



HAL
open science

Anonymizing Speech: Evaluating and Designing Speaker Anonymization Techniques

Pierre Champion

► **To cite this version:**

Pierre Champion. Anonymizing Speech: Evaluating and Designing Speaker Anonymization Techniques. Artificial Intelligence [cs.AI]. Université de Lorraine, 2023. English. NNT : 2023LORR0101 . tel-04218098

HAL Id: tel-04218098

<https://hal.univ-lorraine.fr/tel-04218098>

Submitted on 26 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



**UNIVERSITÉ
DE LORRAINE**

**BIBLIOTHÈQUES
UNIVERSITAIRES**

AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt une poursuite pénale.

Contact bibliothèque : ddoc-theses-contact@univ-lorraine.fr
(Cette adresse ne permet pas de contacter les auteurs)

LIENS

Code de la Propriété Intellectuelle. articles L 122. 4

Code de la Propriété Intellectuelle. articles L 335.2- L 335.10

http://www.cfcopies.com/V2/leg/leg_droi.php

<http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm>

THÈSE DE DOCTORAT

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Lorraine

MENTION : INFORMATIQUE
ÉCOLE DOCTORALE N° 77

Par

Pierre Champion

**Anonymizing Speech: Evaluating and Designing
Speaker Anonymization Techniques**

Anonymisation de la parole : Évaluation et Conception
de Techniques d'Anonymisation du Locuteur

Thèse présentée et soutenue à Nancy, le 20/04/2023

Unités de recherche :

Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA) - UMR 7503
Laboratoire d'Informatique de l'Université du Mans (LIUM) - EA 4023

Encadrants :

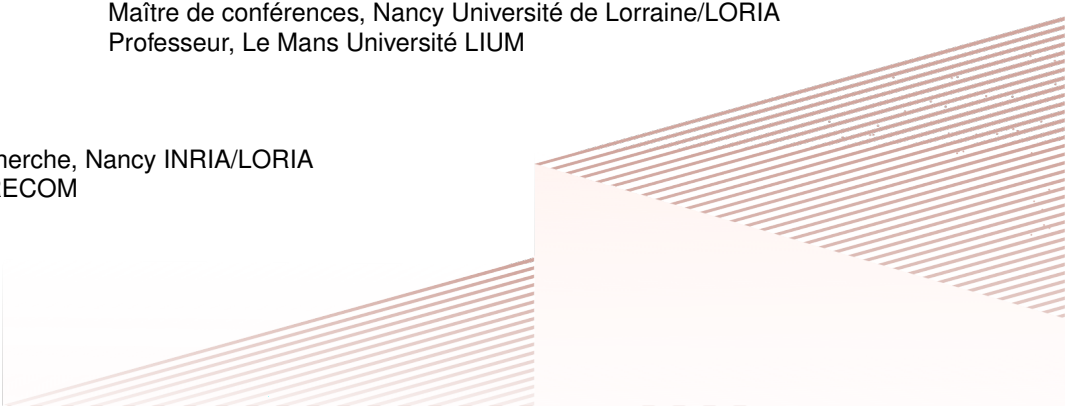
| | | |
|--------------------|-----------------|---|
| Dir. de thèse : | Slim Ouni | Maître de conférences, Nancy Université de Lorraine/LORIA |
| Co-encadrant : | Denis Jovet | Directeur de recherche, Nancy INRIA/LORIA |
| Co-dir. de thèse : | Anthony Larcher | Professeur, Le Mans Université LIUM |

Composition du Jury :

| | | |
|---------------|------------------------|---|
| Président : | Lori Lamel | Directrice de recherche, Université Paris-Scalay |
| Rapporteurs : | Luciana Ferrer | Chargée de Recherche, University of Buenos Aires |
| | Lukáš Burget | Associate professor, Brno University of Technology |
| Examineurs : | Jean-Francois Bonastre | Professeur, Université d'Avignon |
| Encadrants : | Slim Ouni | Maître de conférences, Nancy Université de Lorraine/LORIA |
| | Anthony Larcher | Professeur, Le Mans Université LIUM |

Invités :

| | |
|----------------|---|
| Denis Jovet | Directeur de recherche, Nancy INRIA/LORIA |
| Nicholas Evans | Professeur, EURECOM |



Un hommage chaleureux, six ans plus tard, à la mémoire de Suzanne.

ACKNOWLEDGEMENT

Ce n'est pas tous les jours que nous disposons d'un moment pour remercier les personnes de ce que nous sommes devenus. Je saisis cette occasion pour leur adresser ma gratitude car elles ont joué un rôle dans le fait que je sois arrivé à ce moment de ma vie, à la fois personnelle et scientifique, Merci, je n'en serais pas ici sans vous.

Pour commencer, je remercie Luciana Ferrer et Lukáš Burget, relecteurs de cette thèse, pour le temps qu'ils ont accepté d'y consacrer et pour les remarques pertinentes et précises qui ont été formulées, qui m'ont permises d'affiner davantage ce document et mes connaissances. Je remercie aussi Lori Lamel, Jean-François Bonastre et Nicholas Evans pour avoir accepté de faire partie de mon jury et pour les questions et commentaires au cours de la soutenance.

Je souhaiterais particulièrement remercier Denis Jovet et Anthony Larcher, mes encadrants, de m'avoir donné cette opportunité et de m'avoir prodigué des conseils précieux tout au long de cette thèse. Nos réunions ont toujours été à la fois excessivement bénéfiques pour mon travail et très agréables. Je tiens aussi à remercier tous les enseignants, depuis le début avec Sylviane Viviant en maternelle, qui ont pris le temps de m'apprendre à lire/écrire (tâche qui s'est avérée bien difficile), et à compter, pour plus tard donné le goût à la science.

Faire partie du Loria/Inria-Nancy a été une formidable expérience. Pour les bons moments passés et les interactions enrichissantes, je tiens à remercier les membres de ce centre de recherche. En particulier, mes remerciements les plus sincères vont à Noémie et Adrien, mes amis et collègues de bureau. J'ai eu la chance de tomber sur vous, et je n'aurais pas pu rêver de meilleures personnes avec qui partager mon travail au quotidien, même si nous parlons plus que nous ne travaillons par moment. Je tiens à remercier chaleureusement tout le personnel du Laboratoire d'Informatique de l'Université du Mans pour leur accueil bienveillant, alors même que ma présence était très sporadique.

Je tiens à exprimer ma gratitude à tous mes amis, du Mans de l'escalade et du lycée, de Nancy. Leur présence dans ma vie a été une source de joie et de soutien inestimable. Leurs rires, leurs encouragements et leur amitié sincère ont illuminé mes journées et m'ont aidé à surmonter les défis. Je suis reconnaissant d'avoir des personnes si merveilleuses à mes côtés.

Je tiens également à exprimer ma profonde gratitude envers ma famille, qui est un pilier solide dans ma vie. Leur amour inconditionnel, leur soutien constant et leurs valeurs transmises ont façonné la personne que je suis aujourd'hui. À travers les hauts et les bas, ils ont toujours été là pour moi, prêts à tendre la main et à m'offrir leur épaule réconfortante. Je souhaite qu'ils sachent à quel point ils sont importants pour moi. Merci du fond du cœur.

TABLE OF CONTENTS

| | |
|---|-----------|
| Introduction | 1 |
| Motivation | 1 |
| Scope and objectives | 3 |
| Organization and main contributions of the thesis | 4 |
| | |
| I State of the art | 7 |
| | |
| 1 Speech and neural networks | 9 |
| 1.1 Introduction | 9 |
| 1.2 Speech production | 10 |
| 1.3 Speech processing | 12 |
| 1.4 Artificial neural networks | 14 |
| 1.5 Conclusion | 23 |
| | |
| 2 Speech-centric machine-learning | 25 |
| 2.1 Introduction | 25 |
| 2.2 Automatic speech recognition | 26 |
| 2.3 Automatic speaker recognition | 34 |
| 2.4 Voice conversion | 41 |
| 2.5 Conclusion | 48 |
| | |
| 3 Speaker anonymization | 49 |
| 3.1 Introduction | 49 |
| 3.2 Legal perspectives | 50 |
| 3.3 Threat model | 51 |
| 3.4 Application cases | 55 |
| 3.5 Evaluation methods | 56 |
| 3.6 Current methods for speaker anonymization | 60 |
| 3.7 Conclusion | 66 |

| | |
|--|------------|
| II Contributions | 67 |
| 4 The role of the target speaker | 69 |
| 4.1 Introduction | 69 |
| 4.2 Impact of the target selection algorithm in privacy evaluation | 70 |
| 4.3 A quest for the golden target speaker | 82 |
| 4.4 Conclusion | 88 |
| 5 Voice conversion anonymization with feature-level disentanglement | 89 |
| 5.1 Introduction | 89 |
| 5.2 Fundamental frequency feature transformation | 90 |
| 5.3 Linguistic model transformation | 98 |
| 5.4 Linguistic feature transformation | 105 |
| 5.5 Conclusion | 115 |
| 6 New kinds of privacy and utility measurements | 117 |
| 6.1 Introduction | 117 |
| 6.2 Invertibility evaluation using embedding alignment | 118 |
| 6.3 Subjective mispronunciation evaluation | 127 |
| 6.4 Conclusion | 130 |
| Conclusion and perspectives | 131 |
| Résumé étendu en français | 137 |
| Bibliography | 143 |
| List of abbreviations | 160 |
| English abstract | 162 |
| French abstract | 163 |

INTRODUCTION

Contents

| | |
|--|----------|
| Motivation | 1 |
| Broader impact statement | 2 |
| Scope and objectives | 3 |
| Organization and main contributions of the thesis | 4 |

Motivation

Speaking and listening, or verbal communication is a natural and convenient way for people to interact. During conversations, individuals exchange linguistic information through speech, but they also transmit additional information, such as identity, emotions, age, and gender, through paralinguistic cues. Human-machine interaction can benefit from the richness and convenience of speech allowing the machine to better understand humans and humans to easily share information with machines. However, making the machine understand and reproduce speech remains a complex task that has been the subject of ongoing research for many years. Still, over the last decade, great progress has been made in several speech tasks such as speech-to-text (automatic speech recognition), text-to-speech (speech synthesis), and more. With these advancements came the introduction of a new product to the market: voice assistants.

The goal of many companies is to establish a seamless and natural communication experience between humans and machines powered by the latest developments in speech processing technology. Currently, consumers are embracing various voice assistant devices. The [2022 U.S. Smart Home Consumer Adoption Report](#) revealed that 50-60% of the U.S. population has access to one or many voice assistant devices.

In order for companies to propose competitive services, advanced forms of deep learning techniques are used to power voice assistants. Those algorithms are data-hungry meaning that the best performances are usually achieved when a large quantity of data is used to train the models. This created a necessity for companies to collect, process and store the speech data of their user in centralized servers to continuously improve the proposed services and remain competitive. As speech data contains a lot of personal information such as the identity of the speaker, the process of collecting speech data raises serious privacy concerns.

While the data collection practices of companies were initially unknown or not well understood by the public, this has recently changed. Around 2017, more and more people became aware

of the situation with press headlines¹ disclosing companies usage of the collected speech data. Simultaneously, the General Data Protection Regulation (GDPR) established the strictest privacy and security legislation in the world. The (GDPR 2016) specifically stipulates that European citizens personal information must be handled with the utmost respect and privacy.

Privacy is an individual’s right to keep confidential information or data private. It is the ability to control who can access information about oneself, and for what purpose it is used. In the context of the GDPR, privacy is referred to as a fundamental human right, and it is essential for protecting citizens from having their personal information misused or abused. In today’s digital age, privacy is of utmost importance as the growing use of technology enables massive amounts of personal data to be collected and stored. As technology continues to evolve, so must our understanding of the influence it has on privacy and our commitment to protecting it.

Speech is considered a highly sensitive type of personal data that must be protected. Recent guidelines provided by the French *Commission Nationale de l’Informatique et des Libertés* and European *Data Protection Board* recalls that speech data is inherently biometric data given Article 4(14)² of the GDPR (CNIL 2020; EDPB 2021). As the storage and processing of biometric data are even more regulated than that of personal data, it creates a necessity to develop more private data collection schemes.

With the above context about the state of speech and privacy, this thesis is an effort to propose more private speech collection solutions relying on data anonymization. Data anonymization is the process of removing or altering personal identifying information from data to protect the privacy of individuals. While the GDPR does not impose systematic anonymization to personal data, it is one solution, among others, that enables the processing of personal data in compliance with the rights and privacy of individuals. In this thesis, we work on speaker anonymization methods that aim to remove speaker identity from speech signals while preserving linguistic content and speech quality.

Broader impact statement

We believe that our work in the field of speaker anonymization is a new and important research domain in today’s society. With the increasing use of speech-enabled technology, there is a growing concern about the privacy and security of the speech of each individual. Speaker anonymization has the potential to be used in a wide range of applications, such as in call centers, virtual assistants, or any smart devices that share speech information, to better protect the privacy of individuals.

1. *Yep, human workers are listening to recordings from Google Assistant, including audio recorded by mistake.*
2. Article 4(14) GDPR defines biometric data as “personal data resulting from specific technical processing relating to the physical, physiological or behavioral characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopy data”.

Scope and objectives

In order for the widespread adoption of privacy-preserving techniques to occur in current data processing pipelines, one of the main requirements deals with the ease of use of the technique. Indeed, if the privacy-preserving technique to be implemented requires a complete modification of companies current data processing pipelines, it is likely that it will not be implemented. A few privacy-preserving techniques have been proposed in the last decade. They include encryption, distributed learning, and anonymization, in the following, we briefly present them to justify our choice to work with anonymization-based techniques. Encryption solutions (Brasser et al. 2018) rely on cryptographic methods to transform clear (personal) data before sending it to the server. The transformed data being sent is unreadable to any external observer, addressing security constraints. For privacy constraints, the server should not be able to decrypt the data, which is possible with homomorphic encryption, where for example the only operation possible for the server is a neural network inference (Pathak 2012) on the encrypted data. Distributed learning methods such as federated learning (Leroy et al. 2018) enable multiple data sources to be used in the same learning process, this allows the clear data to remain on the devices instead of being centralized. As a result, the clear data is not shared, allowing for more secure and private machine learning. Anonymization aims to remove personal information from the speech signal to improve privacy while reducing as much as possible the degradation of the other information necessary for other tasks (this is referred to as the utility).

The main advantage of encryption and distributed learning is that they are, today, the most effective techniques to ensure a sufficient level of privacy. However, their main disadvantage is that the data shared between the client and server is no-longer speech. In order for companies to implement such techniques, a complete re-work of the data processing pipeline is needed as there is a paradigm shift. Additionally, it is not practical for companies or research labs to create large speech datasets with those techniques as speech signals are not shared. For those reasons, anonymization techniques which take speech and produce privacy-preserving speech are quite interesting in the fact that they do not have an impact on current data processing pipelines.

Over the last few years, speaker anonymization techniques have recently received interest with the release of the VoicePrivacy Challenge. Most attempts rely on voice conversion systems to transform the identity of the speaker in the clear speech signal to another one in the anonymized signal. The goal is that the anonymized signal can no longer be linked back to the real identity of the speaker. To evaluate the privacy of an anonymization system, automatic speaker verification techniques are used on anonymized speech to assess its degree of linkability, the lower the better. In contrast, the preservation of the linguistic content (utility) is currently evaluated with automatic speech recognition.

The first challenge of speaker anonymization relates to privacy evaluation. Privacy evaluation depends on the target speaker identity parameter of voice conversion. Indeed, voice conversion

can increase linkability if requested by making a speaker’s voice highly different from others speakers. As such, there is a necessity to understand voice conversion such that the anonymized voices are not linkable and that the evaluation performed with the automatic speaker verification system corresponds to a privacy evaluation.

A second challenge of speaker anonymization relates to the voice conversion algorithm. Initially, voice conversions were designed to transform the speaker of a signal, such that subjective (humans) listeners believe a signal was spoken by someone else. While the quality of the transformation can be sufficient to fool humans, it is not the case for machines. As such, more advanced forms of voice conversion systems need to be used for speaker anonymization.

A third challenge relates to the diversity and number of possible evaluations to assess privacy and utility. Indeed there is not a single way to evaluate privacy and utility. For example, for privacy, a game involving an attacker and defender, where the attacker aims to break anonymization and the defender improves it has to take place to continuously adapt the technology to current threats. Whereas for utility, the evaluation always depends on the intended purpose of sharing speech. We focused on the preservation of the linguistic content, but there are many other valid application cases such as emotion recognition.

In summary, those three challenges give the objective of investigating the evaluation and design of speaker anonymization systems

Organization and main contributions of the thesis

After this introduction, three state-of-the-art chapters present the current level of development in the field of machine learning applied to the speech, then follows three chapters detailing our contributions. The remaining of the thesis is structured as follows:

- In Chapter 1 of the thesis, we introduce the fundamental of speech and how it can be categorized at different levels (e.g., semantic, prosodic, etc.). We also discuss various methods for numeric speech processing, and present artificial neural network architectures that recently revolutionized the field of speech processing.
- In Chapter 2, we focus on various speech-centric machine-learning disciplines that are necessary for building a speaker anonymization system. This includes automatic speech recognition, voice conversion, and objective evaluation methods using automatic speech recognition and automatic speaker verification.
- In Chapter 3, we delve into the specific topic of speaker anonymization including the current state-of-the-art techniques for removing personal information from speech signals, and the evaluation of these techniques. We also discuss the broader context of biometrics security and legislation.
- In Chapter 4, we focus on the role of the target speaker identity parameter in voice

conversion-based speaker anonymization. Through our analysis, we found that the way current voice conversion-based anonymization systems are parameterized with target speakers creates a bias in privacy evaluation. We also analyze the effect of the target speaker on privacy and utility, to answer the question of whether there is a golden target speaker parameter that maximizes performance. Overall, our chapter highlights the importance of the target for proper privacy evaluation and utility preservation.

- In Chapter 5, we delve deeper into voice conversion techniques for speaker anonymization and carefully analyzed the input features. We found through in-depth analysis that traditional methods for disentangling linguistic and fundamental frequency features from speaker information were not fully effective. We experimented with adversarial learning to improve the level of disentanglement, however, this method showed limitations when it comes to removing speaker information. As a solution, we introduce a new approach using vector quantization to disentangle these features. Our results indicate that this method significantly improved privacy performance while producing anonymized speech that sounded more natural compared to noise-based alternatives.
- In Chapter 6, we present new measurement metrics and techniques for evaluating the privacy and utility of speaker anonymization systems. Specifically, we propose an invertibility attack and measurement that assesses the ability to reconstruct clear x-vectors from anonymized ones. We also propose a new utility measurement that aims to better evaluate the preservation of linguistic content after anonymization by taking into account only mispronunciation errors, rather than inherent recognition decoding errors.

This thesis was funded by the ANR project DEEP-PRIVACY, which promotes the development of distributed, personalized and privacy-preserving approaches for speech recognition and Région Grand Est. We thank them for their financial support, which allowed us to conduct the research and experiments necessary to complete this work.

PART I

State of the art

SPEECH AND NEURAL NETWORKS

Contents

| | | |
|------------|---|-----------|
| 1.1 | Introduction | 9 |
| 1.2 | Speech production | 10 |
| 1.2.1 | Speech production mechanism | 10 |
| 1.2.2 | The different components and aspects of speech | 11 |
| 1.3 | Speech processing | 12 |
| 1.3.1 | Time-frequency processing | 13 |
| 1.3.2 | Short-term features | 13 |
| 1.3.3 | Long-term features | 14 |
| 1.4 | Artificial neural networks | 14 |
| 1.4.1 | Bottleneck layer | 16 |
| 1.4.2 | Residual network | 16 |
| 1.4.3 | Time delayed neural and convolutional neural networks | 16 |
| 1.4.4 | Transformer | 19 |
| 1.4.5 | Adversarial network | 20 |
| 1.4.6 | Generative adversarial network | 22 |
| 1.5 | Conclusion | 23 |

1.1 Introduction

This chapter introduces the very basic notions of speech, how it is produced and how it can be categorized at different levels ranging from low-level acoustics to high-level semantics. As this thesis aims to remove the personal speaker characteristics from the lower levels, we will discuss methods for numeric speech processing.

In speech processing, acoustic features are numerical representations of speech signals that can be extracted using algorithms hand-crafted based on years of research, or automatically using machine learning techniques. These features provide the foundation for further analysis and processing.

For this thesis, manual analysis of the acoustic features to determine and transform those that reveal personal speaker characteristics is impractical due to the convoluted nature of the

representations. Some works have experimentally identified acoustic features that are highly correlated with speaker identity (Memon 2020) or emotions (Eyben et al. 2015; Arias et al. 2020) but due to the high variability and convoluted properties of speech, making mechanistic conclusions is difficult. As such, we rely on machine learning-based analysis using artificial neural networks to determine and transform the acoustic features that reveal personal speaker characteristics. This chapter introduces the key concepts of artificial neural networks, deep learning and relevant network models for speech processing.

1.2 Speech production

Human communication is both a social and a cognitive process in which individuals exchange information through a standard system of codes and signs. Humans communicate to ask for help, inform others, and share attitudes to integrate within a group of individuals. The primary mode of communication among Humans is speech. Humans have evolved a complex set of voice, hearing, and cognitive abilities allowing them to express a sophisticated natural language (Hauser et al. 2002). In this section, we introduce the different components and aspects of human speech and detail some key notions associated with speech production and their respective properties.

1.2.1 Speech production mechanism

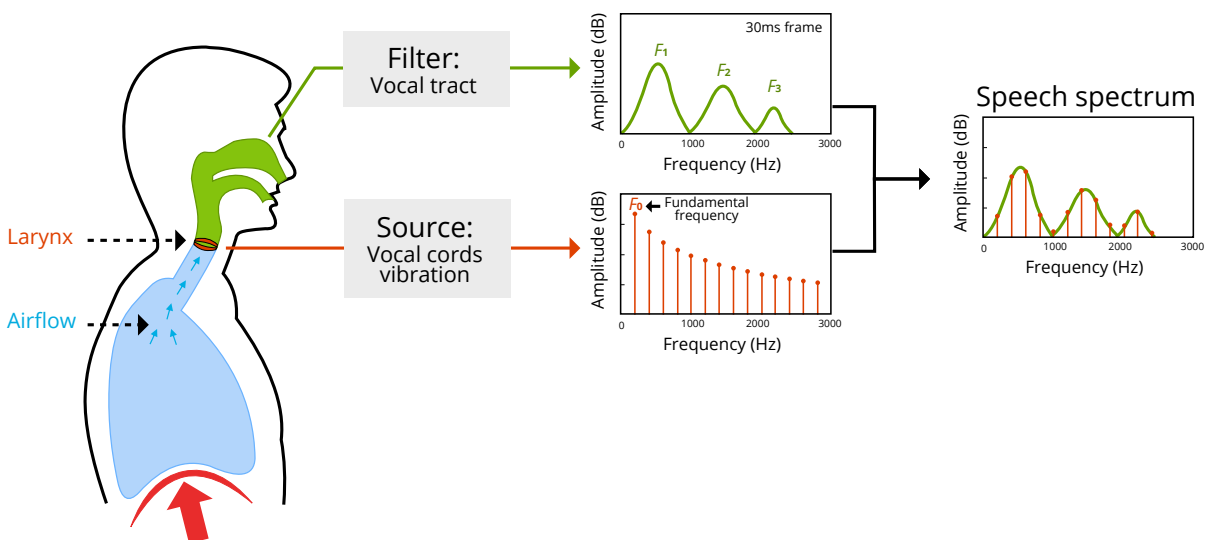


Figure 1.1 – The production mechanisms of voiced sounds as a source-filter model.

By definition, speech is an acoustic signal described by the modification of acoustic pressure over time. Figure 1.1 illustrates a schematic view of the primary organs involved in creating speech. Speech generation starts with an airflow from the lungs, which goes through the larynx and the vocal tract to exit the lips. The larynx contains two vocal cords, which vibrate rapidly

with the airflow to create the source of acoustic vibration. Sounds created with the vibration of the vocal cords are called *voiced* sounds, while sounds made without the vocal cords engaged are called *unvoiced* sounds. The vocal cord's vibrations are periodic and determine the sound's fundamental frequency (F_0). F_0 is expressed in Hertz (Hz), and the range of F_0 correlates with the speaker's gender. For the American population, the average F_0 values are around 120 Hz for men and around 210 Hz for women. (Hillenbrand et al. 2009).

Once the airflow exits the larynx, it is modulated by many organs in the vocal tract (for example, the tongue, the nasal cavity, etc.) to pronounce a multitude of speech sounds. The shape of the vocal tract acts as a resonator and behaves as a frequency filter on the source signal. This model of sound production is referred to as a source-filter model. The frequency peaks from the vocal tract's shape are called formants and are denoted as F_1 , F_2 , F_3 , etc.

1.2.2 The different components and aspects of speech

The capability of a Human to generate an acoustic signal is not the only requirement that defines speech. Speech uses acoustic signals as a transmission medium. Other aspects are essential to speech communication.

The comprehension process between a speaker and a listener requires both sides to have shared knowledge. First, a common understanding of the semantics of the words in their context must be known to deduce the meaning of the communication. Then, a mapping between words and their corresponding acoustic production and sounds has to be known. Lastly, words should be arranged to match a defined syntax using the correct grammar. In order to study speech, it is essential to break it down into several levels what constitutes speech. The following list inspired by (Hickey 2010; Schreiber 1991) decomposes speech into different aspects:

- **Semantics** concerns the content and meaning in language. Semantics analysis can be applied to entire texts or single words.
- **Syntax** refers to the arrangement of words, phrases, clauses, and punctuation, in a specific order defined by a set of rules.
- **Phonetics and phonology** are the study of speech sounds and their function in a given language. The production and classification of speech sounds according to their properties is the subject of phonetics, while their functions in a language are concerned with phonology.
- **Prosody** deals with how the semantic content is delivered. It is important to structure the message (similar to the role of punctuation in writing), to emphasize certain terms or expressions, and to indicate the type of the sentences (statement, question, orders, exclamations). (Mary 2018). It helps to make speech more natural by carrying information about emotions, speaker intention, etc.
- **Acoustics** deals with the physical properties of the speech signal, including F_0 , etc.

These levels allow humans to communicate with a very high degree of flexibility. However, Personally Identifiable Information (PII) can be extracted at different levels. At the acoustic level, the voice characteristics of a speaker vary with the shape and size of the organs of the vocal tract allowing recognition of physical traits, and voice disorders (e.g., vocal cords paralysis, laryngeal cancer) (Cummins et al. 2018). At the prosody level, a speaker’s manner of expression contains a rich array of PII, including cues to the identity of the speaker, personality, emotions, age, gender, and many others (Jain et al. 2000; Kröger et al. 2019). The syntax level contains information about potential language disorders (e.g., dysphasia, underdevelopment of vocabulary, or grammar). At the semantic level, speakers may disclose private information, such as phone number, person’s name, etc., in the linguistic content.

This thesis focuses on removing speakers PII from the prosody and acoustic levels while keeping the linguistic content intact. The other privacy aspects also need to be considered for complete privacy, such as removing phone numbers, person names, etc., from the linguistic content. Still, they are out of the scope of this thesis. The following section explains the basic processing blocks used to remove and evaluate the PII contained in the prosody and acoustic levels.

1.3 Speech processing

Speech or any other sound is composed of acoustic waves traveling through the air. When recorded with a single microphone, acoustic waves are represented as the modification of pressure at a single point in space over time. The microphone transforms acoustic waves into an electrical signal that is sampled and converted to a digital signal in order to be processed by computers. This electrical signal is called a speech signal, and can also be referred to as a waveform or a time-domain signal. A speech signal is, by nature, a convoluted, redundant, and highly variable signal, which makes it difficult to process directly. Therefore, extracting a more compact representation that reduces redundancy is necessary.

The focus of this section is to introduce the process of creating traditional hand-crafted speech features. Recent trends in research show the progressive abandoning of this step thanks to the increase in computational power and a new research field called *Self-supervised Learning* (Liu et al. 2022). This aspect will later be explored in Chapter 5 of this thesis.

The ideal properties of the speech features should be:

- relatively easy to extract
- resilient to background noise and transmission channel
- adapted to the downstream task, they should embed the information required while removing other information and noise

Satisfying all of these conditions simultaneously is difficult to accomplish in practice. However, these criteria are considered idealistic design goals for speech processing.

1.3.1 Time-frequency processing

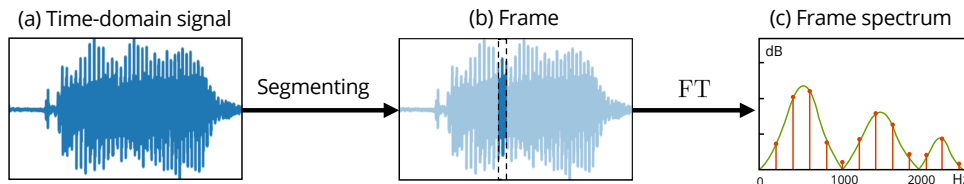


Figure 1.2 – The process of extracting a speech frame spectrum.

Traditional speech features are extracted by converting the time-domain signal representation to its frequency-domain representation with the help of the Fourier transform. The Fourier transform ([Fourier et al. 1822](#)) decomposes a complex, non-sinusoidal speech signal into a series of sinusoidal sub-signals. Figure 1.2 shows an example of the application of the Fourier transform on a small speech segment (called a frame) illustrated by a spectrum.

A time-frequency representation of the whole speech signal is obtained by segmenting the source signal into overlapping frames of fixed length (usually 10 to 30 milliseconds). Then, for each frame, a window function is applied that brings down, close to zero, the values near the edge of the window. Lastly, a Fourier transform is applied on each frame independently, revealing their corresponding spectrum response. This analysis is known as the Short-time Fourier Transform (STFT). Plotting the changing spectrum response as a function of time generates a spectrogram plot. An example of a power spectrogram is shown in Figure 1.3 (b).

1.3.2 Short-term features

Short-term features were initially developed in the 1980s for speech recognition. They describe traits of human speech that are assumed to be stationary inside very short intervals. At the time, the modeling approaches and available computational power were much more limited. As such, low dimensional features were interesting as they require less computational effort for the subsequent processing, usually based on statistical models that could not handle very well high-dimensional data ([Zeghidour 2019](#)).

One of the most popular low-dimensional short-term features are the Mel Frequency Cepstral Coefficients (MFCCs). To compute MFCCs, first, a power spectrogram is obtained by taking the squared magnitude of the STFT of the signal. Having the STFT power spectrogram, the next step is to compute the energy in predefined frequency bands. For MFCCs, the bands are arranged according to the Mel scale that mimics the frequency resolution of the human ear ([Shannon et al. 2003](#)). The filterbank coefficients can be visualized on a linear scale; however, this is not very informative for the human eye. A logarithmic scale improves the quality, as

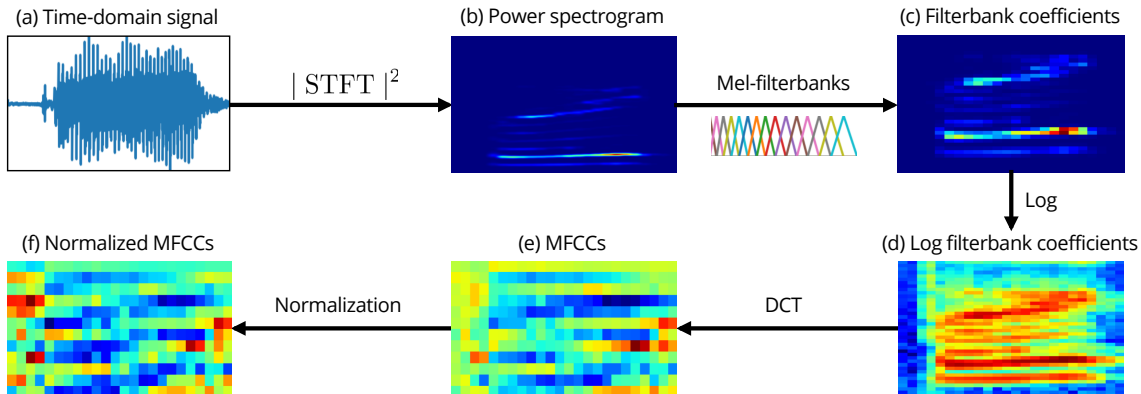


Figure 1.3 – The process of computing Mel Frequency Cepstral Coefficients (MFCCs).

observed in Figure (d). The filterbank coefficients computed in the previous step are highly correlated, which could be problematic when modeling their behavior statistically. Therefore, the Discrete Cosine Transform (DCT) function is applied to decorrelate them and returns the MFCCs representation. Finally, the MFCCs are normalized to lessen the influence of noise. Cepstral mean normalization compensates for convolutional noise, while variance normalization (over time) compensates for additive noise. With the advent of deep learning-based modeling systems, MFCCs are not the primary features anymore. Deep neural networks are less susceptible to highly correlated input, making the log filterbank coefficients or even the time-domain signal, which contains more information, a more relevant feature today.

1.3.3 Long-term features

Short-term features in speech or audio signals are gleaned from a brief interval of 10 to 30 milliseconds, while long-term features are drawn from a much longer period, ranging from 100 milliseconds to the full duration of an utterance - which can be a single word, a phrase, or an entire sentence. They represent voice characteristics over longer intervals, for example, the mean and standard deviations of the F_0 , or the F_0 trajectory. When the F_0 is analyzed jointly with a spoken word, its variation models the prosody. Long-term features encode speaker-related information as the speaker identity information in a signal is considered stationary for the utterance.

1.4 Artificial neural networks

Previously, we introduced hand-crafted features as a means of representing speech through numerical values. In this section, we will delve into the fundamental machine learning concept that utilizes these features as inputs to perform the analysis. The field of artificial intelligence (AI), and more specifically machine learning, aims at teaching a computer to learn to perform a

task for which it is not directly programmed. It relies on mathematical approaches and data exploitation. In this thesis, we utilize machine learning approaches to let the computer find adequate speech transformations that anonymize speech. Interestingly, and more importantly, the evaluation procedure comprises a set of other machine learning methods that objectively assess the privacy and utility of the transformed (e.g., anonymized) speech signal.

Machine learning encompasses a multitude of methods, and among them are Artificial Neural Networks (ANN), on which deep learning is based. In recent years, ANN have become an indispensable tool for tackling a wide array of real-world problems. At a fundamental level, ANN comprise many interconnected units called neurons. The first formal neuron model was proposed in 1943 by Warren Sturgis McCulloch and Walter Pitts, inspired by the working of biological neurons. It was then in 1958 that Frank Rosenblatt set up a learning algorithm applicable to an artificial neuron, leading to the creation of the perceptron (Rosenblatt 1958).

The concept of the Multilayer perceptron (MLP) (Rumelhart et al. 1987) is to organize several simple perceptrons in such a way that the output of one neuron is the input of one or several other neurons. Multiple neurons can be arranged side by side and connected to multiple perceptrons in a subsequent layer. For example, the MLP in Figure 1.4 has two hidden layers. In this example, the outputs of the neurons of a given layer are connected to the inputs of the neurons of the next layer. The information propagates from the input layer to the output layer; therefore, it is called a feedforward neural network.

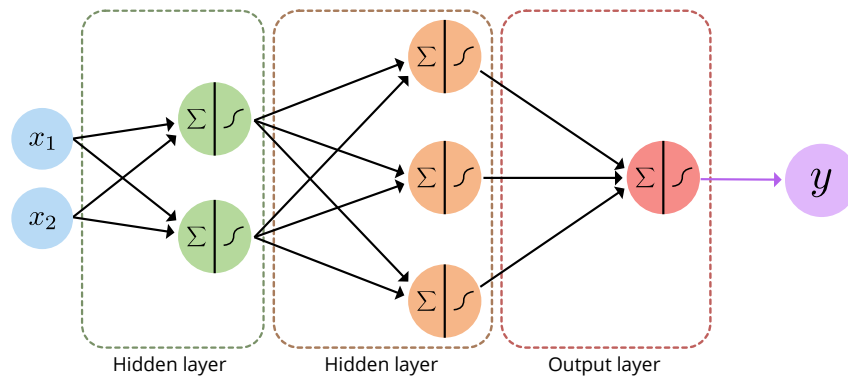


Figure 1.4 – Feedforward multilayer perceptron.

The non-linear activation functions of the neurons are critical. They allow the network to learn a complex mapping between the input and the output. Without non-linearity, the outputs would be linear function of the input, resulting in an inability to learn patterns that are more complicated and potentially more meaningful.

A Deep Neural Network (DNN) is a neural network architecture having many hidden layers. The architecture in Figure 1.4 can be considered a simple DNN, more complex architectures usually have more than a dozen layers. The number of hidden layers defines the depth of the network.

In this thesis, we employ DNN models to process acoustic features for various tasks. For a long time, the training methods for this type of neural network did not allow to achieve good performance. However, major advances in training methods and network structures have enabled the use of bigger and larger neural networks. This section introduces fundamental neural network architectures and concepts that allow efficient modeling for many speech-related tasks.

1.4.1 Bottleneck layer

A bottleneck layer in a neural network is a layer with fewer neurons than the layers before or after. Such layers encourage the network to compress its internal representations of the input (He et al. 2016b). A bottleneck representation can be extracted from the layer activations and used as a high-level representation of the input information.

1.4.2 Residual network

Upon training, the backpropagation calculates the partial derivatives of the error from the output to the first hidden layer. Because of the chain rule, the first layers have much more non-linear activation functions on the way to compute their derivatives than the output layer. If the derivatives are small the gradient will decrease exponentially as the error propagates backward until it eventually vanishes. Vanishing gradients creates an issue as some weights of the network will not change throughout training. As the number of layers in a deep neural network increases, the weights in the earlier layers are not effectively used as they are not effectively updated because of the vanishing gradient issue. Reinforcing the error gradient of the first layers would solve this problem. As such, a simple solution is to add skip connections between layers to allow the gradients to flow backward through the network easily. This solution is referred to as residual or skip connections and was first proposed for image recognition (He et al. 2016a), where the (residual) network had hundreds of layers for the first time. The output of the residual connection can either be added or concatenated to the current layer.

1.4.3 Time delayed neural and convolutional neural networks

For many sequence classification tasks, it is interesting to include the context of the sequence to generate an output at a given time. For example, to determine which phoneme is spoken at time t , it is helpful to know some parts of the sequences that precede and follow it. In many speech-processing tasks, capturing long-term dependencies between acoustic events is helpful. Many methods exist for capturing long-term sequences, and recurrent neural networks (Graves et al. 2013) are one of them. However, they are sensitive to the gradient vanishing problem and are hard to parallelize for efficient training.

In contrast, the Time-Delay Neural Network (TDNN) (Waibel et al. 1989) is a type of architecture that can efficiently model temporal dependencies without suffering from the same problems. It uses multiple stacked layers to encode an input sequence. Each layer encodes a couple of spatially expanded elements from the previous layer. The first layers learn a small context, while the higher layers learn a wider context. The context resolution increases as the network gets deeper.

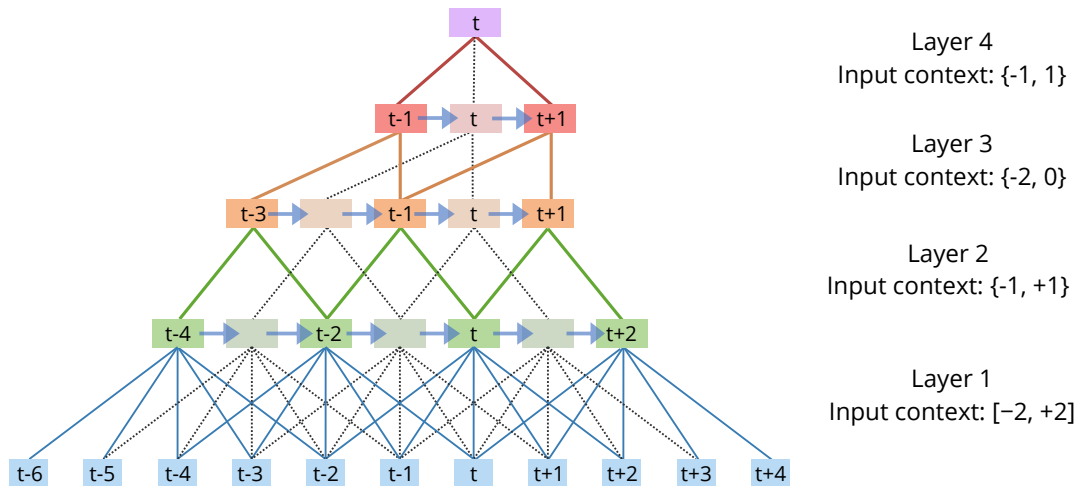


Figure 1.5 – Computations for a TDNN in dotted and solid lines, and sub-sampled TDNN in solid lines only. Frames $t - 2$ through $t + 2$ are spliced together at the input layer (which can be written as context $\{-2, -1, 0, 1, 2\}$ or more compactly as $[-2, 2]$); and then at the upper three hidden layers with splice offsets of $\{-1, 1\}$, $\{-2, 0\}$ and $\{-1, 1\}$.

The hyperparameters of a TDNN layer are defined by the length of the input context required to compute an output activation at a time step. Figure 1.5 shows the receptive field for each layer in dotted and solid lines. The notation $\{-2, 0\}$ for the third layer means that only the input at $t - 2$ and t will be used to compute the activation. In the example, the network does not take symmetrical past and future contexts to generate the final output at time t . A smaller context on the right side of the network lowers the latency for online decoding use as fewer future elements are to be awaited before emitting an output.

TDNN are, in fact, 1D Convolutional Neural Network (CNN) (LeCun et al. 1998) with specific settings for kernel size, dilatation, and strides. Both network architectures process signals in a way that considers temporal dependencies between elements. Similar to the convolution operation in CNN, TDNN models nearby dependencies using a fixed-length context window to learn local correlations in the signal. Additionally, both TDNN and CNN can be stacked (like in Figure 1.5) to form deep networks, allowing for complex feature extraction and modeling of hierarchical relationships in the data.

In (Peddinti et al. 2015), it is proposed to use subsampling to reduce the number of hidden activations. This method avoids unnecessary redundancy as large contexts overlap leading to highly similar subsequent activations. A computation path of a subsampled TDNN is shown by the solid line connections in Figure 1.5. With subsampling, the overall computation is reduced leading to faster training and inference time.

Factorized version In (Povey et al. 2018), the author proposed a method to reduce the number of weights in a TDNN model by using matrix decomposition. The weight matrix \mathbf{W} of each layer is approximated as a product of two low-rank matrices: $\mathbf{W} = UV$, where $U \in \mathbb{R}^{u \times k}$ and $V \in \mathbb{R}^{k \times v}$ are obtained using Singular Value Decomposition (SVD) of $\mathbf{W} \in \mathbb{R}^{u \times v}$ (Golub et al. 1970) while discarding the basis corresponding to the smallest singular values.

Selecting a rank value k smaller than u and v effectively inserts a bottleneck layer within a conventional TDNN layer. This bottleneck can be relatively strong, as it can reduce the dimension by a large factor. For example, in real-world applications, the 1536x1536 weight matrix can be factorized as a product of 1536x160 and 160x1536 matrices. Figure 1.6 displays the differences between the architectures of the TDNN and Factored Time-Delay Neural Network (TDNN-F) layers; in the TDNN-F layers, a skip connection is added between the input layer and the output projection.

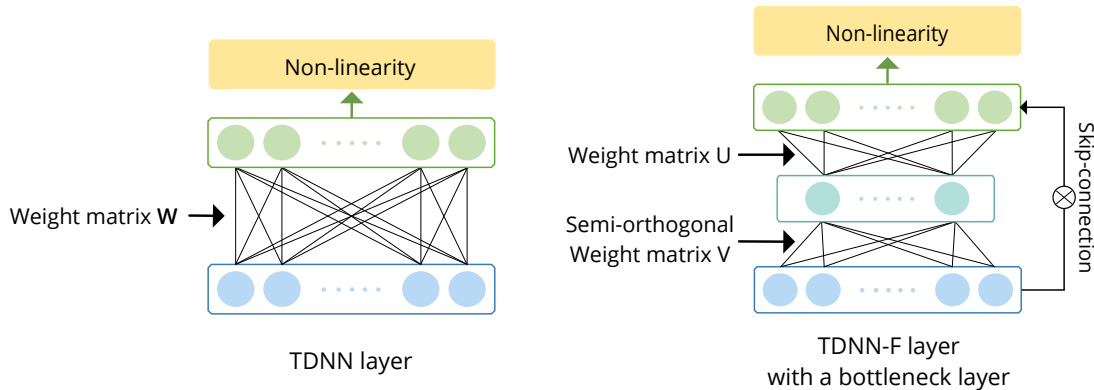


Figure 1.6 – TDNN and TDNN-F layers.

In order to make the training converge, it is necessary to constrain the matrix V to be semi-orthogonal, meaning that it must respect one of the following conditions $VV^T = I$ or $V^TV = I$, where I is the identity matrix. To ensure the V matrix remains close to the semi-orthogonality, a constraint is applied every four backpropagation updates (Povey et al. 2018).

In this thesis, we use both the Kaldi implementation (Povey et al. 2011) and a PyTorch implementation (Paszke et al. 2019; Madikeri et al. 2020) of the TDNN-F network. The trick of factorizing the weight matrices has other valuable properties than just minimizing the number of weights. As observed by (Ryffel et al. 2019) smaller bottleneck dimensions encode less personal information.

1.4.4 Transformer

The transformer model proposed by (Vaswani et al. 2017) for sequence-to-sequence machine translation has received tremendous interest in recent years in many fields (Lin et al. 2021), including speech processing. This model consists of an encoder and a decoder, each with L stacked processing blocks consisting of similar architectures. In contrast with the recurrent neural network, the transformer architecture takes the whole input sequence at once rather than one symbol or frame at a time during training. At inference, the transformer decoder is an autoregressive model¹. The decoder always generates the entire sequence from y_1 up to $y_{current}$, given the encoder state and the previous output sequence [$\langle sos \rangle, y_1 \dots y_{current-1}$] (starting with a start-of-sequence $\langle sos \rangle$ symbol for the first prediction). At each step, the current output sequence is fed to the bottom right decoder block (Figure 1.7) in the next time step until the $\langle eos \rangle$ (end-of-sequence) symbol is generated by the network. Interestingly, during training, the transformer model is not computationally autoregressive as it processes the input all at once instead of processing it one step at a time in a loop. This is achieved by using masking techniques that prevent the network from depending on future information and allows it to be used during inference exactly like an autoregressive model.

The transformer model is agnostic to the order of the input (and output) sequence unless explicit positional embeddings are added to each input (and output). Positional encoding allows to retain the position information necessary to keep the meaning of the sequence. The encoder includes **multi-head self-attention** modules and a MLP network. Residual and normalization layers are employed around each module to facilitate backpropagation of gradients through the network during training. Compared to the encoder blocks, the decoder blocks have a **multi-head attention** layer called **cross-attention module**. The overall architecture of the transformer is well described in (Alammar 2018) and displayed in Figure 1.7.

Transformer architectures are very efficient at encoding the temporal dependency of a sequence with attention. Additionally, this architecture is very versatile as it can be used in multiple manners:

- Encoder-Decoder: The entire architecture is used. This is typically used in sequence-to-sequence modeling (e.g., neural machine translation).
- Encoder only: Only the encoder is used, and the outputs of the encoder are utilized as a high-level representation of the input sequence. This encoder can be pretrained and fine-tuned to perform better on specific downstream tasks (for example, Wav2Vec-2.0, GPT-3 (Baevski et al. 2020; Brown et al. 2020)), or, jointly trained with different decoders (for example, ASR models with CTC decoder).

1. Autoregressive models rely on previously generated output to predict the current ones.

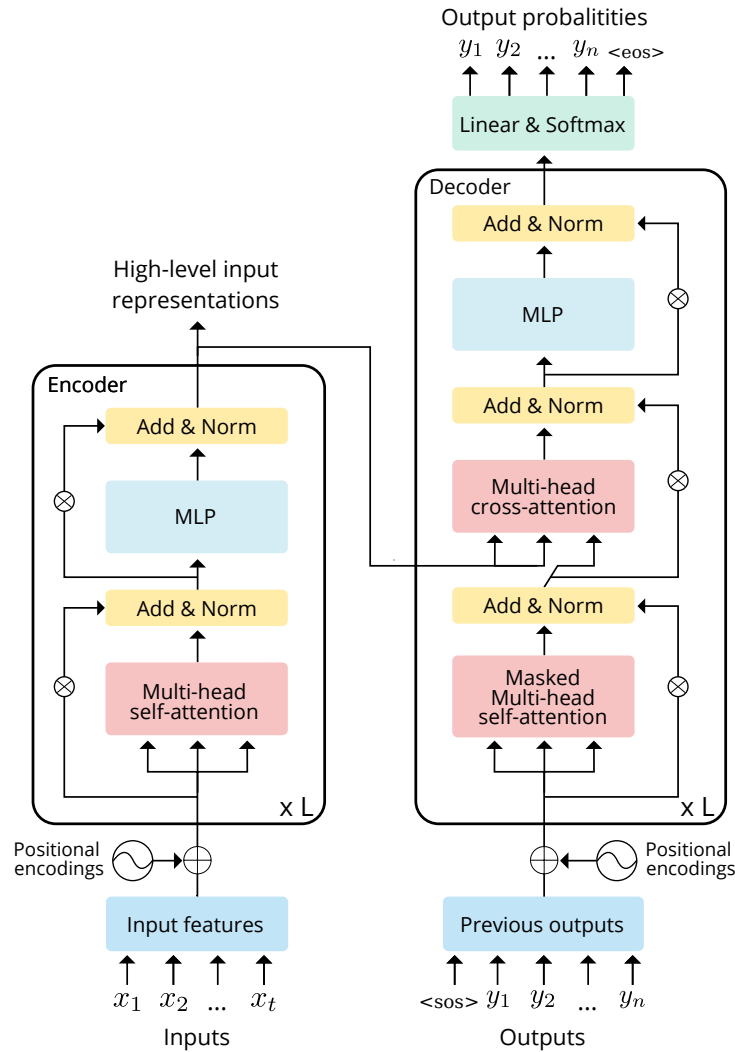


Figure 1.7 – Overview of the transformer architecture, “x L” corresponds to the number of stacked encoders and decoders.

