p-value, e.g. the opso test in Δ time-freq or the dna test in Eig Ctx. Such outliers indicate a weakness towards specific statistical patterns. An attacker with knowledge about such bias would be advantaged to predict the fingerprint used in the authentication process. In our case, where we assume continuous generation of fingerprints, and thereby frequently renew the key for the authentication process, an attacker would still be challenged to produce a series of consecutive matching keys in order to be able to eavesdrop on the network. For practical application, we therefore conclude that the proposed fingerprint generation methods are suitable for a secure authentication process.

COMPLEXITY

The proposed fingerprint generation and authentication will be an ad-hoc process between the involved devices and all the processing will be done within each device. We need to consider that mobile devices have limited computing power and periodically generating an audio fingerprint using complex methods would affect their performance and limit their battery life. For this reason, the proposed methods were designed to perform in mobile devices with low computational complexity. The computational load of the proposed methods are primarily due to the sub-process of 1) windowing of the audio signal and STFT, 2) energy band grouping and time averaging, 3) the decorrelation transform and 4) quantization of the fingerprint values.

The complexity of the STFT process is proportional to the number of time frames required and their length. This means that, if the number of windows is too great, the system will need to perform a great number of FFTs, and the number of samples of the transformation has to be chosen according to the length of these windows. Table 2 shows the number and length of the extracted windows for a single fingerprint.

While the FFT is an efficient algorithm with a complexity of $\mathcal{O}(n \log n)$ [31], [32], the calculations can become a burden if the size of the transformation is too large or it has to be repeated very frequently. For instance, using 0.375 second windows and a sampling rate of 16 kHz, each time frame will contain 6000 samples. The authentication system will then need to continuously calculate FFTs with a minimum of 6000 samples.

The proposed methods aim to adapt the pre-processing of the signal to typical speech processing applications such as speech codecs or voice assistants, which mainly analyze the speech signal in the frequency domain. This means that, by adapting the parameters of the fingerprint to the application that is going to use it, it is possible to share the windowing and STFT steps, thus reducing the overall complexity.

The energy grouping and time averaging of the frequency bands comprises a series of multiplications and additions proportional to the number of FFT samples and time frames. However, these operations are simple enough not to pose a significant difference in the complexity between the different configurations of the proposed algorithms.

The main difference in complexity for the proposed methods with respect to the previous ones can be observed in the fingerprint extraction. In this case, we can divide the fingerprint extraction methods into three categories: 1) Deltas over time and frequency [13], 2) Context-based transformations and 3) 2D-DCT based fingerprints. For this comparison we will use the numbers defined in Table 2, which correspond to the methods evaluated in this section.

Table 4 shows an estimation of the number of operations required by each method. We can observe that the delta over time and frequency is the most efficient method as it requires roughly two operations for every component of the energy spectrum. The context-based transformations, as it is explained in Section V, can be reduced to a series of array multiplications using a previously trained transformation array. The resulting number of operations for a 3×3 context is then 9 times the total number of components. Finally, we observe that the 2D-DCT is the most complex method. This method requires a DCT transformation for every frequency and time band. Similarly to the FFT, a DCT transformation can be performed with $\mathcal{O}(n \log n)$ complexity and, in order to compare it to the previous methods, the number of operations required for an n-sample DCT is considered $n \log n$.

TABLE 4. Number of operations required for the proposed fingerprint generation methods.

Method	Data dimensions	Operations
Δ Time-Frequency	33×17	1056
Context-based transform	34×18	4608
2D-DCT	32×32	10240

While the 2D-DCT looks like the most complex of the proposed methods, the components of this transformation will be quantized. Figs. 8 and 7 show that less than half of the 2D-DCT components are necessary for the final calculation of the fingerprint, as most of them are represented using 0 bits. This means that the complexity of this method can be reduced by calculating only the components that will be quantized and ignoring those that do not have any bit assigned to them [33], [34].

The values estimated in Table 4 represent the total number of operations required for the calculation of one fingerprint. These values can then be normalized by the duration of the respective audio recording to obtain the number of operations per second that a device would need to perform. Following the data from Table 2, the normalized number of operations per second for each of the methods is 1) 166 for the reference Δ time-frequency method, 2) 2194 for the context-based methods and 3) 3984 for the 2D-DCT transformations. Overall, the complexity of the proposed methods is higher than existing ones. However, the amount of operations required to estimate each fingerprint does not entail a significant burden even for current low-resource devices.

VIII. CONCLUSION

Service quality can be improved by using multiple distributed devices instead of a single device, collecting and processing information. However, it is important to define access management rules to ensure that our private information is not shared with unauthorized devices. In speech systems, the access to a distributed network can be defined based on the reach of one's voice. In a discussion between people, those who can hear the discussion are implicitly included in it. Similarly, if our electronic devices could hear the same speech signal, that would mean that they are located in a trusted area and they would be allowed to collaborate.

