Preserving Speech Privacy in Interactions with Ad Hoc Sensor Networks

Pablo Pérez Zarazaga
Preserving Speech Privacy in Interactions with Ad Hoc Sensor Networks

Pablo Pérez Zarazaga

A doctoral thesis completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall F239a of the school on 28 October 2022 at 13:00.

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics
Speech Interaction Technology
Supervising professor
Professor Tom Bäckström, Aalto University, Finland

Preliminary examiners
Professor Nicholas Evans, EURECOM, France
Associate Professor Adriana Stan, Technical University of Cluj-Napoca, Romania

Opponent
Professor Dr.-Ing. Rainer Martin, Institute of Communication Acoustics Ruhr-Universität Bochum, Germany
Abstract

Speech is our main method of communication that allows us to intuitively communicate complex ideas and provide our messages with deeper meaning than the lexical content of such messages. For example, we can stress specific words to emphasise or subtract significance from different sections of a sentence. Considering this, the increasing popularity of voice user interfaces is only natural and expected to keep growing in the following years, as they allow us to interact with our electronic devices using our speech. Any device with which we can interact using our voice can be considered a voice user interface, and among them we can find a great variety of services, from telecommunication applications like Zoom or Skype, to virtual assistants like Alexa or Siri.

However, in order to provide better services and more natural interactions, voice user interfaces require the gathering of a great amount of our speech data and transmitting it usually without us being aware of it. If that data is misused or an unauthorised user manages to obtain it, it would cause a grave violation of the user's privacy.

In an environment where multiple electronic devices can provide a voice user interface, collaboration between them as a wireless acoustic sensor network can improve the services that they provide individually. It is important then to study those applications that require sending our voice to a remote party in order to provide their services, and more specifically, in a scenario where multiple devices can pick up the voice of multiple users, it is crucial to define which of these devices are actually allowed to record the user’s speech. For example, if a user's voice leaks into another user's interaction, and is therefore transmitted to a destination that they have not specifically authorised, the privacy of the users is violated. As a solution, if our devices could perceive our privacy the same way as we do, they could adapt the information they shared to protect the personal data of the users. For that reason, we need to analyse how users perceive privacy in their spoken interactions, based on which we can devise rules that our devices can follow when they provide a voice user interface.

In this thesis we study methods to recognise when two devices are located in the same acoustic space based on the audio signals that they record. We show how acoustic fingerprints can be used to securely share the audio information from a device and estimate the physical proximity of devices. We also generated a speech corpus in conversational scenarios to analyse the effect that the acoustic properties of the environment have on the level of privacy that we perceive. Finally, we developed source separation methods to remove the voice of interfering speakers in a multi-device scenario, thus protecting the privacy of external users.

Keywords Voice User Interface, Experience of Privacy, Audio Fingerprint, Acoustic Sensor Networks
Preface

First of all, I want to thank my supervisor, Tom, for the continuous support and motivation before, and during this thesis. When we first met in 2015 in the Speech Processing course that Tom was teaching, he convinced me to go to Germany to work on my master’s thesis. Working with him encouraged me to continue doing research. And when I told Tom that I was interested in continuing this path, he accepted me in his group and has always been providing helpful insights into my work throughout my thesis. But Tom’s support also extends further than the office. He cares about the well-being of the team and has always been available to listen, and understanding, whenever I brought personal matters to him. For these reasons, I see him not only as a supervisor, but also as a friend, and I hope our paths keep crossing in the future.

I also had the chance to share my time in the office with some amazing people that deserve a mention here. To Sneha, sharing an office for several years we have seen each other go through all kinds of situations. I will always cherish our discussions, about our research and all other kinds of topics, that sometimes distracted us a bit too much, but from which we always got very interesting insights. To Mohammad, who has also seen me through the last year of this thesis, and was always there to hear me ramble about ideas and give his valuable opinion. And Abraham, who joined the team as a postdoc in the middle of this four year period and has given me great advice about how to complete this stage of my life. Unfortunately, having to work remotely, I could only interact with the rest of the team on Zoom but they also deserve a mention here. Thank you, Mariem, for your collaboration and enthusiasm in the work that we shared, I hope that we get to meet in person soon. And last but not least, Silas and Esteban, even though we did not have the chance to interact much apart from our Zoom meetings and an occasional evening out, I really enjoyed our brief discussions before the group’s daily meetings and hope that we can work together again in the near future.

A special mention goes to Ioanna, my partner and soulmate. We met when I was a few months into this thesis, and since then, we have shared
unforgettable experiences, moved in together and finally got engaged. I want to thank her for being by my side through every step of the way, not only as a partner, but also as a researcher that knows and understands what we can go through. For supporting me on every idea and new endeavour that came in front of me. We know we have been through easier and harder times, but she has managed to make these years the best of my life. Thank you for walking by my side, I love you.

Finally, I want to thank some of the people that accompanied me through my life in Finland, even before I started this path. Leo and Janani, we met on our first day studying our master’s in acoustics, we have shared interests, long hours of study, and I got the honour of being a witness at your wedding. Our busy schedules do not allow us to meet very often anymore, but I will always be happy to have you as my friends. And the final mention before I switch to Spanish goes to Jose, who I met at the beginning of our master’s as “the other Spanish student in the programme”, and has grown to be one of my best friends in this country. Thank you.

Now, with the permission of the reader, I will continue writing this section in Spanish.

En esta parte quiero agradecer a mis padres, Joaquín y Mary Carmen, a quienes quiero con todo mi corazón, todo su apoyo a lo largo de mi vida. Ellos me enseñaron el valor de tener claras mis metas y luchar por ellas. Ellos me animaron a perseguir mis sueños fuera de España y han continuado animándome en cada camino que he elegido seguir. Vivir lejos de tu familia es muy duro, especialmente en un camino tan exigente como este, donde tenemos que lidiar continuamente con la presión y el estrés. No es nada fácil tener que perderse eventos familiares porque nos encontramos a cientos de kilómetros y el trabajo apremia. Pero ellos siempre han estado allí, haciéndome saber el orgullo que sienten por mí, y eso me da fuerzas para continuar. Muchas gracias.

Por último, he querido guardar para el final el agradecimiento que mas me cuesta escribir. A mis abuelos, que siempre creyeron en mis capacidades y me empujaron a apuntar alto en la vida. Mis abuelos, Baltasar y Ramón, fallecieron cuando yo todavía era joven, pero siempre estarán en mi corazón dándome fuerzas para continuar. A principio de este año también nos dejaba mi abuela Pilar, tan rápido que ni siquiera tuve oportunidad de pasar ese tiempo con mi familia. Sin embargo, siempre recordare todas esas tardes de viernes, cuando ella me esperaba al salir del instituto con la comida preparada en su casa. Y por último, mi abuela Demetria, que venía todas las mañanas antes de comenzar el día de colegio a repasar la lección de ese día. Ella llegó a verme alcanzar el final de este camino, y solo nos dejó recientemente. Pero nunca olvidaré como atendía ávidamente y con emoción cuando le explicaba en qué iba a consistir mi trabajo después de esta tesis. Escribo estas últimas líneas entre lágrimas, pero antes
de terminar, os quiero dedicar esta tesis a vosotros, Ramón, Demetria, Baltasar y Pilar. Solo espero poder estar a la altura de lo que vosotros veíais en mí, y siempre os llevaré en el corazón.

Helsinki, September 27, 2022,

Pablo Pérez Zarazaga
Contents

Preface 1
Contents 5
List of Publications 7
Author's Contribution 9
List of Figures 11
Abbreviations 15
Symbols 17

1. Introduction 19
   1.1 Thesis Structure .................................. 21

2. Perception of Privacy in Speech Interactions 25
   2.1 Privacy in speech interfaces ...................... 27
   2.2 Perception of privacy in speech interactions ...... 28

3. Device Authentication 29
   3.1 Wireless Acoustic Sensor Networks ................. 29
   3.2 Distributed Voice User Interfaces .................. 31
   3.3 Device Authentication ............................. 32
       3.3.1 Acoustic fingerprint .......................... 33
       3.3.2 Decorrelation methods ......................... 34
       3.3.3 Fuzzy commitment scheme ...................... 36

4. Applications of a Distributed VUI 39
   4.1 Cross-talk cancellation for privacy protection .... 40
   4.2 Single-channel source separation .................... 41
       4.2.1 Time-frequency masking ......................... 41
   4.3 Multi-channel source separation ..................... 43
       4.3.1 Beamforming ................................. 43
4.3.2 Minimum Variance Distortionless Response . . . . 43
4.3.3 Independent Component Analysis ................. 45
4.4 Deep learning for source separation .................. 46
4.5 Distributed Speech Coding ........................... 47
  4.5.1 Traditional speech coding ..................... 48
  4.5.2 Transition to distributed coding .............. 49
  4.5.3 Decoder side ............................... 50

5. Summary of Publications ............................... 53
  5.1 Publication I: Acoustic Fingerprints for Access Management in Ad-Hoc Sensor Networks. .................. 54
  5.2 Publication II: Robust and Responsive Acoustic Pairing of Devices Using Decorrelating Time-Frequency Modelling. 55
  5.3 Publication III: Sound Privacy: A conversational Speech Corpus for Quantifying the Experience of Privacy. . 56
  5.4 Publication IV: Cancellation of Local Competing Speaker with Near-field Localisation for Distributed Ad-Hoc Sensor Network. ........................... 58
  5.5 Publication V: Speech Localization at Low Bitrates in Wireless Acoustics Sensor Networks ........................ 59
  5.6 Publication VI: PyAWNes-Codec: Speech and Audio Codec for Wireless Acoustic Sensor Networks. .............. 60
  5.7 Publication VII: Perception of Privacy Measured in the Crowd - Paired Comparison on the Effect of Background Noises. ............................................. 61
  5.8 Publication VIII: Provable Consent for Voice User Interfaces 62

6. Conclusion .............................................. 65

References ............................................. 69

Publications ........................................... 77
List of Publications

This thesis consists of an overview of the following publications which are referred to in the text by their Roman numerals.


Author’s Contribution


The author designed and implemented the four presented methods, generated the datasets that were used in the different experiments and carried out the experiments analysing the robustness of the generated fingerprints. The author was the main writer of the article.

Publication II: “Robust and Responsive Acoustic Pairing of Devices Using Decorrelating Time-Frequency Modelling”

The author designed and implemented the presented DCT-based method, also carried out the experiments analysing the robustness of the generated fingerprints. The author was the main writer of the article.

Publication III: “Sound Privacy: A conversational Speech Corpus for Quantifying the Experience of Privacy”

The author planned and designed the recording sessions, processed the audio data for the construction of the database and collaborated in the design of the questionnaire and analysis of the received answers. The author wrote sections 1, 2, 5 and 6 of the article, dividing the work with the second author in a ratio of 50%.
Publication IV: “Cancellation of Local Competing Speaker with Near-field Localization for Distributed Ad-Hoc Sensor Network”

The author programmed the training and testing of the proposed methods, designed the dataset used in the model training and testing and analysed the obtained results. The author was the main writer of the article.

Publication V: “Speech Localization at Low Bitrates in Wireless Acoustics Sensor Networks”

The author worked in the conceptualisation and implementation of the presented methods, as well as the generation and pre-processing of the training and test data.

Publication VI: “PyAWNes-Codec: Speech and Audio Codec for Wireless Acoustic Sensor Networks”

The author provided the spatialisation function to simulate sources in different locations for the training and testing of the proposed methods.