1.4.5 Adversarial network

To enhance the capability of a network to encode interesting attributes, one can rely on adversarial machine learning. In adversarial machine learning, a *game theory* (Osborne et al. 1994) with two or more players is considered. The loss of each player must be minimized, but the decision of other players impact their loss. Applied to neural networks, this game enhances feature representation by promoting the emergence of representation relevant to a primary task while also being indiscriminate concerning one (or several) so-called adversarial task(s).

Such kind of models are interesting for training models capable of generating highly realistic data (e.g., image generation) (Goodfellow et al. 2014), adapting models to out-of-domain data (Ganin et al. 2017; Samarakoon et al. 2018), and extracting privacy-preserving features (Feutry et

al. 2018; Srivastava et al. 2019). The architecture comprises three modules: encoder, decoder and the adversarial branch. An encoder takes the input in the first module and generates a representation f . Then the decoder uses this representation f to predict the output y of the primary task. This occurs while an additional adversarial branch, alongside the decoder, uses the representation f to predict the output d of the adversarial task. This architecture is described in Figure 1.8.

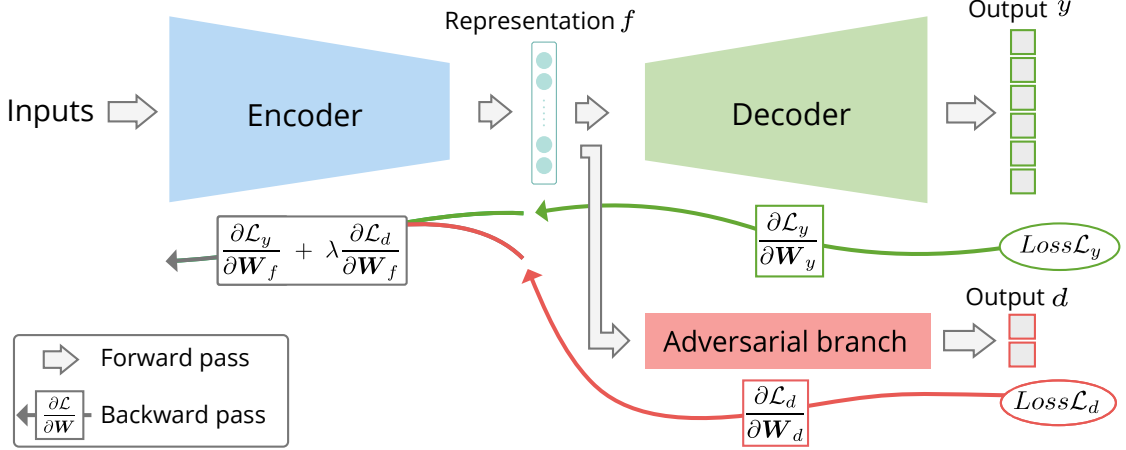


Figure 1.8 – Outline of an adversarial neural network.

During training, the forward pass computes the output of the primary and adversarial tasks. The backward pass first optimizes the decoder and adversarial branch weights to minimize the two losses of primary and adversarial tasks. Then, the encoder's weights are adjusted to maximize the adversarial task loss while minimizing the primary task loss. A gradient reversal function achieves loss maximization by multiplying the gradient with a negative scalar during the backpropagation. The following equation defines the gradient updates for each neural network module.

$$\mathbf{w}_{f_i}' = \mathbf{w}_{f_i} - \alpha \left(\frac{\partial \mathcal{L}_y}{\partial \mathbf{w}_{f_i}} + \lambda \frac{\partial \mathcal{L}_d}{\partial \mathbf{w}_{f_i}} \right) \quad (1.1)$$

$$\mathbf{w}_{y_i}' = \mathbf{w}_{y_i} - \alpha \frac{\partial \mathcal{L}_y}{\partial \mathbf{w}_{y_i}} \quad (1.2)$$

$$\mathbf{w}_{d_i}' = \mathbf{w}_{d_i} - \alpha \frac{\partial \mathcal{L}_d}{\partial \mathbf{w}_{d_i}} \quad (1.3)$$

where λ is the negative scalar coefficient, \mathcal{L}_d the adversarial loss, \mathcal{L}_y the primary loss, \mathbf{w}_{f_i} , \mathbf{w}_{y_i} , \mathbf{w}_{d_i} are weights of the encoder, primary task decoder, and adversarial branch respectively. This training scheme encourages the encoder to be invariant and, to some extent, disentangled from the adversarial task. In this thesis, we will investigate if personal speaker information PII can be removed using an adversarial network similarly as in (Srivastava et al. 2019). However, we

addressed some limitations of their approach by, most notably, using a more advanced training scheme designed for privacy protection (Ryffel et al. 2019). Also, we will use adversarial networks with a generative model to synthesize speech.

1.4.6 Generative adversarial network

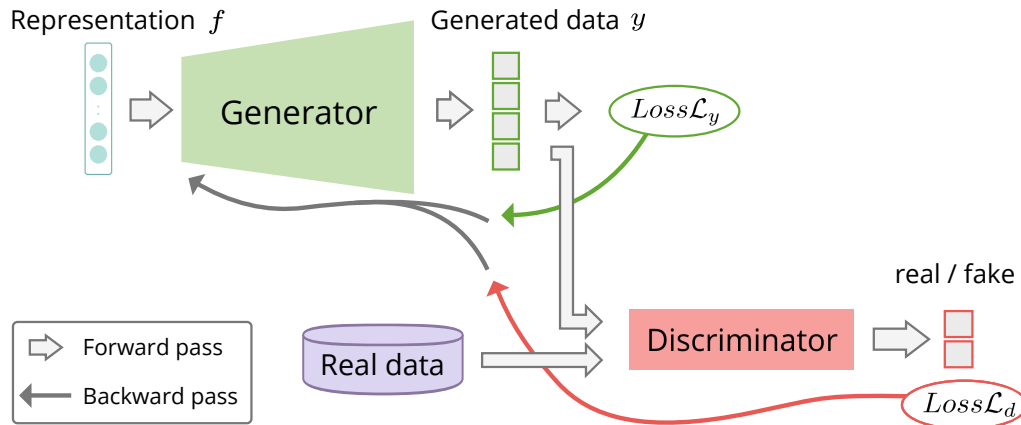


Figure 1.9 – Outline of a Generative Adversarial Network (GAN).

The Generative Adversarial Network (GAN) is a network architecture and training scheme based on the adversarial network (Goodfellow et al. 2014). This type of network has seen broad interest as it is, as of today, the state-of-the-art method for generating realistic data. The two players that compete and feed off each other involves:

1. **A generator model** trained to generate new data aiming to reproduce the same distribution as real training data.
2. **A discriminator model** trained to classify inputs as real or fake, attempting to identify if an input originates from the training dataset or the generator model.

During training, the discriminator is trained to distinguish real data from generated data (or fake in the GAN terminology). On the other side, the generator is trained to generate output that can no longer be distinguished from real data by the discriminator. This adversity between these two players drives both networks to improve until the fake data is indistinguishable from the real one. Figure 1.9 shows an example of a GAN taking as generator input a representation, which can be a latent noise variable or a representation of real data obtained by one or several encoders. For voice conversion tasks, GANs are particularly interesting because they can generate highly realistic speech data.

Each training step contained in the GAN framework consist of two mini-batches, the first coming from the real data \mathbf{x} (labeled as 1) and the second coming from the generator $G(\mathbf{f})$

(labeled as 0). In the original GAN paper (Goodfellow et al. 2014), the generator attempts to minimize the following function, whereas the discriminator attempts to maximize it:

$$\min_G \max_D \mathcal{L}_D(D, G) = \mathbb{E}_x[\log D(\mathbf{x})] + \mathbb{E}_f[\log(1 - D(G(\mathbf{f})))] \quad (1.4)$$

where \mathbb{E}_x is the expected value according to the real distribution, \mathbb{E}_f the expected value with respect to the distribution of the generator, $D(\mathbf{x})$ the discriminator estimate of a probability that a real data instance x comes from the real distribution rather than the fake one, and $D(G(\mathbf{f}))$ the discriminator estimate of the probability that a fake instance is real. This formula is derived from the binary cross-entropy between two distributions. Because the generator cannot directly influence the $\log D(\mathbf{x})$ term in the function, minimizing the loss for the generator is equal to minimizing $\mathbb{E}_f[\log(1 - D(G(\mathbf{f})))]$. In practice, a reformulation of this formula is necessary to make the network converge better while avoiding vanishing gradient or other training failures. Techniques such as the *non-saturating* generator loss (Goodfellow et al. 2014) or the least-square loss (Mao et al. 2017) have proven to avoid vanishing gradients problems. Other losses can be applied to the generator to increase the generator training efficiency and the quality of the produced data; however, this depends on the dataset and the availability of labels.

GANs have been successfully employed by the speech processing community (Wali et al. 2022), in particular for speech synthesis (Kong et al. 2020), voice conversion (Fang et al. 2018), speech enhancement (Pascual et al. 2017). Chapters 4 and 5 of this thesis rely on GAN architectures to generate anonymized, natural, and realistic-sounding speech.

1.5 Conclusion

In this chapter, the process of speech production is introduced and important key concepts in numeric signal processing and machine learning are covered. The focus then shifts to specific network architectures like TDNN, transformer, and GAN, which will play a crucial role in later chapters. All of this knowledge will not only enable us to build speaker anonymization systems but also evaluate their effectiveness in terms of privacy and utility using speech-centric machine-learning disciplines that will be covered in the next chapter.

SPEECH-CENTRIC MACHINE-LEARNING

Contents

| | | |
|------------|---|-----------|
| 2.1 | Introduction | 25 |
| 2.2 | Automatic speech recognition | 26 |
| 2.2.1 | Acoustic modeling | 27 |
| 2.2.2 | N-gram language model | 30 |
| 2.2.3 | End-to-end and self-supervised learning | 31 |
| 2.2.4 | Evaluation | 33 |
| 2.3 | Automatic speaker recognition | 34 |
| 2.3.1 | Speaker embedding extractor | 35 |
| 2.3.2 | Scoring function | 36 |
| 2.3.3 | Evaluation | 38 |
| 2.4 | Voice conversion | 41 |
| 2.4.1 | Linguistic representation | 42 |
| 2.4.2 | Speaker representation | 44 |
| 2.4.3 | Speech synthesizer | 45 |
| 2.4.4 | F_0 conditioning | 48 |
| 2.4.5 | Evaluation | 48 |
| 2.5 | Conclusion | 48 |

2.1 Introduction

The task of speaker anonymization and its evaluation relies on many speech-centric machine-learning disciplines that will be presented in this chapter. Knowledge of acoustic modeling is necessary to build a speaker anonymization system; as such, we will present how traditional Automatic Speech Recognition (ASR) work. Voice Conversion (VC) is another fundamental tool to generate an anonymized voice that will also be introduced. The automatic evaluation of the performance of a speaker anonymization system relies on ASR and Automatic Speaker Verification (ASV) systems to objectively assess the preservation of the linguistic content and the capability of the system to conceal the speaker identity. That is why we will also present the fundamental of ASV.

2.2 Automatic speech recognition

Automatic Speech Recognition (ASR) or speech-to-text aims at transforming an audio speech signal into the sequence of corresponding words. In the scope of this thesis, ASR systems play a crucial role. During the anonymization process, the acoustic model of an ASR system is used to extract a feature representing the linguistic content spoken by the speaker. Precise modeling is necessary as the anonymization system will later use this representation to generate anonymized speech signals without distorting the content as best as possible. Importantly as well, during the evaluation, an ASR system is used to assess unwanted degradation of the linguistic content intelligibility distortion. First, this section defines the task of ASR, details the key models employed for speech recognition, and finally explains how to evaluate the performance of an ASR system.

As mentioned in Section 1.3, speech can be considered as a finite sequence of acoustic features $X = [x_1 \dots x_T]$ where T is the number of audio frames. Performing speech recognition consists in producing a sequence of M words $W^* = [w_1 \dots w_M]$ from the observation X . This problem can be formulated as the maximization of the probability $P(W|X)$ (Jelinek 1976):

$$W^* = \underset{W}{\operatorname{argmax}} P(W|X) \quad (2.1)$$

The equation 2.1 can be simplified since $P(X)$ is independent of W (constant for all possible sequences of words). Following Bayes' rule, the equation is decomposed into:

$$\begin{aligned} W^* &= \underset{W}{\operatorname{argmax}} \frac{P(X|W)P(W)}{P(X)} \\ &= \underset{W}{\operatorname{argmax}} P(X|W)P(W) \end{aligned} \quad (2.2)$$

An ASR system usually relies on two models: an acoustic model, which models the probability $P(X|W)$ of a speech signal X , given a corresponding word sequence W ; and a language model, which models the probability $P(W)$ of the sequence of the words W .

As shown in Figure 2.1, a pronunciation lexicon may be employed to associate multiple discrete spoken sound units (for example phonemes) with a word. In such a case, the acoustic model no longer models $P(X|W)$ but $P(X|Q)$, with Q a sequence of phonemes. The equation can be formulated as follows:

$$W^* = \underset{W}{\operatorname{argmax}} P(X|Q)P(Q|W)P(W) \quad (2.3)$$

where $P(Q|W)$ is the probability of Q , the sequence of phonemes, knowing W , the sequence of words. With a pronunciation lexicon, one or several pronunciations are mapped to each word. This enables the acoustic model to identify spoken units smaller than words, for example,

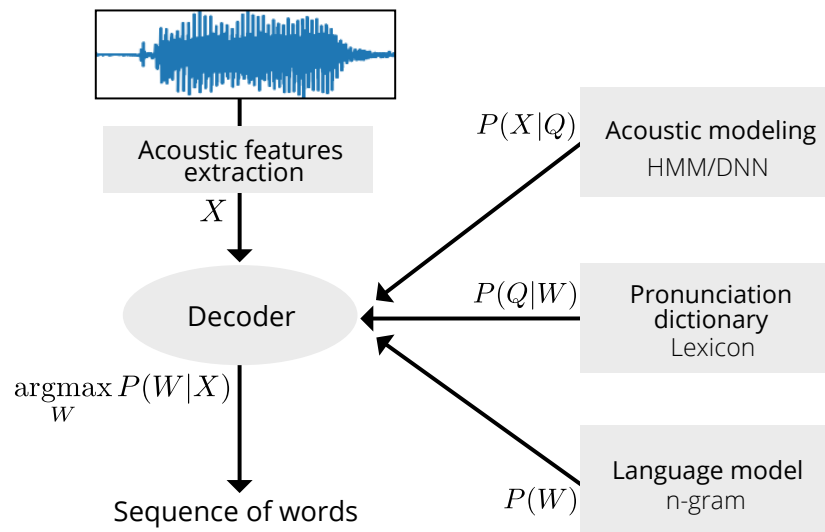


Figure 2.1 – Overview of an ASR system

phonemes, which are the smallest unit of speech that differentiates between words in a language. This mapping of phonemes to acoustic features is what the acoustic model learns. Another benefit of the lexicon is that it allows for more efficient training of acoustic models when data is limited. This is because the number of phonemes is typically much smaller than the number of words in the lexicon, resulting in a larger number of examples per phoneme in the training data.

2.2.1 Acoustic modeling

The acoustic model is the most crucial element of ASR systems since it models the correspondence between the audio input and a sequence of spoken units. This section details the methods based on, generative acoustic modeling combining Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM), hybrid Deep Neural Network (DNN) HMM modeling, and end-to-end and self-supervised learning.

2.2.1.1 Hidden Markov model

Hidden Markov Models were initially proposed in the 1970s for acoustic modeling ([Baker 1975](#); [Rabiner 1989](#)). Today HMMs are still a widespread approach but considered traditional, with the advent of end-to-end models. HMMs are statistical models that are used to model sequential or temporal data, by assuming that the observed data is generated by a hidden Markov process, where the current state only depends on the previous state, and the observation at the current time only depends on the current state. HMMs are composed of emitting states, transition probabilities between states, and probability densities for the emission of observations. The latter can be obtained with GMM or DNN models.

In the context of acoustic modeling, the emitting states represent the probability of a particular sound being emitted from the model at a given time. This sound can be a phoneme, possibly in a given surrounding context. The three main units are the monophone, biphone, and triphone models, each referring to the different contexts that are taken into account when determining the probability of a sound being emitted. A monophone model only reflects a single phoneme, the observation units do not consider the context in which it is being pronounced. Monophones do not allow modeling co-articulation, which is how the configuration of the vocal tract gradually changes from one phoneme to another, causing a distortion of these phonemes in the process. For example, the sound "a" (/æ/) in the word "bat" (/bæt/) will have a different probability of being emitted than the sound "a" in the word "bad" (/bæd/), because the surrounding context is different. A biphone model better captures co-articulation by taking into account the one phoneme that is before (or after) a given phoneme. This allows the model to better capture the context in which a phoneme is being pronounced, and can improve the overall accuracy of the ASR system. A triphone model goes even further, by considering the two phonemes that surround a given phoneme. This provides even more contextual information, potentially improving the accuracy of the ASR system.

In practice, for each phonetic unit, the HMM usually uses three-states corresponding to the beginning, middle, and end of a unit. The following figure shows an example using monophones and triphones to represent the word "cup". Both the monophone and triphone are modeled using three-states models. If we consider 40 possible phonemes for monophones represented with

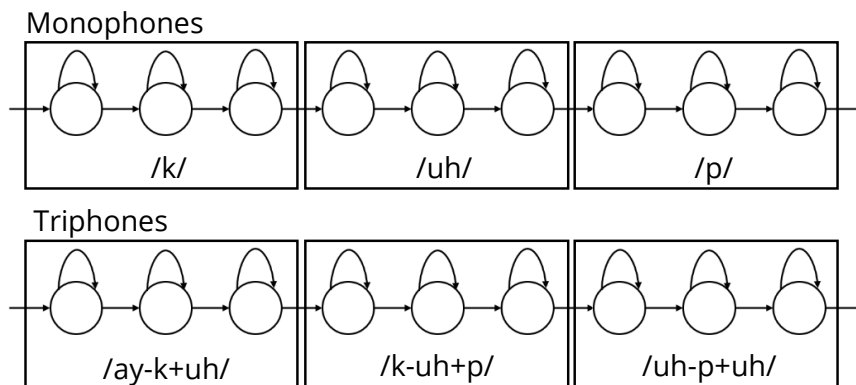


Figure 2.2 – Monophone and triphone three-state HMM for the word "cup" in the expression of "a cup you".

three-state HMM, there is a total of 120 possible emitting states. When triphones are considered, the number of units increases to 40x40x40x3, resulting in 192 000 emitting states. This explosion of number emitting states leads to two issues: insufficient data to train each triphone, and as such, some triphones will be appearing in testing but not in training. One solution is clustering,

where similar triphones are grouped into "senones" (reducing the emitting states to about 10 000). This is done using a decision-tree clustering process where a decision tree is built for every state of every context-independent phone. In general, the number of possible emitting states in an HMM is a trade-off between the accuracy of the model and the computational resources/amount of data required to train it.

2.2.1.2 Generative acoustic models GMM-HMM

Traditionally, the probability densities of emitting observation of HMM models were represented by generative Gaussian Mixture Models (GMM). The idea is to associate each HMM state with a weighted sum of Gaussian probability densities. However, GMM has some limitations, such as its tendency to overfit training data and difficulty to deal with high-dimensional data.

2.2.1.3 Hybrid neural network models DNN-HMM

The usage of neural networks as an alternative to GMMs was first proposed with MLPs (Bourlard et al. 1987). However, it is not until 2012 that Deep Neural Network (DNN) have outperformed GMMs at estimating the emission probabilities of HMM states (Hinton et al. 2012). The TDNN (Waibel et al. 1989) deep learning architecture presented in section 1.4, which is particularly well-suited for speech processing, has been widely adopted in the community after the work of (Peddinti et al. 2015), gradually replacing GMMs. DNNs have the advantage over GMMs of being able to model more complex and non-stationary representations thanks to how they are trained with their loss function. As discriminative models, they are directly trained to model the posterior probabilities, allowing them to better distinguish classes.

The model DNN-HMM used in this thesis is a Kaldi (Povey et al. 2011) "chain" TDNN-F model, trained with the Lattice-Free Maximum Mutual Information (LF-MMI) cost function, the next paragraph explains into further details Maximum Mutual Information (MMI) based cost functions.

MMI training The MMI discriminative training method computes the cost function at the sequence level by maximizing the MMI between the distribution of a predicted sequence and the distribution of the expected correct transcription. This is equivalent to maximizing the likelihood of the input sequence given (the state sequence corresponding to) the correct transcription, while simultaneously minimizing the marginal likelihood for any possible state sequence/transcription. This approach allows the model to better distinguish between correct and incorrect transcriptions, improving the overall performance of the ASR system (Veselý et al. 2013).

The loss function is defined as:

$$\begin{aligned}\mathcal{L}_{\text{MMI}} &= \sum_{r=1}^R \log p(W_r | X_r) \\ \mathcal{L}_{\text{MMI}} &= \sum_{r=1}^R \log \frac{p(X_r | W_r) P(W_r)}{\sum_W p(X_r | W) P(W)}\end{aligned}\tag{2.4}$$

where R is the total number of training segments, W_r is the correct transcription of the r^{th} speech segment X_r , $P(W_r)$ is the probability given by the language model for the sentence W_r . For any sequence, W , $P(W)$ is estimated with a language model. The numerator indicates the likelihood of the input sequence for the reference word sequence. In contrast, the denominator indicates the total likelihood of the input sequence for all possible word sequences, which is equivalent to the sum of all possible word sequences estimated by the acoustic and a phoneme language model. This cost function is optimized by maximizing the numerator (i.e., increasing the probability that the model predicts a sequence similar to the reference) and minimizing the denominator (decreasing the probability of other non-valid sequences).

LF-MMI training The Lattice-Free Maximum Mutual Information (LF-MMI) cost function is an extension of MMI to make the computation of the numerator and denominator graphs faster and manageable in size (Povey et al. 2016). Other implementation optimizations are also made, such as subsampling the output frame rate to one frame every 30ms instead of every 10ms which makes the denominator graph smaller and speeds up the forward/backward computation. Consequently, the HMM topology differs from other common three-state HMM. The HMM has two states per phoneme unit and can be traversed in a single frame, instead of three, improving decoding speed.

In order to train the acoustic DNN model with the LF-MMI loss, a phonetic alignment between the transcriptions and audio is required (Peddinti et al. 2015). This alignment is usually estimated with a GMM-HMM system. Flat-start training of the DNN in one stage (i.e., without using any previously trained model, alignment, or performing prior estimation) is possible, as shown in (Hadian et al. 2018a) with the use of biphones. This form of training can be assimilated into end-to-end training and is usually called E2E-LF-MMI.

2.2.2 N-gram language model

The language model estimates the probability $P(W)$ that a sequence of words W corresponds to a sentence. The goal is to estimate whether a sequence of words conforms to a particular grammar, specific to the language studied. The probability of observing a sequence of words

$W = [w_1 \dots w_M]$ in a statistical language model is expressed in the equation below:

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \dots P(w_i|w_1, w_2, \dots, w_{m-1})$$

$$P(W) = \prod_{m=1}^M P(w_m|w_1, \dots, w_{m-1}) \quad (2.5)$$

where M is the length of the sequence W and $P(w)$ is the probability of observing the word w . This equation indicates that the probability of observing a sequence of words is determined by the probability of observing each individual word given the past words in the sequence.

The most common n-gram models are 2-gram and 3-gram, which require a history of one or two words respectively (Brown et al. 1992). For example, the probability of a word knowing the preceding word in a 2-gram model can be calculated using the following formula:

$$P(w_j | w_i) = \frac{\text{count}(w_i, w_j)}{\sum \text{count}(w_i, w)} \quad (2.6)$$

The probability $(w_m | w)$ is given by the number of occurrences of word w_m followed by word w_j , divided by the number of occurrences of the same word w_m , followed by any other word. The 2-gram model is then:

$$P(W) \approx P(w_1) \times \prod_{m=2}^M P(w_m | w_{m-1}) \quad (2.7)$$

In speech recognition, language models are generally n-gram models whose order is bounded between 2 and 4.

2.2.3 End-to-end and self-supervised learning

With the increase of computational power, new methods to build ASR systems have emerged, called end-to-end automatic speech recognition (Li 2022). End-to-end architectures eliminate the need to pre-align the data with an additional GMM model for example. They unify the training process with a single model. Unlike models based on HMM, language modeling is inherent to the architecture as end-to-end models can directly convert acoustic features into a sequence of words or subword units (e.g., byte-pair-encoding (Sennrich et al. 2016) and unigram language model (Kudo 2018)), avoiding the need for a pronunciation lexicon. In practice, an external language model is often used to improve recognition accuracy.

One paradigm from which end-to-end training benefits, is self-supervised learning for acoustic feature extraction. Self-supervised learning enables to easily create systems that can, on one end, input raw speech signals and, on the other end, output words without necessitating large text-annotated corpora. Here we will focus on the task of extracting speech features from raw

audio data without using traditional hand-crafted features. To accomplish this, large neural networks, usually based on the transformer architecture (see Section 1.4.4), learn from a very large quantity of unlabeled data the structure of the speech signal itself. By doing so it replaces traditional hand-crafted features with models that can be fine-tuned later on for specific tasks. One example of it is Wav2Vec-2.0 (Baeovski et al. 2020), which is learned with a training objective similar to BERT (Devlin et al. 2019) masked language modeling loss. In the paper of Wav2Vec-2.0, the authors showed that fine-tuning the self-supervised pre-trained transformer with only one hour of labeled data outperforms existing state-of-the-art ASR systems trained on 100 times more labeled data.

The Wav2Vec-2.0 model is pre-trained to take raw audio data as input and generate latent speech representations using a multi-layer 1-d convolutional neural network. Then, the model learns quantized vectors of the latent representations, where continuous latent representations are matched with discrete representations from a fixed-sized dictionary containing similar representations. Then, the model selects the closest quantized representation from the dictionary for each latent representation. During training, about half of the latent representations are masked before being fed to the transformer. For each masked vector, 100 other quantized vectors from the same utterance are randomly selected as negative distractors (in red in Figure 2.3). The model is trained to identify the real quantized vector (in green in Figure 2.3) among negative distractors from the output of the transformer at the masked positions, this kind of training is called contrastive learning (Oord et al. 2018). Figure 2.3 presents this pre-training process.

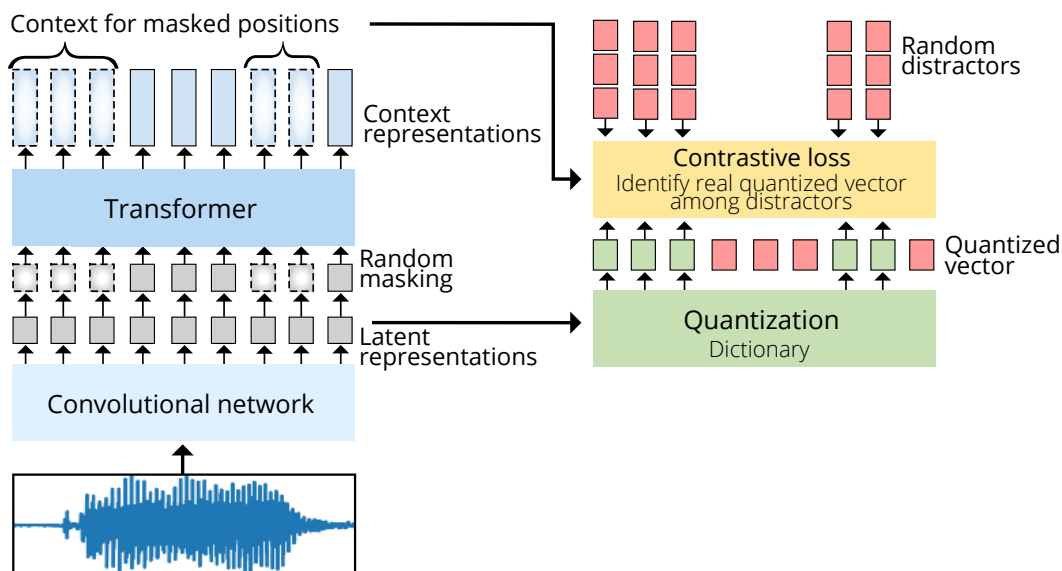


Figure 2.3 – Overview of the Wav2vec 2.0 architecture and its pre-training process. (Modified from: *Illustrated Tour of Wav2vec 2.0*)

2.2.4 Evaluation

Speech recognition is the process of transcribing an acoustic signal into words. In order to evaluate the correctness of a predicted transcript, it is compared to a reference transcript. The Word Error Rate (WER) metric can be employed to measure the number of errors between two transcripts. The errors taken into account are substitution, deletion, and insertion errors. Substitution errors are words that are incorrectly transcribed. Insertions errors are words added during transcription, and deletion errors are words omitted. The word error rate is computed as follows:

$$\text{WER} = \frac{\text{count}(\text{substitutions}) + \text{count}(\text{deletions}) + \text{count}(\text{insertions})}{\text{count}(\text{words in the reference})} \quad (2.8)$$

The WER is a useful metric for comparing the performance of different speech recognition systems, as well as for tracking the improvements made within a single system over time. Additionally, the WER can be used to diagnose the specific types of errors that a speech recognition system is making, which can help identify areas for improvement.

2.3 Automatic speaker recognition

Speaker recognition systems identify or verify the identity of a speaker based on a speaker’s speech characteristics. A distinction must be made between the automatic speaker identification task and the Automatic Speaker Verification (ASV) task. Speaker identification aims to find the most probable speaker among a set of known speakers (close-set), and speaker verification aims to assess whether a test sample matches a claimed speaker’s identity. This use case is more representative of real-world applications, where new identities are constantly being encountered. As such, it is the one primarily used in this thesis.

In a *text-dependent* scenario, the speaker must utter a predefined prompt. This helps to remove/control one of the major sources of acoustic variability, the linguistic content. Having the user always utter similar passphrases allows better speaker representation, which is beneficial when high accuracy is required. In contrast, in the *text-independent* scenario, the linguistic content does not require to be imposed. The *text-independent* scenario is more generic and user-friendly. In research, there is a larger amount of data available in the *text-independent* scenario (Nagrani et al. 2017), and overall more research has been done in this scenario. For this thesis, the application context chosen to evaluate how much the identity of a speaker is recognizable before and after anonymization is *text-independent* speaker verification. This context is also the de facto standard when it comes to evaluating biometric information protection (Jain et al. 2008; ISO-24745 2011; Gomez-Barrero et al. 2018).

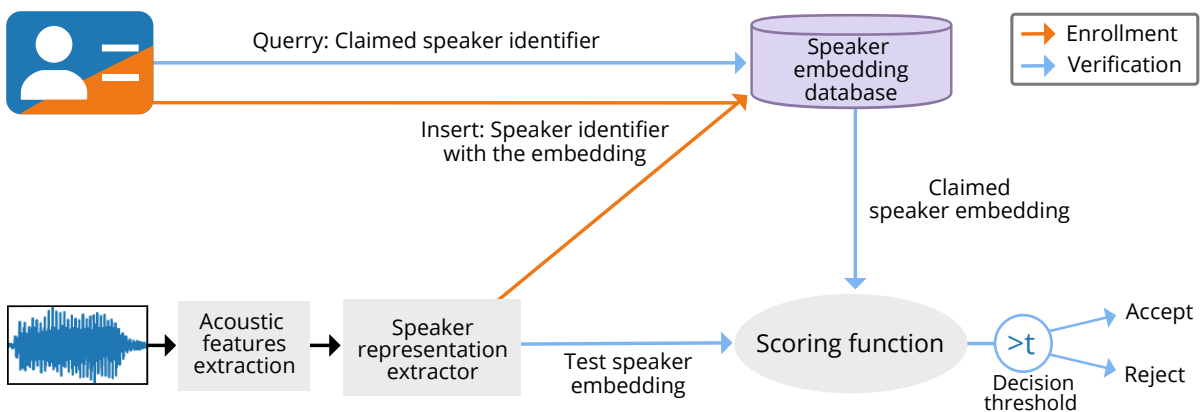


Figure 2.4 – Overview of the enrollment and verification steps in ASV.

Speaker verification pipelines are composed of three main components, depicted in Figure 2.4. It starts with acoustic features extraction, as presented in chapter 1.3. The speech signal first needs to be converted to a time-frequency representation to obtain a sequence of acoustic features. In contrast to traditional ASR systems where the 12 first MFCCs coefficient are used (Shannon et al. 2003), traditional ASV requires more coefficients, above 24 (Espy-Wilson et al. 2006), to

better capture speaker characteristics. Then, given a sequence of multiple acoustic feature frames, a speaker embedding extractor transforms this acoustic sequence into a fixed and compact vector called a speaker embedding. An embedding is a high-level representation usually extracted from a bottleneck layer (see Section 1.4.3). However, in contrast to a bottleneck that can be extracted for each audio frame, speaker verification typically requires an embedding representation (speaker embedding) that describes the full audio segment. Speaker embeddings should primarily encode speaker discriminative information while discarding incidental information not relevant to the speaker (e.g., microphone, recording session, etc.). Finally, the speaker embeddings are compared to obtain a comparison score indicating how similar two speakers are for the system. Given a comparison score, a binary decision can be made to confirm or deny the correspondence.

Figure 2.4 summarizes how a user interacts with the system during the enrollment and verification phases. During enrollment, speaker embeddings are extracted from each user’s speech sample and stored in a database indexed by a unique user identifier. During verification, the user provides the system with a test speech sample and a unique user identifier which is used to get the associate speaker embeddings in the database. If the database contains several speaker embeddings for the user (multiple samples), an average operation can be used to aggregate them (which increases robustness) to create a claimed speaker embedding. A test speaker embedding is extracted from the test speech sample and used to compute a similarity score s with the claimed speaker embedding. The score is compared against a decision threshold t . If $s > t$, then the verification trial is accepted; otherwise, rejected.

The following sections detail one method based on deep learning to train a speaker embedding extractor (x-vector), scoring function (PLDA or cosine similarity), and metrics used to evaluate ASV systems performances (Equal Error Rate, Linkability).

2.3.1 Speaker embedding extractor

The sequence of acoustic features contains a good deal of information about the signal. They encode the spoken linguistic content, speaker voice characteristics (F_0 , volume, timbre, speaking rate, etc.), speech characteristics (rhythm, accent), ambient noise, and background sounds. Extracting a speaker embedding from the acoustic features invariant to the non-speaker-related characteristics is a complicated task (Hansen et al. 2015). In recent years, statistical and DNN-based speaker models have primarily been used. This section discusses the latter, as it has superior speaker modeling capabilities due to the advancements made in deep learning.

One of the first DNN-based speaker embedding is called x-vector (Snyder et al. 2017; Snyder et al. 2018). The model used to extract x-vectors is trained in a supervised manner to classify the speakers of a training dataset given the sequences of acoustic features. As shown in Figure 2.5, the model used to extract x-vectors can be separated into two main components: frame-level and utterance-level computations. The network takes as input frame-by-frame acoustic features

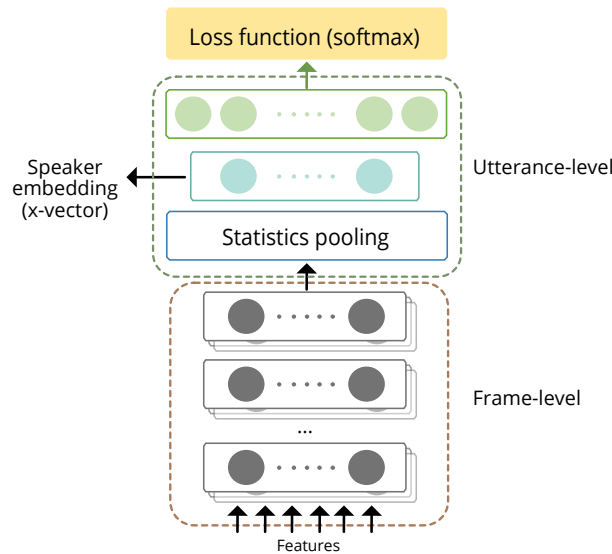


Figure 2.5 – The architecture of the model used to extract x-vector speaker embeddings.

(e.g. MFCCs) which are transformed by multiple frame-level TDNN layers. The most common architecture is composed of 5 TDNN layers, having respectively the following input contexts: $[t-2, t+2]$, $\{t-2, t, t+2\}$, $\{t-3, t, t+3\}$, $\{t\}$, $\{t\}$ (refer to section 1.5 for the notations). After the frame-level layers, a statistic pooling layer computes the mean and standard deviation of intermediate features. Then, five layers (including a bottleneck layer) are incorporated after the pooling layer to classify the speakers during training. The cross-entropy loss function is employed to train the network to classify speakers. During verification, the speaker embedding is typically extracted from the pre-final bottleneck layer of the network and usually has a dimension of 512. After the extraction, the subsequent classification and softmax layers are not used. Before scoring, the speaker embedding is length-normalized.

Although the cross-entropy loss aims to train the network to separate speakers, researchers have sought other loss functions that improve the speaker embedding representation. Recent advancements in model training include the incorporation of angular/cosine-margin-based losses, which have the objective of minimizing intra-variance and maximizing inter-variance. Among these losses, a particularly noteworthy one is the ArcFace loss, also known as the Additive Angular Margin (AAM) loss (Liu et al. 2019). This loss function has gained significant attention and has shown promising results in speaker verification.

2.3.2 Scoring function

The scoring function goal is to compare how close two speaker embeddings are. Since it is almost impossible to have the exact same vector between the claimed (also called enrollment) and test embeddings (which would result in a binary yes/no same speaker), the scoring methods

provide a score indicating how much the two vectors correspond to the same speaker. If this score is higher (or lower) than a predefined threshold, the system accepts (or rejects) the trial. The most straightforward comparison is to compute the score with the cosine similarity (Dehak et al. 2010) function. Alternatively, the more sophisticated Probabilistic Linear Discriminant Analysis (PLDA) (Ioffe 2006; Prince et al. 2007a) can provide better scores on cross-entropy trained systems.

2.3.2.1 Cosine Similarity

Cosine similarity scoring is a computationally efficient method in many verification tasks. The cosine similarity is a measure of the angle between the claimed (\mathbf{x}^C) and test (\mathbf{x}^T) embeddings. This technique has the advantage of not requiring any training. Scoring is performed directly in the speaker embedding space.

2.3.2.2 Probabilistic Linear Discriminant Analysis

Unlike the cosine similarity, PLDA is a supervised method where speaker labels are necessary to estimate the PLDA parameters. Several PLDA variants exist (Prince et al. 2007b; Vaquero et al. 2011; Ramoji et al. 2020). In this thesis, the same one as in (Vaquero et al. 2011) is employed. It is known as the two-covariance PLDA variant and is implemented in Kaldi. The two-covariance PLDA is used to model the distribution of speaker embeddings in a multidimensional space. In the two-covariance PLDA, the embedding \mathbf{x}_i from the i -th speaker is assumed to be generated from a linear Gaussian model $p(\mathbf{x}_i | \mathbf{y}_i) = \mathcal{N}(\mathbf{x}_i | \mathbf{y}_i, \mathbf{W}^{-1})$, where \mathbf{y}_i represents the mean of the i -th speaker and is also assumed to follow a Gaussian distribution $p(\mathbf{y}_i) = \mathcal{N}(\mathbf{y}_i | \boldsymbol{\mu}, \mathbf{B}^{-1})$. \mathbf{W}^{-1} and \mathbf{B}^{-1} are the within-speaker and between-speaker covariance matrices that are used to capture the variation in speaker embeddings within and between different speakers, respectively. $\boldsymbol{\mu}$ is the global mean in the speaker space. The iterative E-M algorithm (Dempster et al. 1977) is usually used to estimate the \mathbf{W} , \mathbf{B} and $\boldsymbol{\mu}$ PLDA parameters. The score between two embeddings x^C and x^T is computed as a likelihood ratio based on two same/different class hypotheses as follows:

$$s(\mathbf{x}^C, \mathbf{x}^T) = \frac{\text{likelihood (same speaker)}}{\text{likelihood (different speaker)}} = \frac{p(\mathbf{x}^C, \mathbf{x}^T)}{p(\mathbf{x}^C)p(\mathbf{x}^T)} \quad (2.9)$$

where $p(\mathbf{x}^C, \mathbf{x}^T) = \int p(\mathbf{x}^C | \mathbf{y}) p(\mathbf{x}^T | \mathbf{y}) p(\mathbf{y}) d\mathbf{y}$ and $p(\mathbf{x}^C) = \int p(\mathbf{x}^C | \mathbf{y}) p(\mathbf{y}) d\mathbf{y}$. One advantage of using PLDA is that it explicitly takes into account speaker variability, which can improve the accuracy of speaker verification for the chosen system (TDNN). In contrast, cosine distance only measures the similarity between two vectors and does not consider the distribution of speaker embeddings.

In general, both PLDA and cosine distance are useful for speaker verification, but the appropriate method depends on the speaker embedding extractor and the available data.

2.3.3 Evaluation

Evaluating the performance of ASV systems is a delicate process, as many parameters can influence the reliability of the system. Those factors, listed in (Bimbot et al. 1997), can range from the quality of the speech signals to the quantity of speech for a given speaker or the size of the population of speakers. Evaluation campaigns have been introduced to produce standardized datasets and establish assessment procedures that enable an objective comparison.

Speaker verification systems are evaluated using a test dataset containing multiple trials. Each trial consists of a claimed identity and test audio segment. If the claimed identity and the audio segment point to the same speaker, the trial is considered *genuine*, else if they belong to different speakers, it is considered as *impostor*. The ASV systems generate a score for each trial; high scores values reflect a high similarity, and low values indicate a difference. An example of the score distributions for genuine and impostor trials is illustrated in Figure 2.6. In order to make a decision, a decision threshold t must be set. This threshold directly affects the two errors made by the ASV systems. A low threshold value will accept too many impostor trials, while a high threshold increases the chance of rejecting genuine trials. The two types of errors can be measured with two metrics:

- The False Acceptance Rate (FAR) measures how many impostor trials are accepted.
- The False Rejection Rate (FRR) measures how many genuine trials are rejected.

Increasing the detection threshold t decreases the FAR while increasing the FRR, and vice versa.

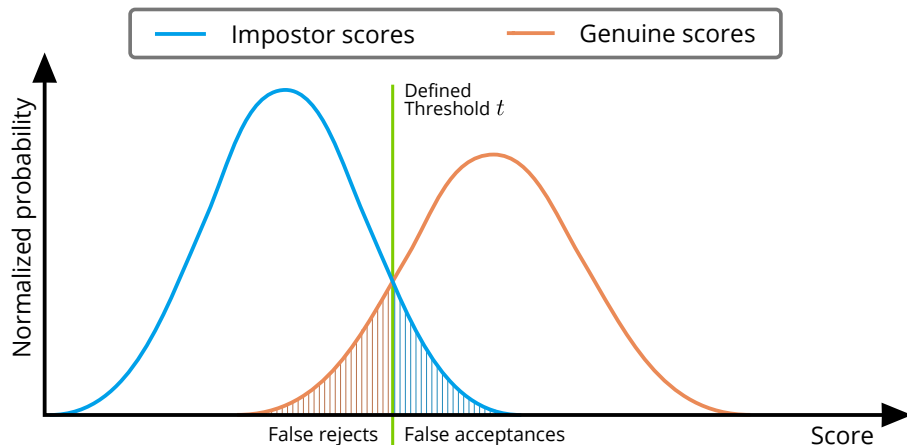


Figure 2.6 – Threshold-based decision-making on impostor/genuine scores

2.3.3.1 Equal Error Rate metric

The Equal Error Rate (EER) metric is one of the most popular metrics in speaker verification as it compares two systems based on a single threshold (t_{EER}). The t_{EER} is defined as the point

where the FAR and FRR values are equal. This configuration is reached by varying the threshold until the two areas corresponding to the false acceptances and false rejections become equal (see example in Figure 2.6).

$$\text{EER} = \text{FAR}(t_{\text{EER}}) = \text{FRR}(t_{\text{EER}}) \quad (2.10)$$

The EER values range from 0% to 50%. The best value of EER is 0% of error, meaning the impostor scores are entirely separated from the genuine one at threshold t . Whereas 50% of EER means the system’s performance equals random decision.

Limitation of the EER metric and uncalibrated scores

A limitation of threshold-based metrics is that they evaluate speaker verification given a decision threshold. The EER is considered to be an accurate indicator when the scores behave as they are expected, i.e. the higher score, the most likely the trial is genuine (like in Figure 2.6). However, when the scores are not as well distributed, evaluations based only on a single decision threshold will not necessarily be reflective of real-world application. Other decision-based threshold metrics, in particular, the log-likelihood-ratio cost function (Cllr) (Brummer et al. 2006), evaluate every possible threshold value (considering all FAR / FRR scores) and also considers calibration.

As shown in (Maouche et al. 2020), against uncalibrated scores, the EER fails to evaluate the discriminative potential of ASV systems. An extreme kind of uncalibrated score distributions is shown in Figure 2.7, where the genuine score distribution is located between the two modes of the impostor scores distribution. In this example, without calibration, the EER equals 50%, indicating the system performs random decisions. A solution to better evaluate this example is to calibrate the scores to be proper log-likelihood ratio scores with a non-monotonic transformation (Leeuwen et al. 2014). After calibrating the scores, the EER equals 15%, a much lower value that better represents the discriminative potential of the system. Such a situation is unlikely to occur in speaker verification assessment, where unmodified speech data is used. However, in the field of speaker anonymization, where speech is anonymized, unexpected, more overlapping score distribution becomes frequent (Maouche et al. 2020), as the ASV system will have more difficulty recognizing the speaker’s identity.

2.3.3.2 Linkability metric

The linkability metric introduced in (Gomez-Barrero et al. 2018) for biometric template protection solves threshold-based metric limitations by analyzing the overlap between the impostor/genuine distributions (like the Cllr). However, in contrast to the Cllr which usually performs calibration with a Pool Adjacent Violators (PAV) (Leeuwen et al. 2007), the linkability metric uses histogram binning which works better with extremes uncalibrated score that requires a non-monotonic transformation.

The local linkability metric is based on a score-wise measure that depends on the likelihood

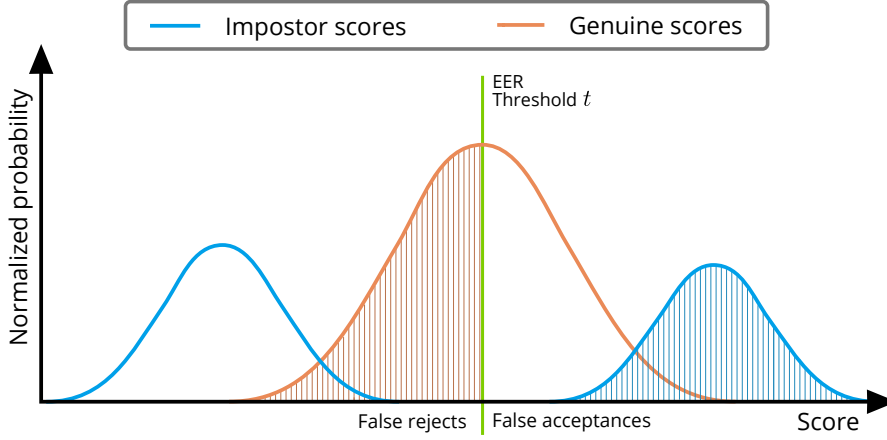


Figure 2.7 – Simulated uncalibrated impostor and genuine scores where the genuine scores are between the two modes of the impostor scores.

ratio between scores distributions: $p(H|s) - p(\bar{H}|s)$, where s is the score, H is the binary variable for a genuine trial and \bar{H} for an impostor trial. When the local linkability is negative, it indicates with high confidence that the score corresponds to different speakers. However, as this measure is targeted to describe the strength of linkability rather than the strength of unlinkability (different speakers), negative values are clipped:

$$D_{\leftrightarrow}(s) = \max(0, p(H|s) - p(\bar{H}|s)) \quad (2.11)$$

The global linkability measure $D_{\leftrightarrow}^{\text{sys}}$ is the average value of D_{\leftrightarrow} over all genuine scores:

$$D_{\leftrightarrow}^{\text{sys}} = \int p(s|H) \cdot D_{\leftrightarrow}(s) ds \quad (2.12)$$

In practice, $D_{\leftrightarrow}(s)$ is rewritten as $(2 \cdot \omega \cdot \text{lr}(s)) / (1 + \omega \cdot \text{lr}(s)) - 1$ where the likelihood ratio $\text{lr}(s)$ is $p(s | H) / p(s | \bar{H})$ and the prior probability ratio ω is $p(H) / p(\bar{H})$, and $p(s | H)$ and $p(s | \bar{H})$ are computed via one-dimensional histograms.

The $D_{\leftrightarrow}^{\text{sys}}$ linkability values range from 0.00 to 1.00, and in opposition to the EER, the higher the $D_{\leftrightarrow}^{\text{sys}}$, the better the ASV system can accept or reject trials.

The advantage of the $D_{\leftrightarrow}^{\text{sys}}$ metric over the EER is the non-single-threshold-based evaluation that it makes. An advantage that is also shared with the Cllr. The $D_{\leftrightarrow}^{\text{sys}}$ advantage over the Cllr is the default calibration function that it uses. As against scores non-monotonically related to the likelihood ratio, histogram binning is better than PAV. It is for that reason that the $D_{\leftrightarrow}^{\text{sys}}$ is strongly advocated by the authors of (Gomez-Barrero et al. 2018; Maouche et al. 2020) and used in this thesis for robust ASV evaluation under adversary and anonymization conditions. It is worth noting that in practice, on real data, the Cllr and $D_{\leftrightarrow}^{\text{sys}}$ follow a clear relation (Maouche et al. 2020).

2.4 Voice conversion

Voice Conversion (VC) is a discipline where the goal is to modify the voice characteristics of a source speaker’s speech to match those of a target speaker without changing the linguistic content. The principle of VC is to define a transposition function that converts the speech of one (or more) source speakers to the voice of one (or more) target speakers.