In this paper we have presented methods to generate audio fingerprints that would be suitable for authentication of devices in a distributed acoustic sensor network. The presented methods outperform previous solutions, increasing the robustness and responsiveness of the acoustic fingerprint generation. Additionally, the generated fingerprints meet the basic entropy requirements to be used a cryptographic keys. We can observe that the best performing methods are the eigenvalue extraction over a spectral context and the 2D-DCT with mutual information based quantization. While the former presents a higher robustness, the latter has better statistical properties.

The proposed methods are intended to be used in mobile devices with low computational capabilities and low power consumption, for example, with a limited battery life. Therefore, the fingerprint generation algorithms need to remain efficient to be periodically performed in this type of device. We observe that the proposed methods do not entail a big computational load, even on mobile devices, as the maximum frequency of the fingerprint generation process is limited by the length of the recorded audio, i.e., to 2.5 seconds. We have thus demonstrated that the proposed fingerprint methods are robust, low in complexity and provide high-entropy keys for cryptographic applications. The proposed methods can thus be widely applied in privacy-aware applications for distributed sensor networks for speech and audio.

REFERENCES

[1] M. Cobos, F. Antonacci, A. Mouchtaris, and B. Lee, "Wireless acoustic sensor networks and applications," *Wireless Commun. Mobile Comput.*, vol. 2017, Dec. 2017, Art. no. 1085290.

[2] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *Proc. 18th IEEE Symp. Commun. Veh. Technol. Benelux (SCVT)*, Nov. 2011, pp. 1–6.

[3] D. Cherkassky and S. Gannot, "Blind synchronization in wireless acoustic sensor networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 3, pp. 651–661, Mar. 2017.

[4] S. Markovich-Golan, A. Bertrand, M. Moonen, and S. Gannot, "Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks," *Signal Process.*, vol. 107, pp. 4–20, Feb. 2015.

[5] S. Markovich-Golan, S. Gannot, and I. Cohen, "Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 2, pp. 343–356, Feb. 2013.

[6] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. New York, NY, USA: Springer, 2007.

[7] T. Bäckström, *Speech Coding with Code-Excited Linear Prediction*. Cham, Switzerland: Springer, 2017. [Online]. Available: http://www.springer.com/gp/book/9783319502021

[8] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop Autom. Speech Recognit. Understand.*, 2011, pp. 2–42.

[9] M. B. Hoy, "Alexa, siri, cortana, and more: An introduction to voice assistants," *Med. Reference Services Quart.*, vol. 37, no. 1, pp. 81–88, Jan. 2018.

[10] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320310004619

[11] K. Levy and B. Schneier, "Privacy threats in intimate relationships," *J. Cybersecur.*, vol. 6, no. 1, pp. 1–13, Jan. 2020.

[12] P. Perez Zarazaga, S. Das, T. Bäckström, V. V. Raju, and A. Vuppala, "Sound privacy: A conversational speech corpus for quantifying the experience of privacy," in *Interspeech*. Kolkata, India: ISCA, 2019.

[13] D. Schürmann and S. Sigg, "Secure communication based on ambient audio," *IEEE Trans. Mobile Comput.*, vol. 12, no. 2, pp. 358–370, Feb. 2013.

[14] N. Nguyen, C. Y. Kaya, A. BrUsch, D. SchUrmann, S. Sigg, and L. Wolf, "Demo of BANDANA–body area network device-to-device authentication using natural gAit," in *Proc. IEEE Int. Conf. Pervas. Comput. Commun. Workshops (PerCom Workshops)*, Mar. 2018, pp. 190–196.

[15] A. Brusch, N. Nguyen, D. Schurmann, S. Sigg, and L. Wolf, "Security properties of gait for mobile device pairing," *IEEE Trans. Mobile Comput.*, vol. 19, no. 3, pp. 697–710, Mar. 2020.

[16] M. Wang, W.-T. Zhu, S. Yan, and Q. Wang, "SoundAuth: Secure zero-effort two-factor authentication based on audio signals," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, May 2018, pp. 1–9.

[17] A. Kalia, S. Sharma, S. K. Pandey, V. K. Jadoun, and M. Das, "Comparative analysis of speaker recognition system based on voice activity detection technique, mfcc and plp features," in *Intelligent Computing Techniques for Smart Energy Systems*. Singapore: Springer, 2020, pp. 781–787.

[18] B. Agüera y Arcas, B. Gfeller, R. Guo, K. Kilgour, S. Kumar, J. Lyon, J. Odell, M. Ritter, D. Roblek, M. Sharifi, and M. Velimirović, "Now playing: Continuous low-power music recognition," 2017, *arXiv:1711.10958*. [Online]. Available: http://arxiv.org/abs/1711.10958

[19] A. Wang, "An industrial strength audio search algorithm," in *Proc. 4th Int. Conf. Music Inf. Retrieval (ISMIR)*, Washington, DC, USA, 2003, pp. 7–13.