Publication VII: “Perception of Privacy Measured in the Crowd - Paired Comparison on the Effect of Background Noises”

The author provided the audio dataset and prepared the real-life recording data used in the study.

Publication VIII: “Provable Consent for Voice User Interfaces”

The author prepared the recording scenario, organised the recording sessions and pre-processed the recorded data, cleaning and cutting the audio signals.
List of Figures

1.1 Architecture of a distributed voice user interface, Multiple devices capture the voice of different speakers, which is processed by a speech front-end and sent to the corresponding service. .................................................. 20

2.1 Scenario where the same interaction is carried out between humans (a) and with a voice user interface (b). ................. 27

3.1 Comparison between a scenario containing a microphone array and a WASN. While in (a) a single 3-microphone array points at the user or sound source, (b) shows the users surrounded by simple sensors forming a WASN. . . . 30

3.2 (a) Structure of an ad hoc WASN, where devices communicate with each other directly without a managing device; and (b) a centralised WASN, where all the devices communicate with a central node. ............................... 31

3.3 Ideal behaviour of an access management system for distributed VUIs. Devices within the reach of Bob’s voice will be allowed to collaborate, while those outside will be rejected. 32

3.4 Basic workflow of acoustic fingerprint used for authentication. The devices record the audio signal, calculate a binary fingerprint and compare it to allow collaboration. ........... 34

3.5 Schematic of the Error Correcting Code process. Data in a space \( \mathbb{Z}^p \) is encoded into a higher dimensionality space \( \mathbb{Z}^q \), and returned to the original space in decoding. ........ 36

4.1 Multi-speaker scenario. The signal from each speaker to their device (blue line) is, simultaneously, an interference to other devices (red lines). ................................. 40
4.2 Multi-speaker mixture followed by a blind source separation (BSS) system. The convolution between signals $S_n$ and the RTF $H_{m,n}$ is gathered by the $m$-th microphone. The BSS then estimates the separated signals $\hat{S}_n$. ........ 42

4.3 Schematic of microphone array beamforming in far-field conditions. The blue line represents the wavefront of a plane wave arriving with a DOA of $\theta$. The wave reaches each sensor with a delay of $\Delta_t$. ................. 44

4.4 DNN based blind source separation schematics. (a) Time-frequency mask for a source $i$ is estimated from the mixture and (b) The separated sources are estimated directly as output of the network. ......................... 46

4.5 Structure of a frequency domain codec. Reproduced under Creative Commons Attribution-ShareAlike 4.0 International License from: https://wiki.aalto.fi/display/ITSP/Frequency-domain+coding .................. 48

4.6 Structure of a distributed speech codec. Independent encoders send the recorded signal through the transmission channel to a single decoder that merges the received signals into the target output. ......................... 50

5.1 Boxplot of the similarity distribution for fingerprint pairs in matching scenarios, expressed as a percentage of the total number of bits of the fingerprint. ................. 54

5.2 Similarity of fingerprints between distinct microphones with varying delay in an office environment. Original, proposed fingerprinting methods with 200 ms and 50 ms overlapping windows, as well as comparison to microphones that do not match for original and proposed. ........ 55

5.3 Results from the onsite questionnaire conducted in Finland and India to questions Q1, Q2 and Q3. ................. 57

5.4 Architecture of the soft-mask calculation for cross-talk cancellation. ........................................ 58

5.5 System structures for DOA estimation in a distributed scenario. ........................................ 59

5.6 Single channel quality across bitrates (8, 9.6, 13.2, 16.4, 24.4 and 32 kbps); dashed red and dotted blue lines express the median of EVS and proposed codecs, respectively, and the corresponding filled areas their 95% quantiles. ........ 60

5.7 Frequency responses of each of the analysed environments. 61
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.8</td>
<td>Schematic of the key exchange in an audio-based authentication process. Devices 1 and 2 are both within the area of trust, and their generated keys match, while device 3 is outside of the trusted area and its key will differ from the other devices.</td>
<td>62</td>
</tr>
</tbody>
</table>
Abbreviations

AEC  Acoustic Echo Cancellation
ASR  Automatic Speech Recognition
BSS  Blind Source Separation
CELP Code-excited linear prediction
CPS  Cross-Power Spectrum
CTC  Cross-Talk Cancellation
DNN  Deep Neural Network
DoA  Direction of Arrival
DCT  Discrete Cosine Transform
DFT  Discrete Fourier Transform
ECC  Error Correcting Code
GAN  Generative Adversarial Network
GDPR General Data Protection Regulation
ILD  Inter-device Level Difference
IoT  Internet of Things
ITD  Inter-device Time Difference
LPC  Linear Prediction Coefficients
MVDR Minimum Variance Distortionless Response
NLP  Natural Language Processing
RNN  Recursive Neural Network
Abbreviations

**RTF** Relative Transfer Function

**STFT** Short-Time Fourier Transform

**TCX** Transform Coded Excitation

**TDoA** Time Delay of Arrival

**VAE** Variational Auto Encoder

**VUI** Voice User Interface

**WASN** Wireless Acoustic Sensor Network
Symbols

List of Latin symbols

\( a_n \)  Steering vector of a microphone array with respect to source \( n \).

\( A \)  Matrix containing the steering vectors of all the sources in the system.

\( \mathbb{E} \)  Expectation operator.

\( F_0 \)  Fundamental frequency.

\( H_{m,n} \)  Frequency response of the relative transfer function from source \( n \) to sensor \( m \).

\( H \)  Matrix form of the relative transfer functions in a signal mixture.

\( (\cdot)^H \)  Transpose complex-conjugate of a vector or matrix.

\( s_n \)  Clean signal from source \( n \) in time domain.

\( \hat{s}_n \)  Estimated enhanced signal from source \( n \).

\( x_m \)  Signal recorded by sensor \( m \).

List of Greek symbols

\( \alpha \)  Scalar coefficients that define the signal mixture in the ICA process.

\( \Delta_t \)  Time delay between sensors

\( \lambda \)  Wavelength of a specific frequency band in a signal

\( \Lambda \)  Diagonal matrix of eigenvalues

\( \Sigma_{uu} \)  Covariance matrix of the noise signal

\( \theta \)  Direction of arrival of target signal
1. Introduction

With the growing popularity of voice user interfaces (VUI) and the easy access to electronic devices that can provide such services (e.g. smartphones, tablets or laptops), it is likely that several of these devices can be found around us in our everyday lives. Collaboration between these devices could then improve the quality of the services that they provide individually. However, if multiple devices share the user’s voice in order to process it jointly, it is necessary to consider the implications of this collaboration on the user’s privacy. By studying the speech interactions between humans and between humans and their electronic devices, we can develop methods that will protect the privacy of the users in collaborative multi-device scenarios.

The continuous advances of voice technologies are allowing the integration of voice interfaces in more devices everyday. Among the most popular examples of VUIs are virtual assistants, like Alexa and Siri [Hoy, 2018], which record and process the user’s voice, interpret the content of their queries and respond, combining automatic speech recognition (ASR), natural language processing (NLP) and speech synthesis techniques, allowing for a natural interaction with the user [Polyakov et al., 2018]. Nevertheless, any device with an integrated microphone and the ability to process audio can provide a VUI [Cohen et al., 2004], and communication applications like Skype, or even voice activated coffee machines can be considered VUIs.

A problem of VUIs, however, is that they require physical proximity between the user and the device in order to capture sound with good enough quality, binding the user to the specific device that provides such a VUI. Considering that most electronic devices also have the ability to transmit data, they could all collaborate, sharing their own recorded signals, providing a joint distributed VUI, as depicted in Fig. 1.1. This would not only detach the user from a specific device, but could also improve the quality of services with respect to what each device could provide independently by processing the information of multiple recorded signals. Collaboration between devices could also reduce the computational load on one single device, sharing it among multiple devices, hence allowing for an optimisa-
Figure 1.1. Architecture of a distributed voice user interface, Multiple devices capture the voice of different speakers, which is processed by a speech front-end and sent to the corresponding service.

Designing user interfaces that mimic human behaviour is beneficial for their adoption, as allowing the users to adopt a natural interaction makes them easier to learn [Song et al., 2022]. In the case of VUIs, our devices need to analyse how we share information in our spoken interactions and adapt according to our actions. For example, in a conversation where we do not trust one of our conversational partners, we limit the amount of information that we share, or modify the way we talk so that the corresponding individual does not receive our voice [Elkins and Derrick, 2013]. Similarly, if our devices collaborated together in a distributed VUI, it would be necessary to establish which of them could be trusted in the VUI, thus preventing unauthorised devices from receiving the user’s voice.

This thesis aims to study spoken interactions between humans and their devices, in such a way that criteria can be defined for voice user interfaces to protect the privacy of the users. We analyse how humans perceive their privacy when they interact in different environments, and the intuitive cues that are used in order to grant permission to other individuals into a conversational interaction. Our goal is to exploit these cues and define access management rules for devices located in the area of trust of the user, allowing them to collaborate by sharing their recorded audio. Once the devices are authorised to collaborate as a distributed VUI, several multi-channel signal processing methods are studied to improve the quality of the recorded signals and protect the privacy of external speakers in a specific speech interaction.
1.1 Thesis Structure

The research in this thesis can be classified into three categories, which study the implementation of a multi-device collaborative user interface from the perspective of protecting the privacy of the user. The following sections explore the questions of 1. when and why collaboration between devices might be desired, 2. how to securely decide which devices meet the criteria for collaboration, and 3. which methods can be implemented in a collaborative environment to ensure the privacy of the users. Therefore, the three categories in which the publications of this thesis are divided are:

- Design of an authentication method for electronic devices based on the intuitive consent granted in human spoken interactions.
- Analysis of the effect that the acoustic environment has on the users’ perception of their privacy.
- Design of multi-channel signal processing methods that would work in a distributed VUI improving the quality of the recorded signals and protecting other users’ privacy.

In a scenario where multiple devices record and share the voice of a user, it is crucial to ensure which devices are actually allowed to receive and process the user’s voice. Following the assumption that people modify the volume of their voice depending on how private the interaction is (e.g. whispering to tell a secret), we consider that devices that can hear the user’s voice will be implicitly authorised to record and process such voice signal. Methods combining acoustic fingerprints with cryptography have been proposed to securely evaluate the similarity of two recordings. In Publication II, we study the effect of delays between recordings in the fingerprint comparison and propose a method to reduce the length of the required audio recordings. We observed that additional decorrelation in the time-frequency transformation provides higher robustness to the fingerprint. This compensates for the delay degradation introduced by the shorter windows, providing higher accuracy for small delays between the audio samples. The proposed method is improved in Publication I, where different solutions are proposed to decorrelate the spectrogram of the recorded signals in order to generate a robust acoustic fingerprint with short audio recordings. The authentication process is then tested in Publication VIII with speakers at multiple distances in several noise conditions. This article shows the influence of distance between recording devices in the authentication process.

We consider that individuals modify the way they talk depending on the level of privacy of the conversation, and therefore aim to study the perception that users have in such spoken interactions. In Publication III,
a series of recordings were organised in Finland and India. Conversational data was gathered in diverse locations with different privacy levels, such as an office, cafeteria and the street. We then used a questionnaire to collect information about the perception of the subjects. In this paper, we observe the effect of the environment in the perception of privacy. This perception, however, is affected by the variability introduced by the subjects’ personalities and cultural background. Additionally, in order to obtain more statistically robust results and remove the bias of visual feedback, a similar evaluation was implemented in a crowd-sourcing platform, where multiple users rated their perception of privacy in a mixture of the previously recorded samples and synthetically generated ones. The results of this study are presented in Publication VII, showing that the acoustic environment does indeed affect our perception of privacy.