Traditional VC systems require parallel speech corpus, in which speech recordings come in pairs by the source speaker and the target speaker. Direct relations between source and target pairs enable the creation of the transposition function. Numerous approaches have been proposed, such as GMM VC (Stylianou et al. 1998; Kain et al. 1998; Toda et al. 2007a), frequency warping VC (Erro et al. 2010; Godoy et al. 2012), DNN VC (Desai et al. 2010; Nakashika et al. 2013). All the above approaches provide reasonably good results, the best being DNN-based. However, the requirement of a parallel corpus causes limitations as, in practical applications, parallel data is not easily available. Hence, lately, the VC research community has mainly focused on approaches where non-parallel data is used to build VC systems.

Non-parallel trainable VC systems are more valuable as training data acquisition is much easier, despite the fact that training is more complex. Most non-parallel methods rely on separating the linguistic and speaker-related representations carried out by acoustic features. During training, the model is asked to reproduce the speech of the source (same identity). At the conversion stage, the linguistic content of the source speaker utterance is extracted by a linguistic encoder and kept unmodified. In contrast, the speaker representation (usually a one-hot or x-vector embedding) is derived from the target speaker’s speech. A synthesis model is then used to generate speech with the target speaker’s characteristics with the linguistic content of the source speaker (example in Figure 2.8).

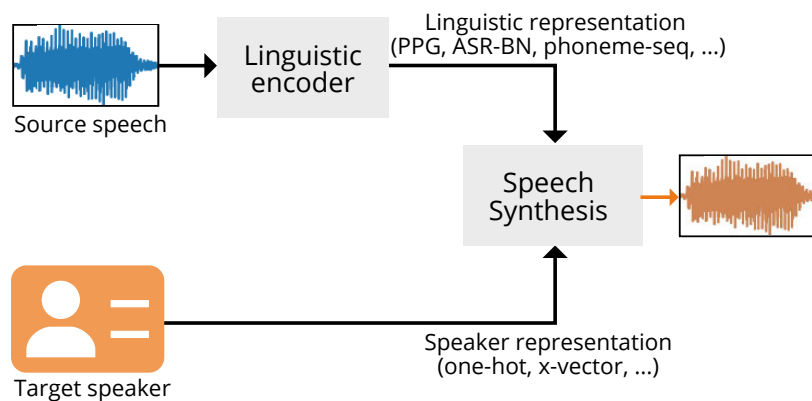


Figure 2.8 – A typical flow of a voice conversion system.

The following sections present some methods to separate linguistic and speaker information of speech signals. Then we categorize voice conversion approaches based on the numbers and origins of the source and target speakers. Last, we present the current state-of-the-art speech synthesis model.

2.4.1 Linguistic representation

Linguistic information extraction can be divided into two categories, supervised and self-supervised. Supervised representation typically uses an Automatic Speech Recognition (ASR) model to extract the linguistic representation. The linguistic representation of speech signals can be expressed in a variety of forms, such as a sequence of discrete tokens (i.e., phonemes) as described in (Huang et al. 2020a), a sequence of Phonetic Posteriorgrams (PPG) as described in (Sun et al. 2016; Liu et al. 2018), or a sequence of bottleneck features from an acoustic model as described in (Zhang et al. 2020). This category requires phoneme (or text) supervision during the training as an ASR model is necessary. The second category does not require any label as it relies on self-supervised objectives. Self-supervised objectives allow learning from unlabeled data by using the inherent structure of the data to generate labels for itself. Models implicitly separate linguistic and speaker features in a single model using auto-encoder architectures (Hsu et al. 2016; Chou et al. 2019; Qian et al. 2019; Wu et al. 2020) or generative adversary networks (Gao et al. 2018).

For this thesis, the text supervision required to train the linguistic encoder is a realistic application setting. Extracting linguist representation using ASR is an easier framework than letting an auto-encoder find the separation on its own while also having high-converted voice quality (Qian et al. 2019). Text supervision helps when training the linguistic encoder as a clear indication of what constitutes an appropriate linguist representation is available. Additionally, this method of converting voices has been proven relevant for speaker anonymization (Fang et al. 2019). The following section presents the main methods to obtain supervised linguistic representations for voice conversion.

2.4.1.1 Discrete token sequence as linguistic features

One of the most straightforward methods to implement VC is to extract a sequence of discrete tokens (e.g., phonemes, text) from an ASR system and feed it to a Text-To-Speech (TTS) system, this method is referred to as cascade ASR+TTS (Huang et al. 2020b). The major advantage and drawback of cascade ASR+TTS is the discrete token used to transmit information from the ASR to the TTS system (example: “b ey s b ə l” for the word “baseball”). The advantage is that no acoustic speaker information is contained in discrete token representation allowing the TTS more consistency in synthesizing specific target speaker characteristics across different source speakers. The disadvantages are that early stages errors propagate to downstream models, meaning any recognition failure in the first ASR stage will harm linguistic consistency, creating mispronunciation errors in the VC pipeline. Significant degradation of emotions and speaker intention cues are also expected, as the discrete token representation does not encode the acoustic features which enable human understanding of the prosody. Altering those aspects may affect the sentence’s understanding as a statement may be synthesized similarly to a question. Even

if TTS system can be conditioned on with on prosodic (Raitio et al. 2020) features, prosody recognition stays a complicated and very subjective topic of research (Rosenberg 2018).

2.4.1.2 Phonetic posteriorgrams sequence as linguistic features

To mitigate some mispronunciation and prosody transfer errors (like the speech rate, or the articulation) of the aforementioned discrete representation, PPG (Sun et al. 2016) were introduced for voice conversion. A PPG is a time/class matrix representing the posterior probabilities of each phoneme class for each time frame of an utterance. In other words, a soft label is generated for each speech frame, indicating the likelihood of each possible phoneme being uttered (see example in Figure 2.9).

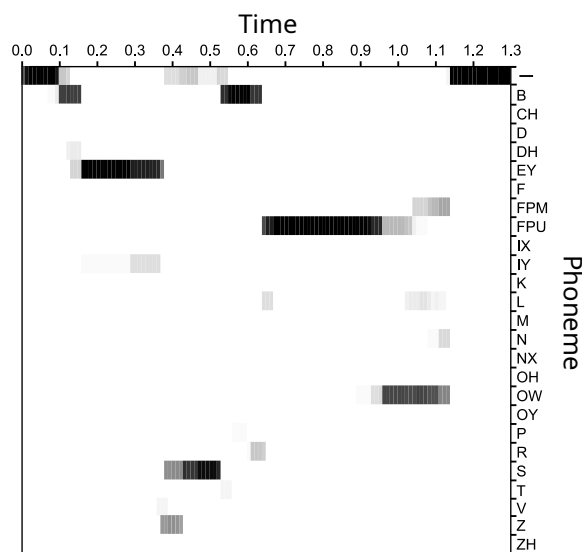


Figure 2.9 – An example Phonetic Posteriorgrams for the spoken word “baseball”. (reproduced from: « [Query by example spoken term detection](#) », for display purposes, not all phonemes are listed on the y-axis).

In contrast to discrete tokens, PPG encodes pronunciation duration as the same consecutive labels are not merged. This helps to keep the source speaker’s emotions and intention information in the converted speech. Furthermore, PPG are less sensitive to modeling error than a sequence of discrete tokens because no decoding is made. Operating at the frame level and extracting soft phoneme labels makes PPG more resilient to the source speaker’s pronunciation, as likelihood scores can encode non-common pronunciations. However, PPG are usually extracted after the softmax layer of a TDNN-F-based acoustic model (see Section 2.2.1). And, as a result of ASR training, the soft label values of PPG are usually bimodal (highest values close to one and other values close to zero). This creates a deficiency where mispronunciations can still occur in the pipeline if the acoustic model does not generalize (due to domain mismatch or noisy recording for example).

2.4.1.3 ASR bottleneck sequence as linguistic features

To address the PPG deficiencies mentioned above, linguistic features can be extracted from the lower intermediate bottleneck layer rather than at the softmax layer (Liu et al. 2021a). This simple modification makes bottleneck linguistic features similar to PPG while greatly helping with noise, accent, and domain mismatch generalization. However, bottleneck from automatic speech recognition (ASR-BN) based linguistic features also encode more speaker-related information (Fang et al. 2019; Adi et al. 2019) than PPG, which can restrict downstream speaker modification in the VC pipeline. Bottleneck-based linguistic features for voice conversion are gaining popularity (Polyak et al. 2020; Zhang et al. 2020; Liu et al. 2021b; Liu et al. 2021a), and in fact, we also use them for the approaches proposed in this thesis.

2.4.2 Speaker representation

Similarly to the linguist representation, speaker representation can be extracted by supervised specialized models (Snyder et al. 2018) or self-supervised in auto-encoder-based VC models (Qian et al. 2019). For this thesis, we focus on supervised methods where the two main speaker representations employed are x-vector (presented in Section 2.3.1) and one-hot embedding.

One-hot embeddings represent categorical variables as binary vectors in a format compatible with machine-learning algorithms. One-hot speaker embedding vectors have all values set at zero except for the speaker’s index. Thus, one dimension is allocated for each speaker in the dataset.

During training, VC learns to synthesize speech from the linguistic and speaker representations. Depending on the speaker representation, a VC system might be constrained to a limited number of speakers (Toda et al. 2007b; Ohtani et al. 2009; Liu et al. 2021a). For one-hot embeddings, the target speaker must be present in the training data, whereas for x-vectors, the target can be arbitrary. One-hot speaker embedding performs the best in terms of synthesis quality because the target speaker is present in the training dataset. However, when an application needs to consider new target speakers, x-vectors or a similar speaker representation is necessary as the target speaker might not be present in the training dataset. In the following section, we describe the different variants that may lead a system to employ either one-hot or x-vector representations.

2.4.2.1 Numbers and origins of the source and target speakers

An essential characteristic of VC methods is the number and origin of source speakers and target speakers that a single system can support. VC approaches can be categorized into *one-to-one*, *many-to-one*, *many-to-many*, *any-to-many* and *any-to-any* VC. The following list briefly presents the differences between them:

- *One-to-one* VC is the easiest model to build as the mapping function is limited to a specific pair of source and target speakers; thus, there is no need to generalize to multiple and unseen speakers (Fang et al. 2018).

- *Many-to-one* (or *one-to-many*) VC approaches extend the versatility by extending VC to multiple source (or target) speakers (Toda et al. 2007b). The source and target speakers must be present in the training dataset.
- *Many-to-many* VC approaches further improve the functionality as a single system can convert a known source speaker into a known target speaker voice (Ohtani et al. 2009). The source and target speakers must still be present in the training dataset.
- *Any-to-many* VC approaches allow converting an unseen source speaker into a known target speaker voice (Liu et al. 2021a).
- *Any-to-any* VC approaches achieve conversion across unseen source-target speaker pairs without prior knowledge of them (Liu et al. 2018). Generalizing to arbitrary sources and target speakers makes this approach more complicated than the others, and usually results in lower synthesized voice quality.

For this thesis, the most suited approaches rely on *any-to-many* (using one-hot embedding) or *any-to-any* (using x-vector embedding) voice conversion as the speech signals to anonymize come from unknown speakers.

2.4.3 Speech synthesizer

Given the linguistic and speaker representations, most pipelines separate speech synthesis into two steps: decoder and vocoder. The decoder’s purpose is to generate an output Mel spectrogram. Then, the vocoder transforms the Mel spectrogram into the output waveform. The linguistic and speaker representations need to be concatenated before being given to the decoder. To do so, the utterance level speaker representation is incorporated into each linguistic frame.

The decoder can be built using recurrent neural networks (Qian et al. 2019), fully convolutional architectures (Wu et al. 2020) or transformer (Huang et al. 2020b). On the other hand, popular vocoders include Griffin-Lim (Perraudin et al. 2013) algorithm and neural network models such as WaveGlow (Prenger et al. 2019) and HiFi-GAN (Kong et al. 2020). State-of-the-art approaches require fine-tuning the vocoder on the decoder’s Mel spectrograms (Lorenzo-Trueba et al. 2019). Training the vocoder this way helps to smooth out some decoder errors in contrast to training the vocoder only on real Mel spectrograms and using it with converted Mel spectrograms.

Recently, with the development of GAN-based techniques, this two-step synthesis pipeline has been compressed into one (Polyak et al. 2021; Kashkin et al. 2022). The efficient HiFi-GAN vocoder framework can be adapted to perform both decoder and vocoder tasks removing the need for the intermediate Mel spectrogram in the synthesis phase. Waveforms can be generated from the linguistic and speaker representation directly. As a result, this technique does not necessitate training a decoder and then fine-tuning the vocoder. The HiFi-GAN framework is presented in more detail in the subsequent section.

2.4.3.1 HiFi-GAN

Compared to other vocoders, the Hi-Fi-GAN (High Fidelity) (Kong et al. 2020) model delivers superior computational efficiency and sample quality. Those results are achieved using a non-autoregressive network architecture and carefully crafted loss functions based on GAN (see 1.4.6). The HiFi-GAN framework consists of a generator, G , and a set of two types of discriminators, D . The generator produces a speech signal from a sequence of features corresponding to linguistic and speaker representations (direct VC with implicit decoder) or Mel spectrograms (vocoder only). Then, various modules generate speech from the features using successive transposed convolutions with residual blocks having dilated layers. The transposed convolutions upsample the features to match the temporal resolution of the corresponding waveform, while the dilated layers increase the receptive field.

Natural speech waveforms contain long-term dependencies. For instance, if a phoneme lasts more than 100 ms, the raw waveform will exhibit a strong correlation between more than 1 600 adjacent samples¹. Furthermore, waveforms consist of sinusoidal signals with various periods. Modeling periodic patterns and long-term interdependence is essential to produce realistic speech audio. Those problems have been addressed in the HiFi-GAN framework using two different types of discriminators. During training, the multiscale and multi-period discriminators receive the generated audio sample \hat{x} to determine if the produced audio is realistic.

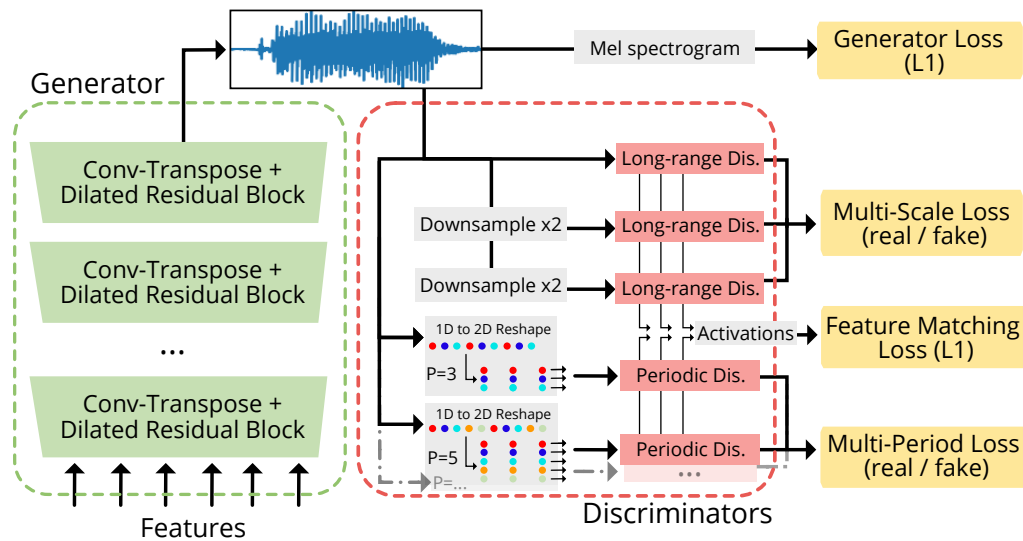


Figure 2.10 – Generator, discriminators, and training losses of a HiFi-GAN model.

The multi-scale discriminators are asked to explore long-range and consecutive audio interactions. As presented in Figure 2.10, three long-range discriminators operating at different audio scales (raw audio, x2 downsampled audio, and x4 downsampled audio) are used to assess audio samples at various ranges. On the other hand, five multi-period discriminators explore multiple

1. Depending on the sampling frequency.

periodic patterns (only two of them are displayed in the figure). All period discriminators differ from each other based on the space between the samples. As shown in Figure 2.10, the audio is first reshaped from a 1D structure to a 2D structure of length T/p and height p , T is the length of the raw audio file and p the period to analyze. The p periods to analyze by each of the period discriminators are [2, 3, 5, 7, 11], they are chosen to avoid overlaps as much as possible. The periodic samples are independently processed by the multi-period discriminators using convolution having a kernel size of one on the height axis (row by row in Figure 2.10). The overall training losses of HiFi-GAN involve a set of adversarial and discriminator losses (\mathcal{L}_{adv} , \mathcal{L}_D), a generator loss \mathcal{L}_G , and a feature-matching loss. For each j discriminator (8 in total), D_j is tasked with minimizing the following losses:

$$\mathcal{L}_{adv}(D_j, G) = \sum_{\mathbf{x}} \|1 - D_j(\hat{\mathbf{x}})\|_2^2 \quad (2.13)$$

$$\mathcal{L}_D(D_j, G) = \sum_{\mathbf{x}} \left[\|1 - D_j(\mathbf{x})\|_2^2 + \|D_j(\hat{\mathbf{x}})\|_2^2 \right] \quad (2.14)$$

where $\hat{\mathbf{x}}$ is obtained from the generator G , \mathbf{x} is a real sample. Equation 2.13 purpose is to encourage the generator to create (fake) audio indistinguishable from genuine (real) audio. Equation 2.14 goal is to train the discriminator to differentiate real and fake audio. The least-square GAN loss variant is used (Mao et al. 2017).

A reconstruction loss between the Mel-spectrogram of the real audio signal and the generated signal is used to train the generator efficiently. The Mel-spectrogram loss is the L1 difference and is described as:

$$\mathcal{L}_{recon}(G) = \sum_{\mathbf{x}} \|\phi(\mathbf{x}) - \phi(\hat{\mathbf{x}})\|_1 \quad (2.15)$$

where ϕ is a function that computes a Mel-spectrogram from a waveform.

Finally, a feature-matching loss (Larsen et al. 2016) measures the distance between the discriminators internal activation for real audio and fake audio to help the generator better create fake signals.

$$\mathcal{L}_{fm}(D_j, G) = \sum_{\mathbf{x}} \sum_{i=1}^R \frac{1}{M_i} \|\psi_i(\mathbf{x}) - \psi_i(\hat{\mathbf{x}})\|_1 \quad (2.16)$$

where ψ is the operator extracting the discriminator activations at layer i , M_i the number of weighs in layer i , and R the total number of layers in D_j .

The final generator and discriminators losses to minimize can be written as:

$$\begin{aligned} \mathcal{L}_G^{\text{multi}}(G, D) &= \sum_{j=1}^J [L_{adv}(D_j, G) + \lambda_{fm} L_{fm}(D_j, G)] + \lambda_r L_{recon}(G), \\ \mathcal{L}_D^{\text{multi}}(G, D) &= \sum_{j=1}^J L_D(D_j, G) \end{aligned} \quad (2.17)$$

where J is the number of discriminators (here 8), λ_{fm} is set to 2 and λ_{recon} to 45 to balance the adversarial losses, Mel-spectrogram loss, and the feature matching loss.

2.4.4 F_0 conditioning

Speech converted using the above pipeline, where the two main features are the bottleneck linguistic and speaker representations, exhibit inconsistent F_0 distribution compared to real speakers distribution (Huang et al. 2019; Qian et al. 2020a). These inconsistencies come from the bottleneck linguistic representation that encodes some prosody information (mainly F_0). As a result, the synthesis model generates speech with unnatural F_0 values based on the linguistic bottleneck and target speaker representation. Inadequate encoding of the target speaker’s prosodic style is the main limitation of speaker representations. A straightforward method to address this problem is to have a normalized or transformed F_0 sequence added to the linguistic and speaker representations mix to help the synthesis model modify the F_0 trajectory given the target speaker representation.

2.4.5 Evaluation

The quality of a voice conversion system (or any speech synthesis system) resides in its ability to convey a linguistic message by speaking the correct sequence of words while having suitable prosody and a similar speaking style as the target speaker. Objective and subjective evaluations are necessary to assess the synthesized speech quality in all aspects. For objective evaluations, a ASR system with the WER metric can be a good proxy metric to assess if the linguistic message is still intelligible (understanding what is being said). Concerning the target speaker identity, the use of a ASV can give some indication of whether or not the synthesized speech can be associated with the actual speech of the target speaker. Those two evaluations are interesting because they provide a score without having human listeners. However, human listeners must be present in the evaluation loop to measure the more subjective aspects of speech. For instance, humans are needed to assess factors like naturalness (similarity to a natural voice), intonation, and other prosody-related features. Even the intelligibility, which the ASR system can estimate, needs to be evaluated subjectively because a human’s understanding of speech differs from that of an ASR model.

Many scoring exists for subjective evaluation, most notably the Mean Opinion Score (MOS) score is the most commonly used in the literature. The MOS score evaluates a system by candidates ranking from 1 (poor) to 5 (excellent) each sample in different dimensions (listening effort, intelligibility, naturalness, quality, rhythm, intonation, etc.).

2.5 Conclusion

This chapter presented the speech-centric machine learning disciplines crucial for the task of speaker anonymization. It covers the basics of acoustic modeling, Automatic Speech Recognition (ASR), Automatic Speaker Verification (ASV) and Voice Conversion (VC). In the next chapter, we delve into the topic of speaker anonymization and see how these disciplines combine together.

SPEAKER ANONYMIZATION

Contents

| | | |
|------------|--|-----------|
| 3.1 | Introduction | 49 |
| 3.2 | Legal perspectives | 50 |
| 3.3 | Threat model | 51 |
| 3.3.1 | Attacker capabilities | 52 |
| 3.4 | Application cases | 55 |
| 3.5 | Evaluation methods | 56 |
| 3.5.1 | The VoicePrivacy challenge requirements | 56 |
| 3.5.2 | Privacy model and metrics | 59 |
| 3.5.3 | Utility model and metric | 59 |
| 3.6 | Current methods for speaker anonymization | 60 |
| 3.6.1 | Signal processing anonymization | 60 |
| 3.6.2 | Voice conversion anonymization | 61 |
| 3.6.3 | Adversarial ASV attack anonymization | 65 |
| 3.7 | Conclusion | 66 |

3.1 Introduction

One modern technology that is used by many people is biometrics recognition ([Langenderfer et al. 2005](#)). The most common examples are fingerprint or face recognition technologies used in smartphones. However, online social platforms also used them to better track their users. These technologies aim to extract individual personal characteristics from biometric data into a biometric template used to verify their identity. The characteristics might be physical traits like fingerprints or behavioral traits like a particular method to solve a security-authentication puzzle ([Roth et al. 2004](#); [Hirakawa et al. 2018](#)). Although many end-users have become more familiar with this technology ([Habibu et al. 2022](#)), recent federal regulations strictly restrict the use and storage of biometric and personal data. The nature of data necessitates such laws to prevent biased decisions depending on the gender, origin, and other personal attributes of the user, which would raise privacy and ethical issues. Additionally, users are more vulnerable in the

event of a data breach since, unlike passwords and tokens, compromised unmodified biometric data cannot be revoked and reissued.

By nature, speech falls into the categories of physical and behavioral biometrics because its generation depends on physical traits like the shape of the vocal tract and personality traits like extroversion (Trilok et al. 2004; Polzehl et al. 2010). Moreover, speech encapsulates a large amount of personal data, like age, gender, health and emotional state, racial or ethnic origin, etc. encouraging the need to develop privacy-enhancing solutions for speech technology (Kröger et al. 2019). One modern technology used to enhance the privacy of the user when sharing speech data is anonymization. In this chapter, we present the legal aspects regarding data regulation, the application cases, the threat models and current systems and methods to anonymize speech signals.

3.2 Legal perspectives

The General Data Protection Regulation (GDPR 2016) obligates the entity storing personal data to implement all possible technical measures to enforce the protection of personal data by the principle of data protection by design and by default. More specifically, Article 9 of the GDPR prohibits the processing of biometric data that shows the gender, origin or health indications of a person. This legislation is applicable for authentication platforms relying on fingerprint, face recognition, or other forms of template-based verification (Jain et al. 2008) to avoid as much as possible unfair bias. However, for certain modalities such as speech, this constraint is too restrictive as speech data inherently contains biometric information and additional useful information that does not fall under the same data regulation. For that reason, Recital 26 of the GDPR relaxes this constraint by allowing the processing of anonymous data that can not identify the user. Article 5(c.1) of the GDPR mentions the principle of data minimization which is to only collect data strictly necessary for the usage/improvement of the service. In the case of a speech recognition service, being able to remove the other sensible attributes unnecessary to speech recognition will help to comply with data minimization. To obtain anonymous and minimized data, one can rely on anonymization¹ methods to remove biometric clues from the data. The strength of anonymization depends directly on the technical measures available, which motivates the pursuit of this thesis.

Related to biometric data protection, the European standard ISO/IEC 24745 (ISO-24745 2011) defines multiple criteria regarding the processing of biometric data. The two criteria studied in the thesis that can be borrowed for data anonymization are *invertibility* and *unlinkability*. First, the *invertibility* criterion aims to prevent the use of biometric data for any purpose other than

1. From a legal standpoint, "anonymization" refers to methods that achieve complete concealment. In this thesis the term "privacy enhancement" is more appropriate as we do not fulfill full concealment, however, the term "anonymization" is more broadly used by the community.

the ones originally intended, for that reason, biometric data must be processed by irreversible transformation before storage. Secondly, the *unlinkability* criterion aims to ensure that stored biometric templates can not be linkable across applications or databases. Linkability (the opposite of *unlinkability*) is considered the main threat in this thesis and will be explored in more detail in the following section.

It is worth mentioning two other federal publications, the white paper "Explore legal, technical and ethical issues associated with voice assistant" by (CNIL 2020), which present multiple practical use cases and applications of the GDPR, and the guidelines on virtual voice assistants by (EDPB 2021), which presents general recommendations regarding data retention, user profiling, data protection and more.

3.3 Threat model

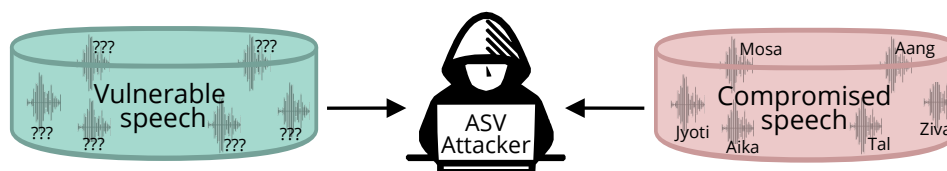


Figure 3.1 – The considered threat model, where the *vulnerable* speech needs to be protected with privacy-enhancing solutions removing biometric Personally Identifiable Information (PII) from the data. Whereas the *compromised* speech is unprotected data full of biometric PII, used by the attacker to find the original speakers of the *vulnerable* speech.

The main threat model in this thesis is a linkability attack which corresponds to multiple tests where, given two speech samples, an attacker can distinguish if they come from the same speaker or different speakers, this is similar to Automatic Speaker Verification (ASV) (see Chapter 2.3). An attacker refers to a formal or informal description of assumed capabilities that a security/privacy mechanism is designed to protect against. In this thesis, we consider multiple users having their speech samples in a compromised/public dataset, where each sample is associated with a speaker identity. The attacker has access to this data, and his goal is to link each known speaker's identity with an unknown speaker's speech coming from another dataset (Figure 3.1).

In this thesis, the goal is to remove biometric Personally Identifiable Information (PII) from a speech signal such that it becomes unlinkable to the speaker's identity, this task is referred to as speaker anonymization (Figure 3.2). This operation is judged to be necessary for publishing useful datasets for scientific and commercial use without compromising the users' privacy (Fung et al. 2010).

The evaluation of this PII removal (privacy performance) relies on the ASV model of the attacker. Until the work of (Srivastava et al. 2020b), most studies assumed the attacker had

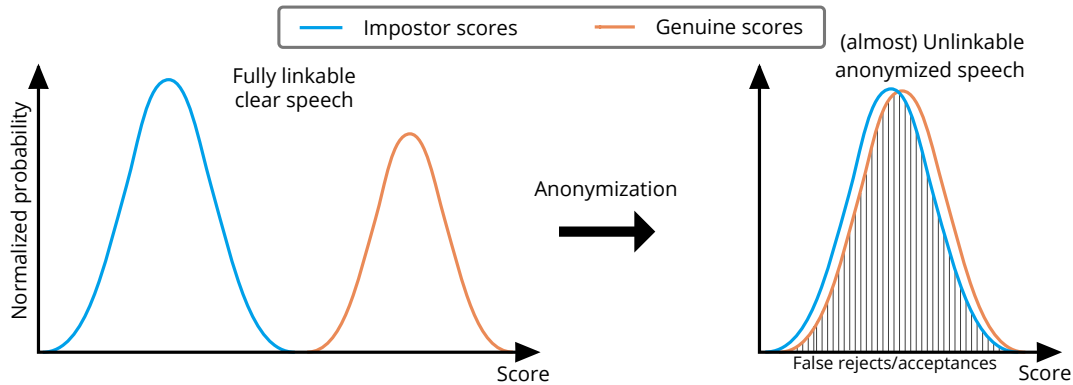


Figure 3.2 – Linkability assessment using an ASV system. In the example, before anonymization (left side), the speech is fully linkable, meaning the identity of the speakers could be verified all the time. To protect the identities, the speech should be as unlinkable as possible after anonymization (right side).

restricted capability i.e., the attacker was unaware of the anonymization. This aspect leads to a high domain mismatch between the *vulnerable* anonymized speech and *compromized* clear speech datasets which the ASV model was not trained to overcome, additionally, no countermeasures were employed to defeat the anonymization system (Hashimoto et al. 2016; Qian et al. 2017; Magarinos et al. 2017; Bahmaninezhad et al. 2018). In security and data protection, “clear data” refers to the original, unmodified and unprocessed data collected. In the security/cryptography communities, evaluation based on “security by obscurity” (Mercuri et al. 2003), where only weak attackers are considered is strongly refuted. In the following section, we present and categorize more effective methods for evaluating privacy performance with various ASV attackers having varying levels of capability.

3.3.1 Attacker capabilities

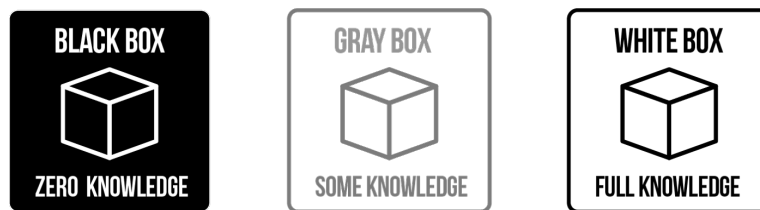


Figure 3.3 – Adversary’s knowledge about the anonymization procedure.

In software quality and security, testing techniques can be classified into three main categories, black-box, white-box and gray-box (see Figure 3.3) (Khan et al. 2012; Guo et al. 2019). Black-box testing is a testing scenario where zero knowledge of the application is required. Black-box testing provides a perspective where the attacker and the application are separated, however, it

is an inefficient testing scenario, and in the case of speaker anonymization assessment can be assimilated to as an evaluation following the “security by obscurity” principle. White-box testing is the case where full knowledge of the application and parameters is required. White-box testing is the most efficient testing scenario to identify the largest amount of vulnerabilities, at the cost of being more expensive and requiring expertise. In the middle, many gray-box testing scenarios can be crafted each having varying amounts of knowledge of the application. Gray-box testing offers assessment from the point of view of the users rather than the application designer, for that reason it is a realistic scenario for speaker anonymization evaluation, however, it may not identify all vulnerabilities. When this categorization is applied to test the privacy performance of speaker anonymization systems, (Srivastava et al. 2020b) multiple attacker schemes can be defined:

○ **Black-box**

- The *ignorant* attacker (II in Figure 3.4) is unaware that speech was anonymized.

○ **Gray-box**

- The *lazy-informed* attacker (III in Figure 3.4) is aware of the anonymization system but unaware of the parameter used to anonymize each utterance. This attacker tries to reduce the domain mismatch between the clear (non-anonymized) *compromized* speech and the anonymized *vulnerable* speech datasets by anonymizing itself the *compromized* speech dataset.
- The *semi-informed* attacker (IV in Figure 3.4) is aware of the anonymization system and some hyperparameters used but unaware of the exact parameters used to anonymize each utterance. This attacker reduces the domain mismatch between the *compromized* speech and the anonymized datasets by anonymizing the *compromized* speech (without using the most appropriate hyperparameters). Additionally, this attacker adapts the ASV evaluation model to work with somewhat relevant anonymized data.

○ **White-box**

- The *informed* attacker (V in Figure 3.4) is completely aware of the anonymization system, hyperparameters, and exact parameters used to anonymize each utterance. This attacker minimizes the domain mismatch between the *compromized* speech and the anonymized datasets by anonymizing the *compromized* speech using the most appropriate hyperparameters. Additionally, this attacker adapts the ASV evaluation model to specifically work against one anonymization system, hyperparameter and exact parameters used to anonymize each utterance.

For an anonymization system implemented through the use of voice conversion (see Chapter 2), being aware of the anonymization system means having access to the VC model and weights. Being aware of the hyperparameters refers to knowing the target selection strategy used to select

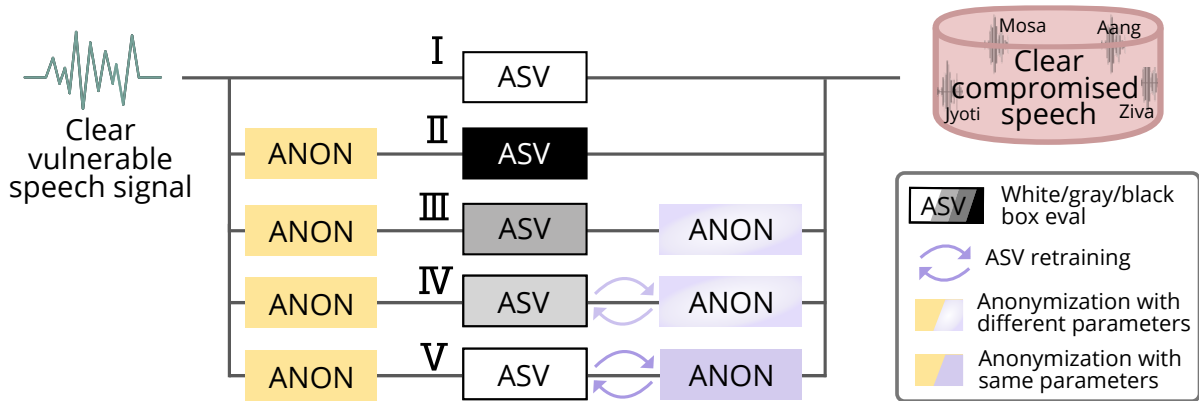


Figure 3.4 – Privacy evaluation in clear scenario I (baseline), *ignorant* II, *lazy-informed* III, *semi-informed* IV, and *informed* V attacker scenario.

the target speaker to anonymize each utterance (but not the exact target speaker). Being aware of the parameters means knowing the exact target speaker used to anonymize each utterance. The creation of the target speaker is explained in more detail in Section 3.6.2.2 with the presentation of an anonymization system.

The semi-informed and informed attackers allow better privacy performance evaluation as adapting (through retraining) the ASV model on anonymized speech avoids a domain mismatch. This process might also modify the main goal of a classical ASV model. Instead of directly recognizing each speaker from the others, the ASV model is trained to, to some extent, invert the anonymization to then identify the underlying speaker.

The semi-informed attacker is interesting in the fact that it is the most realistic attacker. As anonymization should be performed by each user before sharing their data, the anonymization system is open to the public, which means the attacker can gain control over it. With this anonymization system, the attacker can generate anonymized speech in a similar (but not exact) way as the users. With anonymized speech, the attacker can adapt both the *compromized* speech and ASV model to work with such kind of data. However, depending on the hyperparameters used, the attacker-generated anonymized speech might not enable correct retraining of his/her adapted ASV model, this aspect is explored in Chapter 4.

In contrast, the *informed* attacker might be unrealistic as having complete knowledge, even of a per utterance target speaker is impractical. However, as this attacker falls into the category of white-box testing, it enables an evaluation from the point of view of the application designer (the specialist) who has more ideas about the limitation of his/her system and can identify more vulnerabilities. With such capabilities, this attacker can precisely retrain his/her adapted ASV to ensure any improvement of privacy is indeed caused by a better anonymization system rather than a weak/unadapted ASV model.

The previously mentioned attackers rely on making sure the ASV system is properly adapted to recognize the original speaker from anonymized utterances, this process labeled “domain

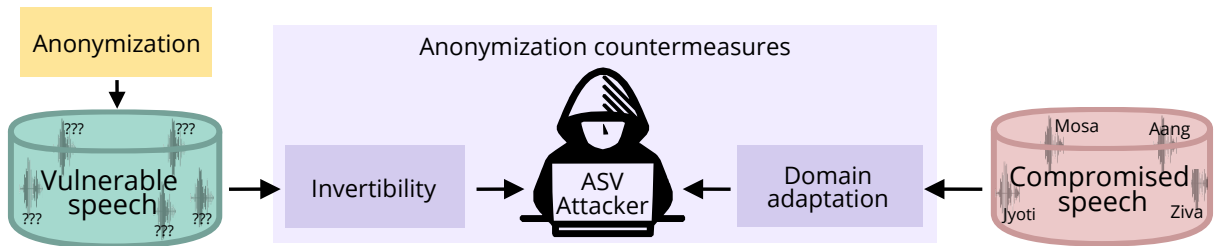


Figure 3.5 – Possible countermeasures an attacker may use to defeat anonymization.

adaptation” in Figure 3.5 does not consider the anonymized *vulnerable* speech as an attack vector. New kinds of “invertibility” attacks recently developed (Champion et al. 2021b; Kai et al. 2022) consider using the anonymized *vulnerable* speech to invert the anonymization procedure, such a form of attack is a contribution of this thesis and will be presented in more details in Chapter 6.

3.4 Application cases

While reducing linkability attacks, the anonymized speech also needs to be useful for other tasks, the terminology used to refer to this aspect is utility. The utility of anonymized speech depends on the performance obtained when used for downstream tasks, i.e., recognizing speech content, emotions, and more. Depending on the task, the anonymized speech utility requirements may differ from application to application. For instance, in the case of a voice assistant, the service provider wants to receive speech where the linguistic content is preserved such that an ASR can decode and comprehend. Additionally, the voice assistant service provider might also want to preserve various usage conditions, i.e, noisy environment, to improve their service under adverse conditions. Other requirements such as preservation of the intonation, naturalness, and voice distinctiveness (Tomashenko et al. 2022a) might be necessary for some applications. Some requirements like gender or emotion recognition are attributes that can fall into both a private aspect to suppress or a utility aspect to keep (Nourtel et al. 2021; Aloufi et al. 2020). Figure 3.6 presents a spectrum of a few privacy/utility requirements. Speaker anonymization has well-defined privacy and utility requirements for some of them, while for others it depends on the targeted application. Given this indecisive aspect, evaluations in this field should be done considering a specific use case.

The use case considered in this thesis corresponds to the one where a service provider wants to collect a large dataset to, improve their service. This thesis focuses on the data collection part, where the aim is to find a privacy-preserving transformation of the speech data that removes the biometric PII speech information that an ASV system may capture while keeping high utility for the linguistic content such that an ASR system can still decode the content. Additional requirements may be added while maintaining this core objective, which can be thought of as the fundamental objective that all speaker anonymization should share.

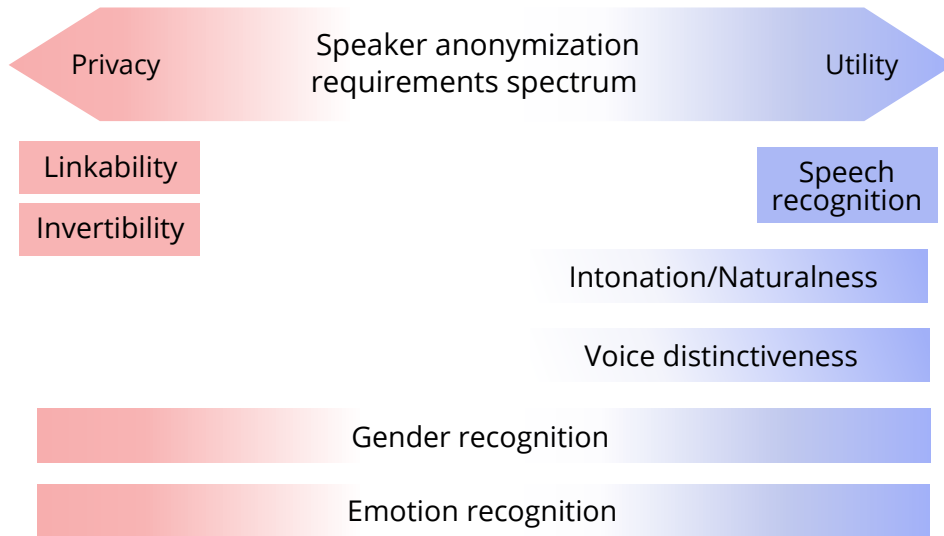


Figure 3.6 – Example of the range of requirements of a speaker anonymization system.

3.5 Evaluation methods

To assess the privacy and utility against ASV linkability attack and ASR decoding, the VoicePrivacy Challenge (VPC) provides a very relevant evaluation framework. As such, in this thesis, the experimental setup used in all experiments comes from the VPC evaluation framework. Established in 2020, the VoicePrivacy initiative ([Tomashenko et al. 2020](#)) is spearheading the effort to develop privacy preservation solutions for speech technology. Up until now, two challenges have taken place one in 2020 and the other in 2022. The most important difference between the two challenges is the use of a stronger semi-informed attacker in VPC 2022 instead of a lazy-informed attacker in 2020.

3.5.1 The VoicePrivacy challenge requirements

The challenge aims to advance progress in the development of anonymization and pseudonymization² solutions that suppress personally identifiable information contained within recordings of speech while preserving linguistic content, paralinguistic attributes, subjective voice distinctiveness (make each individual’s voice easily recognizable and distinguishable from others by human listeners), intelligibility and naturalness properties.

The anonymization goals of the VPC are the following:

- (a) output a speech waveform
- (b) conceal the speaker’s identity against the thread model presented in Section 3.3

². Pseudonymization is the process of replacing PII with a pseudo or an alias, which can be used to link back to the original information if the original/pseudo mapping is available. This differs from anonymization, which aims to completely remove all PII.

- (c) leave the linguistic content and paralinguistic attributes unchanged
- (d) ensure that all trial utterances from a given speaker are uttered by the same pseudo-speaker, while trial utterances from different speakers are uttered by different pseudo-speakers

The requirement (d) is motivated by multi-party conversation application cases, where the anonymized voices of all speakers must be distinguishable from each other and should not change over time. The terminology used to refer to this type of anonymization is *speaker-level* target selection. This is an alternative approach to *utterance-level* target selection where different utterances of the same source speaker are anonymized using different parameters of the anonymization system, so that they may sound as if they were spoken by different speakers. Requirement (d) is assessed subjectively for voice distinctiveness and objectively for pseudonymization.

3.5.1.1 Data

The training, development and evaluation datasets of the VPC consist of several publicly available corpora:

- *LibriSpeech* (Panayotov et al. 2015) is a corpus of read English speech derived from audiobooks and designed for ASR research. It contains approximately 1000 hours of speech sampled at 16 kHz.
- *LibriTTS* (Zen et al. 2015) is a corpus of English speech derived from LibriSpeech and designed for research in text-to-speech (TTS). It contains approximately 585 hours of read English speech sampled at 24 kHz.
- *VCTK* (Veaux et al. 2017) is a corpus of read speech collected from 109 native speakers of English with various accents. It was originally aimed for research in TTS and contains approximately 44 hours of speech sampled at 48 kHz.
- *VoxCeleb-1,2* (Nagrani et al. 2017; Chung et al. 2018) is an audiovisual corpus extracted from videos uploaded to YouTube and designed for speaker verification research. It contains approximately 2770 hours of speech sampled at 16 kHz collected from 7363 speakers, covering a wide range of accents and languages.

Below, we provide a detailed description of the datasets used for training the anonymization system, training the evaluation models (privacy/utility), and testing datasets. Table 3.1 presents statistics about the number of speakers and utterances per dataset, the table also indicates the evaluation usage and potential training usage for each of them.

Training data The data allowed to train an anonymization system consists of *VoxCeleb-1,2*, *LibriSpeech train-clean-100*, *LibriSpeech train-other-500*, *LibriTTS train-clean-100* and *LibriTTS train-other-500*. The *LibriTTS train-other-500* dataset can be used to select the target speaker

to anonymize each utterance. The motivation for using this dataset is to have a wide variety of target speakers, which, at the time, was seen as an asset in order to generate many kinds of voices for the voice distinctiveness requirement. In this thesis, we will challenge whether or not having the voice distinctiveness requirement (as well as the use of the *LibriTTS train-other-500* dataset) is beneficial for speech privacy.

Testing data The evaluation dataset consists of *LibriSpeech test-clean* and *VCTK test*. For each of those two datasets, the challenge organizers divided them into *vulnerable* and *compromised* speech datasets.

Attacker data To match a realistic scenario, the attacker has access to another dataset, here *LibriSpeech train-clean-360*, which he uses to train the ASV linkability attack model. Depending on the attacker’s capability, he might train the ASV (privacy) evaluation model with anonymized data to reduce the domain mismatch as seen in Section 3.3.1. If anonymized data is used, the VPC request that the *LibriSpeech train-clean-360* is anonymized at the *utterance-level* rather than at the *speaker-level*, because it has been observed that training the ASV model with *LibriSpeech train-clean-360* anonymized at *speaker-level* leads to overfitting.

Application data In the VPC this data is the same as the attacker data, it is used to train the ASR model that assesses the preservation of the linguistic content (utility). In the 2022 challenge, the ASR model is trained with anonymized data to maximize the utility as demonstrated in (Tomashenko et al. 2022b).

Table 3.1 – Statistics of the datasets.

| | Dataset | Usage | # Speakers | # Utterances | Avg duration |
|--------------|------------------------------------|-----------------------------------|------------|--------------|--------------|
| Train | <i>LibriSpeech train-clean-100</i> | Linguistic extractor | 251 | 28 539 | 12.70 sec. |
| | <i>LibriSpeech train-other-500</i> | Linguistic extractor | 1 166 | 148 688 | 12.03 sec. |
| | <i>LibriTTS train-clean-100</i> | Speech synthesizer | 247 | 33 236 | 5.82 sec. |
| | <i>LibriTTS train-other-500</i> | Pool of x-vectors | 1 160 | 205 044 | 5.45 sec. |
| | <i>VoxCeleb1,2</i> | Speaker extractor | 7 363 | 1 281 762 | 8.10 sec. |
| Test | <i>LibriSpeech test-clean</i> | Compromised speech | 29 | 438 | 6.18 sec. |
| | | Vulnerable speech | 40 | 1 496 | 8.67 sec. |
| | <i>VCTK test</i> | Compromised speech | 30 | 600 | 3.05 sec. |
| | | Vulnerable speech | 30 | 11 448 | 3.22 sec. |
| Eval | <i>LibriSpeech train-clean-360</i> | Train privacy/utility eval models | 921 | 104 014 | 12.58 sec. |

In this thesis, we only focus our work on the core objective of anonymization, so, we do not present any paralinguistic, or subjective analysis in contrast to the VPC evaluations. For similar reasons, we believe requirement (d) about *speaker-level* target selection falls out of

the scope of the core objective of speaker anonymization evaluation. We believe the way it is objectively evaluated in the VPC (described in (Noé et al. 2020)) misses the opportunity to use watermarking/steganography methods (Cheng et al. 2001; Faundez-Zanuy 2009; Djebbar et al. 2012; Nematollahi et al. 2013; Nematollahi et al. 2017) that could embed a pseudo identifier into anonymized speech to comply with pseudonymization. As for the subjective voice distinctiveness evaluation, it is missing real-world application cases defining the size of the multi-party conversation. We believe having “all” voices distinguishable from each other is impractical knowing it can become difficult for a person to keep track of who is speaking and what was being said by whom, especially when hearing many unfamiliar voices.

3.5.2 Privacy model and metrics

Like the VPC, the primary objective privacy metric of this thesis is obtained through the ASV linkability attack between vulnerable and compromised speech for both *LibriSpeech test-clean* and *VCTK test* datasets. The model used for all experiments in this thesis (involving target VC x-vectors), is based on the Kaldi x-vector model with five TDNN layers as presented in Section 2.3.1. The metrics presented in Section 2.3.3 and used to assess the strength of linkability are the EER, and the linkability $D_{\leftrightarrow}^{\text{sys}}$ metrics. For comparison without anonymization (clear speech), a linkability test is performed without anonymizing the vulnerable speech, (i.e., scenario I (baseline) in Figure 3.4). After anonymization, the higher the EER, or the lower the $D_{\leftrightarrow}^{\text{sys}}$, the greater the privacy.

3.5.3 Utility model and metric

The capability of the anonymization method to maintain the linguistic content is objectively evaluated using an ASR model based on Kaldi. This evaluation model is used for all experiments in this thesis and is a hybrid DNN-HMM triphone acoustic model with a 17 TDNN-F layers architecture presented in Section 2.2.1.3 and a large trigram language model presented in Section 2.2.2. The ASR decodes the word sequence for both *LibriSpeech test-clean* and *VCTK test* anonymized speech datasets. The WER is then calculated. For comparison, an ASR decoding is also performed on clear (non-anonymized) speech. The lower the WER, the greater the utility.

As improvement in privacy performance is usually correlated with utility degradation, good speaker anonymization must achieve a suitable privacy/utility trade-off (Li et al. 2009; Wu et al. 2021). In the next section we present some methods to anonymize speech data, and when relevant, present their privacy/utility trade-off.

3.6 Current methods for speaker anonymization

This section presents several methods to perform speaker anonymization. They can be classified into three main categories: signal processing, voice conversion, and adversarial ASV attack anonymization.

3.6.1 Signal processing anonymization

Signal-processing speaker anonymization aims to shift the perceived original speaker by manipulating speech features such as pitch, rhythm, tempo, pause, etc. In contrast to other methods, this anonymization method does not require any training as only simple signal processing techniques are used.