[20] A. Juels and M. Sudan, "A fuzzy vault scheme," *Designs, Codes Cryptogr.*, vol. 38, no. 2, pp. 237–257, 2006.

[21] P. P. Zarazaga, T. Buackstrom, and S. Sigg, "Robust and responsive acoustic pairing of devices using decorrelating time-frequency modelling," in *Proc. 27th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2019, pp. 1–5.

[22] S. Das and T. Bäckström, "Postfiltering with complex spectral correlations for speech and audio coding," in *Proc. Int. Speech Commun. Assoc. (ISCA)*, Grenoble, France, 2018.

[23] G. H. Golub and C. F. Van Loan, *Matrix Computations*, vol. 3. Baltimore, MD, USA: JHU Press, 2012.

[24] G. Hudson, A. Léger, B. Niss, I. Sebestyén, and J. Vaaben, "JPEG-1 standard 25 years: Past, present, and future reasons for a success," *J. Electron. Imag.*, vol. 27, no. 4, pp. 1–19, 2018, doi: 10.1117/1.JEI.27.4.040901.

[25] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 6, Jun. 2004, Art. no. 066138.

[26] J. S. Garofolo, *Timit Acoustic-Phonetic Continuous Speech Corpus ldc93s1. Web Download*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.

[27] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-noise-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Interspeech*. Makuhari, Japan: Makuhari Messe International Convention Complex, Sep. 2010. [Online]. Available: https://eprints.qut.edu.au/38144/

[28] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 351–355.

[29] J. Walker. (2008). *ENT: A Pseudorandom Number Sequence Test Program*. Software and Documentation. [Online]. Available: www.fourmilab.ch/random/S

[30] R. G. Brown, D. Eddelbuettel, and D. Bauer. (2019). *DieHarder: A Random Number Test Suite*. Accessed: Feb. 28, 2019. [Online]. Available: https://webhome.phy.duke.edu/ rgb/General/dieharder.php

[31] D. G. Manolakis and V. K. Ingle, *Computation of the Discrete Fourier Transform*. Cambridge, U.K.: Cambridge Univ. Press, 2011, p. 434–484.

[32] M. Frigo and S. G. Johnson, "The design and implementation of FFTW3," *Proc. IEEE*, vol. 93, no. 2, pp. 216–231, Feb. 2005.

[33] C. A. Christopoulos, J. Bormans, J. Cornelis, and A. N. Skodras, "The vector-radix fast cosine transform: Pruning and complexity analysis," *Signal Process.*, vol. 43, no. 2, pp. 197–205, May 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/016516849400153Q

[34] A. N. Skodras, "Fast discrete cosine transform pruning," *IEEE Trans. Signal Process.*, vol. 42, no. 7, pp. 1833–1837, Jul. 1994.

PABLO PÉREZ ZARAZAGA (Graduate Student Member, IEEE) was born in Zaragoza, Spain, in July 1993. He received the bachelor's degree in telecommunications engineering from Zaragoza University, Spain, in 2015, and the master's degree in computer, communication, and information sciences with a major in acoustics and audio technologies from Aalto University, Finland, in 2017, where he is currently pursuing the Ph.D. degree with the Department of Signal Processing and Acoustics. His main research interests include speech signal processing as well as privacy in speech interfaces.

TOM BÄCKSTRÖM (Senior Member, IEEE) received the master's and Ph.D. degrees from Aalto University, in 2001 and 2004, respectively, which was then known as the Helsinki University of Technology. He has been an Associate Professor with the Department of Signal Processing and Acoustics, Aalto University, Finland, since 2016. He was a Professor at the International Audio Laboratory Erlangen, Friedrich-Alexander University, from 2013 to 2016, and a Researcher at Fraunhofer IIS from 2008 to 2013. He has contributed to several international speech and audio coding standards and is the chair and co-founder of the ISCA Special Interest Group "Security and Privacy in Speech Communication." His research interests include technologies for spoken interaction, emphasizing efficiency and privacy, and in particular in multi-device and multi-user environments.

STEPHAN SIGG (Member, IEEE) received the M.Sc. degree in computer science from TU Dortmund in 2004 and the Ph.D. degree from Kassel University in 2008. Since 2015, he has been an Assistant Professor with the Department of Communications and Networking, Aalto University, Finland. His research interests include the design, analysis, and optimization of algorithms for distributed and ubiquitous systems, proactive computing, distributed adaptive beamforming, context-based secure key generation, and device-free passive activity recognition. He has served as a TPC member of leading conferences including the IEEE PerCom, ACM Ubicomp, and the IEEE ICDCS. He is an Editorial Board Member of the *Proceedings of the ACM on Interactive, Mobile, Wearable, and Ubiquitous Technologies* (IMWUT).