Once the communication between neighbouring devices has been established, it is possible to apply multi-channel signal processing methods that will take advantage of the wireless acoustic sensor network (WASN) structure. Considering an ad hoc scenario containing multiple edge devices with limited processing capabilities, Publication VI presents a multi-channel speech coding method where every encoding device sends their recording to the receiver, merging the multiple channels of the WASN in the decoding device.

Additionally, we consider the case in which multiple speakers could use different VUIs in the same acoustic space. If a user external to a spoken interaction does not specifically provide their consent, transmitting their voice would be considered a violation of their privacy. When multiple devices are recording speech from multiple speakers, Publication IV presents a cross-talk cancellation method to remove the information of the interfering speaker in the recorded mixture. The presented method shows high accuracy estimating the direction of arrival (DoA) of the target source, which is then used to remove the information from other interfering sources. A crucial point to consider in this architecture is that the signals need to be processed in one device, and therefore, some of them will need to be encoded and transmitted to the processing device. The effect of coding noise can then affect the estimation of the presented model. This effect is analysed in Publication V, evaluating the performance of the proposed method in estimating the DoA of the dominant speech source under coding noise from multiple speech codecs at different bitrate configurations. While we observe a degradation in the DoA estimation due to the reduced coding bandwidth, this estimation provides better results when both input channels have been encoded than when keeping one of them as a clean signal.

The following sections analyse the ideas that motivate these works in a comprehensive manner and describe the background methods on which the proposed solutions are based. Section 2 discusses the perception of
privacy that we have in spoken interactions, with our electronic devices and with other humans. Then, in Section 3, we analyse a scenario where multiple devices with embedded microphones can collaborate providing a distributed VUI. The architecture of a WASN is explained and we present the background methods on which current audio-based authentication solutions are based. Finally, assuming a connected network of devices, Section 4 presents single- and multi-channel source separation methods which could be implemented in a WASN and how they can protect the privacy of their users.
2. Perception of Privacy in Speech Interactions

Privacy is a basic psychological need of humans that applies to many aspects in our lives [van der Sloot and de Groot, 2018]. The general definition of privacy, provided by the Cambridge dictionary as “Someone’s right to keep their personal matters and relationships secret”¹, contains the main idea on which the assumptions presented in this thesis for protecting user’s privacy will be based. Privacy depends on someone’s perception of what they consider personal, and therefore it will be perceived differently depending on the user and the situation. For that reason, in order to define rules that protect the privacy of the users, it is necessary to understand how each user perceives their privacy in each specific situation.

If a service requires information from a user that they would not willingly share, the application then violates the privacy of the users. To avoid the misuse of users’ private data, multiple regulations have been issued around the world, defining how companies should collect and use the data of their users. For example, in the European Union, a set of guidelines is collected in the General Data Protection Regulation (GDPR) [Voigt and Von dem Bussche, 2017], which represents how the EU considers a user’s privacy should be preserved. However, due to the broad set of fields in which private data can be collected and processed, it is necessary to analyse the meaning of privacy for the users in each of these fields in order to develop methods that properly protect their privacy.

Deep learning and other data driven methods have experienced great advances in the last years, allowing us to analyse and process data more efficiently. This has improved the performance of many applications in a wide range of fields. For example, in the field of audio processing, deep learning methods have proven especially useful in ASR and NLP applications, which have allowed the fast development of voice assistants. But as these applications become more demanding, a need for more complex models emerges, which, in turn, require more training data. Additionally, when an application is expected to adapt to a specific user’s preferences,

its performance can greatly benefit from tuning using actual data from the
user. The use that these applications can give to our data raises concern
towards the implications that it can have for our privacy. Collecting and
processing a user's data against their consent would entail a violation of
their privacy.

An extreme case of users' privacy violation can be observed when the
personal data of the users is collected and traded to analyse specific user
behaviours and to target them for commercial purposes [Isaak and Hanna,
2018]. It is a natural practice to remove information that can identify
specific users, however, it has been observed that individual features of
a person's private information that can seem harmless individually, can
lead to the identity of the users when analysed together [Narayanan and
Shmatikov, 2008]. For that reason, data anonymisation techniques are
crucial for the protection of people's privacy [Murthy et al., 2019]. Methods
like differential privacy study how to introduce controlled noise into the
entries of a dataset, so that individual entries cannot be used to trace back
to the specific subjects without compromising the statistical properties of
the original dataset [Abadi et al., 2016, Husnoo et al., 2021].

The growing popularity of the Internet of Things (IoT) has also provided
a great number of devices that collect and process information around
us. Training models to personalise the behaviour of the devices to a user,
typically requires data from the specific user. And in some cases, the
user's collected data can only be processed locally. Federated learning
techniques provide solutions for devices to train models in a distributed
manner using only data that is available locally [Yang et al., 2019, Savazzi
et al., 2020, Wu et al., 2022]. Devices working in a federated learning
structure train their models using their own local data, and then share
the calculated gradients of the models with the rest of the collaborating
devices. The gradients are combined in a central node (e.g. averaged), and
finally, the central node distributes the updated gradients for the devices to
update their models. This allows IoT devices to process data using machine
learning models, while maintaining their collected data private.

Considering how significant privacy is in our lives, it should be a key
element in the design of any application that requires the processing of
user's data. The concept of Privacy-by-Design (PbD) [Cavoukian, 2009]
proposes that privacy should be proactively implemented in the design
of any application, and data handling should be performed from a user-
centric perspective. Additionally, implementing privacy protection as a
default element in the design of an application, can help us avoid the need
of finding "patches" to ensure the privacy of the users afterwards.
Perception of Privacy in Speech Interactions

2.1 Privacy in speech interfaces

Speech is our main and most intuitive method of communication, which we use every day to exchange ideas with other people. It is not surprising then, that VUIs keep growing in popularity. VUIs allow us to interact with our electronic devices in a natural way using our voice, as if we were interacting with another human. For that reason, we can find voice user interfaces in a wide variety of services. These services range from telecommunication applications, such as the basic function of a telephone, to virtual assistants, like Alexa or Siri [Hoy, 2018], which process what we tell them, reacting and answering as if we were interacting with another person.

These applications, however, usually require our voice to be transmitted to a remote party, which can be the receiver on the other side of the call or a server where our voice will be stored and processed [Chung et al., 2017]. We usually give our consent to the application to record our voices, and even to store portions of that audio data to provide better services in the future. However, speech signals contain a great amount of information and sometimes we are not even aware of what information we share when we speak. For that reason, it is important to define rules that enable our devices to protect our speech information or the speech information of a speaker that is not the owner of an electronic device.

Considering the case of virtual assistants, our interaction with such devices has become as natural as interacting with another person. As we can see in Fig. 2.1, it is possible to maintain the same type of dialogue without considering the type of information that is being collected in order to provide such a service. It has been observed how, due to the natural interactions that voice assistants provide, users begin to perceive these assistants as other humans in their interactions [Lopatovska and Williams, 2018]. This personification of virtual assistants, and the shift in trust that it involves, can expose users to bigger privacy violations, as people may willingly share more personal information in an environment of trust [Levy and Schneier, 2020]. It is important to educate people on the implications

Figure 2.1. Scenario where the same interaction is carried out between humans (a) and with a voice user interface (b).
that the use of VUIs have on their privacy. By openly sharing how an application processes the user’s personal data can encourage a faster integration of the service in the user’s environment, while simultaneously ensuring the correct handling of their personal information [Liao et al., 2019, Liao et al., 2020, Shin et al., 2022].

2.2 Perception of privacy in speech interactions

Designing methods that protect the privacy of the users in speech interfaces requires that we understand the meaning of privacy for the users in their own spoken interactions. Several studies have analysed the interactions that we have with other humans and the difference when those interactions happen with electronic devices and VUIs [Yeasmin et al., 2020, Leschanowsky et al., 2021]. In the case of speech applications, it is necessary to analyse how we perceive privacy with respect to our voices and the effect that the environment has on this perception. The effect of privacy in the environment has been studied in applications like, for example, the design of open-plan offices. If the workplace environment transmits the sensation of being private to each employee, higher productivity has been observed as a result [Kim and de Dear, 2013, Park et al., 2020, Di Blasio et al., 2019].

As mentioned above, the understanding of the users about how a service handles their private data results in an easier adoption of the product. Therefore, when a service is open about how they handle the private data of the users, it will be more natural for the users to adopt this service. By analysing how we perceive our privacy in our human-to-human interactions, our devices can understand our perceptions and adapt their behaviour accordingly. If the privacy management of our devices could adapt to what is natural in our interactions and users would understand that their devices behave the same way as another human would, they could more easily manage what personal information they share and how they communicate with their devices.

When we interact with other people, we intuitively modify the way we talk according to how we want the information to be shared. For example, if we want to tell a secret to a friend, we lower the volume of our voice to reduce the distance it reaches. The environment can also affect how we share information, for example, if it feels like our voices will reach farther due to the acoustic properties of the room, or when a crowded area presents more potential listeners. Analysing these interactions and recognising what causes changes in our perception can allow our devices to recognise devices that can be trusted in our area, or adapt the amount of information that is shared depending on how we would do it in a spoken interaction.
3. Device Authentication

We live surrounded by electronic devices that have the ability to record, process and transmit our speech information, and it is common for a single person to own multiple such devices, like smartphones, tablets or laptops. In an environment where multiple devices can provide a VUI, collaboration between these devices could be established in such a way that 1. the recorded signals could be combined to improve the quality of the services that they could individually offer, or 2. share the computational load of multi-channel processing to extend the battery life of mobile devices involved. However, if multiple processing devices shared information about the audio they recorded, the users could be exposed to violations of their privacy; especially considering that the collaborating devices would not necessarily belong to an exclusive user. Access management rules have to be defined to establish which devices can be trusted with the user’s voice.

This chapter presents the structure of a WASN and discusses how our devices can benefit from this architecture to provide a distributed VUI. Additionally, we analyse how different proposed methods exploit features of our intuitive perception of privacy to define the necessary access management rules for such user interfaces to operate in a secure manner. We will present signal processing and cryptography concepts that are combined for a secure authentication of devices.

3.1 Wireless Acoustic Sensor Networks

Considering the growing popularity of VUIs, it is not surprising that the number of devices containing embedded microphones is continuously increasing. This allows us to interact with our devices in a natural way using our voices. However, most VUIs are limited by the hardware of the device that provides it, and they require the user to be in close proximity to capture audio with sufficient quality. Multi-channel audio processing methods have proven that microphone arrays can significantly improve
Figure 3.1. Comparison between a scenario containing a microphone array and a WASN. While in (a) a single 3-microphone array points at the user or sound source, (b) shows the users surrounded by simple sensors forming a WASN.

the quality of a recorded signal. These methods usually take advantage of the different positions of the single microphones in the acoustic scene, exploiting the spatial difference between the sensors. For instance, beamforming techniques consider the delay that a signal has reaching each of the sensors to locate and focus on sources arriving from specific directions, or attenuate sounds from undesired sources.