In (Qian et al. 2017), the authors present VoiceMask, a VTLN-based frequency warping technique to transform the spectral envelope before resynthesizing a waveform. Evaluated with the informed attacker, the relative utility degradation is by 10%, while the privacy does not change, (Srivastava et al. 2020b) argues this result comes from the parameters of the transformation always wrapping the spectra at each frame of the utterance in a single direction.

In (Vaidya et al. 2019), the authors present an “audio sanitizer” that modifies the pitch, tempo and pause features, then, they add white noise after resynthesizing the waveform from MFCCs. In their paper, the utility decreased by 27% (relative), while the privacy gain is not measured with a strong enough attacker to draw any conclusion.

In (Patino et al. 2021), the authors present a speaker anonymization method based on McAdams transformation (McAdams 1984). The McAdams transformation relies on timbre modification using the McAdams coefficient which expands/contracts the frequency of each harmonic. Under semi-informed evaluation, the utility decreased by 8.5% (relative), while the privacy increased by 100% (from 4% to 8% for EER).

In (Kai et al. 2021), the authors propose a cascade of previously mentioned techniques. Multiple combinations are explored, the most notables are a combination of Resampling and Modulation Spectrum smoothing, and a combination of Resampling, Modulation Spectrum Smoothing, McAdams Transformation clipping, and Chorus. The evaluation shows a utility degradation of 36% and 600% for each combination respectively, while privacy only slightly increases (from 4% to 7 and 10% of EER) for both methods when using a semi-informed attacker.

In (Mawalim et al. 2022), the authors proposed a phase vocoder-based time-scale modification anonymization that compresses or stretches audio signals. Very interestingly, they perform multiple evaluations using Ignorant, Lazy-informed and semi-informed attackers. With the Ignorant and Lazy informed attackers, the privacy metric reaches around 44% of EER, while with the semi-informed attacker, the EER obtained is 16%. The utility remains very close to clear speech (10% of degradation). Note that these results are the ones highlighting the biggest overestimation of privacy protection when using under-capable attackers.

3.6.2 Voice conversion anonymization

As presented in Section 2.4, Voice Conversion (VC) is the process of adapting the characteristics of a source speaker’s speech to match those of a target speaker without changing the linguistic content. In contrast to signal-processing speaker techniques, VC needs a linguistic representation and paralinguistic information to work. Section 2.4.1 explains the two main approaches to obtaining the linguistic representation. The supervised approaches typically use specialized models trained with text supervision like an ASR acoustic model and the self-supervised approaches typically have an implicit linguistic representation like auto-encoders for example.

3.6.2.1 Self-supervised linguistic representation

Starting with the latest self-supervised linguistic representation extraction VC approach, works of (Yoo et al. 2020) have presented a CycleVAE-GAN which uses Variational Autoencoders to extract linguistic representation and a one-hot vector for the target speaker. Evaluation with a lazy-informed attacker shows almost perfect privacy under the experiment where the one-hot represented a single speaker which is the farthest from the source in cosine similarity distance. However, the nature of this type of attacker may not accurately reflect the true privacy performance of the system. The relative utility degradation recorded is 22%.

Similarly, the authors of (Aloufi et al. 2020) propose a VAE to extract a representation accordingly to the user’s preference about the characteristics to remove. However, their evaluation lacks a linkability attack. Their evaluation instead focuses on preserving the utility of the linguistic content and emotional state.

Following on a similar VC approach, (Prajapati et al. 2022) proposed a Cycle-GAN and speed perturbation pipeline, where the target speaker is the opposite gender of the source. Results under the semi-informed attacker show a privacy improvement of 360% from (4% of EER to 17.5% of EER) while affecting the utility by less than 10%.

3.6.2.2 Supervised linguistic representation

Using a supervised linguistic representation extractor based on ASR acoustic model for VC, (Fang et al. 2019) proposed the first x-vector-based speaker anonymization system used as a baseline for many following works including the VoicePrivacy Challenge. This baseline and components of this baseline are used in our experiments and proposed systems.

In their model, speaker identity and linguistic content are first extracted from an input speech utterance. Assuming that those features are disentangled, an anonymized speech waveform is generated by altering only the features that encode the speaker’s identity. The anonymization system depicted in Figure 3.7 can be decomposed into three groups of modules. Modules from group A extract different features from the source signal: the fundamental frequency, the linguistic

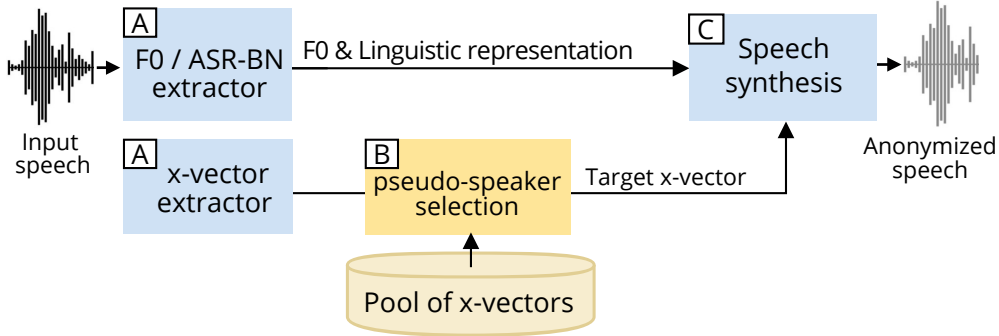


Figure 3.7 – X-vector-based speaker anonymization.

representation and the speaker’s x-vector. Module B derives a new target identity using a target selection strategy. The target selection strategy is the following: the x-vector from each source input speaker is compared to a pool of external x-vectors in order to select the 200 furthest vectors; 100 of them are randomly selected and averaged to create a target x-vector identity. Finally, module C synthesizes a speech waveform from the target x-vector together with the original linguistic representation and F_0 . Speaker anonymization is achieved by selecting a private target x-vector. The fundamental frequency is extracted using the YAAPT algorithm (Kasi et al. 2002). The linguistic representation is obtained using a 17 TDNN-F layers Kaldi triphone ASR-BN extractor (dimension 256 extracted from the final factorized hidden layer) and trained with *LibriSpeech train-clean-100* and *LibriSpeech train-other-500*. The x-vectors are also extracted using Kaldi, which is a five-layers TDNN trained on *VoxCeleb1,2*. The speech synthesis model is HiFi-GAN (see Section 2.4.3.1). Table 3.1 describes the datasets used to train each model and the pool of x-vectors. Evaluated with the informed attacker, the privacy increases by 175% (from 4% to 11% of EER), while the utility is not significantly decreased. In the following sections, we describe some extensions of this system, however, they are not included in our baseline.

X-vector modification For this system, (Srivastava et al. 2020a) experienced many design target selection strategies to generate the target x-vector based on the pool of x-vectors. They observed that randomly picking half of the x-vectors from a random dense cluster and averaging them created the best privacy/utility trade-off under the lazy-informed attacker. Other studies explored this topic of target vector generation for speaker anonymization: (Yoo et al. 2020) randomly modifies each component of a one-hot speaker embedding to generate target vectors, (Turner et al. 2020) uses a GMM coupled with PCA to generate target x-vectors mirroring the distribution of those found in the real world (increasing naturalness), (Mawalim et al. 2020) uses singular value decomposition and statistical regression to generate target x-vectors, (Espinoza-Cuadros et al. 2020) uses adversarial learning to enhance the disentanglement of the x-vectors

from gender and accent, (Meyer et al. 2022b) uses Wasserstein GAN to generate target x-vectors with the same naturalness property as Turner.

ASR-BN modification Research in ASR technology by (Adi et al. 2019; Srivastava et al. 2019) showed that the speaker/linguistic disentanglement property of the ASR-BN representation is limited. This directly impacts the performance of anonymization systems that rely on the ASR-BN representation. Compared to other features (namely the F_0 and target x-vector), the ASR-BN representation is the feature with the highest dimension and is sequential, as such, not disentangled ASR-BN can substantially restrict performance in terms of speaker concealment. In the following, we present valuable research that aims to improve the speaker/linguistic content disentanglement (or privacy) of ASR-BN representation.

Dimensionality Reduction One approach to enhance the privacy of bottleneck representation discussed in (Osia et al. 2017; Ryffel et al. 2019), is to reduce their dimensionality. By reducing bottleneck dimensionality, the encoding capacity is reduced, as such, only the most important information should be kept. Training a network with a bottleneck of very low dimensionality can be challenging, so using Principle Component Analysis (PCA) or Singular Value Decomposition (SVD) methods after training is usually easier. PCA or SVD methods attempt to preserve the primary structure of bottlenecks and remove as many unnecessary components as possible. When the bottleneck is extracted with a Factored Time-Delay Neural Network (TDNN-F) model at the factorized layer, which is the case for the model presented in Figure 3.7, the bottleneck has already been reduced with a technique similar to an SVD as presented in Section 1.4.3.

Adversarial network Another approach to enhance the representation of ASR-BN is through the use of an adversarial network (see Section 1.4.5). Using the ASR-BN representation, the main network is trained for the main task (e.g., triphone classification) while an additional adversarial model is trained alongside the same ASR-BN representation to identify the speaker. To improve the disentanglement of the ASR-BN using this training scheme, the negative gradient of the adversarial model is used to encourage the ASR-BN to encode less speaker information. Work done by (Srivastava et al. 2019), found at the feature level (no anonymized speech produced) that the speaker classification accuracy was reduced, but did translate to reduced linkability. The hypothesis regarding this disparity may be due to the use of an atypical adversarial speaker verification model architecture, as well as the formulation of the adversarial loss. Further research is needed to assess the benefits and limitations of this approach to speaker information removal in an ASR model using adversarial learning. Work of (Ericsson et al. 2020), has generated speech hiding the gender information using adversarial learning while maintaining a high quality of the anonymized audio.

Noise perturbation Motivated by the field of Differential Privacy (Dwork 2006) and the *Laplace mechanism* (Dwork et al. 2014), noise can be used to perturb the ASR-BN representation such that the representation maintain a satisfactory privacy/utility trade-off. Work of (Shamsabadi et al. 2022) has modified the F_0 and ASR-BN representation of the x-vector-based speaker anonymization system presented above with *Laplace* noise. Directly applying noise to the ASR-BN representation (or raw speech) destroys linguistic and prosodic information. Hence, the authors have adapted their model to work with noise, by training them to retain relevant information while adding the required level of noise to get the Differential Privacy guarantees. For their model which adds the highest amount of noise, under semi-informed evaluation, the utility decreased by 48%, while the privacy increased by 650% (from 4% to 30% for EER).

In (Tran et al. 2022), the authors present a method that adds noise to specific regions of the ASR-BN representation (from Wav2Vec-2.0) by using a transformer-based privacy-risk saliency estimator to estimate the regions. Their evaluation is performed at the feature level (no anonymized speech produced), with an ignorant attacker which does not reflect the true privacy performance. For the best noise parameters, the utility degradation recorded is 58% with a privacy improvement of 366%.

Discrete token vs PPG vs ASR-BN Up until now, the considered supervised linguistic representation has been ASR-BN. However, as presented in Section 2.4.1, two other supervised representations, namely Phonetic Posteriorgrams and discrete token, can also be used for voice conversion. In the original paper of the x-vector-based speaker anonymization system (Fang et al. 2019), the authors experimented with extracting both ASR-BN and PPG representations. Accordingly, to the voice conversion literature, the use of PPG results in better speaker transformation, so better privacy results, however, this comes at the cost of more mispronunciations errors. In contrast, ASR-BN preserves more of the speaker’s original pronunciation resulting in significantly better utility, at the cost of lower privacy due to the encoding of more speaker attributes. The ASR-BN version of this system was considered to have a better privacy/utility trade-off and was used in the VoicePrivacy 2020 challenge. This system has received an update in 2022 for the VoicePrivacy 2022 challenge using a HiFi-GAN synthesizer (directly taking linguistic and speaker representations rather than Mel spectrograms) (Tomashenko et al. 2022a) instead of a two-stage spectrogram+vocoder pipeline (Wang et al. 2020). Both versions have been broadly used as baselines for many other works.

The Preech pipeline (Ahmed et al. 2020), performs many steps to ensure anonymization not only at the prosody and acoustic level but also at the semantics, and syntax level. Among segmentation and shuffling, sensitive word scrubbing (requiring ASR), and dummy segment injection (requiring TTS to generate speech from dummy words), they convert voices (dummy and real) to a single target voice using PPG representations. The evaluation provided goes into deep detail about textual privacy, however, proper evaluation of the anonymized speech

unlinkability is missing. There is no guarantee that an adversary could not be able to distinguish dummy segments from real ones (cf the research in (Zhang et al. 2022)) and perform a linkability attack on the real segment.

In (Turner et al. 2022) the authors present AltVoice which uses discrete tokens as linguistic representation. They manage to anonymize speech to the fullest extent possible (fully unlinkable), by producing just text as their linguistic representation, leaving no room for any incidental transfer of the source identity. They are properly outlining the limit of their system, with this solution the privacy/utility trade-off archived is unfavorable for utility. The cascade ASR+TTS leads to the anonymized speech having high WER, limited perceived audio quality, and no intonation and emotion preservation.

Similarly, in (Meyer et al. 2022a) the authors have presented a cascade ASR+TTS using phonemes as supervised linguistic representation. They achieve perfect unlinkability but degrade the utility by 82%.

3.6.3 Adversarial ASV attack anonymization

In contrast to traditional signal processing and voice conversion techniques presented above, very recent kinds of anonymization have been proposed. They rely on adversarial attacks and are inspired by the use of adversarial examples to put into failure the deep neural network ASV models while remaining imperceptible to the human ear.

The authors of (Deng et al. 2022) introduced V-Cloak which modulates and adjusts the bottleneck features of a Wave-U-Net audio source separation model at each frequency level according to the source speaker characteristics and requested constraints on the anonymization perturbations. Their model not only preserves intelligibility, but also naturalness, intonation, and more. They show excellent privacy scores, with EER above 40% on most experiments, while not having the utility significantly decreased. Additionally, they evaluated subjective speaker verification with a limited number of test speakers and asses that the speaker timber is persevered in comparison to McAdams, VoiceMask, and x-vector-based speaker anonymization. The authors also consider three kinds of attacks against their system, ignorant, gray-box with a denoising technique to invert the anonymization, and gray-box by adapting the compromised speech. However, informed (white-box) evaluation where the ASV attack model is trained to accommodate specific modulation is missing.

The authors of (O’Reilly et al. 2022) presented VoiceBlock, which applies a time-varying finite impulse response filter to the speech signal, allowing inconspicuous perturbations. The way they obtain the filter is similar to audio-to-audio tasks such as denoising and speech enhancement. Objective and subjective evaluation of the quality of the signal is very high as well as intelligibility, however, under lazy-informed attack, VoiceBlock “de-identification”³ performance

3. de-identification is the process of removing identifying information, same as speaker anonymization.

suffers significantly”. Similarly to V-Cloak, VoiceBlock does not conceal speaker identity from human listeners.

3.7 Conclusion

In this chapter, we presented numerous techniques to protect the speech of a speaker against linkability attacks. In addition to the technique, we also glanced at the attackers’ capabilities the authors used to evaluate the privacy of their model. The main observation regarding this aspect is that many kinds of attackers are used, all ranging from black-box to white-box, this makes decisive comparison difficult with the scores presented. Besides, electing a unique superior anonymization technique is almost impossible as for each of the ones presented, a hidden privacy/utility trade-off is present, not recorded by the metric. For instance, the cascade of ASR+TTS methods are the ones offering the best trade-off in today’s measurement method, however, we can question if their utility performance generalizes to noisy environments, reverberated speech, strong accents, the latter one being a real issue as strong racial disparities in ASR exist (Koenecke et al. 2020). Additionally, when the application case is to collect large speech corpora that are representative of the various usage conditions to improve a downstream ASR system for all users, the use of discrete tokens for anonymization makes it impossible to annotate downstream ASR error, as the speech will have mispronunciations from the anonymization procedure. Adversarial attacks on ASV models are methods for intentionally altering input data in a way that causes the model to make incorrect predictions. This can be used to reduce linkability attacks, however, as it does not remove PII (human listeners can still link speakers) we do not classify them as anonymization solutions. Finally, depending on the training data available one may be forced to use signal processing methods, or voice conversion with self-supervised linguistic representation if no label is available. In summary, to select the best anonymization technique one has to first identify his application case, and given it, make an informed decision.

For this thesis, we use the evaluation framework of the VoicePrivacy challenge, hence we chose as our baseline system the same as the VoicePrivacy, e.g., the x-vector-based speaker anonymization system. The x-vector-based speaker anonymization technique requires text supervision during training, which is more restrictive than self-supervised approaches but, as discussed in Section 2.4.1, might be an easier framework to separate and modify speaker and linguistic information, providing better anonymization. Additionally, we believe its privacy/utility trade-off, where the utility is very well preserved for objective and subjective linguistic recognition, is a good starting point to enhance the privacy of this system. Overall, we also think the x-vector-based system has the most possible application cases as it uses the less restricted ASR-BN representation and a defined F_0 for intonation.

PART II

Contributions

THE ROLE OF THE TARGET SPEAKER

Contents

| | | |
|------------|---|-----------|
| 4.1 | Introduction | 69 |
| 4.2 | Impact of the target selection algorithm in privacy evaluation | 70 |
| 4.2.1 | Experimental setup | 70 |
| 4.2.2 | Clear speech linkability | 71 |
| 4.2.3 | “Farther 200 random 100” target selection strategy | 71 |
| 4.2.4 | “Dense” target selection strategy | 74 |
| 4.2.5 | “Random speaker” target selection strategy | 76 |
| 4.2.6 | “Random vector” target selection strategy | 77 |
| 4.2.7 | “Constant speaker” target selection strategy | 77 |
| 4.2.8 | <i>Speaker-level</i> target selection | 78 |
| 4.2.9 | Discussion | 79 |
| 4.3 | A quest for the golden target speaker | 82 |
| 4.3.1 | Experimental setup | 82 |
| 4.3.2 | Global privacy results | 83 |
| 4.3.3 | Detailed privacy results | 83 |
| 4.3.4 | Utility results | 85 |
| 4.3.5 | Discussion | 87 |
| 4.4 | Conclusion | 88 |

4.1 Introduction

In this chapter, we are studying the impact of the target speaker in the x-vector-based speaker anonymization system (Fang et al. 2019) presented in Chapter 3. This system works by swapping the source speaker identity (here extracted as the x-vector) to a target identity (selected given a specific strategy) to later generate anonymized speech. As such, the conclusions made in this chapter are potentially applicable to any voice-conversion-based anonymization system which relies on replacing the source speaker identity with a target speaker (see Section 3.6.2). The goal of this chapter is to analyze how this speaker modification influences the performance of the anonymization system in both privacy and utility metrics. In the first section, we will study

the role of the target selection algorithm used to select the target speaker for voice conversion. In particular, we study how different target selection algorithm affects the privacy evaluation. In the second section, we will study how the selected target speaker influences the privacy and utility of each source speaker to anonymize.

4.2 Impact of the target selection algorithm in privacy evaluation

In speaker anonymization, the privacy of a user is evaluated using linkability attacks. Before diving into the target selection algorithm used for anonymization, it is important to understand the relation between the privacy evaluation procedure and the test datasets. Linkability attacks involve determining the extent to which an individual’s speech can be linked to their identity. This evaluation relies on ASV where the test dataset (compromised and vulnerable speech) strongly affects the privacy measurement. For a speech signal to be linked to an identity, the ASV must provide high similarity scores for the considered trial and lower scores for all other impostor trials.

In this section, we aim to study target selection algorithms that were designed to increase privacy/unlinkability, some of them have been presented in Section 3.6.2.2. We propose to analyze target selection algorithms to query how they influence the creation of vulnerable and compromised anonymized datasets specifically for linkability attack evaluation (see Section 3.3). The goal is to better understand the impact of the speaker selection step necessary for VC-based anonymization and its influence against linkability evaluation.

4.2.1 Experimental setup

This experiment mainly follows the VoicePrivacy challenge requirements and uses the x-vector-based anonymization system (Section 3.6.2.2) trained with the dataset of the VPC (Table 3.1). Privacy is evaluated using the gray-box semi-informed attacker, trained on the anonymized version of *LibriSpeech train-clean-360* to reduce as much as possible domain mismatch with the testing data. In contrast to the VPC ASV, where the x-vectors of an enrollment speaker are averaged over his/her utterances before PLDA scoring, we obtain PLDA scores for each possible pair of one vulnerable utterance and one compromised utterance. Additionally, in contrast to the VPC anonymization, anonymization is performed at the *utterance-level* rather than at the *speaker-level*. The difference between the two will be studied in this section. The research will concentrate exclusively on the *LibriSpeech test-clean* dataset and will feature only male subjects. The choice of using only one gender is justified by the fact that some target selection algorithms are gender dependent. It is important to note that the key findings will be relevant to any dataset and any gender, selecting male was arbitrary.

4.2.2 Clear speech linkability

In this section, we present how linkable the clear (non-anonymized) speech is using the baseline linkability evaluation scenario presented in Section 3.3.1, where the vulnerable and compromised speech datasets are clear and evaluated with an ASV system also trained on clear speech. Figure 4.1 displays the PLDA score distributions for genuine and impostor trials, for each vulnerable speaker of our test dataset (right part of the figure), and all speakers together (left part of the figure). The EER is computed over all speakers and the dotted red line represents the corresponding PLDA EER threshold. From the distinct distributions shown and the low 3.1% EER score, we can conclude that clear speech is highly representative of the speaker identity as the linkability attack allow to highly separate genuine and impostor trials. We observe that speaker ID 260 seems to be more difficult than the others to recognize whereas speaker ID 8224 had a distinct voice compared to the others.

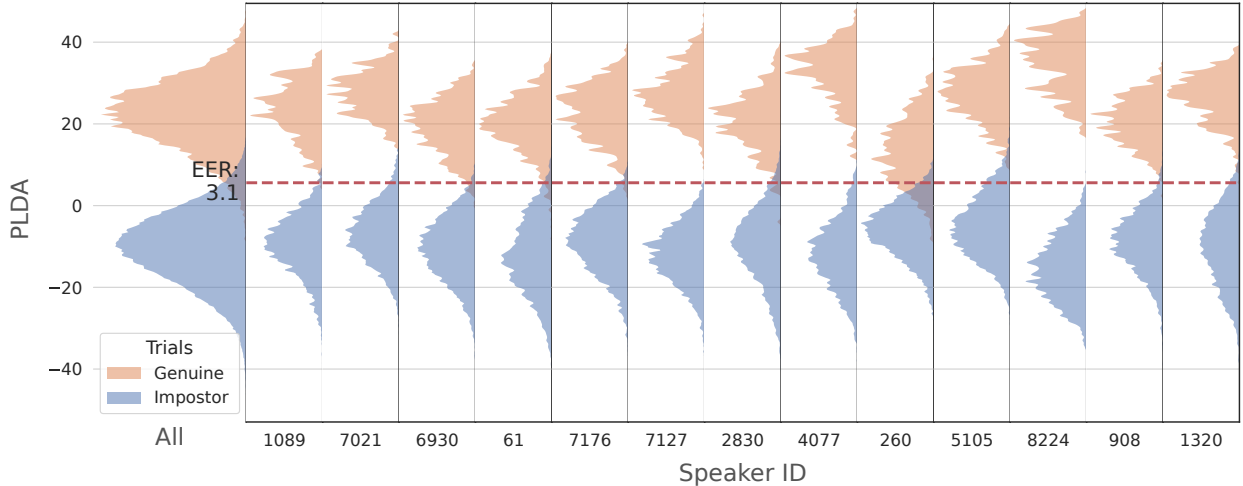


Figure 4.1 – Clear speech linkability.

4.2.3 “Farther 200 random 100” target selection strategy

In this section, we study the target selection strategy introduced by the authors of the x-vector-based system (Fang et al. 2019) and used as the selection strategy for the VPC 2020 and 2022 baseline systems. We will describe how this target selection strategy works, and how it is conceptually and experimentally flawed. The following steps are followed to generate the target x-vector: 1) the process begins by extracting an x-vector from a clear signal; 2) the clear x-vector is then compared to each of the speakers in a pool of external x-vectors with a PLDA scoring function; 3) the 200 x-vectors that are the least similar (farthest) to the clear x-vector are selected from the pool, and 4) 100 of these x-vectors are randomly chosen and averaged to create the anonymized target x-vector. As presented in Table 3.1 of the VoicePrivacy requirements section, the pool of external x-vectors is *LibriTTS train-other-500* and the x-vector extractor is trained on VoxCeleb1,2.

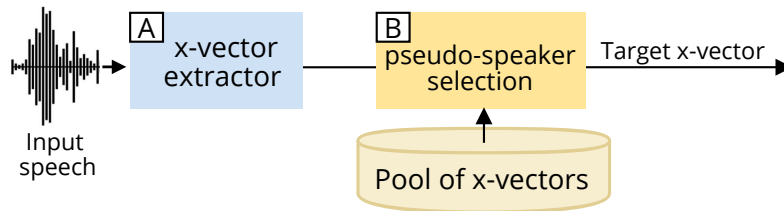


Figure 4.2 – X-vector anonymization defined in (Fang et al. 2019).

To evaluate how robust against linkability attacks this target selection algorithm is, we will start by analyzing how linkable the selected target x-vectors are by themselves, without using speech synthesis. ASV is performed with the target x-vector obtained from the selection algorithm, see Figure 4.2. Figure 4.3 shows the PLDA scores distributions for the selected target x-vectors for each speaker, the first thing that stood out is the low 4.8% EER value which indicates a strong link between the clear x-vector and anonymized target x-vector. If we were to suppose the anonymization system can completely remove the source speaker identity and replace it with one of the targets, effectively performing the complete PII removal needed for anonymization, this target selection strategy prevents a linkability attack as the anonymized speech will have the same EER as the target x-vector (4.8%). Hence, even the most perfect anonymization system fails as privacy is evaluated through linkability. We can consider the ASV linkability results (EER here) of target selection algorithms as the expected linkability results that anonymization systems should produce.

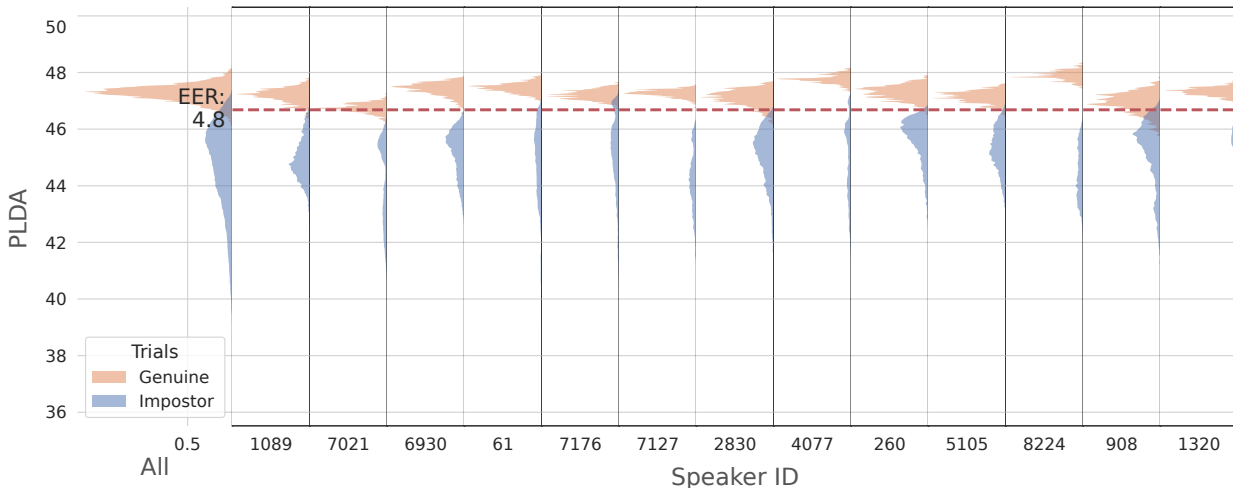


Figure 4.3 – PLDA genuine and impostor scores distributions for the “farther 200 random 100” target selection strategy. Scores are computed directly from target x-vectors, no speech synthesis is done. The PLDA model is trained on clear data (*VoxCeleb1,2*, see Table 3.1).

Limitations of this approach can be explained by the target x-vector selection process. Indeed, the attacker in the semi-informed scenario has access to the anonymization pipeline, hence a

similar x-vector will be extracted for the vulnerable and compromised speech of a speaker as the x-vector extractor (of the anonymization pipeline) is the same. Then given similar clear x-vectors from the vulnerable and compromised speech, the 200 farther x-vectors selected as candidates will also be similar or at least distributed in the same x-vector space as the pool of speakers (of the anonymization pipeline) is the same. In the fourth stage, differences between vulnerable speech anonymization and compromised speech anonymization occur, the randomly 100 x-vectors selected from the 200 candidates differ and are very unlikely to be the same. However, the last stage of this target selection strategy is to average the 100 x-vector together to generate the anonymized target x-vector. This average operation cancels the fourth stage, as it is very likely that the average of the 100 x-vector is very close to the average of the 200 candidate x-vectors, explaining the very high PLDA scores in Figure 4.3. This means that the anonymized vulnerable speech and anonymized compromised speech for a speaker will have similar target x-vectors, while also being different from other speakers due to the farther 200 selection. Still, the PLDA EER threshold score is very high at around 46, with most scores in the interval of 42 to 48, this indicates that all selected target x-vectors lay in a confined region space.

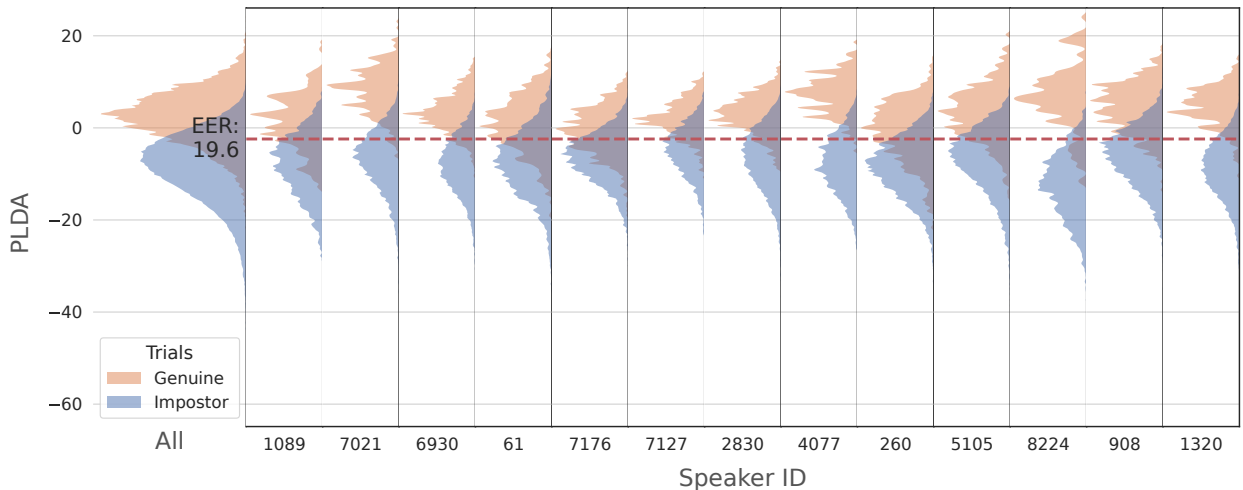


Figure 4.4 – PLDA genuine and impostor scores distributions of the semi-informed attacker ASV model on anonymized speech for the “farther 200 random 100” target selection strategy.

Figure 4.4 displays the attacker PLDA scores distribution for anonymized speech (anonymized with x-vector-based speaker anonymization system). The attacker ASV model is trained with anonymized speech as required by the semi-informed attacker definition. The x-vector-based anonymization system was not able to effectively replace the source speaker identities with those of the target, as indicated by the EER value of 19.6%, higher than the expected EER value of 4.8% of the target x-vector. Multiple factors can explain this disparity. The first one is the ineffectiveness of the synthesis system to perform speaker replacement, resulting in a speech signal having an imprecise speaker identity (maybe it has a mixture of clear and target

identities). The second one is related to the target x-vectors. The target x-vectors only occupy a confined region space indicated by the high PLDA scores in Figure 4.3, and the cause of this is the average operation. This means that the target x-vectors are all close to each other, still, they are preserving inter-speaker and intra-speaker separation for ASV. We hypothesize that for the synthesis system, the inter-speaker and intra-speaker x-vectors are so close to each other that after synthesis, the resulting speaker’s identities are considered similar. This aspect would lead to anonymized voices being less linkable.

Overall, linkability attack evaluations with the “farther 200 random 100” is confusing, we would like the PLDA scores distributions to be as overlapping as possible to comply with unlinkability. However, at the target x-vector level, there are no guarantees of unlinkability, and understanding the intended output of the synthesis system is intricate. Hence, are we evaluating how unlinkable anonymized speech is? Or are we evaluating how distinguishable converted speakers are?

We observe that speaker ID 260 seems to be better anonymized than the others whereas speaker ID 8224 is harder to anonymize, this observation is correlated with the one made in Section 4.2.2 on clear speech, where speaker ID 260 was hard to recognize and speaker ID 8224 had a distinct voice.

4.2.4 “Dense” target selection strategy

In this section, we evaluate the “dense” target selection strategy introduced in (Srivastava et al. 2020a) and slightly modified in (Shamsabadi et al. 2022), the first version is used here. To select the target x-vector for each utterance, the following steps are followed: 1) all x-vectors of the pool are grouped into clusters using the Affinity Propagation algorithm and PLDA, resulting in a total of 80 clusters; 2) one cluster is randomly chosen from the 10 largest clusters (the cluster which is closest to the source speaker is filtered out); 3) half of the members of the chosen cluster are randomly selected to introduce more randomness in the selection of the x-vector, and 4) the selected candidate x-vectors are averaged to obtain the target x-vector. This method is similar to the “farthest 200 random 100” strategy described above, but the x-vector chosen is less related to the input utterance so less linked to the original speaker. As far as our understanding, in the modified version, the cluster which is closest to the source speaker is not filtered out, making it completely independent of the source utterance.

Figure 4.5 shows the target x-vector PLDA distributions obtained for this selection algorithm. The first thing that stood out is the seemingly two Gaussians per distribution which we suppose is the result of the cluster selection in the algorithm. The EER value is 43.1% indicating that at the target x-vector level, better unlinkability is guaranteed. However, for specific speakers such as speaker ID 1089, we can observe that his target x-vectors genuine and impostor scores are not overlapping, which does not guarantee unlinkability for this speaker’s speech. This effect might be caused by the nearest cluster being filtered out.

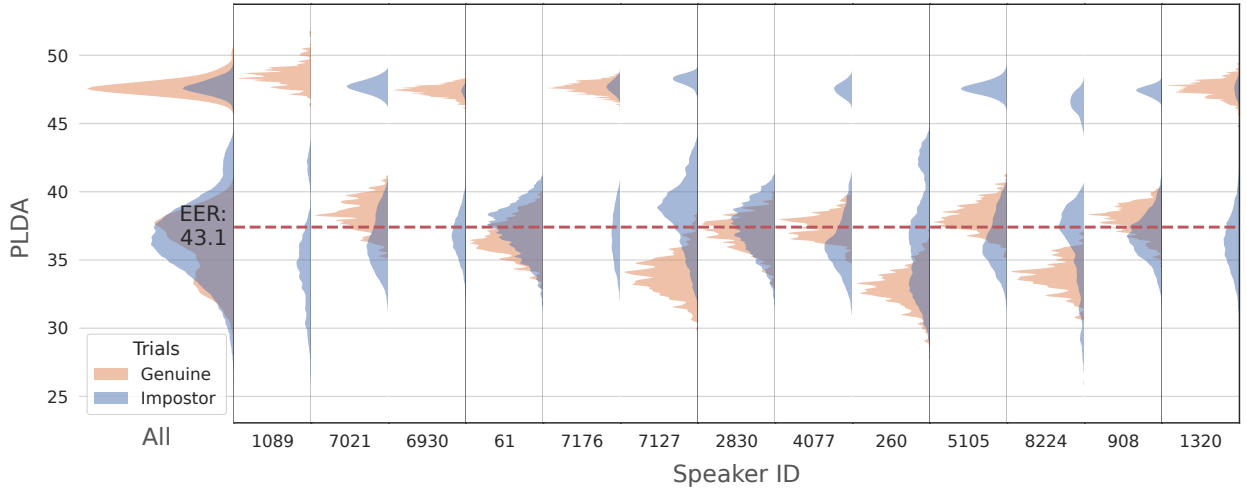


Figure 4.5 – PLDA genuine and impostor scores distributions for the “dense” target selection strategy. Scores are computed directly from target x-vectors, no speech synthesis is done. The PLDA model is trained on clear data (*VoxCeleb1,2*, see Table 3.1).

Figure 4.6 displays the attacker PLDA scores distributions for anonymized speech. The attacker ASV model is trained with anonymized speech data that was anonymized using the same “dense” target selection algorithm as required by the semi-informed attacker. Overall we can observe a high EER value of 39.2%, indicating better performance against linkability attacks.

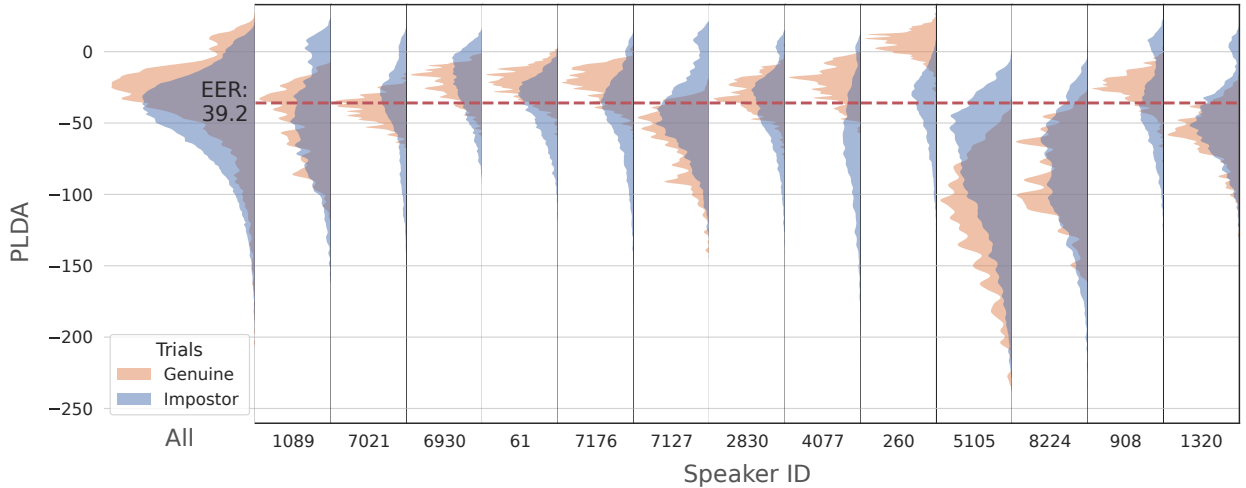


Figure 4.6 – PLDA genuine and impostor scores distributions of the semi-informed attacker ASV model on anonymized speech for the “dense” target selection strategy.

However, the PLDA EER threshold equals -35 , which might indicate that the PLDA produces very miscalibrated scores due to a potential attacker ASV model improperly trained. There is a possibility that this target selection algorithm generates data in a way that makes training the semi-informed attacker fail. To evaluate this hypothesis, we retrained the ASV semi-informed

attacker model with data anonymized with the “random speaker” (presented below) target selection algorithm and obtained an EER of 20.9%.

A similar conclusion has been drawn from (Shamsabadi et al. 2022; Tomashenko et al. 2022a) where training the attacker with *speaker-level* rather than *utterance-level* target selection yielded a training overfit of the ASV model. In this study, we have identified that the design of the target selection algorithm can also contribute to this issue.

4.2.5 “Random speaker” target selection strategy

In this target selection strategy, for each utterance, a random x-vector is sampled from the pool and used as the target identity (Srivastava et al. 2020b). Figure 4.7 shows the target x-vector PLDA distributions, as expected, they follow Gaussian distributions and are completely overlapping, leading to a target x-vector level EER value of 50%. This guarantee that if the anonymization system can completely remove the source identity and replace it with the target, linkability attacks will fail against properly trained ASV models.

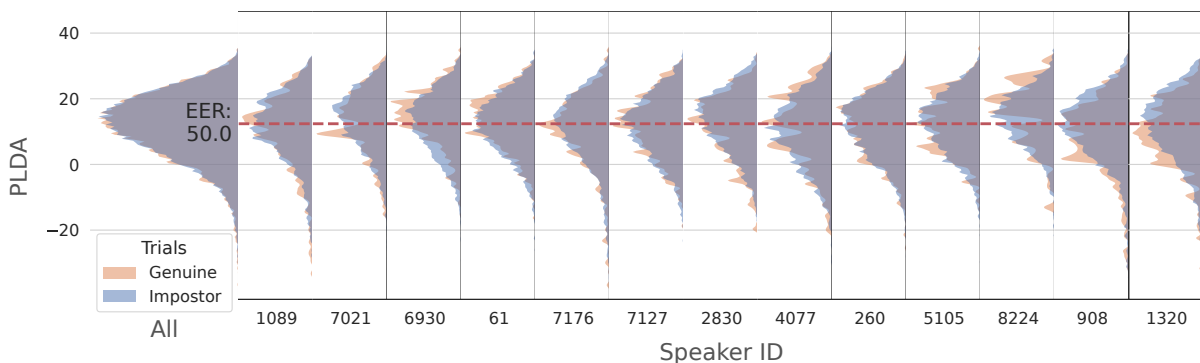


Figure 4.7 – PLDA genuine and impostor scores distributions for the “random speaker” target selection strategy. Scores are computed directly from target x-vectors, no speech synthesis is done. The PLDA model is trained on clear data (*VoxCeleb1,2*, see Table 3.1).

When anonymizing speech and training the ASV attacker model with this target selection algorithm, the EER value obtained is 21.6%. Figure 4.8 displays the corresponding attacker PLDA scores distributions, the PLDA EER threshold is close to 0 indicating the PLDA is correctly calibrated. This time, the EER disparity is primarily caused by the ineffectiveness of the synthesis systems to perform speaker replacement. Overall, in this study, we showed that this target selection strategy has a much better guarantee than the previous ones in terms of expected unlinkability, and, the creation of anonymized data for training the attacker. We observe the same conclusion about speaker ID 260 being easier to anonymize than speaker ID 8224. This disparity will be explored in Section 4.3.

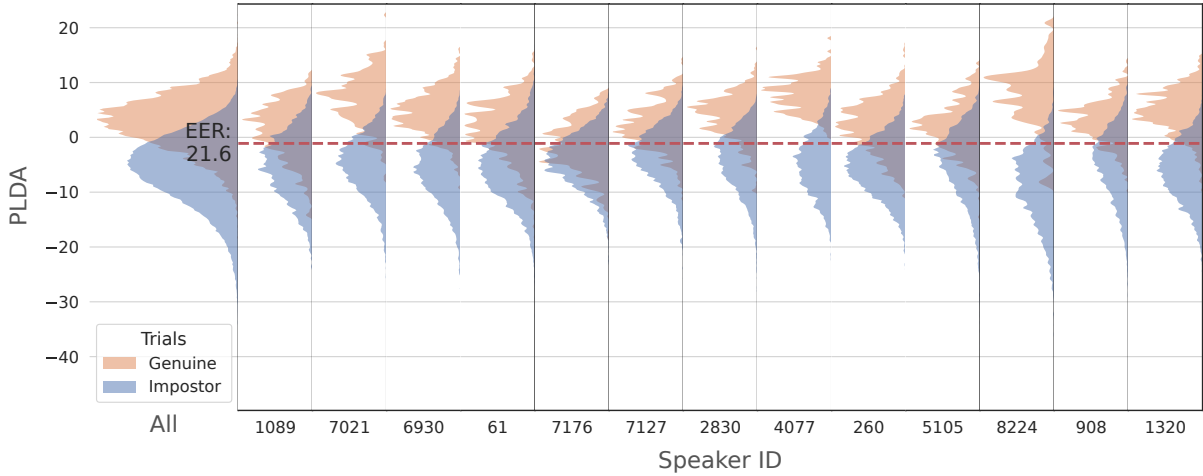


Figure 4.8 – PLDA genuine and impostor scores distributions of the semi-informed attacker ASV model on anonymized speech for the “random speaker” target selection strategy.

4.2.6 “Random vector” target selection strategy

This target selection strategy is similar to the previous one, but instead of sampling for each utterance to anonymize an x-vector from a pool, it samples it from a random Gaussian distribution. At the target x-vector level, the EER value is also 50%. And analogous to the previous strategy, the anonymized speech has an EER value of 23.3%, with a similar distribution to one of the “random speaker” in Figure 4.8. Again, this disparity can be explained by the ineffectiveness of the synthesis systems to perform speaker replacement. The slightly higher EER compared to the “random speaker” target strategy could be explained by a worse speech’s naturalness.

4.2.7 “Constant speaker” target selection strategy

The “constant speaker” target selection strategy might be the most simplistic one. A single target identity is used to anonymize every utterance of every speaker. At the target x-vector level, the EER value is 50% as everyone should sound like a single speaker. Hence, the target level genuine/impostor scores follow a single Dirac-delta distribution (see Figure 4.9). Comparable to the previous strategies, the anonymized speech has an EER value of 22.4%. The attacker PLDA scores distributions are similar to the one of the “random speaker” in Figure 4.8.

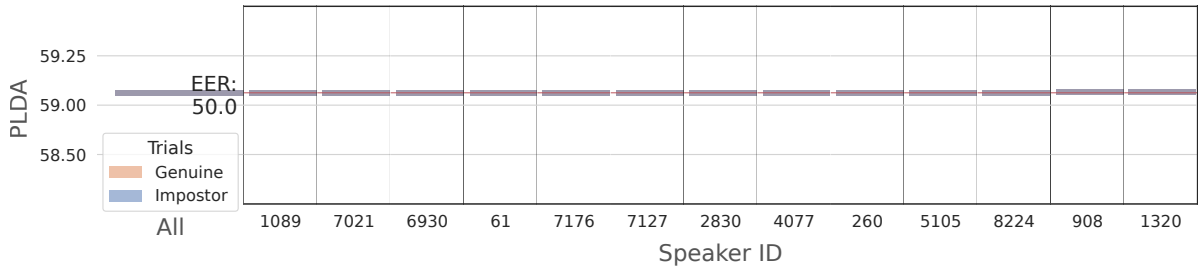


Figure 4.9 – PLDA genuine and impostor scores distributions for the “constant speaker” target selection strategy. Scores are computed directly from target x-vectors, no speech synthesis is done. The PLDA model is trained on clear data (*VoxCeleb1,2*, see Table 3.1).

4.2.8 *Speaker-level* target selection

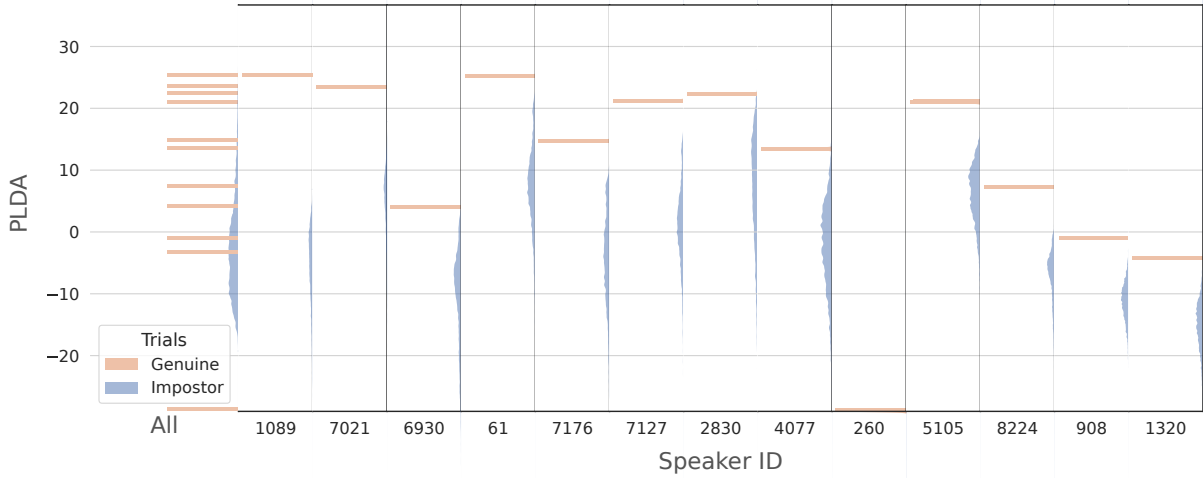


Figure 4.10 – PLDA genuine and impostor scores distributions for the “random speaker” x-vector target selection strategy using *speaker-level* anonymization. Scores are computed directly from target x-vectors, no speech synthesis is done.

The level of anonymization can either be *utterance-level* or *speaker-level* target selection. For the *utterance-level* anonymization, the target x-vector is generated for each utterance that needs to be anonymized. While for *speaker-level* target selection, the target x-vector is generated once for each speaker, and then applied to every utterance made by that speaker. In the above experiments, we only used *utterance-level* as it is the one supposed to have better unlinkability capability (Tomashenko et al. 2022a). However, *speaker-level* target selection is required in the VPC challenge as there is a requirement that the anonymized voices of all speakers must be distinguishable from each other and should not change over time. Something interesting in multi-party conversations application cases.

In this experiment, we study how *speaker-level* affects the creation of anonymized datasets used for privacy evaluation. We chose to use the “random speaker” target selection strategy, as the “Farther 200 random 100” target selection strategy has already been proven to not be a suited one, and the “constant speaker” cannot fulfill the speaker distinctiveness requirement. Figure 4.10 displays the target x-vector PLDA distributions. We can see that each utterance of a speaker has the same target x-vector as the genuine trials have all the same PLDA score indicated by the per speaker Dirac-delta distributions. We observe that the impostor trials are not overlapping the per-speaker Dirac-delta distributions indicating that a given speaker will not get confused by another, this follows the speaker distinctiveness requirement. However, there is an inherent issue, by definition, *speaker-level* target selections do not create datasets where speakers are unlinkable. And this is troublesome as privacy evaluation is based on ASV linkability evaluation. In Figure 4.10, the linkability metric equals 1.0 which corresponds to an EER of 50%. For reference, the same target selection strategy applied at the *utterance-level* has a linkability metric score of 0.0 (full overlap genuine/impostor distributions).