However, as the spatial requirements become more demanding, with more sources in a wider area, the required microphone structures also grow in size and complexity. This can be impractical in our mobile devices, where the space for additional embedded microphones and processing power is limited. A more generalised idea is to have multiple microphones distributed in different locations in an acoustic space, forming a WASN [Bertrand, 2011]. WASN structures have proven useful in monitoring tasks for large areas [Xie and Cui, 2006, Pastor-Aparicio et al., 2020], especially considering that edge computing devices can process the signals that they record and share the data throughout a large network. Figure 3.1 depicts the structure of a WASN compared to a traditional microphone array. This allows us to cover wider areas using simpler sensor structures distributed around the acoustic space, rather than a more complex system concentrated in one location.

An ensemble of devices containing one or more embedded microphones, can work in a WASN to perform speech enhancement and source localisation tasks [de la Hucha Arce et al., 2017, Li et al., 2016], similarly to how traditional microphone arrays do. Most of these methods assume that the devices comprising the WASN are in known locations or the devices themselves are microphone arrays, allowing them to apply spatial audio techniques for signal enhancement [Markovich-Golan et al., 2015, Alexandridis and Mouchtaris, 2018]. Such assumptions are however not possible in a WASN architecture formed by arbitrary mobile devices, and therefore, alternative methods need to be explored.

Depending on the application carried out by the WASN, we can observe two possible architectures that the devices can follow, based on their
Device Authentication

3.2 Distributed Voice User Interfaces

In domestic environments, we are almost constantly surrounded by devices that can record audio and communicate forming a WASN with devices in the same acoustic space. Collaboration between interconnected sensors can then provide better quality services and detach the user from a specific device. For example, if a user wanted to make a phone call without looking for his phone, he would only have to say "Call mum!", and the devices around him would gather the speech signal and transmit it accordingly. Ad hoc or centralised architectures could be adopted depending on the requirements of the application. In the case of distributed speech coding, an ad hoc structure would allow the different sensors to share small amounts of data, while each of them would take on the task of encoding their recorded signal. Alternatively, tasks like speech enhancement or source separation might require the simultaneous processing of multiple communication patterns. Figure 3.2 shows a schematic of the two possible architectures, an ad hoc network where devices communicate arbitrarily with each other while independently processing their recorded signals, and a centralised structure where all the devices communicate with a central node that takes care of the main computational load. On the one hand, an ad hoc network provides a truly distributed solution, where each device can process their recorded signal without depending on a central node, sharing only essential information with the nodes that require it. However, each node requires certain processing power and more demanding multi-channel processing is not possible. On the other hand, if one of the devices has a higher computing power, it can receive the complete signals recorded in each of the other nodes and process them jointly, acting as the central node. This structure, however, may incur problems due to the additional quantisation noise and delays introduced by every single channel encoding and transmitting their signals to the central node.

3.2 Distributed Voice User Interfaces

In domestic environments, we are almost constantly surrounded by devices that can record audio and communicate forming a WASN with devices in the same acoustic space. Collaboration between interconnected sensors can then provide better quality services and detach the user from a specific device. For example, if a user wanted to make a phone call without looking for his phone, he would only have to say "Call mum!", and the devices around him would gather the speech signal and transmit it accordingly. Ad hoc or centralised architectures could be adopted depending on the requirements of the application. In the case of distributed speech coding, an ad hoc structure would allow the different sensors to share small amounts of data, while each of them would take on the task of encoding their recorded signal. Alternatively, tasks like speech enhancement or source separation might require the simultaneous processing of multiple communication patterns. Figure 3.2 shows a schematic of the two possible architectures, an ad hoc network where devices communicate arbitrarily with each other while independently processing their recorded signals, and a centralised structure where all the devices communicate with a central node that takes care of the main computational load. On the one hand, an ad hoc network provides a truly distributed solution, where each device can process their recorded signal without depending on a central node, sharing only essential information with the nodes that require it. However, each node requires certain processing power and more demanding multi-channel processing is not possible. On the other hand, if one of the devices has a higher computing power, it can receive the complete signals recorded in each of the other nodes and process them jointly, acting as the central node. This structure, however, may incur problems due to the additional quantisation noise and delays introduced by every single channel encoding and transmitting their signals to the central node.
channels, in which case the sensors could send their recorded signals to a central device that would jointly process all the channels.

3.3 Device Authentication

In a wireless sensor network, it is crucial to ensure that the collaborating devices are indeed allowed to do so. If the devices are sharing information with each other, a malicious device could extract data from the network that it is not authorised to have, or transmit erroneous data in order to hinder the task of the network. In industrial applications, where wireless networks are defined to perform specific tasks, manually defining the sensor network can ensure its integrity, also providing more information about its structure. However, applications that benefit from more flexible structures, require access methods that can adapt to new devices and automatically remove sensors that exit the network. For that reason, cryptographic methods are widely studied in order to support authentication in wireless sensor networks [Singh and Jain, 2022]. Additionally, as most devices in wireless sensor networks do not have great processing power and tend to be battery operated, these authentication methods need to be efficient enough to run in these specific devices and maintain their battery life.

![Figure 3.3](image.png)

**Figure 3.3.** Ideal behaviour of an access management system for distributed VUIs. Devices within the reach of Bob's voice will be allowed to collaborate, while those outside will be rejected.

Considering the amount of information contained in our speech signals, a robust and responsive decision of which devices would be allowed in a distributed VUI is necessary in order to 1. ensure the correct functioning of the system and 2. protect the private data of the users in the network. First, if a malicious device was allowed access to the network, the VUI would have to process signals from more channels that would not provide any useful information, hence interfering with the application. Second,
Device Authentication

an eavesdropper that has access to the shared signals of the network will have access to the user's private speech data. Chapter 2 describes how humans modify their voices depending on the privacy of the environment, or how far they want their voices to reach. Figure 5.1 shows the ideal behaviour that an authentication method should have in a voice-operated application. We can then assume that if multiple devices are within the range of a user’s voice, they will be allowed to collaborate, and denied access to the WASN otherwise. In this section we will analyse the basic methods behind authentication protocols that take advantage of the user's voice to recognise devices in the same acoustic space.

Audio-based authentication has been studied using acoustic fingerprints to encode the information of the recorded signal [Schürmann and Sigg, 2013]. The authentication process can be observed in Figure 3.4. The audio segments will be simultaneously recorded in both devices, and it is assumed that if the user’s voice is dominant in the recording it will define the calculated fingerprint. The fingerprints are then compared and, if they are similar enough, the devices will be considered to be in the same acoustic space, thus allowing collaboration. The two main steps that define the authentication process are the fingerprint generation and the exchange of fingerprints for comparison. In the following sections we will describe 1. the concept of acoustic fingerprints and their application, 2. the requirements for a secure acoustic fingerprint, and 3. the secure comparison of fingerprints for robust authentication.

3.3.1 Acoustic fingerprint

Acoustic fingerprinting is a popular technique used to easily compare audio segments. Its objective is to compress the significant information of an audio segment to allow robust comparison between audio signals [Grosche et al., 2012, Walczyński and Ryba, 2019]. The compressing properties of acoustic fingerprints also allow for efficient storage of a large number of audio files, making them especially popular in the field of music retrieval. It is important to distinguish between the concepts of audio fingerprinting and audio watermarking, a common misconception caused by the similarity of the two terms. While an audio fingerprint extracts information of an audio segment, audio watermarking aims to encode additional information in a transmitted audio signal without affecting its quality [Arnold, 2000].

A popular example of an acoustic fingerprint applied in audio retrieval can be observed in Shazam [Wang et al., 2003], where the generated fingerprint is based on spectral centroids that are calculated using the relative position of peaks in the signal spectrum. Different methods have also been studied depending on the application requirements, such as wavelet decomposition or direct transformations of the signal spectrum [Chandrasekhar et al., 2011]. The great number of proposed fingerprinting methods show
that a fingerprint needs to be designed for a specific service and its performance might degrade under different requirements. For example, a music retrieval application requires a fingerprint that allows a faster search in a fingerprint dataset, while device authentication requires a robust one-to-one comparison of two audio segments.

### 3.3.2 Decorrelation methods

As it is mentioned before, an acoustic fingerprint needs to contain compressed information of the recorded audio segment. While some methods encode higher level features from the acoustic signals (e.g. variations in pitch or relative positions of spectral peaks), a secure authentication process benefits from shorter audio segments that, additionally, needs to compare two fingerprints without the possibility of optimal alignment. In a real-life scenario, it is likely that synchronisation delays appear between both authenticating devices and the different hardware, and the physical positions of the devices will introduce additional noise in the calculated fingerprint [Schürmann and Sigg, 2013]. For that reason, it is crucial that the generated fingerprints are robust against small delays.

For the generation of an effective acoustic fingerprint, we need to consider that 1. the fingerprint needs to encode the maximum amount of information from the audio signal, and 2. it also needs to resemble a random sequence so that it is not possible to predict sections of the fingerprint if other areas are known. The fingerprint calculation is then similar to coding techniques where redundancies in the data are removed in order to encapsulate as much useful data in the transmitted sequence as possible. Orthogonalising transformations can then provide decorrelation of the audio features for their secure comparison, as well as robustness against noise and delays. Some popular decorrelating transformations are the discrete cosine transform and the eigenvalue decomposition.
Discrete Cosine Transform

The discrete cosine transform (DCT) is a time-frequency transformation similar to the discrete Fourier transform (DFT), in which the signal $x(t)$, where $t$ stands for the time index, is represented as a combination of cosine functions with different frequencies

$$X(k) = \sum_{t=0}^{T-1} x(t) \cos \left( \frac{\pi}{T} \left( t + \frac{1}{2} \right) k \right).$$  \hspace{1cm} (3.1)

Unlike DFT, the coefficients obtained by the DCT are real-valued, and its compressing capabilities have made it a popular method in multiple fields. For example, a two-dimensional version of DCT is used in image compression algorithms like JPEG [Wallace, 1992, Hudson et al., 2018]. In speech coding, a modified version of DCT, the MDCT [Wang and Vilermo, 2003], is used in standardised speech codecs like EVS and LC3 [Dietz et al., 2015, Schnell et al., 2021]. This transformation has become an alternative to DFT in time-frequency transformation, as its orthogonalisation properties avoid the over-coding problem presented by redundant information over consecutive overlapping windows.

Eigenvalue Decomposition

Another popular orthogonalisation transform evaluated for the decorrelation of components in the fingerprint is the eigenvalue decomposition [Golub and Van Loan, 2013]. Assuming a square, diagonalisable matrix $M$, the eigenvalue decomposition can be expressed as

$$M = V \cdot \Lambda \cdot V^H,$$  \hspace{1cm} (3.2)

where $V$ represents a square, orthonormal matrix, containing the eigenvectors of $M$ and $\Lambda$ stands for a diagonal matrix with the corresponding eigenvalues for each eigenvector. The transpose complex-conjugate operation is represented as $M^H$. In the case of an audio signal, we aim to diagonalise the covariance matrix of the input signal, thus reducing the correlation between the signal features. The covariance matrix is defined as $\Sigma_{xx} = E[xx^H]$, which by definition is Hermitian, and hence diagonalisable. Applying the definition of the covariance matrix to eq. 3.2, the diagonalising transformation over $x$ can be defined as

$$y = \Lambda^{-\frac{1}{2}} \cdot V^H \cdot x,$$  \hspace{1cm} (3.3)

so that we obtain $\Sigma_{yy} = I$. While the DFT and DCT have a direct physical interpretation as the different frequency components forming the signal, the eigenvalue decomposition can be seen as a transformation into a space with orthogonal components.
Device Authentication

3.3.3 Fuzzy commitment scheme

Although acoustic fingerprints are designed to compress the information of the audio signal and it is not possible to recover the exact audio information, secure comparison methods have been proposed to further restrict the access to the actual fingerprint. The comparison of audio signals can be seen as a biometric recognition problem, where two matching components can still contain slight discrepancies [Hao et al., 2006]. In this case, two fingerprints can contain several different coefficients due to multiple factors such as synchronisation delay or hardware differences. For that reason, a fuzzy vault scheme is proposed to compensate for possible mismatch between similar fingerprints [Juels and Wattenberg, 1999, Juels and Sudan, 2006]. The generated fingerprints can then be used as cryptographic keys in the communication between the devices [Schürmann and Sigg, 2013, Schürmann et al., 2017], which will only be possible when both devices generate the same key.

Error correcting codes (ECC) are popular methods in communications and storage solutions as they are able to compensate for errors introduced in the data stream. As depicted in Figure 3.5, the encoding algorithm of the ECC transforms the data symbol from a finite field $\mathbb{Z}^p$ into a larger space $\mathbb{Z}^q$, with $q > p$. The encoded symbol then contains additional information, seen as overhead in the communication or consumed memory. If the encoded symbol is then affected by less than a certain number of errors, limited by the chosen code, the decoding algorithm will be able to recognise where these errors were caused and recover the original message.

A fuzzy commitment scheme exploits the decoding capabilities of an error correcting method, thus transforming the corresponding bitstream to a lower dimensionality space where small differences in the original stream are translated into the same symbol. Small differences in the calculated fingerprints due to sensor distance and synchronisation are considered as errors in an encoded message and if they are similar enough they will be
corrected to identical words. This does not only allow to obtain identical fingerprints even when the recordings are affected by noise, but also hides the information of the original fingerprint. As the error correction decoder transforms the fingerprint to a space with lower dimensionality, it is not possible to accurately return to the original space.

While the fuzzy commitment scheme manages to hide the acoustic fingerprint within the lower-dimensional space of the error correction code, past methods propose solutions that avoid the fingerprint being shared, thus removing any possibility of recovering the original audio information [Schürmann et al., 2017]. A Diffie-Hellman key exchange is proposed, and assuming that the fuzzy commitment scheme generates identical sequences in a matching scenario, the fingerprints are used as cryptographic keys.
4. Applications of a Distributed VUI

The speech signals that our devices record in our domestic use are usually mixed with different kinds of noises and other speakers in the same acoustic space. Speech enhancement methods are crucial in providing a good speech quality in these noisy conditions. By recording signals from microphones at different positions in space, multi-channel methods benefit from the additional information available on multiple recorded channels. For example, among popular exploited features in a multi-channel setting, we can find the spatial properties of the environment and the generated signals. Among the tasks where multi-channel speech enhancement has provided great improvement is blind source separation (BSS), where multiple speech signals present in the recorded mixture are extracted into individual tracks. Current mobile devices already include several microphones that allow them to provide a certain level of noise reduction, however, as depicted in 3.1, multiple devices collaborating together can provide coverage over a larger area. If our devices collaborated together in a WASN, they could exploit multi-channel speech processing methods to improve the quality of the services they offered and protect the privacy of the users.

The following sections explore two possible applications of a distributed VUI: 1. blind source separation techniques that can be applied in a centralised approach to remove external speakers that should not be part of a specific interaction, and 2. distributed speech coding using an architecture where devices encode their own signals and send them to the corresponding receiver. We present methods for single- and multi-channel source separation, and speech and audio coding, evaluating the required characteristics to transition them into a distributed setting.
4.1 Cross-talk cancellation for privacy protection

In a communication scenario, cross-talk appears when unwanted speech signals are captured simultaneously with the target speaker's speech. This effect is especially problematic in acoustic echo cancellation (AEC) applications, where both speakers can talk at the same time and the echo signal leaks back into the microphone mixed with the actual target speech. Unlike AEC, in the context of a distributed VUI, multiple devices are located in a relatively wide area and it is possible that multiple speakers coincide in the same acoustic space, hence being recorded as a mixture in different devices, as depicted in Fig. 4.1. In this case, cross-talk does not only degrade the quality of the recordings, but when the speakers are taking part in different interactions, the leakage of another speaker's voice can cause a violation of their privacy. If a user does not explicitly allow an application to record and transmit their voice, the distributed VUI should be able to separate such a user from the interaction.

Cross-talk situations are common in real-life scenarios where speakers do not have a chance to move into a private space. These situations have been analysed in environments like call centre offices [Lombard and Kellermann, 2008] or inside a car [Matheja et al., 2013]. However, these types of scenarios are also becoming more frequent in domestic environments, as working-from-home policies have been intensified by the SARS-COV-2 pandemic situation [Tursunbayeva et al., 2022]. For that reason, cross-talk cancellation (CTC) techniques are crucial in multi-speakers and multi-device scenarios.

The CTC problem can be seen as a specific case of source separation in a multi-speaker scenario, where the separation method aims to extract
only one target signal. Speech proceeding from other sources is considered noise that should be removed. Multiple methods have been widely studied for single- and multi-channel BSS that can be adapted as solutions for the CTC problem. The signal model for a BSS problem with two sources and two microphones is defined in Figure 4.2. This model can be extended to \( N \) sources and \( M \) sensors, where the mixture of the signals from each source \( n \) recorded in a microphone \( m \) can be seen as

\[
X_m(k) = \sum_{n=0}^{N-1} H_{m,n}(k)S_n(k) + U(k), \tag{4.1}
\]

and \( H_{m,n}(k) \) stands for the frequency response of the corresponding relative transfer function (RTF) with \( k \) representing the frequency index. The RTF is applied on the spectrum of the clean signal as a product that would be seen as a convolution in the time domain. In a multi-channel scenario we can stack each vector into a matrix so that \( X(k) = \{X_0(k), \ldots, X_{M-1}(k)\} \), and applying the same procedure to \( H_{m,n}, S_n \) and \( U \), the mixture can be represented as a matrix product:

\[
X(k) = H(k) \cdot S(k) + U(k). \tag{4.2}
\]

The objective of source separation algorithms is to find an estimate of the clean independent signals \( \hat{S}_n \) before the mixture, as it is depicted in Figure 4.2.

### 4.2 Single-channel source separation

Single-channel source separation methods exploit features of the signal in the STFT domain to model the speech signal from the target source. Speech signals have a strong structure both over time and frequency, and they tend to be sparse in the STFT domain. This sparsity causes the so-called W-disjoint orthogonality [Rickard and Yilmaz, 2002, Yilmaz and Rickard, 2004], which means that two signals from independent sources have a considerably low probability of sharing components of meaningful amplitude in the same time-frequency bin.

#### 4.2.1 Time-frequency masking

Considering the W-disjoint orthogonality, source separation methods generate a mask by localising the time-frequency components of the mixture signal that contain the components from the target source. The generated mask is then applied on the corresponding time \( (t) \) and frequency \( (f) \) components of the recorded mixture \( X(f, t) \).

\[
\hat{S}_n(f, t) = W_n(f, t) \cdot X(f, t), \tag{4.3}
\]
where $W_n$ represents the coefficients of the mask generated for the $n$-th source, and $\hat{S}_n$ corresponds to the estimation of the same clean signal. Assuming a signal model where the target speech is affected by additive, uncorrelated noise, different methods have been proposed to estimate the spectral mask, such as spectral subtraction [Boll, 1979] or Wiener filtering [Habets, 2007, Cauchi et al., 2015].

Depending on the requirements of the application or the desired quality of the output signal, different types of masks can be defined:

**Binary mask**
A binary mask is designed under the assumption that only one source dominates in every specific time-frequency bin. Exploiting the sparsity of speech signals, the coefficients of the mask are then only 1 or 0, representing whether the target source can be found in the specified band or not. This approach, however, can introduce a great amount of distortion in the signal due to the spectral gaps, even removing parts of the signal.

**Soft mask**
A soft mask is a less restrictive approach, where signals are assumed to sometimes be simultaneously present in a frequency band. In this case, the mask coefficients are calculated in the range $[0.0, 1.0]$. Maintaining information from the original mixture in each of the bands leads to higher sound quality than binary masks.

**Oracle mask**
Ideally, a time-frequency mask that yields a perfect separation of a source is referred to as an oracle mask. The calculation of this mask requires the information of the original signal and, therefore, it is only used as an upper bound for the quality of mask-based source separation methods.
4.3 Multi-channel source separation

While single-channel source separation and speech enhancement methods use the spectral features of the signal, with multi-channel processing it is also possible to exploit spatial features unavailable to single-channel methods. Recording the acoustic scene from different locations in space provides these spatial features defined by the propagation of the audio signal. Multi-channel source separation methods can be classified into three types: 1. methods based uniquely on the spatial properties of the recording setup, 2. those that combine the spatial information of the signal and its statistical properties, and 3. those only based on the statistical properties of the recorded signal [Vincent et al., 2018].

4.3.1 Beamforming

Among the source separation methods that only use the spatial information of the recording setup, traditional beamforming is probably the most popular one. The simplest example of this type of methods is the delay-and-sum beamformer, depicted in Figure 4.3. Assuming, for example, that we have a linear microphone array and that \( \theta \) is the direction of arrival (DoA) of a signal in far-field conditions, the signal will arrive to each sensor with a time delay of \( \Delta t \). The traditional delay-and-sum beamformer [Dudgeon and Mersereau, 1984] exploits this fact to amplify the signal coming from a specific DoA by matching the corresponding delays between each microphone, generating a constructive interference. It is important to note that information about the position of the source is necessary for these methods to operate correctly.

4.3.2 Minimum Variance Distortionless Response

Other approaches take advantage, not only of the position of the target source, but also on the statistics of the recorded signal. Among these methods we can find the Minimum Variance Distortionless Response (MVDR) beamformer [Haykin, 2008]. Methods based on MVDR have also been combined with other techniques such as time-frequency masking for source separation [Alinaghi et al., 2014, Liu et al., 2018].