4.2.9 Discussion

Table 4.1 – Privacy results for the four main x-vector-based target selection strategies at the target x-vector level and anonymized speech x-vector level for the *LibriSpeech test-clean* male set. The confidence interval stays within ± 0.40 EER for all experiments¹. Evaluation models are retrained on anonymized speech when necessary, and the model architecture is described in 3.5.2

| Target selection algorithm | Target x-vector | | Anonymized speech x-vector | |
|---------------------------------|--|----------------|--|----------------|
| | Privacy $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | EER \uparrow | Privacy $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | EER \uparrow |
| Clear speech | | | 0.92 | 3.1 |
| VPC farther 200 random 100 | 0.87 | 4.8 | 0.54 | 19.5 |
| Dense, Dense ASV model | 0.19 | 43.1 | 0.14 | 39.2 |
| Dense, Random speaker ASV model | " | " | 0.50 | 20.9 |
| Random speaker | 0.00 | 50.0 | 0.49 | 21.6 |
| Random vector | 0.00 | 50.0 | 0.45 | 23.3 |
| Constant speaker | 0.00 | 50.0 | 0.47 | 22.4 |

Table 4.1 summarizes the previous EER results accompanied by their corresponding $D_{\leftrightarrow}^{\text{sys}}$ results. Out of the five target selections presented, only three of them have an EER at the target x-vector level of 50%, and one has a high target level EER value but did not enable proper training of the ASV attacker model. In our experiment, the best target selection algorithms were those that had overlapping impostor and genuine distributions.

The main takeaway of this section is that when building target selection strategies, the main goal should be to ensure unlinkability at the target x-vector level. To do so, the designer has to make sure the target x-vector of an utterance/speaker will be confused by other target x-vector of other utterances/speakers. By doing so, if the anonymization system is capable of completely removing the source identity and replacing it with the target one, the anonymized speech will have the same EER as the target level EER, which is a value of 50% in the best case.

In contrast, if the target x-vector has a link with the source speaker, then the ASV attacker model will evaluate the strength of the link between the target and anonymized speech, rather than evaluating the strength of anonymized speech unlinkability. For example, if the target level EER value is 0%, and the anonymized speech EER value is 0%, this means that we aim to achieve full linkability and that we achieve full linkability. In this case, what is evaluated is the successfulness of speech synthesis to generate distinct voices. Something that can be useful for assessing voice distinctiveness or the preservation of a pseudo-identifier in pseudonymization. However, those voices could very well leak the source speakers’ identities. That is why trying

1. The FEERCI toolkit (Haasnoot et al. 2018) is used to estimate the EER confidence interval with the bootstrapping method.

to produce unlinkable speech at all stages and trying to build the strongest attacker to defeat linkability, is the only way to properly record the strength of the anonymization. We conclude that *speaker-level* target selection can be used for voice distinctiveness evaluation. However, it should not be used for privacy linkability evaluation.

Similarly, in the case of the “farther 200 random 100” target selection strategy, we argue that what is evaluated is the strength of the voice distinctiveness rather than the strength of the anonymization. The EER of the anonymized speech of the “farther 200 random 100” strategic being close to the one obtained with the “constant speaker” algorithm in the *LibriSpeech test-clean* male set does not help to assess the difference. It seems that the strength of voice distinctiveness is equal to the strength of anonymization for this dataset. However, in the *VCTK test* dataset, moving from the “farther 200 random 100” target selection strategy to the “constant speaker” strategy shows an absolute increase by more than 10% of EER, highlighting the evaluation disparity, see privacy results Table 4.2. Another issue of the “Farther 200 random 100” target selection strategy affects fairness. As described in (Turner et al. 2020), male and female are not equally transformed when using the “farther 200 random 100” target selection strategy as a large EER gap exists. Using the “constant speaker” target selection strategy (with the same anonymization pipeline) fixes this issue.

In Table 4.2, the “constant speaker” algorithm has lower utility performance compared to the “farther 200 random 100” one. This is likely to be caused by the selected x-vector having component characteristics different from the ones of the training dataset used to train the synthesis model. The “farther 200 random 100” does not have this issue, as the average of multiple x-vector mitigates this effect. Later in this chapter, we propose and evaluate other solutions to select the constant target speaker.

Table 4.2 – Privacy and utility results disparity for the “farther 200 random 100” and “constant speaker” target selection strategies. Evaluation models are retrained on anonymized speech, and the model architectures are described in 3.5.2 and 3.5.3

| Dataset | <i>VCTK test</i> | | | |
|-------------------------------------|------------------|--|----------------|-----------------------------|
| | Method | Privacy $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | EER \uparrow | Utility WER \downarrow |
| Clear speech | | 0.93 | 2.7 | 12.8 |
| Anonymized “Constant speaker” | | 0.49 | 20.6 | 13.0 |
| Anonymized “Farther 200 random 100” | | 0.73 | 10.2 | 10.7 |

Finally, in this section, we confirmed that the anonymized data used to train the ASV linkability attack matters. More specifically, we showed that not all target selection algorithms generate training anonymized data in a way that makes the semi-informed ASV system learn a linkability attack, in our experiment, the first version of the “dense” algorithm was one of them.

All of those pitfalls are common in today’s speaker anonymization evaluation framework. The study of x-vector generation for anonymization draws many interests and is constantly evolving with new methods to generate target x-vectors, see Section 3.6.2.2. However, to help in making mechanistic conclusions about privacy performance, we recommend using the target selection strategies that have well-defined behavior such as “random speaker” or “constant speaker” at the *utterance-level*. We encourage target speaker strategy developers to make sure their strategy has the correct unlinkability criterion and ability to train ASV attacker. We have found that plotting the target-level genuine/impostor distributions greatly helps. The most important takeaway of this section is that the privacy evaluation of speaker anonymization requires assessing its unlinkability which is an indirect evaluation, a system could very well completely anonymize speech (that is remove PII) but still be linkable to a pseudo-identity.

In this thesis, we focused the rest of our work on the “constant speaker” target selection strategy, as it is the most simplistic one, and allows us to isolate this variable to study it. Interestingly enough, when using the “constant speaker” strategy, the attacker capability is upgraded from a gray-box semi-informed attacker to a white-box informed attacker, as the attacker has complete knowledge of the voice conversion parameters used to anonymize each utterance (see definition in Section 3.3). In practice, this section showed that a well-trained semi-informed attacker measures with the same capability the strength of anonymization as the informed attacker (when the target selection algorithm is appropriate). However, a point can be made about the fact that it is possible to poorly train the semi-informed attacker, whereas, in our experiment, the informed attacker is always well-trained.

4.3 A quest for the golden target speaker

In contrast to the previous section where analysis was mainly conducted at the dataset level. In this section, we are much more interested to analyze the effect of the target speaker on a per-speaker basis. The question that we are aiming to answer is the following: “Is there a source & target speaker combination that maximizes privacy and utility performance?”. If this is to be the case, speaker anonymization could be personalized for each user. (Note that the attacker could use this as an attack vector). By trying to answer this question we will also answer the following question: “Is there a target speaker that maximizes privacy and utility performance for everyone?”.

In order to study the influence of the target speaker, we need to isolate this variable in order to study it. As such, the only target selection strategy that allows specifying the target speaker parameter is the “constant speaker” target selection algorithm. Hence, anonymization will be performed by modifying all utterances of all speakers to sound like a single target speaker.

Experiments are performed multiple times with different target speaker identities to provide averaged global results, and detailed analysis of the target effect on multiple source speakers.

4.3.1 Experimental setup

For this experiment, we run the anonymization and evaluation 40 times, each having a specific target speaker. To cover as best as possible the target speaker space, we identify 20 female and 20 male clusters in the VPC speaker x-vector pool (*LibriTTS train-other-500*) using K-Means. We then pick the speaker x-vector that is the closest to the centroid of each cluster. The performance assessment is carried on for each of those 40 target speakers. The following procedure methodology is used: first, considering each of the 40 target speakers, we anonymize all utterances of *Librispeech test-clean* (compromised and vulnerable speech) and *Librispeech train-clean-360* using the “constant speaker” target selection strategy using the x-vector-based anonymization system (Section 3.6.2.2) trained with the dataset of the VPC (Table 3.1). In contrast to the previous experiment where we obtained PLDA scores for each possible pair of one vulnerable utterance and one compromised utterance, here we follow the VPC evaluation method where the x-vectors of an enrollment speaker are averaged over his/her utterances before PLDA scoring. As such the results presented here slightly differ from the previous sections. Then, we train the ASV model on the anonymized *Librispeech train-clean-360* dataset to comply with the white-box informed attacker. Lastly, we evaluate the privacy performance for each of the speakers of *Librispeech test-clean* using the specially trained ASV attacker. As a result, we obtain EER and $D_{\leftrightarrow}^{\text{sys}}$ results for each compromised speaker and target x-vector, that is a total of 29×40 results. In this section, the primary metric used to draw conclusions is the $D_{\leftrightarrow}^{\text{sys}}$ metric. To evaluate the quality of the conversion process in terms of utility, we use the pre-trained (non-adapted) ASR model released by the VPC 2020 organizers.

In the following sections, we start by presenting the privacy metric at the dataset level, much like previously. This provides the average privacy performances for all the speakers of the *Librispeech test-clean* dataset for all the target x-vectors. Then we present privacy performances for a chosen clear speaker, which gives a more fine-grained insight into the role of the target x-vector. We finish, by classifying the population of the compromised speakers into two groups, depending on their privacy performance.

4.3.2 Global privacy results

Table 4.3 compares the anonymization performance on a global scale. The first line presents the linkability when no anonymization is performed (i.e., on clear speech data). Clear speech encapsulates the speaker’s information to a high degree as the $D_{\leftrightarrow}^{\text{sys}}$ scores are very high > 0.90 . For clear speech, a disparity between male and female linkability performance is observed. Females seem to be harder to recognize. This observation may be a result of the training dataset having more male variety or the inherent differences in pitch and spectral content of female and male speech, which may make female speech more difficult to differentiate from one another.

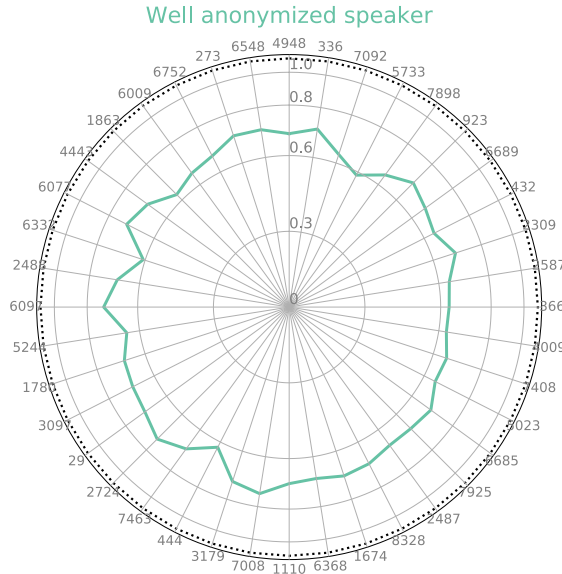
Table 4.3 – Linkability ($D_{\leftrightarrow}^{\text{sys}}$) and EER scores for clear and anonymized speech. For the anonymized data, the mean and standard deviation values are calculated over the 40 experiments (i.e., one for each target speaker).

| | Female speakers | | Male speakers | | Average | |
|--------------|---|----------------|---|----------------|---|------------------|
| | $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | EER \uparrow | $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | EER \uparrow | $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | EER \downarrow |
| Clear speech | 0.90 | 7.7 | 0.96 | 1.1 | 0.93 | 4.4 |
| Anonymized | 0.74 ± 0.01 | 11.6 ± 0.6 | 0.75 ± 0.01 | 10.6 ± 0.8 | 0.74 | 11.1 |

For anonymized speech, the privacy metrics come from 40 ASV evaluations, each using a different target speaker identity and ASV model. The mean and standard deviation values are calculated from the 40 evaluations. From the linkability score difference between clear and anonymized lines of Table 4.3, we can conclude that speakers are less linkable to their true identity after applying the x-vector-base anonymization system. Interestingly, the female and male linkability performance disparity is not present anymore after anonymization, privacy performance is the same regardless of the speaker’s gender. Linkability scores drop by 0.19, meaning the anonymization system has some effectiveness. The rather low standard deviation values across all scores show that there is no large variation when changing the target speaker. This suggests that a given target is not more suited than another to anonymize the whole dataset.

4.3.3 Detailed privacy results

We conducted a detailed analysis to check whether a specific target identity is more suited to anonymize one or more speakers of our test dataset. Figure 4.11 illustrates the visualization used for this study in the case of a single source speaker. The linkability $D_{\leftrightarrow}^{\text{sys}}$ scores are computed for a speaker whose speech was anonymized 40 times, with different target x-vectors.



Results on a single test speaker (speaker ID 5105)

Figure 4.11 – Linkability scores ($D_{\leftrightarrow}^{\text{sys}}$) obtained using 40 informed ASV attackers and baseline ASV, for each of the 40 target speakers (colored solid line) and on clear speech (dotted black circle). Each spoke corresponds to a target speaker (and ASV attacker model).

The dotted black circle indicates the linkability of this speaker on clear speech (note that the clear speech evaluation does not depend on the target speakers, hence the circle). And, for each of the 40 target speakers, the linkability is presented by the colored solid line. After transforming the speech with 40 target speakers, we can observe that none of the 40 targets are significantly better to anonymize this speaker’s voice. The variation between the anonymized linkability scores is more likely due to the difference between the ASV attacker model rather than a better target choice. This observation also applies to the 29 other speakers of our test dataset. It is also noteworthy that, out of the 40 target identities, 20 of them induce cross-gender voice conversion, as we have for the targets, 20 male centroids and 20 female centroids. Results for same-gender and cross-gender voice anonymization were found similar.

Out of this experiment, we observed that some speakers had their privacy increased (e.g., speaker ID 5105), while others had not. In the latter case, the colored solid line in Figure 4.11 would be close to the dotted one. The following figure presents this disparity.

Figure 4.12 shows the linkability $D_{\leftrightarrow}^{\text{sys}}$ scores of two groups to illustrate an anonymization disparity behavior on two groups of speakers: one for which the anonymization system did not conceal PII, and the other for which the anonymization did conceal some PII. Interestingly, both groups are of similar size. For the poorly anonymized speakers, we observe that the distribution of linkability scores on anonymized speech (colored area) completely overlaps the distribution of linkability scores on clear speech (gray area). The anonymization system did not

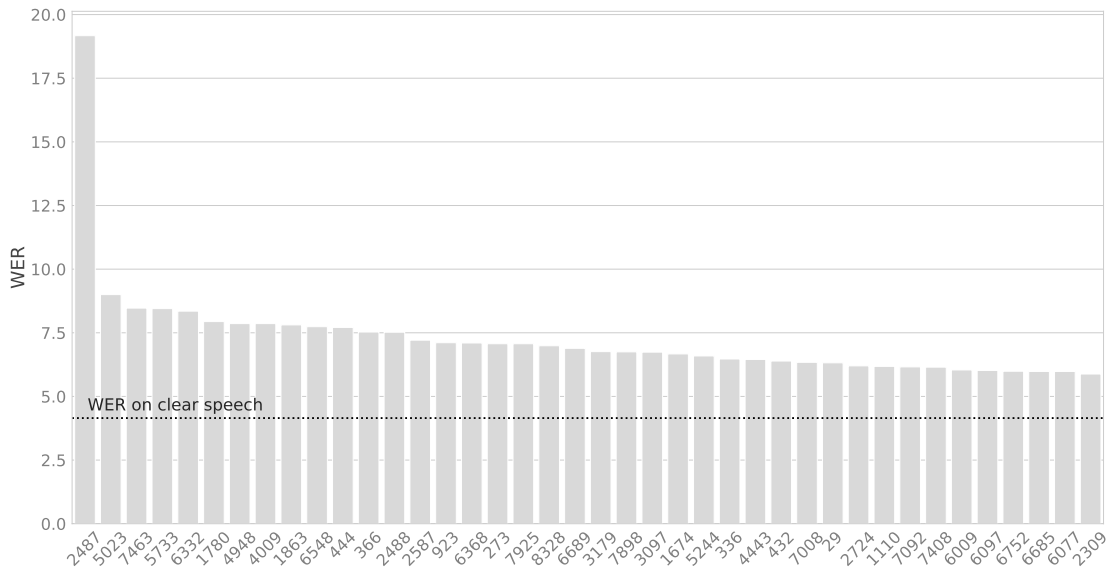


Figure 4.13 – WER scores obtained by the VoicePrivacy ASR evaluation system, trained once on clear speech, for each of the 40 target speakers and on clear speech (dotted black line).

The very high utility loss yielded when using target speaker ID 2487 is due to a generalization issue of the anonymization pipeline, not the ASR model, in this case, non-intelligible speech was generated at the beginning of some segments. We conducted an additional test using 100 randomly selected target speakers from the pool, and were able to find 3 target speaker x-vectors that have similar behavior, in the worse case the WER reached a value of 59.37. Informal listening tests reported that the target speaker’s speech contained singing segments, producing faulty x-vectors. Further analysis needs to be conducted. Interestingly, the target speakers IDs 2487, 7463 and 6332, which were slightly better in terms of privacy protection are all worse in terms of utility preservation as they are in the top 5 worse target speakers in Figure 4.13. Making mechanistic conclusions about privacy/utility trade-offs relative to the target speaker parameter is challenging due to the dependency on the speech synthesis system. However, we observed that some target x-vectors are better suited to generate intelligible speech than others.

In the previous section, we outline that using the “constant speaker” target selection strategy was less effective in terms of utility compared to the “Farther 200 random 100” target selection strategy (see Tables 4.2 and 4.4). The hypothesis explaining this disparity might be a generalization issue of the speech synthesis against unseen (during training) x-vector, the average operation performed in the “Farther 200 random 100” seems to mitigate this issue. Backing up this argumentation, in (Shamsabadi et al. 2022), the authors identified that using too few speakers before the averaging in the “dense” strategy negatively affects utility. The average of many target speakers appears to decrease the specificity of the x-vector, making it a better-suited target that preserves utility. From Table 4.4, the same conclusion is made. The WER is lower for the “*Farther 200 random 100*” from external pool compared to the “*Constant*

speaker” from external pool for both *LibriSpeech* and *VCTK* datasets. To obtain the same WER for the “Constant speaker”, we suggest sampling the target speaker from the dataset used to train the speech synthesis model rather than an external one. Table 4.4 shows that performing this simple modification, yields the same WER for the “Constant speaker” and “Farther 200 random 100”. One may argue that the “Constant speaker” strategy, using a real speaker’s voice as the target, raises ethical concerns, compared to more random target selection strategies, however, there is no guarantee that the “Farther 200 random 100” does not also generate speech similar to the one of a real identity. Overall, for research anonymization evaluation, the ethical concerns are nonexistent, whereas, in real applications the use of a pitch morphing algorithm (that most TVs/media already use) can be applied on top of the anonymized speech generated by the pipeline, to address any eventual ethical concern.

Table 4.4 – Privacy and utility performance metrics of target speaker selection strategies. ASV and ASR models are trained on anonymized speech. Refer to Table 3.1 for the dataset/pool.

| Dataset | LibriSpeech test-clean | | | <i>VCTK test</i> | | |
|---|---|----------------|------------------|---|----------------|------------------|
| Method | Privacy | | Utility | Privacy | | Utility |
| | $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | EER \uparrow | WER \downarrow | $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | EER \uparrow | WER \downarrow |
| Clean speech | 0.93 | 4.1 | 4.1 | 0.93 | 2.7 | 12.8 |
| “Farther 200 random 100” from external pool | 0.78 | 8.3 | 4.4 | 0.73 | 10.2 | 10.7 |
| “Constant speaker” from external pool | 0.67 | 13.5 | 5.1 | 0.49 | 20.6 | 13.0 |
| “Constant speaker” from training dataset | 0.73 | 11.2 | 4.4 | 0.52 | 19.8 | 10.2 |

4.3.5 Discussion

The conclusions presented in this section are multifaceted. To answer the questions, “Is there a source & target speaker combination that maximizes privacy and utility performance?”, and “Is there a target speaker that maximizes privacy and utility performance for everyone?”. In our experiment, privacy performance was not significantly increased by a particular choice of a target speaker. There is no source & target speaker combination that maximizes the privacy, as well as a target speaker that better conceals PII for everyone. We conclude that the linkability ASV informed attacker was robust to different target speakers. As for utility, our experiment shows that there are target speakers’ identities that better preserve objective intelligibility measured by ASR systems.

Hence, we further analyzed the “constant speaker” target selection strategy to optimize its utility performance. We observed that using target speakers that were seen during the training of the synthesis system, generated anonymized speech better preserving the utility. As such, the main conclusion about the influence of the target speaker relates to utility, not privacy. We believe that using the “constant speaker” and cherry-picking a target speaker from the synthesis training dataset in such a way that the anonymized voice produced has good intelligibility is the most effective and secure method to maximize the utility. This form of VC is referred as to

any-to-many VC. As presented in Section 2.4.2.1, *any-to-many* VC approaches cannot target *any* target speaker, only the *many* ones seen during training can be used as target speaker, this is not a problem for the “*constant speaker*” *from training dataset* target selection strategy but rather an advantage. As in the voice conversion field, it is well known that the *any-to-many* VC approaches are simpler than the *any-to-any* one, in terms of designing and training, additionally, they usually produce better speech quality. By using *any-to-many* VC, the use of x-vector as the encoding medium of the target speaker is no longer justified. A one-hot encoding can very well be used to target *many* speakers, this simplifies the target selection procedure as the x-vector extractor and pool of x-vectors are no longer necessary.

In this section, large privacy performance disparities between different speakers of the test dataset are exhibited, indicating an issue about fairness! Tracing the cause of these disparities is a topic of research that has never been explored, however, we believe it is an essential topic. One first step in this research could be to identify how individuals behave differently regarding linkability attacks on their anonymized speech. The concept of the biometric menagerie in the literature formally recognizes the idea of categorizing and labeling user groups with animal names based on their characteristics when interacting with biometric systems (Doddington et al. 1998; Houmani et al. 2016). Extending the menagerie to the field of speaker anonymization could be interesting. The privacy performance gaps between speakers also raise the question of the use of speaker average-based metrics such as the $D_{\leftrightarrow}^{\text{sys}}$ and EER metrics. While not used in this thesis, the ZEBRA (Nautsch et al. 2020) metric seems to be a compelling solution to better evaluate privacy in worst-case privacy disclosure.

4.4 Conclusion

In this chapter, we have challenged the target speaker attribute of voice conversion-based speaker anonymization. We found that existing target selection strategies have two major issues: 1) no target-level unlinkability guarantees that compromise privacy evaluation and 2) potential difficulties to generate appropriate training data for the attacker. Additionally, in our experiment, the target selection algorithm and target speaker do not affect the strength of PII concealment, there is no golden target speaker for privacy. However, in our experiment, we showed that there are multiple golden and bad target speakers when the utility is considered. For those reasons, we promote the use of the “*constant speaker*” *from training dataset* target selection strategy which provides many characteristics: 1) target-level unlinkability guarantees, 2) appropriate data generation for attacker training because of a 3) white-box informed attacker scenario, 4) cherry-picking targets that preserve utility, 5) simplicity and straightforwardness. Ensuring proper privacy evaluation in VC-based anonymization is a challenge. And we conclude that the main role of the target speaker in VC-based anonymization is firstly to ensure proper evaluation.

VOICE CONVERSION ANONYMIZATION WITH FEATURE-LEVEL DISENTANGLEMENT

Contents

| | | |
|------------|---|------------|
| 5.1 | Introduction | 89 |
| 5.2 | Fundamental frequency feature transformation | 90 |
| 5.2.1 | Linear shift transformation | 91 |
| 5.2.2 | Additive white Gaussian noise transformation | 91 |
| 5.2.3 | Quantization transformation | 91 |
| 5.2.4 | Wrapping up F_0 transformations | 92 |
| 5.2.5 | Experimental setup | 93 |
| 5.2.6 | Experimental results | 94 |
| 5.2.7 | F_0 isolated privacy evaluation | 95 |
| 5.2.8 | Discussion | 97 |
| 5.3 | Linguistic model transformation | 98 |
| 5.3.1 | ASR-BN Isolated privacy evaluation | 98 |
| 5.3.2 | Adversarial learning model transformation | 99 |
| 5.3.3 | Experimental setup | 100 |
| 5.3.4 | Experimental results | 102 |
| 5.3.5 | Discussion | 105 |
| 5.4 | Linguistic feature transformation | 105 |
| 5.4.1 | Laplace noise transformation | 105 |
| 5.4.2 | Vector quantization transformation | 106 |
| 5.4.3 | Wrapping up ASR-BN feature transformation | 107 |
| 5.4.4 | Experimental setup | 108 |
| 5.4.5 | Experimental results | 109 |
| 5.4.6 | Discussion | 114 |
| 5.5 | Conclusion | 115 |

5.1 Introduction

Speaker anonymization aims to transform a speech signal to remove the source speaker’s identity while leaving the spoken content unchanged and potentially other information. Our baseline,

the x-vector-based method performs the transformation by relying on the disentanglement of linguistic and F_0 information from speaker information and voice conversion. Usually, an acoustic model from an automatic speech recognition system extracts the linguistic representation while an x-vector system extracts the speaker representation. Additionally, the fundamental frequency is also extracted for intonation preservation. In this chapter, we identify the degree to which the linguistic and F_0 features are disentangled from the speaker, and propose modification methods to improve the disentanglement, and thus the privacy of the converted anonymized speech.

5.2 Fundamental frequency feature transformation

Previously, under the x-vector-based methods, we have assumed that only the original x-vector extracted by the anonymization framework would be transformed into the target speaker x-vector while the other two sets of features (ASR-BN linguistic representation and F_0 trajectory) would remain unchanged. However, intonation features contribute to speaker identity and the F_0 trajectory contains some Personally Identifiable Information (PII), in (Hillenbrand et al. 2009) the authors were capable to distinguish female and male only with the F_0 .

Additionally, maintaining the original F_0 while potentially changing the gender of the x-vector can result in inconsistent features and affect the naturalness of synthesized speech. Multiple works have investigated F_0 conditioned voice conversion (Huang et al. 2019; Qian et al. 2020a) (see Section 2.4.4) and concluded that F_0 conversion effectively improves the naturalness of the output speech. Motivated by those results, we propose to modify the F_0 values of a source utterance (cf. module D in Figure 5.1) in the x-vector-based voice conversion system to see if it also improve privacy.

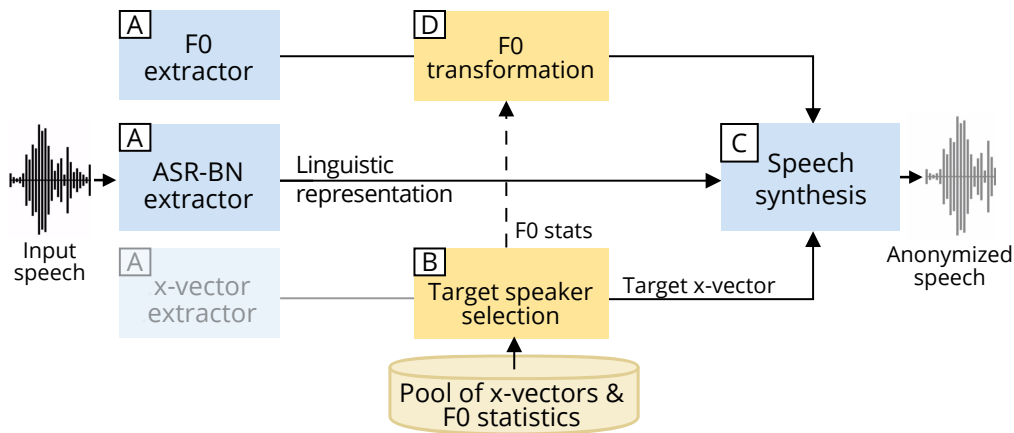


Figure 5.1 – X-vector-based speaker anonymization system with module D added. The original x-vector of the input utterance might, or might not be used by the target selection strategy.

5.2.1 Linear shift transformation

To convert the source F_0 trajectory to be similar to the target speaker, we propose to use a simple linear transformation:

$$\hat{f}_t = \mu_y + \frac{\sigma_y}{\sigma_x} (f_t - \mu_x) \quad (5.1)$$

where f_t represents the log-scaled F_0 of the source speaker at frame t , μ_x and σ_x represent the mean and standard deviation of the log-scaled F_0 for the source speaker. μ_y and σ_y represent the mean and standard deviation of the log-scaled F_0 for the target-speaker. The linear transformation and statistical calculation are only performed on voiced frames. When the target speaker x-vector is a combination of multiple speakers like in the ‘‘Farther 200 random 100’’ target selection strategy, the mean and standard deviation for the target speaker is calculated by averaging information from the same 100 speakers selected to derive the target x-vector. For the target selection strategies, where a unique known speaker is selected as the target x-vector like in the ‘‘Constant speaker’’ strategy, the F_0 statistics come from the same selected speaker. These statistics are first extracted and stored for the pool of speakers when developing the pipeline and passed to the F_0 conversion module (module D in Figure 5.1) before speech synthesis.

5.2.2 Additive white Gaussian noise transformation

Inspired by the work of (Gaznepoglu et al. 2021; Shamsabadi et al. 2022), we also experiment with noise based F_0 modification. Here the goal is not to improve naturalness as the noise will disturb the F_0 , however, it will increase privacy. Noise addition is one of the most used data privacy protection techniques. When noise is added to data, it introduces random variations that can make it more difficult for an attacker to extract sensitive information from the data. This is because the added noise can obscure or mask the underlying patterns in the data that might reveal sensitive information. Work on additive noise was first published in (Kim 1986) with a method named Additive White Gaussian noise (AWGN). Noise is additive because it is added to any noise that might be intrinsic to the information system, white because it has uniform power across the frequency band for the information system (an analogy to the color white which has uniform emissions at all frequencies in the visible spectrum), and Gaussian because it has a normal distribution in the time domain with an average time domain value of zero. For each frame f_t , and a target noise power D in dB, the transformation can be written as:

$$\hat{f}_t = f_t + \mathcal{N}(0, \sqrt{10^{(D/10)}}) \quad (5.2)$$

5.2.3 Quantization transformation

Quantization is a technique for representing a continuous signal as a discrete set of values. It works by mapping a set of input values to a set of output values, with each input value being replaced by the nearest output value from the set. The quantization method that we propose to

use to modify the F_0 is the simple linear quantization method that reduces the values used to represent a signal using the equation described in (Oppenheim et al. 1999):

$$\hat{f}_t = \text{round} \left(2^{B-1} \frac{f_t - f_{min}}{f_{max} - f_{min}} \right) \quad (5.3)$$

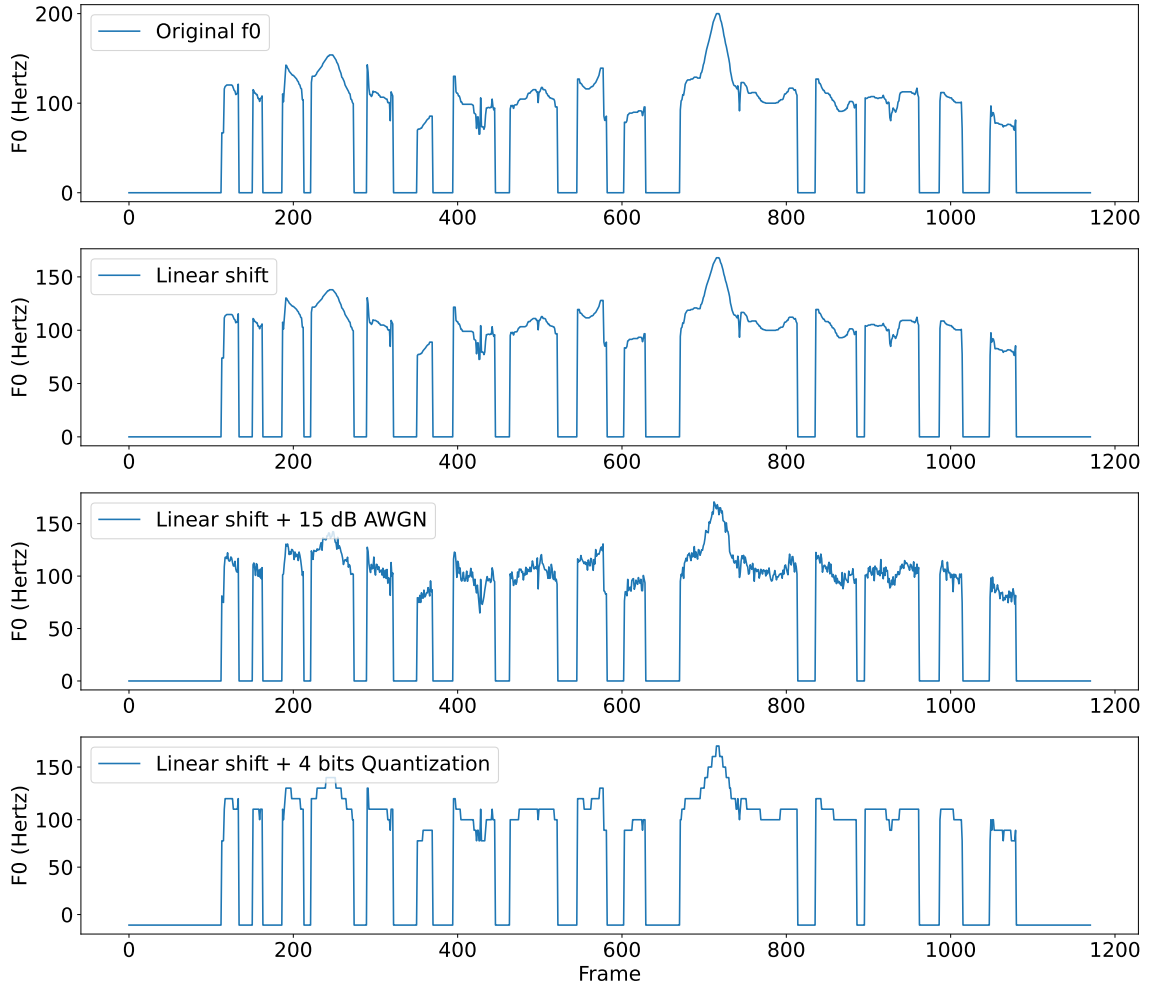
where B is the number of quantization bits. In contrast to AWGN transformation, which adds random noise, quantization smooths out variations in the signal. This can result in the same output property as the AWGN modification that makes attackers extract less sensitive information.

However, it is worth noting, that quantization is not a foolproof method for anonymization and has seen less research interest in contrast to noise-based anonymization. It may be possible for an attacker to extract some speaker-specific information from the quantized F_0 signal. To the best of our knowledge, only the work of (Miche et al. 2016) specifically used quantization methods for data anonymization, whereas much more work has been done with noise-based anonymization methods. We consider this framework to be a good option for voice anonymization because the transformed speech using quantization-based methods will have less audible noise than speech transformed using noise-based methods. By minimizing the perceptible noise in anonymized speech, its usefulness can be improved.

5.2.4 Wrapping up F_0 transformations

An intuitive comparison between the types of target speaker selection strategies presented in Chapter 4 and the F_0 transformation presented here can be made. The “random speaker” target selection strategy introduces random speaker modification for each utterance of a speaker such as it ends up being confused with other speakers. Similarly, after shifting and scaling, the AWGN method, described here has the same goal: modify the input F_0 of a speaker such as it is more likely to be confused with other speaker’s F_0 . In contrast, the “constant speaker” target selection strategy forces everyone to be modified to match a single target, aiming to reduce the number of perceived anonymized speakers to one. Quantization works similarly, after shifting and scaling, the local change of the F_0 trajectory that makes a F_0 linked to an identity/group of speakers is smoothed out by the quantization process reducing the number of perceived F_0 dependent speaker characteristics. Figure 5.2 shows transformation examples.

Noise-based anonymization works by adding noise in a transmission channel such that only the most relevant information stands out from the noise, whereas quantization-based anonymization works by reducing the transmission channel capacity such that the irrelevant information is not encoded.

Figure 5.2 – Example of the F_0 transformation.

5.2.5 Experimental setup

The baseline anonymization system is the x-vector-based provided by the VoicePrivacy Challenge. With this system, two sets of experiments are performed in this section. One set that follows all the VoicePrivacy Challenge 2022 requirements with the “farther 200 random 100” target selection strategy that allows some voice distinctiveness. For those experiments, the gray-box semi-informed attacker is used to evaluate the privacy protection of the anonymization. The other set of experiments does not comply with the voice distinctiveness requirement as it uses the “constant speaker” target selection strategy. As shown in Chapter 4, with the “constant speaker” strategy, a superior white-box evaluation with better guarantees can be used to evaluate privacy protection. Utility evaluation is done with an ASR model trained on anonymized speech. The datasets used for evaluation/training/testing are the same as in both 2020 and 2022 VPC. Results are presented for the *LibriSpeech test-clean* and *VCTK test* datasets. The performances are assessed in terms of $D_{\leftrightarrow}^{\text{sys}}$ and EER scores for privacy and WER for utility.

With the “farther 200 random 100” target selection strategy evaluations, the linear F_0 transformation are evaluated against the same anonymization system that does not have it. For the “constant speaker” target selection strategy evaluations, the transformations being evaluated include the linear shift, a combination of the linear shift and the AWGN transformation with a noise value of 15dB, and a combination of the linear shift and the quantization transformations on 4 bits.

5.2.6 Experimental results

Table 5.1 compares the anonymization performance across F_0 modification techniques. The first line presents the linkability and utility of clear speech. Clear speech encapsulates the speaker’s information to a high degree as the $D_{\leftrightarrow}^{\text{sys}}$ scores are very high > 0.90 for the *LibriSpeech test-clean* and *VCTK* datasets. The WER of this first line corresponds to our baseline utility performance, we observe that the WER score for the *VCTK* dataset has a higher value than on *LibriSpeech*. This utility disparity is explained by the ASR evaluation model which is only trained on *LibriSpeech*, making it more relevant for audiobooks.

When comparing lines 2 and 3, which involve a gray-box semi-informed attacker, we observe a privacy improvement when using the linear shift F_0 transformation for both the *LibriSpeech* and *VCTK* datasets as the $D_{\leftrightarrow}^{\text{sys}}$ score decrease from 0.78 to 0.63 and 0.73 to 0.66 respectively. However, when compared to the “constant speaker” strategy with the white-box informed attacker, no privacy improvements are recorded by the linear shift F_0 transformation on the *LibriSpeech* dataset. With those experiments, we reinforce the conclusion of Chapter 4 by showcasing how a misunderstanding of the role of the target speaker can impact evaluation. With the “farther 200 random 100” target selection strategy and the gray-box semi-informed attacker,

Table 5.1 – Privacy and utility results for the chosen target selection strategies and F_0 transformation. The interval of confidence stays within $\pm 0.40\%$ EER for all experiments. Lines 2 and 4 do not involve F_0 transformations.

| Dataset | <i>LibriSpeech test-clean</i> | | | <i>VCTK test</i> | | |
|--|--|---------------------------|-----------------------------|--|---------------------------|-----------------------------|
| | Privacy $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | Utility EER \uparrow | Utility WER \downarrow | Privacy $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | Utility EER \uparrow | Utility WER \downarrow |
| 1. Clear speech | 0.93 | 4.1 | 4.1 | 0.93 | 2.7 | 12.8 |
| 2. Anon. “Farther 200 random 100” | 0.78 | 8.3 | 4.4 | 0.73 | 10.2 | 10.7 |
| 3. Anon. “Farther 200 random 100” + Linear | 0.63 | 16.5 | 4.4 | 0.66 | 13.0 | 10.4 |
| 4. Anon. “Constant speaker” | 0.73 | 11.2 | 4.4 | 0.52 | 19.8 | 10.2 |
| 5. Anon. “Constant speaker” + Linear | 0.73 | 11.6 | 4.4 | 0.46 | 22.2 | 10.1 |
| 6. Anon. “Constant speaker” + Linear + AWGN | 0.73 | 11.2 | 4.4 | 0.47 | 21.5 | 10.2 |
| 7. Anon. “Constant speaker” + Linear + QUANT | 0.73 | 11.3 | 4.4 | 0.47 | 21.7 | 10.1 |

one might believe that a linear shift F_0 transformation increases the privacy performance of an anonymization system as the $D_{\leftrightarrow}^{\text{sys}}$ decreases from lines 2 and 3 of Table 5.1. With proper white-box evaluation, privacy improvement is almost nonexistent (lines 4 and 5). Due to this evaluation misunderstanding, we made inaccurate claims about the privacy improvement obtained by linear F_0 modification in the paper: « [A Study of F0 Modification for X-Vector Based Speech Pseudonymization Across Gender](#) » (Champion et al. 2020a). That is why, for the rest of this thesis, we will only use a white-box attacker with the “constant speaker” strategy.

Regarding the experiments with the “constant speaker” anonymization strategy, the results show that for all experiments, the WER utility measure is the same whether the F_0 was modified or not. As for privacy, using the AWGN or quantization transformations on top of the linear shift does not improve the performance compared to only using the linear shift in the *VCTK* dataset as $D_{\leftrightarrow}^{\text{sys}}$ is around 0.47. And for in the *LibriSpeech* dataset, none of the transformations improves privacy. This conclusion goes against the main hypothesis that intonation features contribute to the speaker’s identity. This disparity can be explain by the fact that the x-vector system used for finding the linkage is not heavily prosody based. If the attacker were using a prosody based system it is more likely that we would find that the shift helps. However, in an anonymization pipeline, privacy leakage can occur from multiple parts. If the linguistic representation feature is the main source of speaker information leakage, modifying the F_0 to improve F_0 /speaker disentanglement will not result in a significative privacy performance improvement. In the next section, we aim to measure the F_0 /speaker disentanglement.

5.2.7 F_0 isolated privacy evaluation

In this section, our focus is to measure the F_0 /speaker disentanglement only on the F_0 trajectory extracted from the speaker’s speech, no speech transformations are applied beforehand. By isolating the F_0 feature alone and not using the full anonymization pipeline, we can assess and measure if any speaker information leakage occurs from the F_0 in the pipeline. We also put to evaluation the F_0 transformations presented above¹.

5.2.7.1 Experimental setup

The experimental setup is similar to the one in the previous experiment. However, the ASV is trained to identify speakers only using the one-dimensional F_0 features instead of the MFCCs. This model is a five TDNN x-vector model and is trained on the *LibriSpeech train-clean-360* using the Python speaker identification sidekit tool (Larcher et al. 2016). For reference, an experiment is also done with MFCCs as input. When a F_0 transformation is used, the ASV model is retrained to accommodate the transformations, complying with white-box attacks. Evaluation is done on the *LibriSpeech test-clean* dataset and the EER metric is used to evaluate

1. The experiments presented in this section are a collaborative effort, where the experiments were done by Shalini Priya during her internship.

the disentanglement property of the F_0 , a value of 50% indicating perfect disentanglement which would provide the best downstream anonymization guarantees.

5.2.7.2 Experimental results

Table 5.2 – Results in EER of the privacy of the isolated F_0 feature, with and without transformations. The confidence interval is provided for each experiment.

| F_0 transformation | <i>LibriSpeech test-clean</i> EER \uparrow |
|--|---|
| 1. Clear MFCCs | 4.3 \pm 0.5 |
| 2. Clear F_0 | 19.6 \pm 1.1 |
| 3. Linear shift | 19.9 \pm 1.1 |
| 4. QUANT _{2bit} | 19.2 \pm 1.3 |
| 5. AWGN _{15dB} | 30.9 \pm 1.4 |
| 6. AWGN _{30dB} | 36.0 \pm 1.5 |
| 7. Linear shift + AWGN _{15dB} | 44.2 \pm 9.1 |
| 8. Linear shift + AWGN _{30dB} | 46.0 \pm 7.3 |
| 9. Linear shift + QUANT _{4bits} | 26.7 \pm 1.3 |

Table 5.2 summarizes the results, and lines 1 and 2 show the EER obtained with MFCCs and F_0 input features respectively. We observe that for the *LibriSpeech* dataset, speakers were somewhat linkable only with their F_0 trajectories, as the EER equal 19.6%. As expected, the F_0 trajectories do not enable the same degree of linkability performance as MFCCs. This is expected because the MFCCs contains much more information. However, an EER of 19.6% is still too high and indicates that the F_0 is not completely disentangled from the speaker, meaning it could impact the performance of an anonymization pipeline.

Lines 3 to 6 of the table, display the EER when the F_0 is modified with either linear shift, quantization, or AWGN transformations. Out of them, the AWGN with a noise level of 30dB shows the best disentanglement performance with a EER of 36.0%. However, this kind of transformation might be a bit too strong and impact any potential downstream synthesis. The AWGN_{15dB} transformation has a lower disentanglement capability, with an EER of 30.9%, but might be more suited for downstream synthesis. For the 4 bits quantization transformation, the EER is unchanged compared to the clear F_0 , indicating that the local (per-frame) smoothing of the F_0 does not make the speaker less linkable in this dataset. Finally, for the linear shift, no improvement in disentanglement is observed when all F_0 utterances are mapped to a single F_0 mean and variance statistics.

Lines 7 to 9 of the table, show interesting transformations combination. It appears that combining the globally (per utterance) applied F_0 mean and variance linear shift and locally (per

frame) applied noise based AWGN or quantization transformations improve the disentanglement. The best of them is linear shift + AWGN-based having EER above 40.0%. While linear shift + quantization has an EER of 26.7%.

5.2.8 Discussion

In this section, which is dedicated to the F_0 feature, we evaluated the level of disentanglement between the F_0 and the speaker. We concluded that the F_0 can be used to perform speaker linkability attacks, even though its performance is inferior to that of MFCCs. This conclusion applies to the *LibriSpeech* dataset used for evaluation and training the white-box ASV model, the dataset is based on audiobook reading, which may have strong biases in the way readers (and so speakers) read their chapter to match a defined prosodic style.

We presented common signal-processing-based transformations to improve the F_0 disentanglement. In our isolated experiment, we showed that to maximize F_0 disentanglement, a combination of F_0 linear shift and additive white Gaussian noise is the best. However, applying this modification to the baseline x-vector-based anonymization system provided by the VPC organizers does not increase the recorded privacy performance at all, implying that the main source of speaker leakage in this system is not the F_0 feature. As the main features used to perform anonymization are the target x-vector that we tackled in Chapter 4 and the F_0 feature studied in this section, this only leaves the linguistic representation features as the main contributor of speaker information leaker. In the following sections, the linguistic representation features will be our subject of experiments.

Being able to evaluate the disentanglement of the features used by the speech synthesis before generating anonymized voices is a compelling option to guarantee better anonymization. However, do all features need to be disentangled before being used by the speech synthesis to produce the most anonymized speech? The speech synthesis itself could anonymize speech on its own, even if the input feature is not completely disentangled. The quantized F_0 may very well be one type of non-perfectly disentangled representation, but still improve privacy, as quantized values are easy to process. As suggested by (Qian et al. 2020a; Qian et al. 2020b) it appears that quantized F_0 trajectories improve naturalness.

5.3 Linguistic model transformation

Until now, we have assumed that the linguistic representation features used in the baseline x-vector-based anonymization system (which is a sequence of ASR-BN, see Section 2.4.1.3), do not convey speaker information or if so, the speech synthesis system does not rely on it to transform voices. However, in Sections 4.3 and 5.2, we identified that out of the three features used by speech synthesis, modifying the target x-vector identity and the F_0 does not improve privacy, indicating that the linguistic ASR-BN representation is the primary source of speaker PII leakage in this pipeline. This could be explained by the fact that ASR-BN is the feature with the highest dimension and is sequential. In this section, we start by performing an isolated privacy evaluation of the ASR-BN representation to challenge the ASR-BN/speaker disentanglement assumption. Then, in this section, we propose using adversarial learning to transform the ASR-BN model to improve the disentanglement of the features it generates.

5.3.1 ASR-BN Isolated privacy evaluation

In this section, our focus is to measure the content/speaker disentanglement only on the ASR-BN representation, no speech transformation is applied beforehand. The objective is to assess and measure if this representation alone leaks speaker PII, if it is the case, this means the ASR-BN could restrict the speaker concealment performance of speaker anonymization systems. Additionally, as ASR-BN representation also conveys the content of the message, we also evaluate utility using an ASR decoder to obtain the predicted sequence of words which can be compared to a reference.

5.3.1.1 Experimental setup

This experiment uses a PyTorch implementation, based on pkwrap (Madikeri et al. 2020), rather than the VPC Kaldi (Povey et al. 2011) implementation for the acoustic model ASR-BN extractor. The differences between our implementation and the one used in the VPC do not affect conceptually the extraction. Our model is trained only on *LibriSpeech train-clean-100* to reduce the computation time needed for this and the following experiments. The cost function used is the E2E-LF-MMI (Hadian et al. 2018b), allowing flat-start training without pre-training or prior alignment from a GMM model (see Section 2.2.1.3). According to (Madikeri et al. 2020), the outputs of the model are left-biphones rather than triphones and the model is composed of 15 TDNN-F layers, the ASR-BN is extracted from the 13th.

The ASV model directly takes the ASR-BN representation (of 256 dimensions) as input instead of MFCCs. This model is a five TDNN x-vector model and is trained on the *LibriSpeech train-clean-360* using the Python sidekit ASV toolkit (Larcher et al. 2016). For reference, an experiment is also done with MFCCs as input. Evaluation is done on the *LibriSpeech test-clean* dataset and the EER metric is used to evaluate privacy, and the WER is used to evaluate utility.