An MVDR beamformer assumes the noisy signal as an additive mixture between the target source and noise, similarly to the mixture defined in eq. 4.2. For simplicity, an anechoic situation is considered, where the transfer function between a source and the corresponding microphone \( m \) can be represented as a steering vector \( \mathbf{a}(m, \theta, \lambda) = e^{2\pi dm \sin(\frac{\theta}{\lambda})} \). Note that the steering vector depends on the positions of each microphone, as \( d \) represents the distance between the microphones in a linear array. The DoA of the signal is then represented by \( \theta \), and \( \lambda \) stands for the wavelength.
of the signal. The mixture expressed in a similar manner as eq. 4.2 can be seen as

$$X(k) = A(k) \cdot S(k) + U(k).$$

where $A(k)$ stands for a matrix that contains the steering vectors of every source to the microphone array in the $k_{th}$ frequency bin. The goal of MVDR is to define a filter with coefficients $W$, so that $\hat{S} = W \cdot X$, and which follows eq. 4.5. The coefficients of the MVDR filter are defined under the constraint of a distortionless response, $W \cdot a_n = 1$, so that the information proceeding from the direction of source $n$ remains unmodified. The filter then aims to minimise the energy of signals from any other directions, which are considered as noise. The energy of the noise component is defined as $\Sigma_{uu} = \mathbb{E}[uu^H]$.

$$\min_W W \cdot \Sigma_{uu} \cdot W^H$$

$$W_{MVDR} = \frac{\Sigma_{uu}^{-1} \cdot a_n}{a_n^H \cdot \Sigma_{uu}^{-1} \cdot a_n}$$

The coefficients of the MVDR filter are represented in eq. 4.6, defined by solving eq. 4.5 under the distortionless response constraint. As it is mentioned before, this approach relies both on the statistics of the signal, encompassed in the matrix $\Sigma_{uu}$, and the spatial characteristics of the microphone array, which are represented in the definition of the steering vector. Due to the distortionless response forced by the MVDR filter, it is normal that residual noise remains in the output signal. For that reason, MVDR beamformers are usually followed by post-filters for further noise reduction.
4.3.3 Independent Component Analysis

The third type of source separation method only considers the statistics of the signals in order to separate the individual components of the mixture. Among these methods, a popular approach is independent component analysis (ICA). ICA is a popular BSS method that exploits the statistics of the recorded signals to extract single sources from a multi-channel mixture by assuming that the different signals are statistically independent from each other [Hyvärinen et al., 2001]. The mixture model assumed by ICA can be seen as

\[
\begin{align*}
x_1(t) &= a_{1,1}s_1(t) + a_{1,2}s_2(t) + a_{1,3}s_3(t) \\
x_2(t) &= a_{2,1}s_1(t) + a_{2,2}s_2(t) + a_{2,3}s_3(t) \\
x_3(t) &= a_{3,1}s_1(t) + a_{3,2}s_2(t) + a_{3,3}s_3(t)
\end{align*}
\]

(4.7)

where each recorded signal \(x_m(t)\) is represented as an instantaneous mixture between the clean signals \(s_n(t)\) and a real coefficient \(a_{i,j}\) acting as a scaling factor that represents the distance between the source and the corresponding microphone.

The problem that ICA aims to solve, is finding the mixing matrix \(\alpha\). To achieve this, ICA is based on two main assumptions:

1. The separated components are statistically independent. That is, the joint probability density function of the sources is factorisable:
\[
p(s_1,\ldots,s_n) = p_1(s_1)\ldots p_n(s_n).
\]
2. The independent components cannot be normally distributed.

Exploiting assumption number 2, the methods proposed to solve the mixing problem in ICA aim to maximise the non-gaussianity of the independent components. Some of these methods use, for example, a metric like kurtosis, which measures how much the values of the output distribution concentrate in the peak or tails of the distribution [Hyvärinen et al., 2001], or negentropy, which measures how the output samples become more in order, i.e. less random [Novey and Adali, 2008].

As it can be seen in eq. 4.7, the mixture of the signals is considered an instantaneous mixture, which is not a realistic approach for audio signals, as the mixtures of recorded acoustic signals have a convolutive nature. Considering that a convolution is represented as multiplication in the DFT domain, frequency domain variants of ICA have been proposed for better performance on convolutive mixtures [Ikeda and Murata, 1999, Nishikawa et al., 2002, Benzvi and Shafir, 2018]. Additionally, the ICA methods are limited in cases where the number of sources is uncertain or it is greater than the number of recording channels. ICA, nevertheless, introduces the notion of independence between the separated signals, which has afterwards been combined with other techniques such as non-negative matrix...
Applications of a Distributed VUI

4.4 Deep learning for source separation

With the growing availability of training data, deep learning models have also shown great potential in BSS applications, and speech enhancement in general. Deep neural networks (DNN) allow us to evaluate nonlinear relationships in the analysed data, learning hidden features that can be more easily optimised for the desired application. The set of techniques that allow us to learn the hidden features of the data is known as representation learning [Bengio et al., 2013, Yu and Deng, 2011]. The performance of deep learning methods in speech enhancement has been demonstrated with popular models like SEGAN [Pascual et al., 2017], an end-to-end approach based on generative adversarial networks (GAN) and PercepNet [Valin et al., 2020], which combines a recurrent neural network (RNN) with a speech perceptual model for more effective noise reduction.

Similarly to the methods presented in Section 4.3, BSS methods can follow two main approaches. A time-frequency mask can be estimated from the input features, or the trained model can directly output the separated audio signal [Wang and Chen, 2018], as depicted in Figure 4.4. Among mask-based solutions we can observe popular models like TasNet [Luo and Mesgarani, 2018, Luo and Mesgarani, 2019], which processes signals in
the time domain, using an encoder-decoder architecture to estimate the corresponding mask from higher level features, even though the majority of methods work in the frequency domain [Huang et al., 2014]. Alternatively, methods that directly output the audio features, also known as mapping-based methods, have been widely studied with great performance using generative models such as variational auto-encoders (VAE) [Karamathli et al., 2019, Pi et al., 2021, Neri et al., 2021].

4.5 Distributed Speech Coding

While source separation techniques allow us to enhance the recorded signal or remove interfering sources in the acoustic environment, multi-channel solutions require that the signals recorded in each of the devices are sent to a central node to be processed together. This presents several potential problems that could degrade the quality of these services: 1. the additional coding noise on the transmitted signals, 2. the total bandwidth consumed by the WASN to transmit all the signals to the central node, and 3. the reduction of the battery life of the multiple sensors in the network that need to share their recorded signals. For that reason, it is important to study distributed speech coding techniques that can exploit the structure of a distributed VUI. This would allow the user to move freely around the acoustic space without interacting with a specific device, as the whole WASN could capture their speech signal.

Distributed source coding has been widely studied, as a means to efficiently transmit information when multiple sensors are available [Wang et al., 2017]. Due to the proximity of the sensors, the recorded information will be highly correlated, utilising unnecessary bandwidth to transmit the redundant data. Distributed source coding exploits the high correlation between channels to avoid transmitting redundant information, thus considerably reducing the required transmission bandwidth. The decoder then models the correlated components for the reconstruction of the signal. This produces a shift in the complexity from encoders to decoder, which makes distributed coding a suitable application for sensor networks, where the majority of sending devices are power-limited or battery operated. Applying distributed coding to speech then requires to study how to model the correlated components between channels and model them in the decoder [Bäckström et al., 2016].

In the following sections we are going to present an overview of the functioning of traditional speech and audio codecs, discuss the ideas that need to be considered to use similar architectures in a distributed coding system, and analyse different methods such as post-filtering that can be used to enhance the performance of a distributed codec.
4.5.1 Traditional speech coding

State-of-the-art codecs are designed to address both speech and audio signals, which due to their characteristics, usually require different approaches to be efficiently encoded. While music and audio signals benefit from frequency domain coding techniques, time domain coding provides better results on speech signals. For that reason, it is common that state-of-the-art codecs such as EVS [Dietz et al., 2015] and Opus [Valin et al., 2012], provide a unified approach by switching between both working modes depending on the encoded signal, allowing efficient encoding of speech at low bitrates as well as high quality music.

**Time domain codecs**

The code-excited linear prediction (CELP) codec, is a time domain codec especially efficient in encoding speech signals, assuming that the input signal contains only one fundamental frequency (F0) [Bäckström, 2017]. CELP-based methods exploit the structure of the speech generation system, extracting the spectral envelope of the speech signal using linear prediction. The linear prediction coefficients (LPC) can then be used as a filter to reconstruct the signal’s envelope. The harmonic structure of the speech signals caused by the F0 is then encoded. Finally, an analysis-by-synthesis approach is used to efficiently quantise the remaining residual. In this approach, the signal is reconstructed after encoding until the set of parameters that provide the lowest error with respect to the original signal are found [Benesty et al., 2008].

**Frequency domain codecs**

Frequency domain codecs, on the contrary, transform the speech signal into the frequency domain, quantising and encoding the spectrum of each time frame. As it is explained in Section 3.3.2, the modified discrete cosine transform (MDCT) is a preferred time-frequency transform, because it
addresses the over coding problem between overlapping windows that the DFT presents [Wang and Vilermo, 2003, Fuchs et al., 2015]. The orthogonalising properties of the MDCT allow for the encoding of smooth overlapping windows without spending additional bits in the quantisation of redundant information.

The structure of a frequency domain codec is depicted in Figure 4.5. First of all, the input signal is segmented using overlapping windows and transformed to the frequency domain using an MDCT. After the time-frequency transform, a perceptual model is used to estimate frequency masking effects. The masking estimation allows to define the quantisation accuracy for each frequency band, so that low energy bands next to other louder ones are not quantised as accurately because they will not be heard. The quantised information is then encoded using an arithmetic encoder that distributes the bits according to the set bandwidth limit. The perceptual model and the quantised spectrum are then encoded and sent to the receiver. On the receiver side, the quantisation levels are recovered from the perceptual model, and the spectrum of the signal is reconstructed and transformed back to the time domain. The signal is then reconstructed using the overlap-add method. A particular variant of a frequency domain codec, known as transform coded excitation (TCX), can be seen in the EVS codec [Fuchs et al., 2015], characterised by the fact that the spectral envelope is extracted using LPC.

Frequency domain coding allows for high quality music and audio coding. Additionally, the lower computational complexity of frequency domain codecs compared to CELP-based ones, facilitates their integration in low-power and battery-operated mobile devices. For this reason, low delay codecs like LC3 [Schnell et al., 2021], designed for Bluetooth communication between mobile devices, choose frequency domain coding.

4.5.2 Transition to distributed coding

In a distributed speech coding scenario, where the sensors comprising a WASN will work as encoders, we assume that these devices will have a limited performance and they will be battery operated. For that reason, the design of a codec for this architecture needs to consider an encoder structure as simple as possible to preserve the battery life of the sensor devices and ensure their correct performance. All the encoding devices will independently send their information to the receiver, as we can see in Figure 4.6. The decoder then collects the encoded data from all the channels and merges them to decode the actual speech signal.

Note that all the recording devices are in close proximity, therefore, encoding the signal with a traditional coding scheme would require the transmission of a great amount of repeated information at the expense of each device’s available bandwidth. Distributed coding methods need to
account for this redundant data and remove it from the communication. Considering a frequency domain codec, for instance, the perceptual models would be a highly correlated component between devices, which share part of the available bits with the spectrum information. Efficiently encoding the perceptual models over multiple devices could significantly reduce the required bandwidth in each encoding device. Methods have been proposed to efficiently encode the frequency domain perceptual models and recover them at the decoder side in a distributed scenario [Bäckström et al., 2019, Bäckström et al., 2016].

Additionally, the quality of a distributed codec in single channel mode needs to be considered in its design. The usefulness of a distributed codec can be debatable if it is not able to perform with a similar quality than a codec that would not require the additional complexity of the WASN. For that reason, additional complexity is required on the decoder, as post-filtering techniques will be required to compensate for the additional quantisation noise [Das, 2021].

4.5.3 Decoder side

The receiver of a distributed coding architecture is assumed to receive and decode the data streams from all the encoding devices. As the complexity of the codec will lay on the decoding stage, it is expected that the receiver will be connected to a power supply and will have higher processing capabilities than the encoding devices. After the decoding of all the received data streams, the decoder will need to merge the received data to reconstruct the target output signal.

*Time-delay-of-arrival estimation*

The multi-channel processing techniques presented in Sec. 4.3 usually require a set of microphones in static known positions for their optimum performance. In the case of a WASN composed of mobile devices this will not be the case. However, the merging of the signal can be achieved in a similar fashion as in the delay-and-sum beamformer. Assuming the pres-
ence of a dominant signal in the mixture, several methods have explored
the estimation of the time-delay-of-arrival (TDoA) in WASNs [Bouafif and
Lachiri, 2012, Bouafif and Lachiri, 2016, Dang et al., 2022]. The esti-
mated TDoA can then be compensated to merge the signals, similarly to a
delay-and-sum beamformer, to recover the dominant speech.