5.3.1.2 Experimental result

Table 5.3 – Privacy and utility of the linguistic ASR-BN representation and MFCCs. The ASR-BN is the one being used in x-vector-based speaker anonymization systems.

| | <i>LibriSpeech test-clean</i> | |
|-----------------------|-------------------------------|------------------|
| | EER \uparrow | WER \downarrow |
| Clear MFCCs features | 3.7 ± 0.4 | 5.8 ± 0.3 |
| Clear ASR-BN features | 6.7 ± 0.8 | 5.8 ± 0.3 |

Table 5.3 shows the privacy and utility results of the experiment, we observe that the ASR-BN feature EER score is rather close to the MFCCs feature EER, indicating that the transformation used to go from the clear raw speech signal to the ASR-BN does not remove speaker information. This conclusion was also observed in other studies by (Adi et al. 2019; Srivastava et al. 2019) where they showed that the disentanglement property of ASR-BN representation is limited and even nonexistent. This indicates that, much like for the F_0 , modifications of the extractor are needed to disentangle speaker information. The WER for utility is the same in both experiments as the ASR-BN extractor is part of the ASR model.

In the next section, we introduce some transformation techniques applied to the ASR-BN extractor and ASR-BN feature to improve disentanglement.

5.3.2 Adversarial learning model transformation

Adversarial learning is an interesting approach for extracting disentangled representations (Kazemi et al. 2019; Huang et al. 2020c). It has also been used for anonymization purposes (Feutry et al. 2018), where the feature to disentangle is PII. For our purpose, we apply this framework to increase the content/speaker disentanglement of the ASR-BN feature. Based upon the definition of adversarial learning defined in Section 1.4.5, we add to the ASR-BN extractor an ASV speaker adversarial branch. This adversarial branch adds a negative loss to the extractor during training, inducing it to remove speaker information. As such, the loss function of the ASR acoustic model is expressed as:

$$\min_{\theta_{bn}} \left[\min_{\theta_{asr}} \mathcal{L}_{\text{mmi}}(\theta_{bn}, \theta_{asr}) - \alpha \min_{\theta_{spk}} \mathcal{L}_{\text{aam}}(\theta_{bn}, \theta_{spk}) \right] \quad (5.4)$$

where θ_{bn} denotes the weights of the ASR-BN extractor, θ_{asr} refers to the weights of the ASR decoder, θ_{spk} the weights of the speaker identification model, which is a five TDNN sidekit x-vector speaker classifier trained with the Additive Angular Margin (AAM) loss (see Section 2.3.1), and α a trade-off parameter between ASR and speaker loss, empirically set to 1.0.

This type of learning scheme applied to ASR models has been studied many times, in (Adi

et al. 2019), the authors slightly improved the ASR performance, in (Srivastava et al. 2019), the authors showed a very small improvement in unlinkability privacy performance at the cost of a small ASR performance degradation.

Recently to improve adversarial training for feature disentanglement, a “semi-adversarial” training scheme was introduced by (Ryffel et al. 2019) and has shown great success when applied on a written digit dataset. The authors succeeded in reducing the performance of an unwanted inference, i.e., the classification of fonts from a bottleneck, from a digit classifier. As described by the “semi-adversarial” scheme, training is performed in 3 steps:

1. optimizes the ASR and speaker identification models independently until both converge.
2. optimizes with the adversarial objective described above, inducing the ASR-BN extractor to discard speaker information.
3. optimizes both ASR decoder and speaker identification models while freezing the ASR-BN extractor, such that both ASR and speaker models have the time to accommodate the adversarial training phase.
4. perform step 2 and 3 until \mathcal{L}_{aam} does not evolve anymore and the \mathcal{L}_{mmi} is minimized.

Our work was the first one to apply adversarial training to a traditional acoustic model instead of end-to-end (feature to text) ASR model, to apply the “semi-adversarial” training scheme for an ASR-BN extractor, to synthesize anonymized speech given this ASR-BN extractor. Given the anonymized speech that this pipeline generates, we then followed with the same analysis of the privacy/utility as in the VPC. In the following sections, we present the experimental setup used.

5.3.3 Experimental setup

For evaluation, we compare the utility and privacy scores across many anonymization pipelines. First, the baseline pipeline of the VPC 2022 is compared to our implementation pipeline of the ASR-BN extractor without adversarial training. Then, we retrain another anonymization pipeline with the ASR-BN extractor trained with the “semi-adversarial” scheme and compare our privacy/utility results. In this section, we start by presenting the implementation details of the ASR-BN extractor and synthesis models and later present the specificity of the adversarial learning experiment.

For the ASR-BN extractor, the experiments use the same PyTorch TDNN-F-based ASR-BN model as described in the “ASR-BN isolated privacy evaluation” in Section 5.3.1. This is because the ASR-BN extractor needs to be trained with multiple losses and for that purpose, it is easier to work with PyTorch than Kaldi. Similarly, as in Section 5.3.1, we only trained ASR-BN extractors on *LibriSpeech train-clean-100*, which is six times less data than what was used for training in the VPC. Due to this aspect, we expect our model to have somewhat lower performance in utility compared to the VPC baseline.

The Python code of the experiments is available at <https://github.com/deep-privacy/SA-toolkit>

We also reworked the speech synthesis system inspired by the HiFi-GAN-based voice conversion model presented in (Polyak et al. 2021). This model directly generates the speech signal without having to first output a Mel spectrogram (see Section 2.4.3 for more details). The training of the speech synthesis is done with the same dataset as the one defined in the VPC evaluation plan, hence *LibriTTS train-clean-100*. The conclusion drawn from Chapter 4, about the fact that the “constant speaker” target selection strategy allows better privacy evaluation, motivated us to change the speaker representation for the speech synthesis system to use a one-hot speaker embedding. The use of one-hot embedding simplifies the anonymization pipeline, and as discussed in the conclusion of Chapter 4, one-hot-based *any-to-many* voice conversion is usually easier to train than x-vector-based *any-to-any* voice conversion. As for *any-to-one* voice conversion, this framework is usually less robust than *any-to-many* as less training data is used (the training data only comes from the *one* target speaker). During the anonymization, the same target speaker is always used. The target speaker selected is ID 6081, as we found through empirical analysis that the converted speech exhibited favorable utility and naturalness characteristics. In our HiFi-GAN model, the F_0 is mean and variance normalized, as such, it is the one-hot speaker representation that conditions the target F_0 frequencies, the model performs the F_0 linear shift transformation presented in Section 5.2.1 on its own. If wanted, the other F_0 transformations presented in 5.2 can be done on the normalized F_0 . Figure 5.3 present our modified anonymization pipeline.

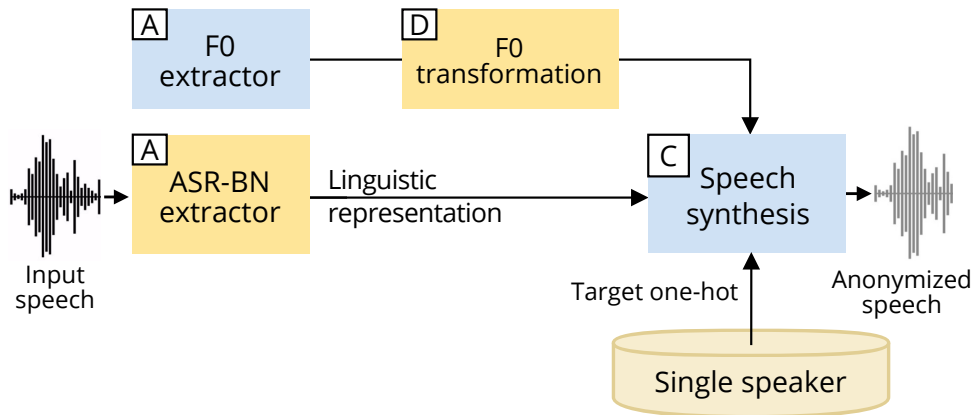


Figure 5.3 – Our speaker anonymization pipeline with module D added and module A adjusted to generate disentangled representation. A one-hot speaker embedding is used for speech synthesis instead of x-vector.

For evaluation, the VPC toolkit is used under the white-box informed ASV attacker, and privacy results are presented in terms of EER and $D_{\text{sys}}^{\text{sys}}$. As for utility, the ASR model is trained on anonymized speech and outputs WER scores. The datasets used for the training of the ASV and ASR evaluation models are the same as in the VPC (*LibriSpeech train-clean-360*). The results are presented for the *LibriSpeech test-clean* and *VCKT test* datasets.

5.3.3.1 Adversarial training specific experimental setup

Our speaker model is a five-layers TDNN sidekit x-vector model trained with an extension of the cross entropy cost function that encourages the network to output large angular margin between speakers (Liu et al. 2019). In our experiment, this cost function showed superior performance than the cross entropy. We also identified that the TDNN architecture was better suited than more recent ResNet architectures, this is probably due to the difference in how information is structured in bottleneck features in contrast to MFCCs or filterbank features.

Additionally, during our experiment, we found out that properly training the ASV speaker adversarial branch was challenging. First, we identified that a larger number of speakers in the training dataset was necessary, as such, we trained the x-vector model on a combination of *LibriSpeech train-clean-100* and *LibriSpeech train-other-500*, for a total of 1417 speakers (whereas the ASR model is trained only on *LibriSpeech train-clean-100*, for a total of 251 speakers). Second, we also identified that the way the mini-batches are created for ASR training does not allow proper ASV training. As such, for steps 1 and 3 of the “semi-adversarial” training scheme, the ASV model is trained using large mini-batches having three seconds of audio randomly sampled per utterance, and an equal distribution of speakers in a mini-batch. The ASR model is trained traditionally. Last, during step 2, the network optimizers (Adam here (Kingma et al. 2015)) conditioning the backpropagation are separated for the ASV and ASR models to avoid the ASV completely collapsing during the application of the gradient reversal function. The mini-batches are designed for ASR training, strongly impacting the adaptation of the ASV system to the ASR-BN extractor modification during this stage. This is supposed to be overcompensated in the “semi-adversarial” training scheme by the multiple repetitions of stages 2 and 3.

5.3.4 Experimental results

Table 5.4 presents the privacy and utility performances of the adversarial system compared to its non-adversarial counterpart and the VPC 2022 baseline on *LibriSpeech test-clean* and *VCTK test* datasets. The first line shows results on clear speech, where speaker verification can be addressed with very high accuracy.

The second line, for the VPC 2022 baseline system provides a strong baseline, keeping the spoken content easily recognizable (absolute degradation of less than 1% WER, compared to clear speech on both datasets) while significantly increasing privacy protection. On the *LibriSpeech* dataset, privacy was increased, as the $D_{\leftrightarrow}^{\text{sys}}$ metric lowered from 0.93 down to 0.67. On the *VCTK* dataset, privacy was even more improved as the $D_{\leftrightarrow}^{\text{sys}}$ dropped from 0.93 to 0.49. This overall trend of seeing the *VCTK* dataset more unlinkable than the *LibriSpeech* one can be explained by the dataset’s nature and the data used to train the ASV attack model. *LibriSpeech* does not offer much variability within a single speaker due to the long recording sessions of audiobook chapters. In addition, book reading speech differs from spontaneous speech, which impacts speech rate and

overall intonation. Those biases are captured by ASV systems (Ajili et al. 2018; Raj et al. 2019). Additionally, the white-box attacker is trained on *LibriSpeech train-clean-360*, which is of the same nature as *LibriSpeech test-clean* meaning that if any of the previous biases were to be captured by the ASV model, they are applied to the *LibriSpeech test-clean* dataset and the *VCTK* one. A similar conclusion about the ASR performances can be drawn, the WER for *LibriSpeech test-clean* is always lower than the one on *VCTK*. In the following, we primarily focus on the *VCTK* results because the *VCTK* dataset is less sensible to audiobook ASV evaluation biases.

Table 5.4 – Privacy and utility results for the adversarially trained of the ASR-BN extractor[§].

| Dataset | <i>LibriSpeech test-clean</i> | | | <i>VCTK test</i> | | |
|----------------------|--|---------------------------|-----------------------------|--|---------------------------|-----------------------------|
| | Privacy $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | Utility EER \uparrow | Utility WER \downarrow | Privacy $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | Utility EER \uparrow | Utility WER \downarrow |
| 1. Clear speech | 0.93 | 4.1 | 4.1 | 0.93 | 2.7 | 12.8 |
| 2. VPC 2022 baseline | 0.67 | 13.5 | 5.1 | 0.49 | 20.6 | 13.0 |
| 3. TDNNF | 0.81 | 8.7 | 6.9 | 0.73 | 10.8 | 19.1 |
| 4. ADVERSARIAL TDNNF | 0.83 | 7.8 | 5.3 | 0.70 | 11.8 | 14.4 |

Experiment with our baseline ASR-BN extractor (denoted as “TDNNF”, line 3) shows a very high degradation of utility compared to clear speech. Compared to the VPC baseline, the WER increases by a large margin on *LibriSpeech* and *VCTK* datasets. This is because the ASR-BN model was not trained with the non-clean speech of *LibriSpeech train-other-500*. Interestingly, with the degradation of utility, privacy is not significantly improved.

On *VCTK*, the $D_{\leftrightarrow}^{\text{sys}}$ dropped from 0.93 for clear speech to 0.73 for anonymized speech, a slight improvement but far smaller than the VPC baseline. This disparity can be explained as we extracted the ASR-BN from the 13th layer while the VPC baseline extracted it from the 17th layer. As shown in the “ASR-BN isolated privacy evaluation”, the ASR-BN contains a lot of speaker information which the speech synthesis model does not remove.

Before describing the result of the table, we discuss the intermediate indicators that we obtained during the training procedure of the adversarially trained ASR-BN. In the first step of adversarial training, the ASR-BN extractor and the adversarial speaker model are optimized independently (by freezing a branch or the another.). As a result, the ASR model (composed of the ASR-BN extractor) outputs a WER on *LibriSpeech test-clean* of 5.15%, while the speaker identification accuracy of the adversarial branch reaches 95.8% on the ASR-BN. Then, in the second stage, the ASR-BN extractor is trained to modify its weights with the negative loss of the adversarial speaker identification branch and the normal ASR decoder loss. The third stage retrains both ASR decoder and speaker identification models on frozen ASR-BN extractor. In

§. Audio samples can be extracted from the PDF by clicking (or double-clicking) tables rows index.

our experiment, we run the second and third stages twice, as after the second run, the speaker identification models did not converge anymore, indicating the ASR-BN extractor representation could not be used by the speaker identification model to identify the speaker. This could indicate that the ASR-BN representation does not encode speaker information. At the end of the training, the ASR decoder has a WER of 5.38% while the speaker recognition accuracy of the adversarial is 4.2%. The results of seeing the WER slightly increased is similar as in (Srivastava et al. 2019). However, as suggested by (Adi et al. 2019), using an adversarial branch trained on more data than the ASR decoder should provide a small WER improvement as it falls into the category of semi-supervised learning where most of the labels are about the speakers rather than the text. Overall, there is a lack of agreement in the literature for this form of training applied to this field.

Using the adversarially trained ASR-BN extractor, we trained the same HiFi-GAN speech synthesis model and obtain the anonymization pipeline denoted as “ADVERSARIAL TDNNF” in Table 5.4. Presented in line 4 of the table, we observe that the use of adversarial learning to remove speaker information from the ASR-BN representation is not effective to make the speech less linkable as the $D_{\leftrightarrow}^{\text{sys}}$ on *LibriSpeech test-clean* and *VCTK test* are the same with and without adversarial learning. This result is the same as in (Srivastava et al. 2019), even though we used additional training data, the more advanced “semi-adversarial” training scheme, a more known x-vector architecture for the adversarial model, and different optimizer and mini-batches creation to obtain the best adversarial model as possible. One reason for this discrepancy could be inherent to adversarial training. Before being able to properly remove information from the bottleneck, the adversarial branch must be able to characterize the information to remove. And the model-based approach leaves room for the ASR-BN encoder to find a type of representation that is still relevant for its primary task while adding some sort of noise that makes the particular adversarial architecture fail, similarly as in model adversarial attack (Guo et al. 2019). One solution to fix this could be to use more than one adversarial network/architecture which would each capture the speaker information differently and maybe provide a better negative gradient.

While the privacy performance results are not satisfactory, the utility results are interesting. In the table, we observe a utility improvement, the WER dropped on both datasets from 6.9% to 5.3% for the *LibriSpeech test-clean* and from 19.1% to 14.4% for the *VCTK test* datasets when adversarial training is applied. This conclusion is unexpected as decoding the speech directly from the adversarially trained ASR-BN representation did not yield a WER improvement (actually the WER increases from 5.15% to 5.38%), however, when using the ASR-BN representation to synthesize speech, the adversarially trained ASR-BN is superior. We hypothesize that adversarial training allows the ASR-BN representation to be structured in a better-formatted/easy-to-process way that allows for the speech synthesis system to generate higher utility quality speech.

5.3.5 Discussion

We conducted an experiment using adversarial training to reduce speaker information encoding from the ASR-BN extractor. The reverse gradient of a speaker identification model was applied to the ASR-BN extractor during training. The results showed that this improved the utility of the ASR-BN for speech synthesis, but not for privacy protection. To further enhance privacy, possible extensions of this work include using a triplet loss in the adversarial model to better match unlinkability attacks and using multiple adversarial models to reduce the potential ASR-BN specialization to one adversary.

5.4 Linguistic feature transformation

In this section, we evaluate other kinds of ASR-BN disentanglement techniques than adversarial training. The kind of transformation that we will be studying does not primarily rely on backpropagation to induce the network to discard speaker information from the ASR-BN. Instead, we will focus on transformations applied directly on the ASR-BN features to remove speaker information, hence the name of this section. However, it is worth noting that for the ASR-BN to properly capture the linguistic content information the transformations need to be applied during training to ensure the bottleneck is compatible with a given transformation.

In the following, we will start by briefly explaining an existing method based on noise perturbation used to transform the ASR-BN feature and then propose our alternative method based on vector quantization.

5.4.1 Laplace noise transformation

The Laplace noise is a type of noise that can be added to a signal or a dataset to protect the privacy of the individuals associated with that data. It is a type of Differential Privacy (DP) (Dwork 2006) technique that is based on the Laplace distribution (Dwork et al. 2014).

The idea behind using Laplace noise for anonymization is to add a small amount of random noise to the data in order to make it difficult for an attacker to infer the true values of the data. Usually, Laplace noise is added to data in a controlled manner, with the amount of noise determined by a parameter known as the privacy budget. The privacy budget is a measure of the amount of noise that can be added to the data without compromising too much its utility for the intended application.

The Laplace noise to add to a sample of data is defined by: $\text{noise} = \text{Lap}(0, \frac{1}{\epsilon})$, where Lap is the Laplace distribution with mean 0 and scale parameter $\frac{1}{\epsilon}$, ϵ is the privacy budget. The value of ϵ is typically chosen based on the desired level of privacy protection, with smaller values resulting in more noise and greater privacy protection.

The application of Laplace noise transformation for speaker anonymization was first proposed

by « [Differentially private speaker anonymization](#) », (Shamsabadi et al. 2022), where the authors suggested adding noise to the bottleneck of two models. The first is the bottleneck of a F_0 auto-encoder model that transforms the F_0 while still generating plausible trajectories. The second is the bottleneck of the VPC TDNN-F-based ASR-BN extractor. Note that in both cases, they normalize the bottleneck representation before and after applying the noise, as they found this improves training convergence.

5.4.2 Vector quantization transformation

Instead of adding noise to the ASR-BN representation, we propose constraining the layer that generates the ASR-BN by using Vector Quantization (VQ).

Vector quantization approximates a continuous vector by another vector of the same dimension, but the latter belongs to a finite set of vectors, called prototype vectors, and is contained in a dictionary. In the self-supervised framework of linguistic representation for voice conversion (see Section 2.4.1 and 3.6.2.1), it has been observed that the prototype vectors learned from vector quantization primarily capture information related to the phonemes and discard some speaker information (Oord et al. 2017; Chorowski et al. 2019; Wu et al. 2020). Similarly, our objective of applying vector quantization in the supervised linguist representation framework is to minimize the encoding of speaker information in ASR-BN.

5.4.2.1 VQ objective

Given the input audio sequence $s = (s_1, s_2, \dots, s_T)$ of length T , the first TDNN-F layers produces a continuous vector $h(s) = (h_1, h_2, \dots, h_J)$ of length J ($J < T$ due to the subsampling performed by the network) where $h_j \in \mathbb{R}^D$ for each time step t , and D is the size of the bottleneck representation ($D = 256$ here). VQ takes as input the sequence of continuous vectors $h(s)$ and replaces each $h_j \in h(s)$ by a prototype of the dictionary $E = \{e_1, e_2, \dots, e_V\}$ of size S , each $e_i \in \mathbb{R}^D$. VQ transforms $h(s)$ to $q(s) = (q_1, q_2, \dots, q_J)$ with:

$$\forall j \in \{1, 2, \dots, J\}, q_j = \arg \min_{e_i} \|h_j - e_i\|_2^2 \quad (5.5)$$

The vector h_j is replaced by its closest prototype vector e_v in terms of Euclidean distance. Since the quantization is non-differentiable (because of the $\arg \min$ operation), its derivative must be approximated. To do this, we use a *straight-through estimator* (Bengio et al. 2013) i.e., $\frac{\partial \mathcal{L}}{\partial h(s)} \approx \frac{\partial \mathcal{L}}{\partial q(s)}$. The prototype vectors are learned to approximate the continuous vectors which they replace by adding an auxiliary cost function:

$$\mathcal{L}_{vq} = \sum_{j=1}^J \|\text{sg}[h_j] - q_j\|_2^2 \quad (5.6)$$

where $\text{sg}[\cdot]$ denotes the stop gradient operation, blocking the update of the weights of the TDNN-F layers for this cost function (only updates the dictionary E). Minimizing \mathcal{L}_{vq} is a

similar operation to a k-means, but applied for each mini-batch during learning, the prototypes correspond to the centroids of a k-means.

Since the volume of the continuous vector space $h(s)$ is boundless, it can grow arbitrarily if the dictionary E does not train as fast as the TDNN-F. Adding a cost function that regularizes the TDNN-F to produce continuous vector $h(s)$ close to the prototypes of E is necessary so that learning does not diverge:

$$\mathcal{L}_{vq_reg} = \sum_{j=1}^J \|h_j - \text{sg}[q_j]\|_2^2 \quad (5.7)$$

5.4.2.2 ASR-BN objective

The cost function of the acoustic model can then be expressed as the sum of the MMI, quantization and regularization functions:

$$\mathcal{L} = \mathcal{L}_{mmi} + \mathcal{L}_{vq} + \beta \mathcal{L}_{vq_reg} \quad (5.8)$$

where β denotes the coefficient of the regularization factor (we used $\beta = 0.25$). In practice, we used the learning rule based on the exponential moving average (EMA) (Kaiser et al. 2018) to update the prototypes. EMA updates the dictionary E independently of the optimizer, so learning is more robust to different optimizers and hyperparameters (e.g., learning rate, momentum).

5.4.3 Wrapping up ASR-BN feature transformation

Figure 5.4 proposes an intuitive comparison example where the Laplace noise and vector quantization transformations are applied to a two-dimensional space. If we consider that the gray dots encode both speaker identity and pronunciation of a particular sound, the goal of ASR-BN transformations is to modify the speaker information from this representation to output a representation that encodes the same sound but with a different speaker identity this transformation is represented by arrows on Figure 5.4. For the VQ example, the size S of the dictionary equals the number of possible (reduced in the example) VQ prototypes. Overall, the lower S is, the more the prototypes should encode the linguistic content and not the speaker. As VQ aims to select the centroid of a region, some input vectors will be more modified than others. The output vector corresponds to the most common way to pronounce a sound. In contrast, for the noise addition, all input vectors are modified with the same amount of noise. Overall the higher the noise (lower ϵ for the Laplace noise sampling) the higher the transformation. The output vector corresponds to how a different speaker would pronounce the sound. The utilization of noise transformations can be associated with a “random speaker” target speaker selection strategy, whereas the VQ can be associated with a “constant speaker” strategy (as outlined in Chapter 4). While not displayed in the figure, the shapes of the decision boundaries are affected as the models are trained with the modifications. It is expected that for the VQ method, clusters of similar abstract meaning (which depends on S) will get closer to each other. Whereas for the

noise addition method, the decision boundary will be extended to cover a large region space to accommodate for the added noise.

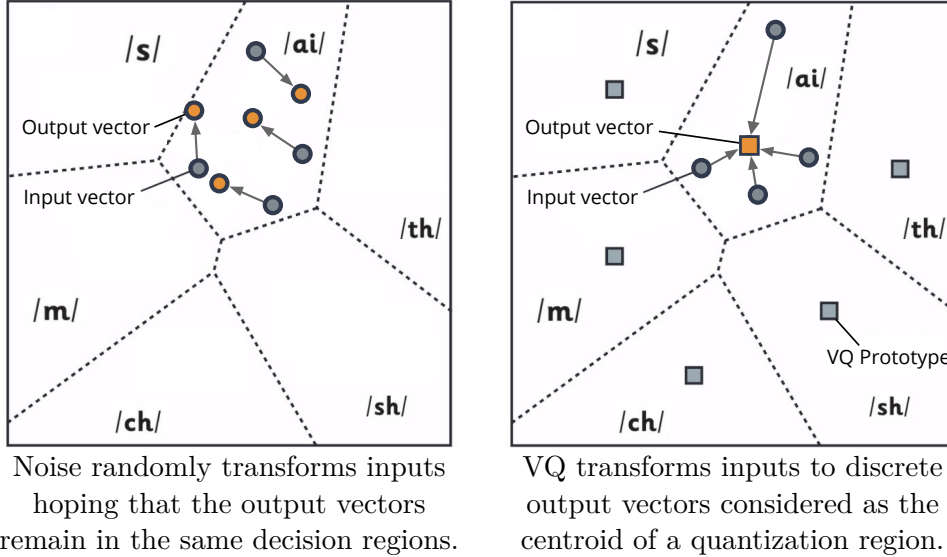


Figure 5.4 – Voronoi diagrams of the noise and VQ based ASR-BN transformations. In the figures, the annotation per region displays the phoneme obtained when decoding the ASR-BN.

5.4.4 Experimental setup

The VQ and noise transformations are applied on the PyTorch ASR-BN extractor and the HiFi-GAN speech synthesis implementation presented in the above Section 5.3.3 is used. In the experiment, we provide different sizes of dictionaries for the VQ transformation and different levels of noise for the Laplace transformation. Our ASR-BN extractors are trained only on *LibriSpeech train-clean-100* to reduce the computation time. While the HiFi-GAN is trained with *LibriTTS train-clean-100*.

For evaluation, the VPC toolkit is used under the white-box informed ASV attacker, and privacy results are presented in terms of EER and $D_{\leftrightarrow}^{\text{sys}}$. As for utility, the ASR model is trained on anonymized speech and outputs WER scores. The datasets used for the training of the ASV and ASR evaluation models are the same as in the VPC (*LibriSpeech train-clean-360*). The results are presented for the *LibriSpeech test-clean* and *VCKT test* datasets.

In the first experiment, we investigated the relationship between the size of the VQ dictionary and the privacy/utility performances obtained by a pipeline. The results are compared to our baseline pipeline already used in Section 5.3.4 which does not use a vector quantization for the ASR-BN. Then, to compare the results with the VPC 2022 baseline which trained the ASR-BN extractor with *LibriSpeech train-clean-100* and *LibriSpeech train-other-500*, we retrain, with the same data, one ASR-BN extractor constrained with a dictionary of 64 prototypes. The speech synthesis model is also retrained.

In the second experiment, we investigated replacing the filterbank coefficients used as input features for the ASR-BN with Wav2Vec-2.0 representation (see Section 2.2.3 for more detail on this architecture). The model topology of the ASR-BN is adjusted, according to (Vyas et al. 2021), we reduced the number of the acoustic model to 9 TDNN-F layers. The ASR-BN is extracted from the 3rd layer, right before the TDNN-F downsampling layer as Wav2Vec-2.0 already downsampled the signal. The ASR-BN extractor is trained with *LibriSpeech train-clean-100* and we fine-tune the Wav2Vec-2.0 model with a learning rate that is 20 times lower than the learning rate of the TDNN-F layers. We used a large Wav2Vec-2.0 model pre-trained on 24.1K hours of unlabeled multilingual west Germanic speech from VoxPopuli (Wang et al. 2021). There is no data overlap between VoxPopuli and the data used by the VoicePrivacy evaluation plan. We also experimented with the use of the F_0 transformations described in Section 5.2, and the one introduced by (Shamsabadi et al. 2022) authors where Laplace noise is added to the bottleneck of an auto-encoder to produce distorted but yet realistic F_0 trajectories¹.

Finally, in the last experiment, we adapted the work of (Shamsabadi et al. 2022), to transform with Laplace noise addition the ASR-BN. In this experiment, we used the same ASR-BN model as previously with the Wav2Vec-2.0 representation. We also experimented with the use of the F_0 transformations.

5.4.5 Experimental results

In the following sections, we present the results of the experiments conducted with the vector quantization approach and outline its effectiveness by comparing it to other methods.

5.4.5.1 ASR-BN vector quantization

Table 5.5 – Privacy and utility results for Vector Quantization-based anonymization. VQ 128 indicates the ASR-BN extractor was constrained with a dictionary of 128 prototypes.

| Dataset | <i>LibriSpeech test-clean</i> | | | <i>VCTK test</i> | | |
|----------------------|--|---------------------------|-----------------------------|--|---------------------------|-----------------------------|
| | Privacy $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | Utility EER \uparrow | Utility WER \downarrow | Privacy $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | Utility EER \uparrow | Utility WER \downarrow |
| 1. Clear speech | 0.93 | 4.1 | 4.1 | 0.93 | 2.7 | 12.8 |
| 2. VPC 2022 baseline | 0.67 | 13.5 | 5.1 | 0.49 | 20.6 | 13.0 |
| 3. TDNNF (NO VQ) | 0.81 | 8.7 | 6.9 | 0.73 | 10.8 | 19.1 |
| 5. TDNNF VQ 256 | 0.62 | 16.2 | 9.9 | 0.46 | 22.9 | 24.1 |
| 6. TDNNF VQ 128 | 0.59 | 17.7 | 10.4 | 0.42 | 24.0 | 26.3 |
| 7. TDNNF VQ 64 | 0.50 | 21.1 | 12.4 | 0.29 | 30.0 | 29.1 |

Table 5.5 shows the privacy and utility performances of the vector quantization system compared to its non VQ counterpart and the VPC 2022 baseline on *LibriSpeech test-clean*

1. We thank the authors for providing their implementation.

§. Audio samples can be extracted from the PDF by clicking (or double-clicking) tables rows index.

and *VCTK test* datasets. By constraining the TDNN-F ASR-BN extractor with the use of vector quantization, the performance of linkability attacks are drastically reduced the as $D_{\leftrightarrow}^{\text{sys}}$ performances for lines 5 to 7 are lower than for line 3 on all datasets. The number S of prototypes in the quantization dictionary constrains the acoustic model. With S prototype vectors, the spoken information of the speech is compressed into a discrete dictionary space of size S . The smaller the dictionary, the more the network must find an efficient transformation to represent the spoken content information, leaving less room to encode the speaker’s information. We tried three dictionary sizes in our experiment: 256, 128, and 64. The most anonymized speech was generated with $S=64$, where the $D_{\leftrightarrow}^{\text{sys}}$ dropped from 0.73 (NO VQ) to 0.29 (VQ 64) on the *VCTK* dataset. However, this privacy improvement comes at a very high utility cost; the WER raises from 19.1% to 29.1%. The other dictionary sizes illustrate well the privacy utility trade-off (Li et al. 2009) this model suffers when lower privacy implies better utility and vice-versa. We hypothesize that the privacy improvement comes from the vector quantization layer, while the major utility loss comes from the small number of layers before the quantization layer and/or the amount of data used to train the model. Constraining the network to such a few discrete vectors could be possible without significant utility loss if the network has the capabilities to transform the speech signal into a compressed high-level linguistic representation.

Table 5.6 presents the experiment where more data is used to train the ASR-BN extractor constrained with 64 prototypes. As a result, it uses the same amount of data as in the VPC evaluation. The outcome is displayed in line 7b of the table and shows that by using more data to train the ASR-BN extractor, higher utility is achieved, on *VCTK test* the WER is reduced from 29.1% to 16.7%; similarly, the WER on *LibriSpeech test-clean* almost divided by two. This improvement in utility does not affect privacy performances, as the $D_{\leftrightarrow}^{\text{sys}}$ on *VCTK test* and *LibriSpeech test-clean* are not significantly different than when the network was trained on six times fewer data. When compared with the VPC 2022 baseline also trained with the same data, our model with VQ improves privacy performances as the $D_{\leftrightarrow}^{\text{sys}}$ is reduced from 0.49 to 0.25 on *VCTK test* and 0.67 to 0.52 on *LibriSpeech test-clean*. However, this privacy improvement comes at a utility cost, as the WER is increased from 13.0% to 16.7% on *VCTK test* and 5.1% to 6.7% on *LibriSpeech test-clean*, this small degradation surely comes from VQ.

Table 5.6 – Privacy and utility results for the VQ 64 configurations trained with 100h (*LibriSpeech train-clean-100*) and 600h (*LibriSpeech train-clean-100* + *LibriSpeech train-other-500*) of data.

| Dataset | | <i>LibriSpeech test-clean</i> | | | <i>VCTK test</i> | | |
|--------------------------------|------------------|---|----------------|------------------|---|----------------|------------------|
| Method: ASR-BN transformations | | Privacy | | Utility | Privacy | | Utility |
| | | $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | EER \uparrow | WER \downarrow | $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | EER \uparrow | WER \downarrow |
| 2. VPC 2022 baseline | <i>train-600</i> | 0.67 | 13.5 | 5.1 | 0.49 | 20.6 | 13.0 |
| 7. TDNNF VQ 64 | <i>train-100</i> | 0.50 | 21.1 | 12.4 | 0.29 | 30.0 | 29.1 |
| 7b. TDNNF VQ 64 | <i>train-600</i> | 0.52 | 20.0 | 6.7 | 0.25 | 32.5 | 16.7 |

5.4.5.2 ASR-BN Wav2Vec-2.0 vector quantization + F_0 transformations

The experiment presented in Table 5.7 evaluates the hypothesis of increasing the network depth by using a large Wav2Vec-2.0 model to replace filterbank. Without vector quantization, our Wav2Vec-2.0 TDNN-F ASR-BN extractor does not significantly improve the privacy protection (line 8); the $D_{\leftrightarrow}^{\text{sys}}$ on the *VCTK* dataset reaches 0.69, far away of the 0.49 score of the VPC baseline. Interestingly, the utility improves compared to the clear speech, the WER drops from 4.1% to 3.8% in the *LibriSpeech* dataset, while in the *VCTK* dataset, it drops from 12.8% to 7.8%. Improvement of utility is achieved because of the Wav2Vec-2.0 module; the ASR-BN is more precise because of the network depth and amount of training data that the Wav2Vec-2.0 was trained on. Applying voice conversion on precise ASR-BN features enhances the speech signal allowing the ASR system to better recognize the linguistic content.

Applying a high vector quantization constraint on this Wav2Vec-2.0 TDNN-F model shows the approach’s potential. With a very small dictionary size of 48 prototypes (line 9), privacy is improved in comparison to the VPC baseline; the $D_{\leftrightarrow}^{\text{sys}}$ on the *VCTK* dataset reaches 0.34 while also having a good utility score of 10.0 WER.

It is important to note that for utility evaluation, the comparison between the WAV2VEC2 TDNNF VQ 48 pipeline (line 9) and the VPC 2022 baseline (line 2) is not fair as the ASR-BN extractor are not trained on the same amount of data. For the WAV2VEC2 TDNNF VQ 48 approach, the Wav2Vec-2.0 model has been pre-trained on 24.1K hours of unlabeled data, and then the WAV2VEC2 TDNNF VQ 48 extractor has been finetuned on 100h of labeled data. Whereas the VPC 2022 baseline is trained on 600h of labeled data. A much more appropriate comparison is between the quantized and non-quantized WAV2VEC2 TDNNF models, which indicates that the utility performance decreased from 7.8% to 10.0% of WER on *VCTK* data, the same can be observed on *LibriSpeech* data. The same observation can be made about the comparison of the utility performance of any anonymization system to the clear speech reference. For clear speech, the ASR evaluation model is only trained on *LibriSpeech train-clean-360* to generate the reference WER score. Whereas for anonymized speech, the speech is first transformed using models trained on much more than *LibriSpeech train-clean-360*, then evaluated with the ASR evaluation model. If the anonymization relies on ASR methods, and it does, the anonymization pipeline can be considered as a ASR enhancing preprocessor. As such, conclusions about a utility improvement from clear speech to anonymized speech should be avoided as they are not evaluated on an even playing field.

The *LibriSpeech* EER of 17.5% in line 9 is close to the one obtained in the F_0 “ F_0 isolated privacy evaluation” of Section 5.2.7 (EER of 19.9%). This indicates that the F_0 trajectory might be the limiting factor of the anonymization. As such for lines 10 to 12 we experimented with F_0 transformations. Line 10 presents the Additive White Gaussian noise (AWGN) noised F_0 transformation (see Section 5.2.2) with a target noise power of 15dB. Line 11 presents the Laplace

noised modified bottleneck auto-encoder transformation presented in (see Section 5.4.1) with an ϵ value of 1 as it appeared to be the most effective in (Shamsabadi et al. 2022). And finally, line 12 presents the quantization-based F_0 transformation (see Section 5.2.3) with four quantization bits.

With the F_0 AWGN transform, the privacy protection increase, as the $D_{\leftrightarrow}^{\text{sys}}$ on the *VCTK* dataset plummeted down to 0.12 while keeping a very high utility with 10.3% of WER, similar behavior can be observed in the *LibriSpeech* dataset. The Laplace noised-based F_0 auto-encoder transformation has similar privacy performance on the *LibriSpeech* dataset as the AWGN. However, it is not the case for the *VCTK* dataset, where the privacy $D_{\leftrightarrow}^{\text{sys}}$ score equals 0.30 instead of being close to 0.12.

One hypothesis for this disparity is that adding noise to the bottleneck of a model to later obtain some features, is different from adding noise directly to the feature. We hypothesize that during training, the model compensates for noise addition with a form of denoising such that for a particular noise level, the model loss still converges. This means that it is complicated to properly define the noise level as the model convergence depends on this noise level but also on the training data. We argue that for Laplace bottleneck noise addition transformations, it is easy to go from an appropriate transformation for a dataset (*LibriSpeech* in line 11) to an inappropriate (too small) transformation for another dataset (*VCTK*). We refer to this as the sensitivity of a transformation to generalize to multiple datasets.

Finally, the F_0 quantization transformation appears to be a very compelling option compared to noised-based transformations. The privacy performances are similar to the AWGN F_0 transformation while being slightly better in terms of utility. Additionally, informal subjective listening tests report that the naturalness of the “WAV2VEC2 TDNNF VQ 48 + F_0 QUANT_{4bits}” pipeline, exceeded the “WAV2VEC2 TDNNF VQ 48 + F_0 AWGN_{15dB}” pipeline[§].

Table 5.7 – Privacy and utility results for Vector Quantization-based anonymization, using a Wav2Vec-2.0 feature extractor. Three kinds of F_0 transformations were also tested, note that the linear shift is included by default in our models.

| Dataset | <i>LibriSpeech test-clean</i> | | | <i>VCTK test</i> | | |
|---|--|---------------------------|-----------------------------|--|---------------------------|-----------------------------|
| | Privacy $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | Utility EER \uparrow | Utility WER \downarrow | Privacy $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | Utility EER \uparrow | Utility WER \downarrow |
| Method: ASR-BN + F_0 transformations | | | | | | |
| 1. Clear speech | 0.93 | 4.1 | 4.1 | 0.93 | 2.7 | 12.8 |
| 2. VPC 2022 baseline | 0.67 | 13.5 | 5.1 | 0.49 | 20.6 | 13.0 |
| 8. WAV2VEC2 TDNNF NO VQ | 0.83 | 7.7 | 3.8 | 0.69 | 12.1 | 7.8 |
| 9. WAV2VEC2 TDNNF VQ 48 | 0.57 | 17.5 | 4.5 | 0.34 | 28.0 | 10.0 |
| 10. WAV2VEC2 TDNNF VQ 48 + F_0 AWGN _{15dB} | 0.44 | 23.4 | 4.6 | 0.12 | 40.8 | 10.3 |
| 11. WAV2VEC2 TDNNF VQ 48 + F_0 LP _{ϵ1} | 0.46 | 22.5 | 4.6 | 0.30 | 30.2 | 9.9 |
| 12. WAV2VEC2 TDNNF VQ 48 + F_0 QUANT _{4bits} | 0.45 | 23.0 | 4.4 | 0.14 | 39.8 | 9.9 |

5.4.5.3 ASR-BN Wav2Vec-2.0 Laplace noise + F_0 transformations

In this experiment, we attempt to reproduce the ASR-BN bottleneck Laplace-noise transformation of (Shamsabadi et al. 2022) with our speaker anonymization pipeline presented in Figure 5.3. We use a Wav2Vec-2.0 feature extractor for a fair comparison with our best “WAV2VEC2 TDNNF VQ 48 + F_0 QUANT_{4bits}” pipeline referred to as “fully quantization-based pipeline”. Another implementation difference to the original paper relates to how the noise is applied to the bottleneck, to match the F_0 LP implementation, noise is sampled and applied for the whole utterance dimension, rather than every time for each frame. Additionally, in contrast to the evaluation performed in the original paper, we use a white-box informed attacker and analyze our results in both *LibriSpeech* and *VCTK* datasets.

Table 5.8 compares the privacy and utility results to our fully quantization-based pipeline. For all Laplace noise (LP) results, we can observe that for the *LibriSpeech* dataset, the pipeline has a lot of trouble preserving the utility. Whereas for the *VCTK* the strength of noise addition seemed to be appropriate as the utility is not as strongly affected. This observation supports our previous hypothesis about the sensitivity of noise addition bottleneck transformations to generalize to multiple datasets.

We can observe that adding more noise to the bottleneck decreases utility preservation as lower ϵ values tend to have higher WER. For an ϵ of 130000 (line 13), the WER is the same as our fully quantization-based pipeline with a value of 9.9%. Increasing the noise with an ϵ of 100000 (line 16) decreases the utility with a WER value of 12.9% in *VCTK*. We notice that compared to the VQ approach, F_0 transformations tend to decrease the utility more significantly. The F_0 LP ϵ 1 transformation results in a higher WER in *VCTK* lines 14 to 15 and 17 to 18, indicating that it performs worse than AWGN.

In terms of privacy results, we also focus on the *VCTK* results as the WER utility performance is similar to our fully quantization-based pipeline in *VCTK*. First, we notice that F_0 transformation is also necessary to generate the most unlinkable speech. The best of them in privacy (and utility) seems to be the AWGN transformations applied with a target noise power of 5dB. The pipeline the most comparable to our fully quantization-based pipeline is the “WAV2VEC2 TDNNF LP ϵ 130000 + F_0 AWGN_{5dB}” one (line 14) as the WER is the same as ours. We observe that this pipeline has a similar privacy performance with a $D_{\text{sys}}^{\text{sys}}$ of 0.12 (compared to 0.14). However, we argue that this fully noised-based pipeline might produce slightly lower naturalness than the fully quantization-based pipeline, and is highly more sensible to the input dataset. Increasing the noise with an ϵ of 100000 (line 17) when the F_0 is transformed does not necessarily increase the privacy, as the privacy results between lines 16 and 13 are the same for the *VCTK* dataset. However, it significantly decreases utility as the WER increases from 10.2% to 13.8%.

Regarding the results from *LibriSpeech*, it is important to note that the decrease in utility performance is not always proportionate to the level of privacy enhancement. The absolute

privacy improvement between lines 12 and 13 is 0.06 of $D_{\leftrightarrow}^{\text{sys}}$, while the absolute utility degradation is by 16.6% of WER.

Table 5.8 – Privacy and utility results for Laplace noise-based ASR-BN transformation and the two noised based F_0 transformations.

| Dataset | <i>LibriSpeech test-clean</i> | | | <i>VCTK test</i> | | |
|--|--|---------------------------|-----------------------------|--|---------------------------|-----------------------------|
| | Privacy $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | Utility EER \uparrow | Utility WER \downarrow | Privacy $D_{\leftrightarrow}^{\text{sys}} \downarrow$ | Utility EER \uparrow | Utility WER \downarrow |
| Method: ASR-BN + F_0 transformations | | | | | | |
| 1. Clear speech | 0.93 | 4.1 | 4.1 | 0.93 | 2.7 | 12.8 |
| 2. VPC 2022 baseline | 0.67 | 13.5 | 5.1 | 0.49 | 20.6 | 13.0 |
| 12. WAV2VEC2 TDNNF VQ 48 + F_0 QUANT _{4bits} | 0.45 | 23.0 | 4.4 | 0.14 | 39.8 | 9.9 |
| 13. WAV2VEC2 TDNNF LP _{e130000} | 0.39 | 25.7 | 21.0 | 0.25 | 33.4 | 9.9 |
| 14. WAV2VEC2 TDNNF LP _{e130000} + F_0 AWGN _{5dB} | 0.25 | 33.3 | 22.7 | 0.12 | 41.4 | 10.2 |
| 15. WAV2VEC2 TDNNF LP _{e130000} + F_0 LP _{e1} | 0.24 | 33.9 | 24.0 | 0.19 | 36.3 | 10.4 |
| 16. WAV2VEC2 TDNNF LP _{e100000} | 0.36 | 27.0 | 27.4 | 0.27 | 32.2 | 12.9 |
| 17. WAV2VEC2 TDNNF LP _{e100000} + F_0 AWGN _{5dB} | 0.16 | 34.5 | 28.9 | 0.12 | 41.5 | 13.8 |
| 18. WAV2VEC2 TDNNF LP _{e100000} + F_0 LP _{e1} | 0.23 | 34.8 | 31.3 | 0.23 | 34.4 | 14.2 |

5.4.6 Discussion

In this section dedicated to the ASR-BN feature, we evaluated the degree of disentanglement between the ASR-BN and the speaker and concluded that the ASR-BN can be used to perform speaker linkability attacker at a very high rate. Indeed the ASR-BN feature, commonly used by the anonymization pipeline, leaks speaker identities as the EER of the ASR-BN in our isolated experiment was 6.7%, very close to the MFCCs EER score of 3.7%. This implies that to better guarantee proper anonymized speech generation by the speech synthesis system, the ASR-BN should be transformed.

We presented common ASR-BN transformation based on bottleneck Laplace noise addition used by differential privacy, and adversarial training before proposing a novel vector quantization bottleneck transformation approach. We showed that the fully quantization-based pipeline, where the F_0 and ASR-BN are quantized achieves the best privacy/utility performance in the VPC testing datasets. In contrast to noise-based pipelines, the quantization-based one did not have the same sensibility issue as, good utility where recorded regardless of the dataset.

During our experiment we played with the Wav2Vec-2.0 feature extractor to replace the traditional filterbank ASR-BN input feature. Using the Wav2Vec-2.0 feature extractor to help process speech and output ASR-BN representation increases the number of layers of the ASR-BN extractor compared to previously. Additionally, the extractor has effectively seen more training samples as Wav2Vec-2.0 requires large unlabeled data to be trained. This new architecture

allows extracting more precise ASR-BN which is beneficial for WER utility, and with such a large network, it begins to be possible to constrain the extractor to only a few discrete quantized vectors because the network has the encoding capacity to transform the speech signal into a compressed high-level representation. In our experiment with the VPC testing dataset, 48 discrete prototypes seemed to be enough with the Wav2Vec-2.0 feature extractor (the fewer the number of discrete quantized vectors, the better the privacy). However, the dataset of the VPC uses clean speech, which is favorable for this approach, under noisier environments the quantization layer has much more trouble identifying the appropriate discrete vector, leading to large utility decreases. This could be compensated by increasing the number of discrete prototypes but at the cost of privacy degradation. The number of discrete prototypes can be adjusted to achieve a suitable privacy/utility trade-off (Li et al. 2009; Wu et al. 2021) depending on the application. Possible extensions of this work relate to how the discrete prototypes are selected, currently, an L2 distance between the continuous and discrete vector is used but a personalized per-speaker function could be used to increase utility. Additionally, for each frame, the number of possible discrete prototypes is currently fixed (e.g., 48), having this number dynamically increased depending on the level of uncertainty of the ASR-BN to correctly capture the acoustic unit would be an interesting research track.

5.5 Conclusion

In this chapter, we have challenged the F_0 and ASR-BN attributes of voice conversion-based speaker anonymization. We found with isolated evaluations, that they are not disentangled from speaker information limiting the performance of speaker anonymization. As such, we presented and proposed transformations to improve disentanglement. Our approach based on quantization has advantages compared to the more traditional one based on noise addition. Quantization-based disentanglement allows us to achieve similar privacy performance as noised-based transformations while having superior WER utility, additionally, we argue that the produced anonymized speech is also more natural which must be assessed through more perceptual tests.

One important observation made concerns the objective utility measurement. We discourage people to claim that an anonymization system improves objective utility even though the WER score might be lower. The main reason why the WER score can be lower compared to a clear speech reference comes from the disparity of training and the amount of data used to train the anonymization pipeline and the evaluation model.

NEW KINDS OF PRIVACY AND UTILITY MEASUREMENTS

Contents

| | | |
|------------|---|------------|
| 6.1 | Introduction | 117 |
| 6.2 | Invertibility evaluation using embedding alignment | 118 |
| 6.2.1 | Evaluation setup | 118 |
| 6.2.2 | Supervised and unsupervised embedding alignment | 119 |
| 6.2.3 | Experimental attack scenarios | 121 |
| 6.2.4 | Experimental results | 123 |
| 6.2.5 | Discussion | 125 |
| 6.3 | Subjective mispronunciation evaluation | 127 |
| 6.3.1 | Real world example | 127 |
| 6.3.2 | Mispronunciation metric proposition | 128 |
| 6.3.3 | Discussion | 129 |
| 6.4 | Conclusion | 130 |

6.1 Introduction

This chapter unveils novel measurement metrics and techniques, offering a fresh perspective on privacy and utility assessment beyond the traditional EER/ $D_{\text{err}}^{\text{sys}}$ and WER metrics. To begin with, we propose a new privacy measurement technique that aims to evaluate the invertibility of anonymization. As such, we will introduce simple inversion attacks which aim to reconstruct clear x-vectors from anonymized ones. If it is possible to invert the anonymization, even partially, it could help the attacker to better perform a linkability attack, but also it opens the room for many other undesirable use such as voice-cloning to impersonate a person’s voice. We finish this chapter by proposing a utility measurement that aims to better assess the degree of linguistic content preservation from a subjective point of view. This measurement aims to overcome the WER-based evaluation limitation, which cannot differentiate between words correctly pronounced and wrongly recognized by the ASR evaluation system and words that are simply not correctly pronounced.