**Post-filtering**

Post-filtering techniques are a popular solution usually applied after speech
processing methods in order to obtain satisfactory performance. Post-filters
aim to compensate for noise or distortions introduced by the system they
follow. For example, speech and audio codecs implement post-filters af-
ter the decoding step to remove quantisation noise. Different popular
methods are usually applied in audio codecs depending on the type of
distortion suffered by the signal. In frequency domain codecs, especially
at low bitrates, isolated peaks appear in the spectrum of the signal due to
quantisation, which is perceived as musical noise. Noise filling methods
introduce random noise in the quantised frequency bins to address this
type of artefacts. In a distributed speech coding scenario, the quantisation
noise added in each of the channels might prove problematic in the merg-
ing step for the target signal reconstruction. Methods have been proposed
to estimate a joint speech model based on observations from the different
individual channels, aiding in the signal recovery [Das and Bäckström,
2021]. Post-filtering techniques are also especially useful in mask-based
signal separation methods. As the T-F mask suppresses frequency bands
that are not considered to contain the target signal, distortions are in-
troduced in the output signal. Gap filling post-filtering techniques have
shown promise in reconstructing the missing suppressed bands on the
target signal [Zhan et al., 2016]. Other post-filtering methods have been
proposed to reconstruct missing parts of the spectrum based on previously
trained speech models [Das et al., 2018].
5. Summary of Publications

Summaries of the publications forming this dissertation are presented in this chapter. The publications have been organised following the structure defined in the previous sections.

Considering the popularity of voice applications and the amount of devices that can process and transmit audio signals, the quality of speech applications could be greatly improved if multiple devices collaborated sharing their recorded signals for a common application. If the signals, however, were shared with devices that weren’t located in the same area, the privacy of the user would be violated. Devices can be authenticated by obfuscating the content of their recorded signals in an acoustic fingerprint and comparing it with another device to distinguish if they are located in the same acoustic space, such that similarity in fingerprints would represent similarity in the recorded signals. Publication I and II present methods to calculate robust acoustic fingerprints with short audio recordings. Publication VIII then analyses the effect that distance between devices and noise in the environment have on the authentication process.

The presented fingerprints assume that speakers can perceive the level of privacy of the environment and modify the way they talk accordingly. To analyse the effect of the environment in this perception of privacy, Publication III presents a conversational dataset recorded to evaluate the perception of privacy of speakers in different environments, and we analyse the effect that different environments have in this perception. The acoustic properties of the different environments are then evaluated in Publication VII, and another questionnaire is performed through a crowdsourcing platform, gathering data from a larger number of listeners.

Assuming that multiple devices have been correctly authenticated and are allowed to collaborate, the remaining publications explore applications of wireless acoustic sensor networks. Publication VI presents a speech codec optimised for distributed processing of the speech signal. In Publication IV a method is developed for multi-channel source separation, based on direction of arrival estimation. This method is then evaluated in Publication V for different speech codecs for a distributed scenario.

With the growing popularity of VUIs, it is common to find devices that provide VUIs everywhere around us. The service provided by each user interface could be improved if all devices in the same acoustic space collaborated together in a distributed manner. For example, distributed coding techniques could be used to provide higher quality in the communication of the speech signal or remove the dependency on only one device, allowing us to have more natural interactions.

Collaboration between multiple devices, however, means that the recorded signals have to be shared among multiple devices that might not all belong to the user. For that reason, it is important to define access management rules to ensure that only authorised devices collaborate in the VUI. In order to adapt the authentication method to the human perception, we propose to authorise devices within the same acoustic space, in such a way that devices that can hear the same voice, will be allowed to collaborate.

In this publication we analyse how acoustic fingerprinting can be used as a measure of proximity between recording devices. An acoustic fingerprint encodes features from recorded audio that allow us to estimate the similarity between the original audio segments. Additionally, the calculation of the fingerprint makes it hard to recover the original audio information, thus protecting the privacy of the user.

It has been shown that a combination of acoustic fingerprinting and fuzzy cryptography provides secure authentication for devices. These methods, however, require that both recorded signals are well synchronised, and delays due to the position of the devices or the network can degrade the performance of the authentication. To compensate for possible delays,
previously presented methods use long non-overlapping windows for the time-frequency transformation of the signal, which translates into very long recordings to achieve a fingerprint with enough data.

We present a set of decorrelating transforms, combined with quantisation methods to calculate the corresponding binary fingerprints. The proposed methods calculate a transformation that diagonalises the covariance matrix of the corresponding audio features. In order to maintain similar dimensions to the methods used as reference, the implemented methods aim to generate a feature matrix with 32 frequency bands and 32 time frames. This data is then quantised to generate a fingerprint with the same length as the reference methods. These methods provide higher accuracy and robustness to delays, while allowing us to significantly reduce the recording time by using shorter overlapping windows. The proposed fingerprints are also evaluated from a cryptographic perspective, as the produced binary sequences need to resemble random sequences of bits.

5.2 Publication II: Robust and Responsive Acoustic Pairing of Devices Using Decorrelating Time-Frequency Modelling.

Figure 5.2. Similarity of fingerprints between distinct microphones with varying delay in an office environment. Original, proposed fingerprinting methods with 200 ms and 50 ms overlapping windows, as well as comparison to microphones that do not match for original and proposed.

Multi-channel audio processing for communications in mobile devices can be achieved with the collaboration of multiple devices located in the same acoustic space, thus improving the transmitted sound quality. Therefore, a secure and responsive method for authenticating devices in the same acoustic space would be necessary to protect the privacy of the users. Previously
proposed methods show that a combination of acoustic fingerprinting and fuzzy cryptography can be used to compress the information of recorded audio in multiple devices and securely allow collaboration when this information is similar enough. While multiple techniques are used to extract acoustic fingerprints in the field of music retrieval, only the presented reference method has been found suitable to be used in an authentication application. For that reason, this method has been chosen as reference on which the proposed improvements are applied.

These methods, however, rely on the synchronisation between the two recording devices to calculate the corresponding fingerprint pair because no additional information should be shared between unauthorised devices. This means that delays in the network or the physical delay of the recorded signals will degrade the estimation of the fingerprints. Previously proposed methods deal with this problem by using long non-overlapping windows, considering that long windows will contain enough information to compensate for small delays. Additionally, the spacing between the windows provides a certain degree of decorrelation that will benefit the cryptographic properties of the fingerprint. This, however, requires very long audio segments to generate a fingerprint with enough information.

In this publication we study the effect of reducing the size of the time-frequency windows and overlap on the quality of the fingerprint. Additionally, we observe how synchronisation delays degrade the performance of the fingerprint, and propose a new fingerprint calculation that provides higher accuracy, allowing for shorter recordings with overlapping windows. We propose a decorrelation transformation over the frequency bands of the signal’s spectrogram, maintaining the difference calculation over time frames. This method provides a lower degradation when delay is introduced between two signals, and manages to obtain a higher accuracy when signals are synchronised even in cases where the window lengths are considerably reduced. The effect of SNR with additive noise is also studied, however, this is not a realistic scenario and we cannot consider it a suitable analysis of the fingerprint’s performance.

5.3 Publication III: Sound Privacy: A conversational Speech Corpus for Quantifying the Experience of Privacy.

In a conversation, humans perceive their surroundings and adapt the way they talk depending on the perceived level of privacy. This perceived level of privacy can be affected by many factors, like the content of the conversation, the conversational partner or the environment in which the conversation is held. If our devices could analyse and estimate how we would perceive the privacy in an environment, they could then adapt elements of their behaviour depending on it.
Analysing the content of a user’s conversation might already represent a violation of privacy, which is also the case for evaluating the relationship between the user and their conversational partner. Among the considered factors that affect the perception of privacy, the environment is the least intrusive for the user. For that reason, we propose to study the environment in which an interaction takes place, hence requiring a minimum use of the subject’s private data.

In this publication, we simulate multiple scenarios in which subjects would have conversations and recorded a speech corpus with segments of these conversations. Additionally, the subjects were asked to pay attention to their surroundings while interacting with their partner, and they answered a set of questions to measure how they perceived the privacy in their environment.

In the recording sessions, the subjects are taken to each of the studied locations and their conversations are recorded with mobile devices to simulate a realistic setup and to facilitate mobility between the scenarios. Considering that the content of the conversation will affect the perception of privacy, and sensitive content depends on the perception of each individual, the simulated conversations are constructed with random sentences that should remove the content from the interactions.

The recording sessions were held in two different countries (i.e. Finland and India), having 10 and 40 subjects respectively. The subjects were asked to simulate a conversation in the proposed environments, using random nonsensical sentences to remove bias due to the speech content. The subjects then answered a series of questions about the level of privacy that they perceived in the environment, and an online questionnaire was built to analyse the perception of subjects only using audio stimuli.

The published recordings contain a set of conversational scenarios in real noisy environments used to study the perceived level of privacy in an envi-
vironment. Additionally, these recordings can be used as conversational data in noisy environments for other applications, such as speech enhancement, as long as they remain within the limits set by the GDPR agreement.

A clear distinction between the different environments has been observed, and we can assume that the environment has an effect on the perception of privacy. However, we can also see a clear difference between countries as a result of the perception of privacy in each specific society. Variability is also observed between the subjects in a specific country, which may be caused by each subject’s personal background. Additionally, we can observe different correlations between online and offline questionnaires depending on the environment. These results show that quieter environments, such as offices, are perceived differently due to the lack of acoustic information, which would favour the influence of other features such as visual cues.

5.4 Publication IV: Cancellation of Local Competing Speaker with Near-field Localisation for Distributed Ad-Hoc Sensor Network.

![Figure 5.4. Architecture of the soft-mask calculation for cross-talk cancellation.](image)

When multiple people interact with their VUIs in the same environment, privacy violations can take place if the voice of a user leaks into the conversation of another user. For that purpose, we propose that devices collaborate to cancel the speech from interfering speakers, hence protecting their privacy.

In this publication we propose a method to reduce the crosstalk in a multi-speaker scenario, where the interfering speaker is removed in order to protect their privacy. This method is based on the DOA estimation of the dominant signal in the system. Traditional microphone arrays assume that the sensors are in a fixed known position with respect to each other and the incoming signals are in far-field conditions. However, this condition cannot be met in a WASN, where the devices are located around a space in arbitrary positions and the speakers’ distances are much smaller, hence
working in near-field conditions.

The experimental setup simulates a pair of microphones in a room. Two speakers are simulated as if they talked directly into each of the devices respectively, being each speaker dominant in their corresponding microphone. The positions of the dominant speaker defines the target DOA for an analysed device, that would use the recordings from both microphones. The range of possible DOAs is then limited by the fact that the speaker needs to remain dominant in the analysed device. This information is used to compensate ambiguities in the DOA estimation.

The proposed method combines a series of sparse autoencoders (AE) that extract high-level features from the input data, then estimates the DOA corresponding to each frequency bin and generates a softmask that removes those frequency components coming from directions not corresponding to the dominant source of the system. In order to improve the efficiency of the system, each autoencoder works on a subset of frequency bands. The proposed system provides higher signal to interference ratio compared to the reference method, however, additional distortion is observed in the signal due to the frequency masking.