6.2 Invertibility evaluation using embedding alignment

In Section 3.2, we presented the *non-invertibility* and *unlinkability* criteria defined by the European standard ISO/IEC 24745 (ISO-24745 2011) that apply to data anonymization. Up until now, the linkability criteria is considered the main threat that speaker anonymization should defend against. The use of the white-box ASV attack allowed us to well evaluate this aspect and output EER/ $D_{\leftrightarrow}^{\text{sys}}$ metrics used up until now.

However, when it comes to invertibility evaluation, to the best of our knowledge, no work has been done to evaluate this aspect. As such, the subject of this section is to propose an invertibility attacker which enables an invertibility evaluation. Conducted with Thebaud Thomas, we proposed to invert voice conversion-based anonymization using embedding alignment techniques. Later on, other invertibility methods were extended to signal processing speaker anonymization (Kai et al. 2022).

6.2.1 Evaluation setup

We propose to invert the VPC 2022 baseline and our fully quantization-based anonymization pipeline (with Wav2Vec-2.0 and a VQ dictionary size of 48, see Section 5.4.5.2), both of them conditioned with the “constant speaker” target selection strategy. We assume that the attacker has access to a compromised speech dataset which can be transformed with the same anonymization pipeline as the one used to anonymize the vulnerable dataset. The attacker also trains a white-box ASV system that is adapted to extract x-vectors from anonymized speech.

As such, it is realistic to assume that the attacker has a dataset of clear x-vectors for which he knows the corresponding anonymized x-vectors. Given the parallel clear/anonymized x-vectors, the attacker can approximate the anonymization function as a rotation from the clear x-vector to the anonymized one. In this study, the anonymization function is approximated using a rotation matrix, taking the anonymized x-vector as input and producing the corresponding estimated clear x-vector. This scenario of having clear/anonymized 1-to-1 mapping allows estimating the rotation matrix using a Procrustes Analysis (Gower 1975). We refer to this scenario as the supervised scenario.

In addition to this supervised scenario, we propose a more restrictive scenario where the attacker does not know the link between the clear x-vector and the anonymized x-vector in the compromised dataset. In this scenario, we use an unsupervised embedding alignment algorithm named Wasserstein Procrustes (Grave et al. 2018).

With the rotation matrix estimated, the anonymized x-vector is inverted to estimate the corresponding clear x-vector. Then, invertibility is evaluated by measuring how well the attacker can accurately (ACC) reconstruct the clear (non-anonymized) vulnerable x-vector. Finally, we can also evaluate linkability between clear compromised and inverted anonymized vulnerable x-vectors. Figure 6.1 summarizes this process.

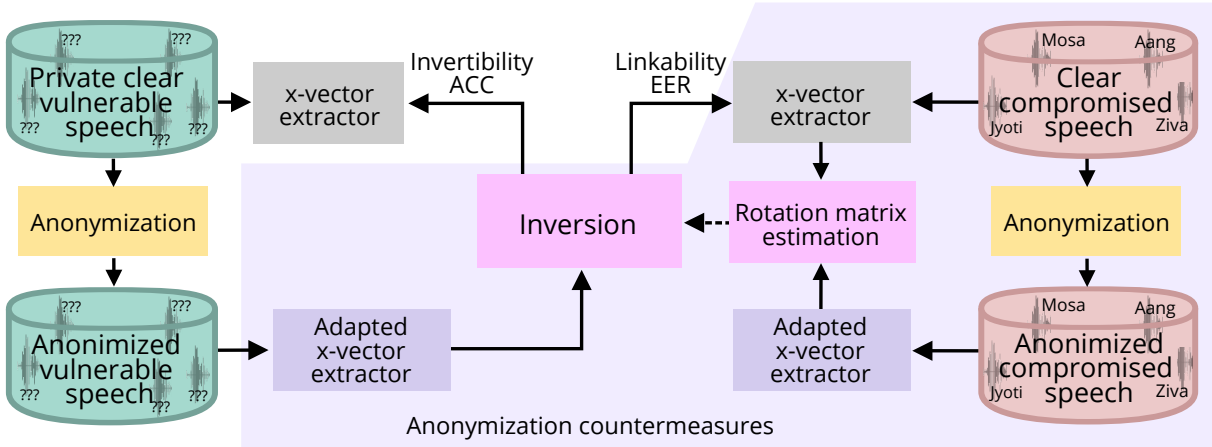


Figure 6.1 – Illustration of training the rotation matrix and inverting anonymized x-vectors. Invertibility and linkability measurements are then performed with clear and inverted x-vectors.

6.2.1.1 Metrics

We use two metrics to evaluate our attack on the different scenarios: the EER and the Top 1 speaker accuracy. For all experiments, the EER is computed by scoring the inverted x-vectors against the clear ones. In contrast with the scoring done in previous chapters, here we use the cosine similarity instead of the PLDA. For this reason, lower linkability performance is expected (higher EER) as the x-vector system that we are using is not optimized with angular/cosine-margin-based losses.

The Top 1 speaker accuracy is computed by comparing the inverted anonymized vulnerable utterances against their clear counterparts. For each inverted x-vector, we looked for the nearest neighbor speaker x-vector from the clear vulnerable x-vector dataset. The Euclidean distance was used to find the nearest neighbor because, during Procrustes analysis, the rotation matrix was estimated for that distance metric. The Top 1 speaker accuracy is the proportion of inverted x-vectors for which the closest clear vulnerable x-vector is from the same speaker. A high Top 1 speaker accuracy means a high success in inverting x-vectors close to their clear counterpart and should raise concerns regarding the *non-invertibility* property of the speaker anonymization system.

6.2.2 Supervised and unsupervised embedding alignment

Computing the alignment of two embeddings of high dimensional real vectors is one of the fundamental problems in machine learning, with applications examples such as unsupervised word and sentence translation (Rapp 1995; Fung 1995; Bojanowski et al. 2017; Grave et al. 2018; Biswas et al. 2020). In this section, we introduce supervised (Procrustes) and unsupervised (Wasserstein Procrustes) embedding alignment algorithms that we will rely on to invert the anonymization.

6.2.2.1 Procrustes Analysis

Let A and B be two sets of N high dimensional real vectors of dimension d . We want to find the optimal rotation $W \in \mathbb{R}^{d \times d}$ that minimizes the squared distance between both sets:

$$\min_{W \in \mathbb{R}^{d \times d}} \|AW - B\|_2^2 \quad (6.1)$$

For correctly parallel sets A and B (the n^{th} element of A corresponds to the n^{th} element B , $\forall n \in \llbracket 1, N \rrbracket$), we can directly use Procrustes analysis (Gower 1975) to compute an optimal W . The operation consists of finding in the space of orthogonal matrices \mathcal{O}_D the singular value decomposition: $U\Sigma V^T = A^T B$, then later obtain W with: $W = UV^T$. This approach is well suited for supervised scenarios as it requires access to the labels of both sets to align them. For two sets where the mapping is unknown, an unsupervised alignment algorithm is required.

6.2.2.2 Wasserstein Procrustes

In the study conducted by (Grave et al. 2018), an unsupervised algorithm was proposed to align sets of language-dependent word embeddings for the purpose of unsupervised translations. The proposed algorithm utilizes a stochastic optimization approach, alternating between minimizing the Wasserstein distance between sets and finding the optimal rotation using Procrustes analysis. This approach enables identifying a rotation matrix that optimally lowers the distance between the two sets of embeddings, while also determining their one-to-one mapping.

Wasserstein distance (Rüschendorf 1985) is a measure of the optimal transportation required to move one set of points to the positions of a second group. If the n^{th} element of A corresponds to the n^{th} element of B , $\forall n \in \llbracket 1, N \rrbracket$, then the distance between A and B is: $\|A - B\|_2^2$. However, if A and B are not ordered, it is necessary to match each point of one set with a point of the other. To do this, we use a permutation matrix. Let \mathcal{P}_N be the set of permutation matrices that ensure a 1 to 1 mapping: $\mathcal{P}_N = \left\{ P \in \{0, 1\}^{N \times N}, P1_N = 1_N, P^T 1_N = 1_N \right\}$. In our problem, we are looking for a permutation matrix $P \in \mathcal{P}_N$ that will minimize the Wasserstein distance between two groups of points:

$$P = \min_{P \in \mathcal{P}_N} \|A - PB\|_2^2 \quad (6.2)$$

Stochastic optimization In our problem, we aim to simultaneously solve equations 6.1 and 6.2 by making use of a stochastic approach for Procrustes analysis in Wasserstein distance:

$$\min_{W \in \mathcal{O}_D} \min_{P \in \mathcal{P}_N} \|AW - PB\|_2^2 \quad (6.3)$$

Described in (Grave et al. 2018), the method alternate between the selection of random sub-sets of A and B , the calculation of permutation matrix P and the update of W .

6.2.3 Experimental attack scenarios

In this section, we first explain the scenarios and then describe the dataset accessibility hypotheses, to end with the implementation details that improve the rotation.

6.2.3.1 Scenarios

In the following, we explain our motivation for experimenting with multiple attack scenarios. First, we will present supervised and unsupervised attack conditions using Procrustes and Wasserstein Procrustes trained on compromised x-vectors. Those two attacks are realistic, but, as they require estimating the rotation with a training dataset, the inversion attack will not be perfect.

To evaluate the privacy protection using a more powerful attack, we use a non-realistic *oracle* scenario, where the attacker has access to clear vulnerable speech to estimate the rotation. This rotation should be close to optimum as estimating the rotation between clear vulnerable speech and anonymized vulnerable speech should give the best rotation matrix to invert the anonymized vulnerable speech. This allows us to better challenge the *non-invertibility* privacy criteria, as it directly challenges the definition requiring that “it should be computationally infeasible to obtain the clear data that led to any given anonymized data”.

For all the following scenarios, we approximate the anonymization function in the x-vector domain. The rotation matrices are estimated to match each x-vector utterance of one dataset to another one. Once the rotation matrix $W \in \mathbb{R}^{d \times d}$ is estimated, the anonymized vulnerable x-vectors dataset is inverted using the transposed W :

Supervised scenario: Procrustes This first scenario follows the dataset requirement of the VPC. To obtain the rotation matrix, a Procrustes analysis is applied to clear and anonymized compromised x-vector datasets, knowing the one-to-one correspondence between them. Then the anonymized vulnerable x-vectors are inverted. The goal of this first experiment is to estimate how well a rotation can approximate the anonymization pipeline in the x-vector domain. During the evaluation we will evaluate:

- How linkable the inverted x-vectors are to clear x-vectors? This will be compared with previous linkability evaluations under similar conditions.
- How many anonymized vulnerable x-vectors can be inverted well enough to recognize their clear source speaker?

Unsupervised scenario: Wasserstein-Procrustes This second experiment explores the performance of an unsupervised algorithm for the invertibility attack. In contrast to the previous one, the clear and anonymized compromised x-vector mapping is unknown. This evaluation is the first step in knowing if unsupervised algorithms can work at all to create a clear/anonymized

inversion attack. This work could be extended into trying to match nonparallel datasets such as the anonymized vulnerable x-vector dataset and the clear or anonymized compromised x-vector datasets.

Oracle scenarios This third and last experiment probes the optimal performances an attacker can get while approximating a speech anonymization system with rotations. In contrast to the two previous scenarios, here the datasets used to estimate the rotations are also the test datasets (the clear and anonymized vulnerable x-vector datasets). The two Procrustes and Wasserstein Procrustes algorithms presented above will be used here. In those scenarios, we are more likely to evaluate the strength of the privacy protection rather than the attack/estimation of the rotation. Here what we query is the degree of the bijection between clear and anonymized x-vectors and if the transformation between clear and anonymized space can be associated with a rotation. If a rotation exists, it indicates an anonymization weakness, as it is possible to identify speakers in the anonymized x-vector space, and a bijective reciprocal function exists to associate each anonymized x-vector to a unique clear x-vector.

As the goal of the “constant speaker” target selection strategy is to generate a unique speaker, the oracle scenarios goal is to control the successfulness of the voice conversion anonymization system to generate a single and unique identity. No bijection between one x-vector space to the other should exist once speakers are anonymized.

6.2.3.2 Dataset accessibility hypotheses

The datasets accessibility hypotheses are summarized in Figure 6.2 for the different scenarios detailed above. Purple boxes show data available to the attacker to train the rotation matrix in a given hypothesis. Black hatched boxes show data inaccessible to the attacker in a given hypothesis. We call supervised the scenarios where labels are available (represented as \leftrightarrow) to align the datasets and unsupervised the ones where the mapping is inaccessible. The scenarios where the attacker has access to clear vulnerable speech allow testing the rotation effectiveness with close to perfect attack (named the *oracle* scenario and presented below). Regardless of the available data, the performances are evaluated using clear vulnerable speech and (inverted) anonymized vulnerable speech for invertibility assessment, and clear compromised speech and (inverted) anonymized vulnerable speech for linkability assessment.

6.2.3.3 Implementation

To improve the rotation performance, we extend the range of our experiments to modified x-vectors domains using principal component analysis and gender-dependent rotation estimation.

The Python code of the experiments is available at <https://github.com/deep-privacy/x-vector-procrustes>

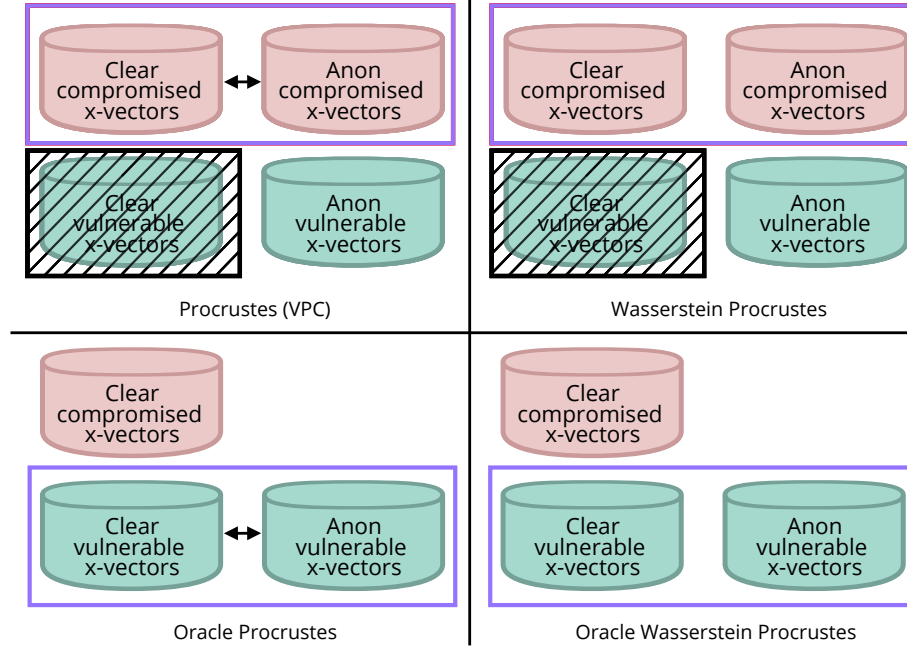


Figure 6.2 – Schematic representation of the datasets used for different scenarios.

Principal component analysis We apply a dimensional reduction technique to the x-vectors datasets using Principle Component Analysis (PCA) (Pearson 1901). Reducing the number of dimensions reduces the candidate rotations manifold, simplifying the search for the optimal one. The PCA also orders the dimensions precisely: the dimensions with the higher variance are placed first. This means that applying PCA on two vector datasets acts as a pre-alignment, easing the following alignment process. We reduced to 70 dimensions the originally 512 dimensions x-vectors using PCA the total explained variance ratio was always above 96.0%.

Gender dependent training The VPC evaluation allows for the attacker to have access to the gender information associated with each clear and anonymized utterance. We utilized this aspect to train two separate rotations, which improve the attack performance.

6.2.4 Experimental results

This section presents the experimental results for the invertibility scenarios presented above. Table 6.1 summarizes the results for the attacks on the VPC 2022 baseline, and table 6.2 the attacks on our fully quantization-based anonymization pipeline.

The first line of each table corresponds to the linkability assessment on clear speech. The clear speech linkability results show a pattern similar to those presented in Table 4.3, although there is a noticeable increase in the EER values, from 7.7% to 10.3% for female speakers and

from 1.1% to 2.9% for male speakers. This EER increase is caused by the use of the cosine similarity instead of the PLDA.

The second line of each table corresponds to the more traditional white-box linkability assessment on anonymized speech. Similarly, as for the clear speech EER, the anonymized EER are degraded compared to when using the PLDA. For Table 6.1 on the VPC baseline, the white-box ASV EER drops of by 14.5% of EER (absolute) compare to Table 5.4. This indicates that, especially for anonymized speech, the x-vector model that we are using indeed requires a PLDA scoring function to better disclose the degree of privacy protection.

For lines 1 and 2, only the EER is computed, because the attackers used cannot invert the anonymization. Lines 3 to 6 explore our invertibility attacks. We can see that for both tables, Procrustes/Wasserstein Procrustes and their oracle version reduce the EER score. This indicates that applying a rotation on the anonymized x-vector before comparing them to clear x-vectors improves the attacker compared to the (cosine similarity-based) white-box ASV attack, where anonymized utterances are compared with anonymized utterances.

Table 6.1 – Experimental results for the rotation-based invertibility attack scenarios on the VPC 2022 baseline. The scoring function for ASV linkability assessment is cosine similarity.

| | | EER \uparrow | | ACC \downarrow | |
|---|-------------------------------|----------------|------|------------------|------|
| | | Female | Male | Female | Male |
| 1 | Clear speech | 10.3 | 2.9 | | |
| 2 | White-box ASV | 27.9 | 27.8 | | |
| 3 | Procrustes | 19.5 | 21.4 | 67.7 | 51.8 |
| 4 | Wasserstein Procrustes | 20.1 | 22.9 | 64.7 | 50.0 |
| 5 | Oracle Procrustes | 17.1 | 12.0 | 98.3 | 97.0 |
| 6 | Oracle Wasserstein Procrustes | 17.7 | 12.8 | 99.0 | 97.2 |

Focusing on Table 6.1, the Procrustes invertibility attack (line 3) manages to achieve a reconstruction accuracy of 67% for females and 51.8% for males speaker meaning that more than half of the time, the anonymized speaker x-vectors can be re-identified. This raises concerns about the *non-invertibility* criteria of the anonymization system. Lines 4 explore the unsupervised scenario. We can see that Wasserstein Procrustes gives slightly worse results than the Procrustes counterpart, as no labels are available in this scenario requiring the attacker to estimate permutation and rotation matrices. We underline that the difference is usually around a few percent in EER and inversion accuracy, so the distribution of x-vectors before and after anonymization is probably quite similar. Similar enough to get close results to when labels are available. Lines 5 and 6 show the results associated with the oracle approaches. We observe that in terms of linkability, the EER are getting closer to the PLDA-based white-box ASV, as from line 2 to line 5, an absolute decrease of 10% of EER is observed. Still, PLDA-based white-box

ASV is a better linkability evaluation, and this is because of the PLDA calibration which relies on more data than the data used to estimate the rotation (360h vs 45min). In terms of invertibility, the almost 100% accuracy means that the anonymized speaker x-vectors can be re-identified using a rotation. As the anonymization pipeline should produce a single unique voice (because of the “constant speaker” target selection), having a 100% of accuracy is completely undesirable as it means that a bijection between clear and anonymized speech exists, and worse, a rotation is enough to estimate this transformation.

Finishing our invertibility attack analysis on Table 6.2 with our fully quantization-based anonymization pipeline. We observe that for most linkability attacks, the EER stays above 30.0%. The oracle Procrustes is the best with an EER of 29.5%. However, this is still far from the EER obtained with the PLDA-based white-box ASV (see Table 5.7 of Chapter 5), where the EER was 17.5%. The interesting results are more about the invertibility, where for this anonymization pipeline the oracle scenarios do not achieve 100% accuracy. Seeing a reduction in invertibility attack confirms the privacy improvement that this pipeline proposes compared to the VPC baseline presented above. However, 85% of accuracy is still too much. For the non-oracle scenarios, the inverting accuracy is drastically reduced compared to the VPC baseline, with more than 2x privacy improvement which is correlated with the linkability improvement. Overall, as we are seeing ASV linkability attack performance decrease, invertibility attacks performance is also decreasing.

Table 6.2 – Experimental results for the rotation-based invertibility attack scenarios on our fully quantization-based anonymization. The scoring function for ASV assessment is cosine similarity.

| | | EER \uparrow | | ACC \downarrow | |
|---|-------------------------------|----------------|------|------------------|------|
| | | F | M | F | M |
| 1 | Clear speech | 10.3 | 2.9 | | |
| 2 | White-box ASV | 37.8 | 36.4 | | |
| 3 | Procrustes | 32.3 | 35.0 | 32.7 | 24.7 |
| 4 | Wasserstein Procrustes | 37.9 | 36.7 | 23.7 | 15.9 |
| 5 | Oracle Procrustes | 29.6 | 29.4 | 85.3 | 84.1 |
| 6 | Oracle Wasserstein Procrustes | 37.8 | 33.4 | 87.8 | 82.7 |

6.2.5 Discussion

In this section, we introduced a new form of attack and privacy evaluation based on the capability of an attacker to reconstruct the clear x-vector from anonymized speech. Our proposed inversion attack is based on estimating a rotation matrix that transforms anonymized x-vectors to clear x-vectors. Our results show that the current anonymization approaches are susceptible to this form of attack, reinforcing the fact that there is room for improvement in current speaker

anonymization systems. The oracle attack scenarios seem to be an interesting framework to evaluate the robustness of future anonymization methods against re-identification attacks as it considers the most powerful attacks. The oracle attack scenarios even if they are unrealistic in terms of knowledge acquisition, allow us to better challenge the *non-invertibility* privacy criteria, as it directly challenges the definition itself “it should be computationally infeasible to obtain the clear data that led to any given anonymized data”.

The simplistic way we implemented our attack by estimating rotation matrices open the possibilities for more sophisticated inversion attacks. The following is a list of improvements that could be made. First, in our experiments, the use of a PLDA scoring instead of a cosine similarity scoring function could improve the linkability results we obtained. Alternatively, a more recent x-vector model trained with more advanced architecture and cost function can also help. Then, to improve the x-vectors transformation from anonymized to clear, the use of other methods such as normalizing flows network (Kobyzev et al. 2020) might also help. Also, training the inversion attack on more data (e.g., *LibriSpeech train-clean-360*) could help for the non-oracle attack scenarios. Finally, the way we record inversion success with accuracy is limited, another metric that better reflects the actual distance from the target x-vector could be interesting. But of course, the better would be to use the inverted x-vector for speech synthesis and compare the de-anonymized voice to the clear voice subjectively.

6.3 Subjective mispronunciation evaluation

In this section, we propose an early prototype to properly evaluate the main utility requirement of speaker anonymization systems: linguistic content preservation. The way it is currently evaluated with an ASR system that outputs a WER score on anonymized speech has one main limitation. The WER score reflects the errors of the ASR evaluation system to recognize correctly pronounced words, and the errors of the anonymization pipelines to correctly pronounce words. To accommodate for this, a WER reference score is computed on clear speech, and the utility performance of anonymized speech is then compared to this reference score. Any degradation is supposed to reflect the anonymization pipeline making errors to correctly pronounce words. The main issue with this approach comes from the fact that the most powerful anonymization pipelines rely on some sort of linguistic feature extraction before synthesizing anonymized speech. We noticed that the linguistic feature extractors are usually better at ASR task than the ASR evaluation model itself because they are trained on more data. For instance, in the VPC plan, the linguistic feature extractor is trained with almost two times more data than the ASR evaluation systems. Overall, it has been shown that the larger the corpora used to train the linguistic feature extractor, the better the anonymization system generalizes. In Chapter 5, the WER is usually improved after anonymization, indicating that the ASR evaluation models makes fewer errors in recognizing properly pronounced words. But what about the mispronunciation errors that the anonymization pipelines make?

6.3.1 Real world example

Below, we show examples of WER computation for clear speech in Figure 6.3, the VPC baseline anonymized speech in Figure 6.4, and our fully quantization-based anonymized speech in Figure 6.5. Here, we are interested in knowing whether the ASR mistakes come from the ASR decoding or mispronunciation errors coming from the anonymization.

| | | | | | | | | | |
|-----|-----|---------|-----|----------|----------|----|-------|-------|-------|
| ref | THE | RAINBOW | IS | A | DIVISION | OF | WHITE | LIGHT | |
| hyp | THE | *** | *** | RAINBOWS | DIVISION | OF | WHITE | LIGHT | [...] |
| op | C | D | D | S | C | C | C | C | |

Figure 6.3 – [Clear speech](#)[§] WER scoring with the ASR evaluation system. **ref** indicates the ground truth, **hyp** the output of the ASR, and **op** the Substitution, Deletion, and Insertion errors or Correctly decoded words.

In Figure 6.3 on clear speech, the ASR makes three errors (two deletions followed by a substitution) in the **RAINBOW IS A** segment. Upon listening to the audio file, it appears that the reference transcript matches what has been said (even though the speech is fast-paced in this specific segment).

§. Audio samples can be extracted from the PDF by clicking (or double-clicking) the text in blue.

| | | | | | | | | | | |
|-----|-----|---------|----|---|----------|----|-------|---------|------|-------|
| ref | THE | RAINBOW | IS | A | DIVISION | OF | WHITE | LIGHT | INTO | |
| hyp | THE | RAINBOW | IS | A | DIVISION | OF | WHITE | LICHENS | OF | [...] |
| op | C | C | C | C | C | C | C | S | S | |

Figure 6.4 – [Speech anonymized with the VPC 2022 baseline](#) WER scoring with the ASR evaluation system. `ref` indicates the ground truth, `hyp` the output of the ASR, and `op` the Substitution, Deletion, and Insertion errors or Correctly decoded words.

In Figure 6.4 on speech anonymized with the VPC baseline, the ASR does not make the same mistakes. The RAINBOW IS A segment is correctly decoded, however, an error occurs for LIGHT INTO segment. Upon listening to the audio file, this error is an ASR decoding error because the segment is correctly pronounced. If we were to evaluate the utility of the anonymized speech against the clear speech using the WER metric, the anonymized speech would be considered better. However, in practice, with subjective listening tests, both anonymized and clear speech might have the same level of preservation of linguistic content.

| | | | | | | | | | |
|-----|-----|---------|---------|---|----------|----|-------|-------|-------|
| ref | THE | RAINBOW | IS | A | DIVISION | OF | WHITE | LIGHT | |
| hyp | THE | RAIN | WITHOUT | A | DIVISION | OF | WHITE | LIGHT | [...] |
| op | C | S | S | C | C | C | C | C | |

Figure 6.5 – [Speech anonymized with our quantization-based anonymization pipeline](#) WER scoring with the ASR evaluation system. `ref` indicates the ground truth, `hyp` the output of the ASR, and `op` the Substitution, Deletion, and Insertion errors or Correctly decoded words.

In Figure 6.5 on speech anonymized with our fully quantization-based pipeline, the ASR shows mistakes in the RAINBOW IS segment. Upon listening, this error comes from the anonymization pipeline, the segment is not properly pronounced. In Chapter 6 we concluded that this pipeline has better or similar utility as clear speech when using the WER metric for utility comparison. However, we believe that with a mispronunciation evaluation, the utility would be decreased instead of being, in the current utility evaluation, better or similar.

6.3.2 Mispronunciation metric proposition

As we believe anonymized speech utility assessment should take into account mispronunciation errors we propose a mispronunciation metric. To compute such a metric we would use the decoded ASR results and ask listeners to annotate if errors in each word come from a mispronunciation on both anonymized and clear speech. If the error associated with a word does not come from a mispronunciation, the error (Substitution, Deletion, or Insertion) is invalidated and replaced with a Correct pronunciation label. Then the WER is calculated, with Substitution, Deletion, or Insertion errors only reflecting the actual preservation of the linguistic content as only mispronunciations errors are taken into account. Utilizing human listeners for this form of measurement can be quite expensive. Nonetheless, the initial error detection performed by the ASR can serve as a useful first step in identifying errors, which can significantly decrease the total costs associated with subsequent subjective evaluations. Human listeners would know where to look for mispronunciations in a speech signal, sparing them the need to annotate everything from

scratch. Furthermore, identifying audio files that are more prone to generating mispronunciations and analyzing only those files can also facilitate the evaluation process.

6.3.3 Discussion

In this section, we identified that automatic linguistic content preservation based on the WER score output by an ASR system might lead to an imprecise evaluation. As the current use of the WER score reflects both decoding and mispronunciations errors, it is complicated to conclude if a particular anonymization system reduces decoding errors so much that the introduced mispronunciations errors are insignificant. We argue that isolating only the mispronunciations to evaluate linguistic content preservation is the best way to evaluate utility. With the help of the initial WER score computation, we believe a subjective evaluation can isolate the actual errors we want to measure.

To conclude this section, we give an overview of the current anonymization processes, with a focus on listing them from the least to the most likely to result in mispronunciations. This will be accompanied by our subjective comments on each pipeline.

1. Signal processing anonymizations such as (Patino et al. 2021) are unlikely to introduce significant mispronunciation errors, as the transformation are usually non-destructive.
2. Voice conversion anonymizations with non-disentangled features such as the baseline of the VPC (Fang et al. 2019) are likely to introduce small mispronunciation errors because the synthesis system can make errors even though the ASR-BN can theoretically encode every possible sound.
3. Voice conversion anonymizations with disentangled features such as fully noised or fully quantized pipelines (Shamsabadi et al. 2022; Champion et al. 2022a) are likely to introduce more mispronunciation errors because the ASR-BN can no longer encode every possible sound.
4. Voice conversion anonymizations with discrete tokens such as phonemes or words representations (Meyer et al. 2022a) are likely to introduce a lot of mispronunciation errors or even completely miss entire words because the linguistic feature extractor does not generate a one-on-one mapping with the real acoustic input.

Interestingly, this ranking from the least mispronunciation errors making systems to the most is reversed when the criteria is privacy, the voice conversion anonymizations with discrete tokens being the best against linkability attacks, and signal processing anonymizations the worse. It is also worth noting that the further we go down on the list, the more amount of training data is required for the model to generalize well. Making them less suitable when the goal is actually to collect training data. Overall we believe a good trade-off between utility/privacy and the

amount of training data required to train the anonymization models can come out of systems in the middle of the list.

6.4 Conclusion

In this chapter, we challenged the *non-invertibility* privacy criteria that speaker anonymization should comply with. We proposed an invertibility attack based on embedding alignment techniques to invert anonymized x-vectors and reconstruct as best as possible their corresponding clear ones. The results show that the successfulness of linkability attacks is correlated with the successfulness of our invertibility attacks. As the main privacy evaluation of anonymization systems already relies on linkability, our invertibility evaluation does not need to be as crucial. Even though the oracle attack is quite interesting from a formal point of view, it theoretically evaluates if the bijection (if there is one) between anonymized and clear x-vectors space corresponds to a rotation. The best anonymization systems should not create bijections that can be estimated by rotations.

This chapter also questioned the method used to measure the preservation of linguistic content after anonymization. We identified from a conceptual point of view that comparing WER scores outputted by ASR evaluation models has major limitations restricting it from properly evaluating the level of linguistic content preservation. We proposed to enhance the computation of the WER to only take into account mispronunciation errors discarding inherent ASR decoding errors. This allows to isolate and measure the errors made by the anonymization systems and have a utility metric that better reflects what we want to measure. The only disadvantage of this approach is that it requires subjective human listening.

CONCLUSION AND PERSPECTIVES

The objective of this thesis is to propose privacy-preserving speech data collection solutions that rely on data anonymization. The motivation behind this research is the increasing concern about privacy and security of speech data, especially with the widespread use of speech technology. This thesis investigated the problems associated with the evaluation and design of speaker anonymization systems. The goal is to develop speaker anonymization methods that can remove speaker identity from speech signals while preserving linguistic content and speech quality.

Summary

In Chapter 4, we assessed the role of the target speaker identity parameter in voice conversion-based anonymization. For this analysis, we made the hypothesis that the voice conversion system is completely capable to remove the clear source identity and replace it with the one of a target. Meaning, that the speaker anonymization is perfect. Under this hypothesis, we evaluated many target speaker identity selection algorithms that are used to parameterize the voice conversion model. We observed that most of the VC parameterization used in today’s speaker anonymization evaluation protocol and challenges create a link between the clear and target speakers which creates a strong bias for privacy linkability assessment. Additionally, as an ASV attacker model is trained on anonymized speech to evaluate linkability, the way the attacker generates anonymized data with the target parameter matters. From this observation, we conclude that the best target selection algorithm for privacy evaluation is the one where all source speakers aim to be converted to a single target identity or completely random identities. With this voice conversion parametrization, there is a guarantee to generate unbiased anonymized speech where linkability assessment is correct, and where the ASV attacker model can be properly trained.

In the second part of this chapter, we queried if the target identity parameter affects the capability of the voice conversion model to generate anonymized voice. In particular, we wanted to know if there is a golden target speaker parameter that maximizes privacy and/or utility performances. Our conclusion showed that there is no particular target identity parameter that will improve privacy. However, there was clear evidence that some target identity parameters were highly destructive of the utility. We recommend manually selecting target identities that lead to good utility performance.

In Chapter 5, we analyzed the effectiveness of a voice conversion system to anonymize speech. The three main features used for voice conversion are the target identity studied in Chapter 4, the linguistic representations and F_0 are studied in this chapter. In order for a voice conversion system to replace a source identity with a target one, the features must be as disentangled as possible. In this chapter, we analyzed the degree of disentanglement of the F_0 and linguistic

representations in order to assess if they contain speaker information. It appeared that both were not freed of speaker information which might limit the performance of the anonymization. The linguistic representation appeared to be the one containing the more speaker information.

We experimented with adversarial training to better disentangle the linguistic representation. Here the goal is to add a negative loss when training the linguistic representation extractor to induce it to remove speaker information. Counter-intuitively, it appears that adversarial training does not improve the privacy of anonymized speech, indicating that the linguistic representation encoded the same amount of speaker information as before. However, we saw that the representation seemed to be more appropriate for speech synthesis as we observed an improvement in utility.

Another well-established method to extract privacy-preserving representation is to add noise to the representation. The noise needs to be applied during training such that the model can learn to properly encode the linguistic representation in a way that only the most relevant information stands out from the noise in a transmission channel. We reproduced this framework and showed that noise addition is very sensible to the source data and can badly impact the utility of the anonymized speech signal.

As an alternative, we proposed to use of a vector quantization technique to improve the disentanglement of the linguistic representations. Vector quantization can improve the privacy of the representation by reducing the transmission capacity so the irrelevant information (speaker identity) cannot be encoded. Similarly, as for noise addition, vector quantization needs to be applied during the training of the model. The results showed that the use of vector quantization is a promising alternative to noise-based transformations, overall generating higher-quality speech, while having similar privacy properties. The F_0 feature can also benefit from vector quantization, indeed, our fully quantization-based anonymization pipeline showed the best performance.

In Chapter 6, we switched from the role of playing the designer of speaker anonymization systems to the role of playing the attacker. We introduced the first invertibility attack of a speaker anonymization system. This invertibility attack aims to reconstruct or decrypt the anonymized speech to obtain the x-vector of the speakers to match the ones extracted on clear speech. The method used for this attacker relies on techniques for aligning two sets of speaker embedding (a clear one, and an anonymized one). In particular, we used Procrustes to estimate a rotation matrix that transforms an x-vector extracted from anonymized speech to the one extracted from clear speech. Our results show that if the attacker has an unrealistic amount of information, a rotation matrix is enough to perform the inversion attack. We must emphasize that under realistic scenarios, the degree of rotation-based inversion is much lower. However, the only possibility that a rotation is enough to invert an anonymization system is alarming, as even if today's realistic scenarios are not successful, tomorrow's attack may get close to today's unrealistic scenario.

Lastly in this chapter, we briefly presented the current limitation of the protocol used to evaluate the preservation of the linguistic content. We propose to perform subjective listening tests to assess if an ASR error comes from the ASR inability to decode the correctly preserved linguistic content or of the anonymization system to pronounce the linguistic content. This would operate by utilizing human listeners to categorize whether a decoding substitution, deletion, or insertion error was produced by the ASR system or a mispronunciation (anonymization error). Therefore, the WER metric only informs of mispronunciation errors, providing an accurate reflection of the utility compared to clear speech, where the same evaluation procedure should be used to obtain the reference clear WER score.

Perspectives and future directions

The primary extension in the field of speaker anonymization that we believe would have the most impact would be to use other datasets. We restricted ourselves to the ones defined in the VoicePrivacy Challenge, where the speech test data for which anonymization was applied is very clean (*LibriSpeech test-clean*, and *VCTK test*). We believe that up until now, the usage of clean data facilitated the early development of this new field, gathering a new community. However, as our understanding of the key challenges of speaker anonymization in terms of evaluation and design evolves, the datasets used should also evolve. Using data that are closer to real-world application cases, (i.e., call center, voice assistant), would allow for providing a better assessment of the current effectiveness of the technology.

Another extension also relates to the datasets, to train the ASV model, that performs the linkability assessment. At the moment the *LibriSpeech train-clean-360* dataset is used. However, by the nature of the dataset (audiobook), a lot of other information may be available in such a way that it helps linkability attacks. For example, across audiobook sections or chapters, the actor might stay consistent with the way he/she narrates the story. This raises questions about what the ASV attack model captures. Are we training an automatic speaker verification system or an automatic narration-style verification system? (or even story verification?). As anonymization techniques will get stronger and stronger, the ASV model is likely to focus on biased attributes to perform linkability. We believe it is currently happening as the attacker used in the thesis is always better for *LibriSpeech test-clean* dataset (i.e., 23% of EER for *LibriSpeech test-clean* and 39.8% for *VCTK test* privacy results for our fully-quantization-based anonymization). The preliminary test that we run by training the ASV attack model on a dataset more suited for speaker recognition (Voxceleb1) showed that the privacy protection for the *VCTK test* dataset is overestimated (updated privacy score of 26.8% of EER), while the privacy protection estimation on *LibriSpeech test-clean* could be biased (updated privacy score of 31.0% of EER). Overall, knowing whether some techniques achieved better anonymization or knowing if the attack model is biased is an inherent problem in this field.

For privacy evaluation, this field should stay informed as much as possible with the advancement done in speaker recognition. As of today, the ASV architecture/loss that we use has five years of delay. We maintained the same ASV model for the purpose of comparison in this thesis, however, we believe it is necessary to adopt a more advanced ASV model for future advancements. Additionally, new forms of attacks should also be designed. Over the last few years, there has been a great number of contributions regarding developing new anonymization systems but only a very few contributions tackled the development of new attacks. We believe that the best attacks are likely to come from the ones who understand speaker anonymization the best, that is to say, the people designing anonymization techniques.

Regarding utility evaluation, a perspective would be to use the mispronunciation WER metric that we defined in Chapter 6 to assess the preservation of the linguistic content. We believe this form of evaluation would be fairer to compare multiple systems against each other. Additionally, depending on the application, other utility metrics could be considered. With Hubert Nourtel and Marie Tahon, we worked on the emotion attribute ([Nourtel et al. 2021](#); [Nourtel et al. 2022](#)) which may be important for the speech of call centers.

Another evaluation extension would be to query the ability of the speaker anonymization to work with multiple languages. In the current state, we observe that the more private the anonymization system is, the less it works with multiple languages. Speaker anonymization has to be designed for a specific language. We do not necessarily consider this an issue as long as the amount of supervised data required to train a speaker anonymization system in a given language is relatively low. Hence, investigating data-efficient model training is of great interest.

Finally, we present future directions that relate to improving the performance of the anonymization systems.

Currently, the trade-off between privacy and the utility of the systems is fixed for a given model. Being able to dynamically adjust the level of transformation, for specific utterances or frames, is likely to improve both privacy and utility performance without requiring a complete paradigm shift. The level of transformation applied could vary dynamically throughout the utterance, rather than being fixed, in response to the fluctuating level of confidence of ASR-BN in correctly capturing the acoustic sound. This approach would allow for more effective adaptation to changing conditions and increase the overall privacy/utility of the system. To further elaborate, a dynamic adjustment of the transformation level could be achieved by utilizing a feedback loop that constantly evaluates the performance of the ASR-BN model and adjusts the transformation accordingly.

Given a model that can dynamically adjust the transformation, being able to adapt the transformation to a speaker/group of speakers could also enable the creation of a fairer algorithm. For example, this could be done in the vector quantization framework where in order to select the prototypes, other functions than the L2 distance could be used.

For the adversarial training approach, potential extensions of this research include incorporating a triplet loss for the adversarial model to more effectively simulate unlinkability attacks, and utilizing multiple adversarial models to counteract the tendency of the ASR-BN models to only deceive a single adversary.

Another extension that could improve privacy, would be to train speech synthesis models to generate anonymized speech. Currently, the speech synthesis model is parameterized and trained to reconstruct a given clear speech of a speaker from the training dataset. This means that the anonymization transformation has to occur before intermediate representations. Using anonymization systems that generate a small amount of mispronunciations errors to anonymize the training dataset used for speech synthesis training, and training another anonymization on that dataset could help to induce the speech synthesis model to also anonymize speech. This step could be performed many times as long as the utility is not too impacted.

For our final words, we want to disclose that although it is rewarding to create an anonymization system that improves privacy and or utility performances in today’s evaluation protocol, this field desperately needs more understandability and explainability regarding privacy and utility evaluation.

Publication list

The work carried out during this thesis led to the following publications:

- [1] (Champion et al. 2020b). Speaker information modification in the VoicePrivacy 2020 toolchain. In: *VoicePrivacy 2020 Virtual Workshop at Odyssey 2020*.
- [2] (Champion et al. 2020a). A Study of F0 Modification for X-Vector Based Speech Pseudonymization Across Gender. In: *The Second AAAI Workshop on Privacy-Preserving Artificial Intelligence*.
- [3] (Champion et al. 2021a). Evaluating X-vector-based Speaker Anonymization under White-box Assessment. In: *23rd International Conference on Speech and Computer (SPECOM)*.
- [4] (Champion et al. 2021b). On the invertibility of a voice privacy system using embedding alignment. In: *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.
- [5] (Champion et al. 2022b). Privacy-Preserving Speech Representation Learning using Vector Quantization. In: *Journées d’Études sur la Parole (JEP, 34e édition)*.
- [6] (Champion et al. 2022a). Are disentangled representations all you need to build speaker anonymization systems? In: *23rd Interspeech Conference*.

Other secondary contributions, were also done during the same period:

- [7] (Nourtel et al. 2021). Evaluation of Speaker Anonymization on Emotional Speech. In: *1st ISCA Symposium on Security and Privacy in Speech Communication (SPSC)*.
- [8] (Nourtel et al. 2022). Analyse de l’anonymisation du locuteur sur de la parole émotionnelle. In: *Journées d’Études sur la Parole (JEP, 34e édition)*.
- [9] (Tomashenko et al. 2022a). The VoicePrivacy 2022 challenge evaluation plan.

RÉSUMÉ ÉTENDU EN FRANÇAIS

1. Introduction

La communication verbale, qu'il s'agisse de parler ou d'écouter, est un moyen naturel et pratique pour nous humains d'interagir. Au cours des conversations, les individus échangent des informations linguistiques par la parole, mais ils transmettent également des informations supplémentaires à travers des signes paralinguistiques qui permettent de reconnaître l'identité du locuteur, mais aussi ses émotions, son âge et genre, etc. L'interaction homme-machine peut tirer parti de la richesse et de la commodité de la parole, permettant ainsi à la machine de mieux comprendre les humains et aux humains de partager facilement des informations avec les machines. Cette approche améliore considérablement l'expérience des utilisateurs et contribue à une utilisation plus inclusive des technologies modernes pour les personnes en situation de handicap. Cependant, la compréhension et la production de parole par les machines restent des tâches complexes qui font l'objet de recherches continues depuis de nombreuses années. Toutefois, au cours de la dernière décennie, d'importants progrès ont été réalisés dans plusieurs tâches liées à la parole, telles que la reconnaissance automatique de la parole (transcription automatique) et la synthèse de la parole (génération de parole à partir de texte), entre autres. Ces avancées ont introduit un nouveau produit sur le marché : les assistants vocaux.

L'objectif de nombreuses entreprises est d'établir une expérience d'interaction naturelle et pratique entre les humains et les machines, alimentée par les derniers développements technologiques en matière de traitement de la parole. Actuellement, les consommateurs adoptent de plus en plus les différents dispositifs d'assistants vocaux. En effet, l'étude intitulée *2022 U.S. Smart Home Consumer Adoption Report* révèle que 50 à 60% de la population américaine a accès à un ou plusieurs dispositifs d'assistants vocaux.

Afin de proposer des services compétitifs, les entreprises ont recours à des techniques avancées d'apprentissage profond pour alimenter les algorithmes des assistants vocaux. Ces algorithmes étant très gourmands en données, les performances optimales sont généralement atteintes lorsqu'une grande quantité de données est utilisée pour les entraîner. Ceci crée une nécessité pour les entreprises de collecter, traiter et stocker les données de parole de leurs utilisateurs sur des serveurs centralisés afin d'améliorer continuellement les services proposés et de rester compétitives. Cependant, les données de parole contenant de nombreuses informations personnelles telles que l'identité du locuteur, leur collecte soulève de sérieuses préoccupations en matière de protection des données personnelles.

Au départ, les pratiques de collecte de données des entreprises étaient inconnues ou mal comprises du public, mais cela a récemment changé. Aux alentours de 2017, de plus en plus de personnes ont pris conscience de la situation grâce à des gros titres de presse révélant l'utilisation des données vocales collectées par les entreprises¹. Parallèlement, l'union européenne a établi la législation sur le traitement de données la plus stricte au monde. Cette législation, la ([GDPR 2016](#)) ou Règlement Général sur la Protection des Données (RGPD), stipule spécifiquement que les informations personnelles des citoyens européens doivent être traitées avec conformité du droit au respect de la vie privée et avec la plus grande confidentialité.

Le droit à la vie privée est un élément juridique qui vise à protéger le respect de la vie privée des individus. C'est la capacité de contrôler qui peut accéder aux informations nous concernant, et dans quel but elles sont utilisées. Dans le cadre du GDPR, le droit à la vie privée est considéré comme un droit fondamental de l'homme, et il est essentiel pour protéger les citoyens contre l'utilisation abusive ou détournée de leurs informations personnelles. À l'ère du numérique, le droit à la vie privée revêt une importance primordiale, car l'utilisation croissante de la technologie permet de collecter et stocker des quantités massives de données personnelles. Alors que la technologie continue d'évoluer, notre compréhension de son influence sur le droit à la vie privée et notre engagement à la protéger doivent également évoluer.

La parole est considérée comme un type de donnée personnelle hautement sensible qui doit être protégé. Les directives récentes émises par la Commission Nationale de l'Informatique et des Libertés (CNIL) et le Conseil Européen de la Protection des Données rappellent que les données de la parole sont intrinsèquement des données biométriques, conformément à l'article 4(14) de la RGPD² ([CNIL 2020](#); [EDPB 2021](#)). Étant donné que le stockage et le traitement des données biométriques sont encore plus réglementés que celui des données personnelles, il est nécessaire de développer des schémas de collecte de données respectant mieux les informations privées.

Cette thèse propose des solutions de collecte de données de parole plus respectueuse en s'appuyant sur l'anonymisation des données. L'anonymisation des données est le processus de suppression ou d'altération des informations d'identification personnelle des données pour protéger les données personnelles des individus. Bien que le RGPD n'impose pas l'anonymisation systématique des données personnelles, l'anonymisation constitue une solution parmi d'autres permettant de traiter les données personnelles en conformité avec les droits et la vie privée des individus. Dans cette thèse, nous travaillons sur des méthodes d'anonymisation de locuteur qui visent à supprimer l'identité du locuteur des signaux de parole tout en préservant le contenu linguistique et la qualité de la parole.

1. *Yep, human workers are listening to recordings from Google Assistant, including audio recorded by mistake* (Oui, des travailleurs humains écoutent les enregistrements de Google, y compris ceux enregistrés par erreur).

2. L'article 4(14) du RGPD définit les données biométriques comme "les données à caractère personnel résultant d'un traitement technique spécifique relatif aux caractéristiques physiques, physiologiques ou comportementales d'une personne physique, qui permettent ou confirment l'identification unique de cette personne physique, telles que les images faciales ou les données dactyloscopies."

1.1 Cadre et objectifs

Ces dernières années, les techniques d’anonymisation de la parole ont suscité un intérêt croissant suite à la publication du VoicePrivacy Challenge. La plupart des approches reposent sur des systèmes de conversion de voix pour transformer l’identité du locuteur dans le signal clair en une autre identité dans le signal anonymisé. L’objectif est que le signal anonymisé ne puisse plus être lié à l’identité réelle du locuteur. Pour évaluer cet aspect, des techniques de vérification automatique du locuteur (ASV - *Automatic Speaker Verification*) sont utilisées sur la parole anonymisée pour évaluer son degré de liaison, plus il est faible, mieux c’est. La Figure 6.6 illustre cet objectif. En ce qui concerne la préservation du contenu linguistique (utilité), elle est évaluée à l’aide de la reconnaissance automatique de la parole.