5.5 Publication V: Speech Localization at Low Bitrates in Wireless Acoustics Sensor Networks

![Figure 5.5. System structures for DOA estimation in a distributed scenario.](image)

In a multi-device scenario, where devices collaborate together forming a WASN, it is likely that the signals recorded in some microphones need to be transmitted to the device that performs the multi-channel processing. The encoding and decoding process could introduce additional noise into the signals, hence degrading the features used by the multi-channel application. This could then translate into a degradation of the method’s performance.

In this publication we propose a CTC scenario based on source separation methods that rely on the estimation of the DoA of a dominant source to separate it from the mixture, as presented in Publication IV. Using this method, we analyse the effect of multiple speech codecs on the estimation of the DoA. The codecs used for the evaluation of the proposed method
are the state-of-the-art EVS and Opus, the PyAWNes-codec presented in Publication VI, which is adapted for WASNs, and Lyra, a codec based on a generative deep learning model.

The DNN-based DoA estimation method is trained using the PyAWNes codec at different bitrates and with clean data without encoding. The effect of quantisation noise in the estimation is then studied with each of the proposed codecs in different bitrate settings, and considering different codecs for each communication channel. Note that Lyra’s only available configuration is 2.3 kbps, and therefore is presented as an extreme use case for this method. We observe that in coding conditions, the optimum solution is to train the estimator with the PyAWNes codec instead of clean data, and as expected, the quality of the DOA estimation is reduced when lower bitrate settings are used in all the codecs. Additionally, the results show a better performance when both channels are encoded rather than maintaining one clean channel.


![Graphs showing PSNR, PESQ, and STOI across bitrates (8, 9.6, 13.2, 16.4, 24.4 and 32 kbps).]  

**Figure 5.6.** Single channel quality across bitrates (8, 9.6, 13.2, 16.4, 24.4 and 32 kbps); dashed red and dotted blue lines express the median of EVS and proposed codecs, respectively, and the corresponding filled areas their 95% quantiles.

When multiple devices record the audio in the same acoustic space, it is possible to combine all the corresponding signals to increase the quality that only one device could achieve. For example, the information from multiple channels can be used to remove interfering speakers from an interaction or efficiently encode the speech of a user. While traditional services require that the user interacts directly with the device that provides such service, speech coding method adapted for WASN architectures would allow the users to be undependable of their devices while the network would take care of sending the recorded signals to the decoder.
This publication proposes a speech coding method for WASNs that encodes multiple channels of audio recorded in multiple devices and combines them to provide the best output quality. Considering that most mobile devices in a WASN would have limited computing power, and they are probably running other applications in the meantime, the proposed method transports part of the computational complexity to the decoder where the signals will be merged, unlike traditional speech codecs. A frequency-domain codec is presented to efficiently encode the signals from the different channels on the recording side, and the decoder is combined with post-filtering methods to recover the signal with satisfactory performance.

Additionally, a distributed speech coding method needs to be able to compete with the quality of single-channel codecs when the multi-device scenario is not available. For that reason, the proposed method is also evaluated in single-channel mode with respect to other state-of-the-art codecs. The presented method provides a similar quality to the compared methods, hence enabling its use in cases where the WASN is not available.

5.7 Publication VII: Perception of Privacy Measured in the Crowd - Paired Comparison on the Effect of Background Noises.

When we interact with other humans, we modify the way we share information depending on multiple factors, such as the content of our message, the trust that we have in our conversational partner or the environment that surrounds us. In other words, our perception of privacy is affected by these factors, and that is reflected in the way we talk. If our electronic devices could analyse our perception of privacy in specific environments,
they could adapt how they share information in a similar way as humans do.

Following the results of Publication III, this publication analyses the acoustic characteristics of the different proposed environments, and extends the study of perception of privacy. Additional samples are generated using the acoustic properties from the different analysed environments and a crowd-sourcing study evaluates the perception of a larger number of users. The results of this study show that the environment does indeed affect the perception of privacy of the speakers in a conversational scenario.

5.8 Publication VIII: Provable Consent for Voice User Interfaces

The authentication of devices using recorded audio data allows VUIs to recognise which devices are located in physical proximity. This will allow us to establish a network of nearby devices in which we can trust to build a WASN. This authentication, however, needs to be performed without disclosing information about the recorded audio and needs to be robust to possible noise added by the different positions of the devices or the specific recording hardware.

By generating a fingerprint of the recorded audio and combining it with a fuzzy commitment scheme, it is possible to generate identical keys in devices that record audio in the same area, allowing a certain number of errors in the acoustic fingerprint. If one of the devices is located outside of this area of trust, the generated fingerprints will significantly differ, hence calculating different keys and failing the authentication.

In this publication we study the effect of proximity between devices, as well as the modifications in the voice level of the speakers. The recording speakers were asked to interact in different positions in a room, with multiple noise conditions. These interactions are recorded simultaneously.
from five different positions. The performance of the authentication process is evaluated with respect to these conditions, showing how the similarity of the fingerprints decreases as the distance between sensors increases.
6. Conclusion

Devices that can provide more natural interactions with their users are more easily adopted. For that reason, user interfaces are in continuous development to provide such natural interactions. It is not surprising then how voice user interfaces have surged in the last years, as they allow users to communicate with their devices in a natural way using their voices. While more devices around us incorporate an embedded microphone and provide some kind of voice-operated user interface, they require the user to stay in close proximity to interact with the user interface. Since a growing number of devices have the ability to provide a voice user interface, it is likely that several of such devices can be found simultaneously in the same acoustic space. These devices could then collaborate, forming a wireless acoustic sensor network, that could improve the individual services provided by each device. For instance, by allowing members of the network to provide the services of individual devices, the user could freely interact with the VUI, without having to focus on a specific sensor. However, multiple devices recording and sharing our voice increases existing, and raises multiple new, privacy concerns that need to be considered in the design of a distributed voice user interface. Speech is our most natural method of communication, and as such, we intuitively modify the way we interact using our voices depending on the level of privacy that we perceive. Analysing the effect of privacy in speech interactions is crucial to develop methods that can adapt to our natural perceptions to protect our privacy.

Privacy is a wide concept that can be applied, and simultaneously defined, differently in every aspect of our lives. The concept of privacy is highly dependent on our perceptions and it is unique for every human. Privacy protection rules can then be seen as the limit where an action would be considered a violation for an individual, however, the perception that a person will have of their own privacy will change depending on the situation, their environment or their interactions. For that reason, in order to design services that protect the privacy of the users, it is necessary to understand the individual perceptions of each user. If our devices recognised the features that modify a user’s perception, they could adapt.
their behaviour according to the user’s natural reactions. The privacy of the user will then be protected as the device can choose to share less information in more private scenarios, while being more accessible when less protection is required.

The first step then to protect the users’ privacy in voice user interfaces is to understand how that privacy is perceived. In a conversational interaction it is possible to define which aspects affect this perception of privacy. From observation of speech interactions we can distinguish three significant aspects that influence the perception of privacy of a speaker. These are the content of the message, the familiarity with the conversational partner and the environment in which the conversation takes place. The studies carried out in Publications III and VII have chosen the analysis of the environment as the least intrusive feature, making it the easiest to analyse from the perspective of preserving the user’s privacy. In these studies, we have observed that the environment does indeed affect the perception of privacy of the speakers and acoustic features of it could be analysed to define such changes in perception.

Open questions remain as to how specific features of the environment affect the perception of privacy that humans have in their conversations. Further analysis would be necessary to define the features that our devices can extract in order to recognise the level of privacy perceived in an environment. Additionally, more intrusive features of spoken communication could be analysed for a more accurate estimation of the privacy in a specific interaction. For example, the prosodic changes in our voices or the content of the conversation could have a great influence in the perceived privacy. It is debatable though, if just collecting these features implies any threat to the privacy of the users, and dealing with them should be done with precaution.

Considering that we live surrounded by devices with the ability to record and transmit our voices, it would be possible for them to collaborate in a wireless acoustic sensor network and provide services in which a single device would be limited. For example, multi-channel speech processing methods could be used to improve the quality of the audio signal, or a service from an individual device could be provided by any member of the network, thus detaching the user from a specific node. Based on the variable perception of privacy that humans have in their spoken interactions and how they modify their voice accordingly, methods have been developed to authenticate devices in close proximity, thus defining an area of “trust” in which a distributed voice user interface could operate. This authentication process is carried out in Publication VIII using a combination of acoustic fingerprints and a fuzzy commitment scheme, generating keys that can be used in secure communication. Publications I and II explore the effectiveness of acoustic fingerprints under the effects of communication delays and propose robust methods based on decorrelating transformations.
that provide an improvement in the authentication performance.

Once devices are connected and working as a wireless acoustic sensor network, we propose different applications that can exploit a multi-device scenario for both improving the quality of existing services and protecting the privacy of the users. Distributed speech coding provides a solution that allows multiple devices to encode the signals they record and efficiently send them to a decoder device that will merge them and reconstruct the target audio signal. Publication VI presents a speech and audio coding architecture optimised for distributed processing on wireless acoustic sensor networks. Additionally, the proposed codec provides competitive audio quality in single-channel mode. In order to preserve the privacy of external speakers in a conversation, the problem of cross-talk cancellation is studied in Publication IV as a source separation problem. The proposed method estimates the direction of arrival of the dominant source, calculating a time-frequency mask to extract the target speaker. In order to estimate the direction of arrival of the dominant source, the signals will need to be transmitted to a central device that processes them jointly. The effect of quantisation noise in this estimation is analysed in Publication V, evaluating the performance of the estimation under the effect of different speech and audio codecs. Methods based only on the statistics of the signals remain to be explored as they could provide a more robust solution against communication delays and quantisation noise. Additionally, methods that exploit the wireless acoustic sensor network architecture can provide more efficient solutions via the distributed processing of the signals and potential protection of the user's privacy by avoiding the sharing of the complete audio signal.

The studied methods have been analysed considering that they could complement each other. Therefore, in future lines of study, the implementation of these methods could be studied as a unified application. Understanding the effect on the environment on our perception of privacy would allow us to define an automatic measure of such privacy. This measure could be used to set a dynamic threshold in authentication task. Additionally, the information shared in the acoustic fingerprints used in the authentication process could be exploited in coding tasks, considering that such information would already be shared and common to multiple devices.

In conclusion, distributed voice user interfaces can provide an improvement in the user experience with our electronic devices by detaching us from the physical sensors and providing unified services with improved quality. This thesis provides an analysis of the privacy concerns that need to be considered when dealing with the users’ voice and proposes methods that exploit this perception of privacy to build a distributed voice user interface. Furthermore, methods for speech coding and speech source separation are proposed for the described multi-device architectures.


Voice user interfaces are increasingly becoming part of our daily lives to interact with electronic devices, from communications on our phones to voice assistants such as Siri or Alexa. Collaboration between devices could improve the services they provided, but if the data gathered by these applications was misused or an unauthorised user managed to obtain it, it would cause a grave violation of the user's privacy. It is then crucial to define which devices are actually allowed to record the user's speech and collaborate in a wireless acoustic sensor network. If our devices could perceive our privacy the same way as we do, they could adapt the information they shared to protect the personal data of the users.

In this thesis, we study methods to recognise when two devices are located in the same acoustic space based on the audio signals that they record. We show how acoustic fingerprints can be used to securely share the audio information from a device and estimate the physical proximity of devices. We also generated a speech corpus in conversational scenarios to analyse the effect that the acoustic properties of the environment have on the perceived level of privacy. Finally, we developed source separation methods to remove the voice of interfering speakers in a multi-device scenario, thus protecting the privacy of external users.