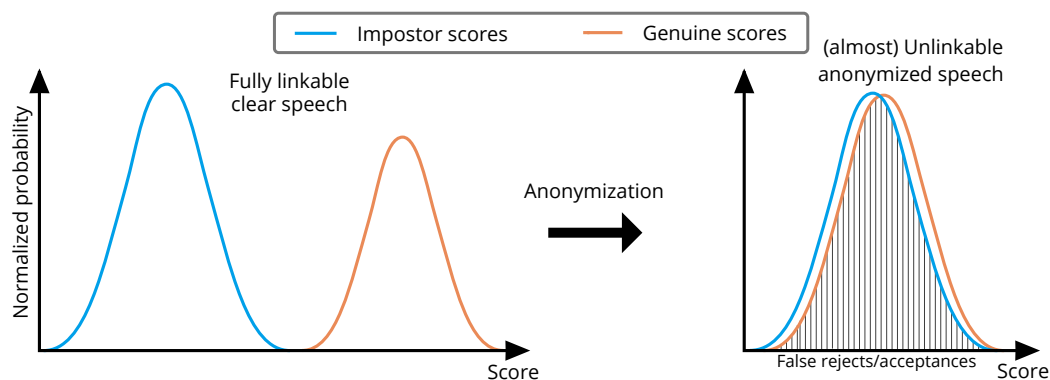


FIGURE 6.6 – Évaluation ASV de la capacité à lier un signal à un locuteur. Dans l’exemple, avant l’anonymisation (côté gauche), les signaux sont entièrement liables, les scores *impostor* (locuteurs différents entre 2 signaux) et scores *genuine* (même locuteur entre 2 signaux) sont disjoints. Après anonymisation (côté droit), les signaux doivent être faiblement liables, c’est-à-dire qu’il ne doit pas être possible de distinguer si deux signaux correspondent (ou pas) au même locuteur.

Le premier défi de l’anonymisation de la parole concerne l’évaluation de la bonne anonymisation des données personnelles (privacité). Cette évaluation dépend du paramètre d’identité de locuteur cible de la conversion de voix. En effet, la conversion de voix peut augmenter le degré de liaison si elle est demandée en rendant la voix d’un locuteur très différente de celle des autres locuteurs. Par conséquent, il est nécessaire de comprendre la conversion de voix de sorte que les voix anonymisées ne soient pas liables et que l’évaluation effectuée avec le système de vérification automatique de locuteur corresponde à l’évaluation de la privacité.

Le deuxième défi de l’anonymisation de la parole concerne l’algorithme de conversion de voix. À l’origine, les systèmes de conversions de voix étaient conçus pour transformer le locuteur d’un signal, de sorte que les auditeurs subjectifs (humains) croient qu’un signal a été prononcé par quelqu’un d’autre. Bien que la qualité de la transformation puisse être suffisante pour tromper les humains, ce n’est pas le cas pour les systèmes ASV. Ainsi, des formes plus avancées de systèmes de conversion de voix doivent être utilisées pour l’anonymisation du locuteur.

Le troisième défi concerne la diversité et le nombre d'évaluations possibles pour évaluer la confidentialité et l'utilité. En effet, il n'y a pas de méthode unique pour évaluer ces deux mesures. Par exemple, pour la confidentialité un jeu impliquant un attaquant et un défenseur où l'attaquant vise à briser l'anonymisation et où le défenseur l'améliore doit avoir lieu afin d'adapter en continu la technologie aux menaces actuelles. Quant à l'utilité, l'évaluation dépend toujours de l'objectif de partage de la parole. Nous nous sommes concentrés sur la préservation du contenu linguistique, mais il existe de nombreux autres cas d'application valables, tels que la reconnaissance des émotions.

Ces trois défis ont pour objectif d'explorer l'évaluation et la conception de systèmes d'anonymisation du locuteur. Ce résumé est organisé de la sorte pour présenter ces trois défis, les parties 2, 3 et 4 présentent respectivement les contributions principales des chapitres 4, 5 et 6 la partie 5 conclut ce résumé.

2. Influence du paramétrage du locuteur cible

Dans cette analyse, notre objectif est d'étudier l'influence du paramétrage du système de conversion de voix, avec les locuteurs cibles, sur l'évaluation de la confidentialité. Pour ce faire, nous proposons de considérer que le système de conversion de voix est parfait, c'est-à-dire, qu'il est capable de remplacer totalement l'identité du locuteur d'entrée avec l'identité choisie pour cible. La capacité de supprimer ou remplacer les données personnelles relatives au locuteur dans un signal de parole est exactement ce qui est recherché pour l'anonymisation du locuteur.

Lors de l'évaluation de la confidentialité avec un système d'ASV (cf figure 6.6), il est totalement possible que les voix anonymisées soient liées à des pseudo-identités qui correspondraient aux locuteurs cibles choisis pour chaque signal. Cela peut poser un problème si un lien existe entre les pseudo-identités et les identités source. Pour évaluer si un lien existe, nous proposons d'étudier les algorithmes de sélection de locuteurs cibles en vérifiant que les locuteurs cibles choisis par ceux-ci ne sont pas liés à l'identité source.

À la suite de nos expériences, nous observons que l'algorithme de sélection utilisé dans le plan d'évaluation du VPC crée une telle sorte de liaison. Ceci est mauvais, car lors de l'évaluation avec l'ASV, ce n'est plus la capacité d'évaluer le lien entre voix anonymisées et locuteur source qui est mesurée, mais le lien entre voix anonymisées et pseudo-locuteur. De ce fait, l'évaluation est faussée en raison du paramétrage du système de conversion de voix. Nous étudions d'autres alternatives d'algorithmes de sélection de locuteurs cibles qui ne créent pas de lien entre locuteur source et cible. La conclusion est que l'algorithme qui consiste à sélectionner une seule cible vers laquelle tous les signaux doivent être convertis est plus simple et conduit à une meilleure anonymisation.

Dans ce cadre d'anonymisation/évaluation où un seul locuteur cible est utilisé pour convertir tous les signaux, nous analysons plusieurs cibles à la recherche de celle qui maximise les performances. Nous concluons que le choix de la cible n'affecte pas la confidentialité mais qu'elle doit tout

de même être choisie soigneusement, car un mauvais choix peut fortement dégrader l'utilité.

3. Analyse des représentations utilisées pour la conversion de voix

Au cours de cette étude notre objectif est de mesurer le degré de liaison des représentations utilisées dans le système de conversion de voix. L'hypothèse est que si les représentations sont démêlées alors cela garanti que le système de synthèse remplace le locuteur source par par le locuteur cible. Les représentations que nous étudions sont la fréquence fondamentale (F0) et la représentation linguistique obtenue à partir d'un modèle acoustique de reconnaissance de la parole (cf figure 6.7). Nous observons que le degré de liaison est élevé entre ces représentations et le locuteur, la représentation linguistique étant celle qui divulgue le plus d'information locuteur.

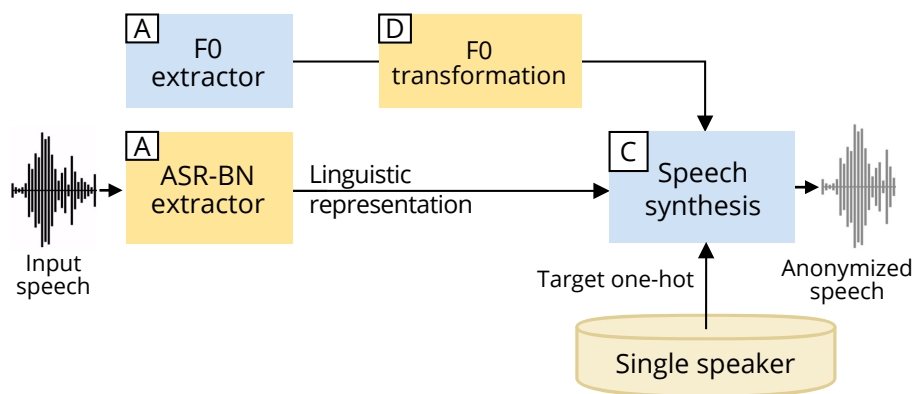


FIGURE 6.7 – Schéma représentatif d'un système de conversion de voix.

Afin d'améliorer le démêlement entre l'information locuteur et la F0 d'une part, et la représentation linguistique d'autre part, nous étudions plusieurs méthodes. Parmi elles, l'une est basée sur l'ajout de bruit et est la méthode la plus utilisée dans le domaine de la *differential privacy*. Cette méthode a été introduite pour l'anonymisation du locuteur par (SHAMSABADI et al. 2022). L'autre méthode, que nous introduisons, est basée sur la quantification vectorielle. L'anonymisation basée sur le bruit consiste à ajouter du bruit dans un canal de transmission de manière à ce que seules les informations les plus pertinentes ressortent du bruit, tandis que l'anonymisation basée sur la quantification consiste à réduire la capacité du canal de transmission de manière à ce que les informations non pertinentes (et personnelles) ne soient pas encodées.

Le résultat des évaluations montre que ces deux méthodes de démêlement ont des performances similaires en termes de privacité et d'utilité. Cependant, nous observons que notre méthode basée sur la quantification est plus robuste que la méthode basée sur l'ajout de bruit, ayant de bonnes performances d'utilité à travers plusieurs jeux de tests. Nous supposons que la quantification vectorielle généralise mieux sur les données bruitées.

4. Un nouveau type d'attaque et mesure d'utilité

Dans cette partie, nous avons remis en question un autre critère de privacité auquel les systèmes d'anonymisation du locuteur doivent se conformer : l'inversibilité. L'inversibilité définit qu'il ne doit pas être computationnellement faisable de reconstruire la donnée source à partir de sa contrepartie anonymisée. Afin d'évaluer ce critère, nous proposons une autre forme d'attaque basée sur des techniques d'alignement de vecteur (*embedding*). Les résultats montrent que le succès des attaques d'inversibilité est corrélé à celles des attaques de liaison.

Nous avons aussi remis en question la méthode utilisée pour mesurer la préservation du contenu linguistique après l'anonymisation. Nous avons identifié d'un point de vue conceptuel que la comparaison des scores (WER - *Word Error Rate*) produits par les modèles d'évaluation de reconnaissance automatique de la parole présente des limitations majeures qui l'empêchent d'évaluer correctement le niveau de préservation du contenu linguistique. Nous avons proposé d'améliorer le calcul du (WER - *Word Error Rate*) pour ne prendre en compte que les erreurs de prononciation en éliminant les erreurs de décodage. Cela permet d'isoler et de mesurer les erreurs faites par les systèmes d'anonymisation et d'obtenir une mesure d'utilité qui reflète mieux ce que nous voulons mesurer. Le seul inconvénient de cette approche est qu'elle nécessite une écoute humaine subjective.

5. Conclusion

L'objectif de cette thèse était de proposer des solutions de collecte de données vocales préservant la vie privée qui reposent sur l'anonymisation des données. La motivation derrière cette recherche est l'inquiétude croissante quant à la collecte de données personnelles de parole, en particulier avec l'utilisation généralisée des nouvelles technologies. Cette thèse a examiné les problèmes liés à l'évaluation et à la conception de systèmes d'anonymisation de locuteur. L'objectif est de développer des méthodes d'anonymisation de locuteur capables de supprimer l'identité des locuteurs des signaux vocaux tout en préservant le contenu linguistique et la qualité de la parole.

Nos contributions sont diversifiées, traitant d'un point de vue conceptuel comment le paramétrage des systèmes d'anonymisation affecte l'évaluation de la privacité, comment améliorer les performances des systèmes d'anonymisation, et comment les évaluations basées à partir d'attaque ou non peuvent être étendues pour mieux refléter les performances des systèmes. Pour conclure, nous tenons à souligner que bien qu'il soit gratifiant de créer un système d'anonymisation qui améliore la privacité et/ou l'utilité dans le protocole d'évaluation actuel, ce domaine a désespérément besoin d'une meilleure compréhension en ce qui concerne l'évaluation de la privacité et de l'utilité.

BIBLIOGRAPHY

1. Adi, Y., N. Zeghidour, R. Collobert, N. Usunier, V. Liptchinsky, and G. Synnaeve, « To Reverse the Gradient or Not: an Empirical Comparison of Adversarial and Multi-task Learning in Speech Recognition », *in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019 44, 63, 99, 104
2. Ahmed, Shimaa, Amrita Roy Chowdhury, Kassem Fawaz, and Parmesh Ramanathan, « Preech: A System for Privacy-Preserving Speech Transcription », *in: 29th USENIX Security Symposium*, 2020 64
3. Ajili, Moez, Solange Rossato, Dan Zhang, and Jean-François Bonastre, « Impact of rhythm on forensic voice comparison reliability », *in: IEEE Odyssey - The Speaker and Language Recognition Workshop*, 2018 103
4. Alammar, Jay, *The Illustrated Transformer*, 2018, URL: <https://jalammar.github.io/illustrated-transformer/> 19
5. Aloufi, Ranya, Hamed Haddadi, and David Boyle, « Privacy-preserving Voice Analysis via Disentangled Representations », *in: ArXiv*, 2020 55, 61
6. Arias, Pablo, Laura Rachman, Marco Liuni, and Jean-Julien Aucouturier, « Beyond Correlation: Acoustic Transformation Methods for the Experimental Study of Emotional Voice and Speech », *in: Emotion Review*, 2020 10
7. Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, « wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations », *in: Advances in Neural Information Processing Systems (NIPS)*, 2020 19, 32
8. Bahmaninezhad, Fahimeh, Chunlei Zhang, and John H. L. Hansen, « Convolutional Neural Network Based Speaker De-Identification », *in: IEEE Odyssey - The Speaker and Language Recognition Workshop*, 2018 52
9. Baker, J., « The DRAGON system—An overview », *in: IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1975 27
10. Bengio, Yoshua, Nicholas Léonard, and Aaron C. Courville, « Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation », *in: ArXiv*, 2013 106
11. Bgn, Jonathan, *Illustrated Tour of Wav2vec 2.0*, 2021, URL: <https://jonathanbgn.com/2021/09/30/illustrated-wav2vec-2.html> 32
12. Bimbot, Frédéric and Gérard Chollet, « Assessment of speaker verification systems », *in: Handbook of standards and resources for spoken language systems*, Mouton de Gruyter Berlin, 1997 38

-
13. Biswas, Russa, Mehwish Alam, and Harald Sack, « Is Aligning Embedding Spaces a Challenging Task? A Study on Heterogeneous Embedding Alignment Methods », *in: ArXiv*, 2020 ¹¹⁹
 14. Bojanowski, Piotr and Armand Joulin, « Unsupervised learning by predicting noise », *in: International Conference on Machine Learning (ICML)*, 2017 ¹¹⁹
 15. Bourlard, Herve and Christian J Wellekens, « Multilayer perceptrons and automatic speech recognition », *in: International Conference on Neural Networks (ICNN)*, 1987 ²⁹
 16. Brassler, Ferdinand, Tommaso Frassetto, Korbinian Riedhammer, Ahmad-Reza Sadeghi, Thomas Schneider, and Christian Weinert, « VoiceGuard: Secure and Private Speech Processing », *in: Interspeech*, 2018 ³
 17. Brown, Peter F., Vincent J. Della Pietra, Peter V. de Souza, Jennifer C. Lai, and Robert L. Mercer, « Class-Based n-gram Models of Natural Language », *in: Computational Linguistics*, 1992 ³¹
 18. Brown, Tom B., Benjamin Mann, et al., « Language Models are Few-Shot Learners », *in: ArXiv*, 2020 ¹⁹
 19. Brummer, N. and J.A. Preez, « Application-independent evaluation of speaker detection », *in: IEEE Odyssey - The Speaker and Language Recognition Workshop*, 2006 ³⁹
 20. Champion, Pierre, Denis Jouvet, and Anthony Larcher, « A Study of F0 Modification for X-Vector Based Speech Pseudonymization Across Gender », *in: The Second AAAI Workshop on Privacy-Preserving Artificial Intelligence*, 2020 ^{95, 135}
 21. — « Speaker information modification in the VoicePrivacy 2020 toolchain », *in: VoicePrivacy 2020 Virtual Workshop at Odyssey*, 2020 ¹³⁵
 22. — « Evaluating X-vector-based Speaker Anonymization under White-box Assessment », *in: International Conference on Speech and Computer (SPECOM)*, 2021 ¹³⁵
 23. — « Are disentangled representations all you need to build speaker anonymization systems? », *in: Interspeech*, 2022 ^{129, 135}
 24. — « Privacy-Preserving Speech Representation Learning using Vector Quantization », *in: Journées d'Études sur la Parole (JEP, 34e édition)*, 2022 ¹³⁵
 25. Champion, Pierre, Thomas Thebaud, Gaël Le Lan, Anthony Larcher, and Denis Jouvet, « On the invertibility of a voice privacy system using embedding alignment », *in: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021 ^{55, 135}
 26. Cheng, Qiang and J. Sorensen, « Spread spectrum signaling for speech watermarking », *in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001 ⁵⁹
 27. Chorowski, Jan, Ron J. Weiss, Samy Bengio, and Aäron van den Oord, « Unsupervised Speech Representation Learning Using WaveNet Autoencoders », *in: IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 2019 ¹⁰⁶

-
28. Chou, Ju-chieh and Hung-Yi Lee, « One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization », *in: Interspeech*, 2019 ⁴²
 29. Chung, J. S., A. Nagrani, and A. Zisserman, « VoxCeleb2: Deep Speaker Recognition », *in: Interspeech*, 2018 ⁵⁷
 30. CNIL, « ON THE RECORD Exploring the ethical, technical and legal issues of voice assistants », *in: Commission Nationale de l'Informatique et des Libertés*, 2020 ^{2, 51, 138}
 31. Cummins, Nicholas, Alice Baird, and Bjoern W Schuller, « Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning », *in: Methods*, Elsevier, 2018 ¹²
 32. Dehak, Najim, Réda Dehak, James R. Glass, Douglas A. Reynolds, and Patrick Kenny, « Cosine Similarity Scoring without Score Normalization Techniques », *in: IEEE Odyssey - The Speaker and Language Recognition Workshop*, 2010 ³⁷
 33. Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin, « Maximum likelihood from incomplete data via the EM - algorithm plus discussions on the paper », *in: 1977* ³⁷
 34. Deng, Jiangyi, Fei Teng, Yanjiao Chen, Xiaofu Chen, Zhaohui Wang, and Wenyuan Xu, « V-Cloak: Intelligibility-, Naturalness- and Timbre-Preserving Real-Time Voice Anonymization », *in: 32nd USENIX Security Symposium*, 2022 ⁶⁵
 35. Desai, Srinivas, Alan W. Black, B. Yegnanarayana, and Kishore Prahallad, « Spectral Mapping Using Artificial Neural Networks for Voice Conversion », *in: IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 2010 ⁴¹
 36. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, « BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding », *in: Association for Computational Linguistics*, 2019 ³²
 37. Djebbar, Fatiha and Baghdad Ayad, « Comparative study of digital audio steganography techniques », *in: EURASIP Journal on Advances in Signal Processing*, 2012 ⁵⁹
 38. Doddington, George R., Walter Liggett, Alvin F. Martin, Mark A. Przybocki, and Douglas A. Reynolds, « SHEEP, GOATS, LAMBS and WOLVES: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation », *in: 5th International Conference on Spoken Language Processing (ICSLP)*, 1998 ⁸⁸
 39. Dwork, Cynthia, « Differential Privacy », *in: Encyclopedia of Cryptography and Security*, 2006 ^{64, 105}
 40. Dwork, Cynthia and Aaron Roth, « The Algorithmic Foundations of Differential Privacy », *in: Found. Trends Theor. Comput. Sci.* 2014 ^{64, 105}
 41. EDPB, « Guidelines on Virtual Voice Assistants », *in: European Data Protection Board*, 2021 ^{2, 51, 138}
 42. Ericsson, David, Adam Östberg, Edvin Listo Zec, John Martinsson, and Olof Mogren, « Adversarial representation learning for private speech generation », *in: ArXiv*, 2020 ⁶³

-
43. Erro, Daniel, Asunción Moreno, and Antonio Bonafonte, « Voice Conversion Based on Weighted Frequency Warping », *in: IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 2010 ⁴¹
 44. Espinoza-Cuadros, Fernando M., Juan M. Perero-Codosero, Javier Antón-Martín, and Luis Alfonso Hernández Gómez, « Speaker De-identification System using Autoencoders and Adversarial Training », *in: VoicePrivacy 2020 Virtual Workshop at Odyssey*, 2020 ⁶²
 45. Espy-Wilson, Carol Y., Sandeep Manocha, and Srikanth Vishnubhotla, « A new set of features for text-independent speaker identification », *in: Interspeech*, 2006 ³⁴
 46. Eyben, Florian, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al., « The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing », *in: IEEE transactions on affective computing*, IEEE, 2015 ¹⁰
 47. Fang, Fuming, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-Francois Bonastre, « Speaker Anonymization Using X-vector and Neural Waveform Models », *in: 10th ISCA Speech Synthesis Workshop*, 2019 ^{42, 44, 61, 64, 69, 71, 72, 129}
 48. Fang, Fuming, Junichi Yamagishi, and Isao Echizen, « High-Quality Nonparallel Voice Conversion Based on Cycle-Consistent Adversarial Network », *in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018 ^{23, 44}
 49. Faundez-Zanuy, Marcos, « Digital watermarking: new speech and image applications », *in: International Conference on Nonlinear Speech Processing*, 2009 ⁵⁹
 50. Feutry, Clément, Pablo Piantanida, Yoshua Bengio, and Pierre Duhamel, « Learning Anonymized Representations with Adversarial Neural Networks », *in: ArXiv*, 2018 ^{20, 99}
 51. Fourier, Jean Baptiste Joseph, Gaston Darboux, et al., *Théorie analytique de la chaleur*, Didot Paris, 1822 ¹³
 52. Fung, Benjamin C. M., Ke Wang, Rui Chen, and Philip S. Yu, « Privacy-Preserving Data Publishing: A Survey of Recent Developments », *in: ACM computing surveys (CSUR)*, 2010 ⁵¹
 53. Fung, Pascale, « Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus », *in: Third Workshop on Very Large Corpora*, 1995 ¹¹⁹
 54. Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, « Domain-Adversarial Training of Neural Networks », *in: Domain Adaptation in Computer Vision Applications*, 2017 ²⁰
 55. Gao, Yang, Rita Singh, and Bhiksha Raj, « Voice Impersonation Using Generative Adversarial Networks », *in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018 ⁴²

-
56. Gaznepoglu, Ünal Ege and Nils Peters, « Exploring the Importance of F0 Trajectories for Speaker Anonymization using X-vectors and Neural Waveform Models », *in: Workshop on Machine Learning in Speech and Language Processing*, 2021 ⁹¹
 57. GDPR, « Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC », *in: General Data Protection Regulation*, 2016 ^{2, 50, 138}
 58. Godoy, Elizabeth, Olivier Rosec, and Thierry Chonavel, « Voice Conversion Using Dynamic Frequency Warping With Amplitude Scaling, for Parallel or Nonparallel Corpora », *in: IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 2012 ⁴¹
 59. Golub, Gene H. and Christian H. Reinsch, « Singular value decomposition and least squares solutions », *in: Numerische Mathematik*, 1970 ¹⁸
 60. Gomez-Barrero, Marta, Javier Galbally, Christian Rathgeb, and Christoph Busch, « General Framework to Evaluate Unlinkability in Biometric Template Protection Systems », *in: IEEE Transactions on Information Forensics and Security*, 2018 ^{34, 39, 40}
 61. Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, « Generative adversarial nets », *in: Advances in Neural Information Processing Systems (NIPS)*, 2014 ^{20, 22, 23}
 62. Gower, John C, « Generalized procrustes analysis », *in: Psychometrika*, 1975 ^{118, 120}
 63. Grave, Edouard, Armand Joulin, and Quentin Berthet, « Unsupervised alignment of embeddings with wasserstein procrustes », *in: ArXiv*, 2018 ^{118–120}
 64. Graves, Alex, Abdel-rahman Mohamed, and Geoffrey E. Hinton, « Speech recognition with deep recurrent neural networks », *in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013 ¹⁶
 65. Guo, Chuan, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q. Weinberger, « Simple Black-box Adversarial Attacks », *in: ArXiv*, 2019 ^{52, 104}
 66. Haasnoot, Erwin, Ali Khodabakhsh, Chris G. Zeinstra, Luuk J. Spreeuwiers, and Raymond N. J. Veldhuis, « FEERCI: A Package for Fast Non-Parametric Confidence Intervals for Equal Error Rates in Amortized $O(m \log n)$ », *in: 2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2018 ⁷⁹
 67. Habibu, Taban, Edith Talina Luhanga, and Anael Elikana Sam, « Assessment of How Users Perceive the Usage of Biometric Technology Applications », *in: Recent Advances in Biometrics*, 2022 ⁴⁹
 68. Hadian, Hossein, Hossein Sameti, Daniel Povey, and Sanjeev Khudanpur, « End-to-end Speech Recognition Using Lattice-free MMI », *in: Interspeech*, 2018 ³⁰
 69. — « Flat-Start Single-Stage Discriminatively Trained HMM-Based Models for ASR », *in: IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 2018 ⁹⁸

-
70. Hansen, John H. L. and Taufiq Hasan, « Speaker Recognition by Machines and Humans: A tutorial review », *in: IEEE Signal Processing Magazine*, 2015 ³⁵
 71. Hashimoto, Kei, Junichi Yamagishi, and Isao Echizen, « Privacy-preserving sound to degrade automatic speaker verification performance », *in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016 ⁵²
 72. Hauser, Marc D., Noam Chomsky, and W. Tecumseh Fitch, « The faculty of language: what is it, who has it, and how did it evolve? », *in: Science*, 2002 ¹⁰
 73. Hazen, Timothy J., Wade Shen, and Christopher White, « Query by example spoken term detection », *in: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2009 ⁴³
 74. He, Kaiming, X. Zhang, Shaoqing Ren, and Jian Sun, « Deep Residual Learning for Image Recognition », *in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016 ¹⁶
 75. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, « Deep residual learning for image recognition », *in: IEEE International Conference on Computer Vision (ICCV)*, 2016 ¹⁶
 76. Hickey, Raymond, « Language contact: Reconsideration and reassessment », *in: 2010* ¹¹
 77. Hillenbrand, James M. and Michael J. Clark, « The role of f0 and formant frequencies in distinguishing the voices of men and women », *in: Attention, Perception, and Psychophysics*, 2009 ^{11, 90}
 78. Hinton, Geoffrey, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., « Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups », *in: IEEE Signal Processing Magazine*, 2012 ²⁹
 79. Hirakawa, Yutaka, Ayaka Shimoda, Isao Sasano, and Kazuo Ohzeki, « Improvements in a Puzzle Authentication Method », *in: Journal of Computational Chemistry*, 2018 ⁴⁹
 80. Houmani, Nesma and Sonia Garcia-Salicetti, « On Hunting Animals of the Biometric Menagerie for Online Signature », *in: PLoS ONE*, 2016 ⁸⁸
 81. Hsu, Chin-Cheng, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, « Voice conversion from non-parallel corpora using variational auto-encoder », *in: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016 ⁴²
 82. Huang, Wen-Chin, Tomoki Hayashi, Shinji Watanabe, and Tomoki Toda, « The Sequence-to-Sequence Baseline for the Voice Conversion Challenge 2020: Cascading ASR and TTS », *in: ArXiv*, 2020 ⁴²

-
83. Huang, Wen-Chin, Tomoki Hayashi, Yi-Chiao Wu, H. Kameoka, and Tomoki Toda, « Voice Transformer Network: Sequence-to-Sequence Voice Conversion Using Transformer with Text-to-Speech Pretraining », *in: Interspeech*, 2020 ^{42, 45}
 84. Huang, Wen-Chin, Haiyan Luo, Hsin-Te Hwang, Chen-Chou Lo, Yu-Huai Peng, Yu Tsao, and Hsin-Min Wang, « Unsupervised Representation Disentanglement Using Cross Domain Features and Adversarial Learning in Variational Autoencoder Based Voice Conversion », *in: IEEE Transactions on Emerging Topics in Computational Intelligence*, 2020 ⁹⁹
 85. Huang, Wen-Chin, Yi-Chiao Wu, Chen-Chou Lo, Patrick Lumban Tobing, Tomoki Hayashi, Kazuhiro Kobayashi, Tomoki Toda, Yu Tsao, and H. Wang, « Investigation of F0 conditioning and Fully Convolutional Networks in Variational Autoencoder based Voice Conversion », *in: Interspeech*, 2019 ^{48, 90}
 86. Ioffe, Sergey, « Probabilistic Linear Discriminant Analysis », *in: European Conference on Computer Vision*, 2006 ³⁷
 87. ISO-24745, « Information technology — Security techniques — Biometric information protection », *in: ISO/IEC JTC1 SC27 Security Techniques*, 2011 ^{34, 50, 118}
 88. Jain, Anil K, Karthik Nandakumar, and Abhishek Nagar, « Biometric template security », *in: EURASIP Journal on Advances in Signal Processing*, Hindawi Limited London, UK, United Kingdom, 2008 ^{34, 50}
 89. Jain, Anil K., Lin Hong, and Sharath Pankanti, « Biometric identification », *in: Commun. ACM*, 2000 ¹²
 90. Jelinek, F., « Continuous speech recognition by statistical methods », *in: Proceedings of the IEEE*, 1976 ²⁶
 91. Kai, Hiroto, Shinnosuke Takamichi, Sayaka Shiota, and Hitoshi Kiya, « Lightweight Voice Anonymization Based on Data-Driven Optimization of Cascaded Voice Modification Modules », *in: IEEE Spoken Language Technology Workshop (SLT)*, 2021 ⁶⁰
 92. — « Robustness of Signal Processing-Based Pseudonymization Method Against Decryption Attack », *in: The Speaker and Language Recognition Workshop*, 2022 ^{55, 118}
 93. Kain, A. and M.W. Macon, « Spectral voice conversion for text-to-speech synthesis », *in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998 ⁴¹
 94. Kaiser, Łukasz, Aurko Roy, Ashish Vaswani, Niki Parmar, Samy Bengio, Jakob Uszkoreit, and Noam M. Shazeer, « Fast Decoding in Sequence Models using Discrete Latent Variables », *in: International Conference on Machine Learning (ICML)*, 2018 ¹⁰⁷
 95. Kashkin, A., I. A. Karpukhin, and Sergei L. Shishkin, « HiFi-VC: High Quality ASR-Based Voice Conversion », *in: ArXiv*, 2022 ⁴⁵

-
96. Kasi, Kavita and Stephen A. Zahorian, « Yet Another Algorithm for Pitch Tracking », *in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002 ⁶²
 97. Kazemi, Hadi, Seyed Mehdi Iranmanesh, and Nasser Nasrabadi, « Style and content disentanglement in generative adversarial networks », *in: IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019 ⁹⁹
 98. Khan, Mohd. Ehmer and Farmeena Khan, « A Comparative Study of White Box, Black Box and Grey Box Testing Techniques », *in: International Journal of Advanced Computer Science and Applications*, 2012 ⁵²
 99. Kim, Jay J, « A method for limiting disclosure in microdata based on random noise and transformation », *in: Proceedings of the section on survey research methods*, American Statistical Association Alexandria, VA, 1986 ⁹¹
 100. Kingma, Diederik P. and Jimmy Ba, « Adam: A Method for Stochastic Optimization », *in: CoRR*, 2015 ¹⁰²
 101. Kobzyev, Ivan, Simon Prince, and Marcus A. Brubaker, « Normalizing Flows: An Introduction and Review of Current Methods », *in: IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020 ¹²⁶
 102. Koenecke, Allison, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel, « Racial disparities in automated speech recognition », *in: National Academy of Sciences*, 2020 ⁶⁶
 103. Kong, Jungil, Jaehyeon Kim, and Jaekyoung Bae, « HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis », *in: Advances in Neural Information Processing Systems (NIPS)*, 2020 ^{23, 45, 46}
 104. Kröger, Jacob Leon, Otto Hans-Martin Lutz, and Philip Raschke, « Privacy Implications of Voice and Speech Analysis - Information Disclosure by Inference », *in: Privacy and Identity Management*, 2019 ^{12, 50}
 105. Kudo, Taku, « Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates », *in: Annual Meeting of the Association for Computational Linguistics*, 2018 ³¹
 106. Langenderfer, Jeff and Stefan Linnhoff, « The Emergence of Biometrics and Its Effect on Consumers », *in: Journal of Consumer Affairs*, 2005 ⁴⁹
 107. Larcher, Anthony, Kong-Aik Lee, and Sylvain Meignier, « An extensible speaker identification sidekit in Python », *in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016 ^{95, 98}
 108. Larsen, Anders Boesen Lindbo, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther, « Autoencoding beyond Pixels Using a Learned Similarity Metric », *in: International Conference on Machine Learning (ICML)*, 2016 ⁴⁷

-
109. LeCun, Yann and Yoshua Bengio, « Convolutional Networks for Images, Speech, and Time Series », *in: The Handbook of Brain Theory and Neural Networks*, 1998 ¹⁷
110. Leeuwen, David van, Niko Brummer, and Albert Swart, « A comparison of linear and non-linear calibrations for speaker recognition », *in: Proc. The Speaker and Language Recognition Workshop (Odyssey 2014)*, 2014 ³⁹
111. Leeuwen, David A. and Niko Brümmer, « An Introduction to Application-Independent Evaluation of Speaker Recognition Systems », *in: Speaker Classification I: Fundamentals, Features, and Methods*, 2007 ³⁹
112. Leroy, David, Alice Coucke, Thibaut Lavril, Thibault Gisselbrecht, and Joseph Dureau, « Federated Learning for Keyword Spotting », *in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018 ³
113. Li, Jinyu, « Recent Advances in End-to-End Automatic Speech Recognition », *in: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2022 ³¹
114. Li, Tiancheng and Ninghui Li, « On the Tradeoff between Privacy and Utility in Data Publishing », *in: ACM Special Interest Group for Knowledge Discovery from Data*, 2009 ^{59, 110, 115}
115. Lin, Tianyang, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu, « A Survey of Transformers », *in: ArXiv*, 2021 ¹⁹
116. Liu, Shuo, Adria Mallol-Ragolta, Emilia Parada-Cabeleiro, Kun Qian, Xingshuo Jing, Alexander Kathan, Bin Hu, and Björn Schuller, « Audio Self-supervised Learning: A Survey », *in: ArXiv*, 2022 ¹²
117. Liu, Songxiang, Yuewen Cao, Disong Wang, Xixin Wu, Xunying Liu, and Helen M. Meng, « Any-to-Many Voice Conversion With Location-Relative Sequence-to-Sequence Modeling », *in: IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 2021 ^{44, 45}
118. Liu, Songxiang, Jinghua Zhong, Lifa Sun, Xixin Wu, Xunying Liu, and Helen Meng, « Voice Conversion Across Arbitrary Speakers Based on a Single Target-Speaker Utterance », *in: Interspeech*, 2018 ^{42, 45}
119. Liu, Yi, Liang He, and Jia Liu, « Large Margin Softmax Loss for Speaker Verification », *in: Interspeech*, 2019 ^{36, 102}
120. Liu, Yufei, Chengzhu Yu, Wang Shuai, Zhenchuan Yang, Yang Chao, and Weibin Zhang, « Non-Parallel Any-to-Many Voice Conversion by Replacing Speaker Statistics », *in: Interspeech*, 2021 ⁴⁴
121. Lorenzo-Trueba, Jaime, Thomas Drugman, Javier Latorre, Thomas Merritt, Bartosz Putrycz, Roberto Barra-Chicote, Alexis Moinet, and Vatsal Aggarwal, « Towards Achieving Robust Universal Neural Vocoding », *in: Interspeech*, 2019 ⁴⁵

-
122. Madikeri, Srikanth, Sibongwe Sibongwe, Juan Zuluaga-Gomez, Apoorv Vyas, Petr Motlicek, and Hervé Bouchard, « Pkwrap: a PyTorch Package for LF-MMI Training of Acoustic Models », *in: ArXiv*, 2020 ^{18, 98}
 123. Magarinos, Carmen, Paula Lopez-Otero, Laura Docio-Fernandez, Eduardo Rodriguez-Banga, Daniel Erro, and Carmen Garcia-Mateo, « Reversible speaker de-identification using pre-trained transformation functions », *in: Computer Speech Language*, Elsevier, 2017 ⁵²
 124. Mao, Xudong, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley, « Least Squares Generative Adversarial Networks », *in: IEEE International Conference on Computer Vision (ICCV)*, 2017 ^{23, 47}
 125. Maouche, Mohamed, Brij Mohan Lal Srivastava, Nathalie Vauquier, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent, « A comparative study of speech anonymization metrics », *in: Interspeech*, 2020 ^{39, 40}
 126. Mary, Leena, « Significance of Prosody for Speaker, Language, Emotion, and Speech Recognition », *in: Briefs in Speech Technology*, 2018 ¹¹
 127. Mawalim, Candy Olivia, Kasorn Galajit, Jessada Karnjana, and Masashi Unoki, « X-Vector Singular Value Modification and Statistical-Based Decomposition with Ensemble Regression Modeling for Speaker Anonymization System », *in: Interspeech*, 2020 ⁶²
 128. Mawalim, Candy Olivia, Shogo Okada, and Masashi Unoki, « Speaker anonymization by pitch shifting based on time-scale modification », *in: 2nd ISCA Symposium on Security and Privacy in Speech Communication*, 2022 ⁶⁰
 129. McAdams, S., « Spectral fusion, spectral parsing and the formation of the auditory image », *in: Ph. D. Thesis, Stanford*, 1984 ⁶⁰
 130. Memon, Shahan Ali, « Acoustic Correlates of the Voice Qualifiers: A Survey », *in: ArXiv*, 2020 ¹⁰
 131. Mercuri, Rebecca T. and Peter G. Neumann, « Security by Obscurity », *in: Commun. ACM*, 2003 ⁵²
 132. Meyer, Sarina, Florian Lux, Pavel Denisov, Julia Koch, Pascal Tilli, and Ngoc Thang Vu, « Speaker Anonymization with Phonetic Intermediate Representations », *in: Interspeech*, 2022 ^{65, 129}
 133. Meyer, Sarina, Pascal Tilli, Pavel Denisov, Florian Lux, Julia Koch, and Ngoc Thang Vu, « Anonymizing Speech with Generative Adversarial Networks to Preserve Speaker Privacy », *in: IEEE Spoken Language Technology Workshop (SLT)*, 2022 ⁶³
 134. Miche, Yoan, Ian Oliver, Silke Holtmanns, Aapo Kalliola, Anton Akusok, Amaury Lendasse, and Kaj-Mikael Björk, « Data Anonymization as a Vector Quantization Problem: Control Over Privacy for Health Data », *in: International Conference on Availability, Reliability, and Security (CD-ARES)*, 2016 ⁹²

-
135. Nagrani, Arsha, Joon Son Chung, and Andrew Zisserman, « Voxceleb: a large-scale speaker identification dataset », *in: ArXiv*, 2017 ^{34, 57}
136. Nakashika, Toru, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Arika, « Voice conversion in high-order eigen space using deep belief nets », *in: Interspeech*, 2013 ⁴¹
137. Nautsch, Andreas, Jose Patino, N. Tomashenko, Junichi Yamagishi, Paul-Gauthier Noé, Jean-François Bonastre, Massimiliano Todisco, and Nicholas Evans, « The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment », *in: Interspeech*, 2020 ⁸⁸
138. Nematollahi, Mohammad Ali and Syed Abdul Rahman Al-Haddad, « An overview of digital speech watermarking », *in: International Journal of Speech Technology*, 2013 ⁵⁹
139. Nematollahi, Mohammad Ali, Chalee Vorakulpipat, and Hamurabi Gamboa Rosales, « Speech Watermarking », *in: Digital Watermarking : Techniques and Trends*, Springer Singapore, 2017 ⁵⁹
140. Noé, Paul-Gauthier, Jean-François Bonastre, Driss Matrouf, N. Tomashenko, Andreas Nautsch, and Nicholas Evans, « Speech Pseudonymisation Assessment Using Voice Similarity Matrices », *in: Interspeech*, 2020 ⁵⁹
141. Nourtel, Hubert, Pierre Champion, Denis Jouvét, Anthony Larcher, and Marie Tahon, « Evaluation of Speaker Anonymization on Emotional Speech », *in: 1st ISCA Symposium on Security and Privacy in Speech Communication*, 2021 ^{55, 134, 135}
142. — « Analyse de l’anonymisation du locuteur sur de la parole émotionnelle », *in: Journées d’Études sur la Parole (JEP, 34e édition)*, 2022 ^{134, 135}
143. O’Reilly, Patrick, Andreas Bugler, Keshav Bhandari, Max Morrison, and Bryan Pardo, « VoiceBlock: Privacy through Real-Time Adversarial Attacks with Audio-to-Audio Models », *in: Advances in Neural Information Processing Systems (NIPS)*, 2022 ⁶⁵
144. Ohtani, Yamato, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, « Many-to-many eigenvoice conversion with reference voice », *in: Interspeech*, 2009 ^{44, 45}
145. Oord, Aaron van den, Oriol Vinyals, and koray kavukcuoglu koray, « Neural Discrete Representation Learning », *in: Advances in Neural Information Processing Systems (NIPS)*, 2017 ¹⁰⁶
146. Oord, Aäron van den, Yazhe Li, and Oriol Vinyals, « Representation Learning with Contrastive Predictive Coding », *in: ArXiv*, 2018 ³²
147. Oppenheim, A.V., R.W. Schafer, and J.R. Buck, *Discrete-time Signal Processing*, Prentice Hall international editions, 1999 ⁹²
148. Osborne, Martin J and Ariel Rubinstein, *A course in game theory*, MIT press, 1994 ²⁰
149. Osia, Seyed Ali, Ali Shahin Shamsabadi, Sina Sajadmanesh, Ali Taheri, Kleomenis Katevas, Hamid R. Rabiee, Nicholas D. Lane, and Hamed Haddadi, « A Hybrid Deep Learning Architecture for Privacy-Preserving Mobile Analytics », *in: IEEE Internet of Things Journal*, 2017 ⁶³

-
150. Panayotov, V., G. Chen, D. Povey, and S. Khudanpur, « Librispeech: An ASR corpus based on public domain audio books », *in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015 ⁵⁷
 151. Pascual, Santiago, Antonio Bonafonte, and Joan Serra, « SEGAN: Speech Enhancement Generative Adversarial Network », *in: Interspeech*, 2017 ²³
 152. Paszke, Adam, Sam Gross, et al., « PyTorch: An Imperative Style, High-Performance Deep Learning Library », *in: Advances in Neural Information Processing Systems (NIPS)*, 2019 ¹⁸
 153. Pathak, Manas A., « Privacy-Preserving Machine Learning for Speech Processing », *in: Ph. D. Thesis, Carnegie Mellon University*, 2012 ³
 154. Patino, Jose, Natalia Tomashenko, Massimiliano Todisco, Andreas Nautsch, and Nicholas Evans, « Speaker Anonymisation Using the McAdams Coefficient », *in: Interspeech*, 2021 ^{60, 129}
 155. Pearson, Karl, « LIII. On lines and planes of closest fit to systems of points in space », *in: The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1901 ¹²³
 156. Peddinti, Vijayaditya, Daniel Povey, and Sanjeev Khudanpur, « A time delay neural network architecture for efficient modeling of long temporal contexts », *in: Interspeech*, 2015 ^{18, 29, 30}
 157. Perraudin, Nathanaël, Peter Balazs, and Peter L. Søndergaard, « A fast Griffin-Lim algorithm », *in: 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013 ⁴⁵
 158. Polyak, Adam, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux, « Speech Resynthesis from Discrete Disentangled Self-Supervised Representations », *in: Interspeech*, 2021 ^{45, 101}
 159. Polyak, Adam, Lior Wolf, and Yaniv Taigman, « TTS Skins: Speaker Conversion via ASR », *in: Interspeech*, 2020 ⁴⁴
 160. Polzehl, Tim, Sebastian Möller, and Florian Metze, « Automatically Assessing Personality from Speech », *in: IEEE Fourth International Conference on Semantic Computing*, 2010 ⁵⁰
 161. Povey, Daniel, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur, « Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks », *in: Interspeech*, 2018 ¹⁸
 162. Povey, Daniel, Arnab Ghoshal, et al., « The Kaldi Speech Recognition Toolkit », *in: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011 ^{18, 29, 98}
 163. Povey, Daniel, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, « Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI », *in: Interspeech*, 2016 ³⁰

-
164. Prajapati, Gauri P., Dipesh K. Singh, Preet P. Amin, and Hemant A. Patil, « Voice privacy using CycleGAN and time-scale modification », *in: Computer Speech and Language*, 2022 ⁶¹
 165. Prenger, Ryan J., Rafael Valle, and Bryan Catanzaro, « Waveglow: A Flow-based Generative Network for Speech Synthesis », *in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019 ⁴⁵
 166. Prince, Simon and James H. Elder, « Probabilistic Linear Discriminant Analysis for Inferences About Identity », *in: 2007 IEEE 11th International Conference on Computer Vision*, 2007 ³⁷
 167. Prince, Simon J.D. and James H. Elder, « Probabilistic Linear Discriminant Analysis for Inferences About Identity », *in: 2007 IEEE 11th International Conference on Computer Vision*, 2007 ³⁷
 168. Qian, Jianwei, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, Xiangyang Li, Yu Wang, and Yanbo Deng, « VoiceMask: Anonymize and Sanitize Voice Input on Mobile Devices », *in: ArXiv*, 2017 ^{52, 60}
 169. Qian, Kaizhi, Zeyu Jin, Mark Hasegawa-Johnson, and Gautham J. Mysore, « F0-Consistent Many-To-Many Non-Parallel Voice Conversion Via Conditional Autoencoder », *in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020 ^{48, 90, 97}
 170. Qian, Kaizhi, Yang Zhang, Shiyu Chang, David Cox, and Mark A. Hasegawa-Johnson, « Unsupervised Speech Decomposition via Triple Information Bottleneck », *in: International Conference on Machine Learning (ICML)*, 2020 ⁹⁷
 171. Qian, Kaizhi, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, « Autovc: Zero-shot voice style transfer with only autoencoder loss », *in: International Conference on Machine Learning (ICML)*, 2019 ^{42, 44, 45}
 172. Rabiner, L.R., « A tutorial on hidden Markov models and selected applications in speech recognition », *in: Proceedings of the IEEE*, 1989 ²⁷
 173. Raitio, Tuomo, Ramya Rasipuram, and Dan Castellani, « Controllable neural text-to-speech synthesis using intuitive prosodic features », *in: Interspeech*, 2020 ⁴³
 174. Raj, D., D. Snyder, and D. Povey, « Probing the Information Encoded in X-Vectors », *in: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019 ¹⁰³
 175. Ramoji, Shreyas, V PrashantKrishnan, and Sriram Ganapathy, « NPLDA: A Deep Neural PLDA Model for Speaker Verification », *in: IEEE Odyssey - The Speaker and Language Recognition Workshop*, 2020 ³⁷
 176. Rapp, Reinhard, « Identifying word translations in non-parallel texts », *in: ArXiv*, 1995 ¹¹⁹
 177. Rosenberg, Andrew, « Speech, Prosody, and Machines: Nine Challenges for Prosody Research », *in: Speech Prosody*, 2018 ⁴³
 178. Rosenblatt, Frank, « The perceptron: a probabilistic model for information storage and organization in the brain. », *in: Psychological review*, 1958 ¹⁵

-
179. Roth, Volker, Kai Richter, and Rene Freidinger, « A PIN-entry method resilient against shoulder surfing », *in: Conference on Computer and Communications Security*, 2004 ⁴⁹
180. Rumelhart, David E. and James L. McClelland, « Learning Internal Representations by Error Propagation », *in: Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, 1987 ¹⁵
181. Rüschemdorf, Ludger, « The Wasserstein distance and approximation theorems », *in: Probability Theory and Related Fields*, 1985 ¹²⁰
182. Ryffel, Théo, David Pointcheval, Francis Bach, Edouard Dufour-Sans, and Romain Gay, « Partially Encrypted Deep Learning using Functional Encryption », *in: Advances in Neural Information Processing Systems (NIPS)*, Curran Associates, Inc., 2019 ^{18, 22, 63, 100}
183. Samarakoon, L., B. Mak, and A. Y. S. Lam, « Domain Adaptation of End-to-end Speech Recognition in Low-Resource Settings », *in: IEEE Spoken Language Technology Workshop (SLT)*, 2018 ²⁰
184. Schreiber, Peter A., « Understanding Prosody's Role in Reading Acquisition », *in: Theory Into Practice*, 1991 ¹¹
185. Sennrich, Rico, Barry Haddow, and Alexandra Birch, « Neural Machine Translation of Rare Words with Subword Units », *in: Annual Meeting of the Association for Computational Linguistics*, 2016 ³¹
186. Shamsabadi, Ali Shahin, Brij Mohan Lal Srivastava, Aurélien Bellet, Nathalie Vauquier, Emmanuel Vincent, Mohamed Maouche, Marc Tommasi, and Nicolas Papernot, « Differentially private speaker anonymization », *in: Privacy Enhancing Technologies*, 2022 ^{64, 74, 76, 86, 91, 106, 109, 112, 113}
187. Shannon, Benjamin J. and Kuldip K. Paliwal, « A Comparative Study of Filter Bank Spacing for Speech Recognition », *in: Microelectronic engineering research conference*, 2003 ^{13, 34}
188. Snyder, D., D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, « X-vectors: Robust DNN Embeddings for Speaker Recognition », *in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018 ^{35, 44}
189. Snyder, David, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, « Deep Neural Network Embeddings for Text-Independent Speaker Verification », *in: Interspeech*, 2017 ³⁵
190. Srivastava, Brij Mohan Lal, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent, « Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion? », *in: Interspeech*, 2019 ^{21, 63, 99, 100, 104}
191. Srivastava, Brij Mohan Lal, Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Junichi Yamagishi, Mohamed Maouche, Aurélien Bellet, and Marc Tommasi, « Design Choices for X-vector Based Speaker Anonymization », *in: Interspeech*, 2020 ^{62, 74}

-
192. Srivastava, Brij Mohan Lal, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, « Evaluating Voice Conversion-Based Privacy Protection against Informed Attackers », *in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020 ^{51, 53, 60, 76}
 193. Stylianou, Y., O. Cappe, and E. Moulines, « Continuous probabilistic transform for voice conversion », *in: IEEE Transactions on Speech and Audio Processing*, 1998 ⁴¹
 194. Sun, L., K. Li, H. Wang, S. Kang, and H. Meng, « Phonetic posteriorgrams for many-to-one voice conversion without parallel data training », *in: IEEE International Conference on Multimedia and Expo*, 2016 ^{42, 43}
 195. Toda, Tomoki, Alan W. Black, and Keiichi Tokuda, « Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory », *in: IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 2007 ⁴¹
 196. Toda, Tomoki, Yamato Ohtani, and Kiyohiro Shikano, « One-to-Many and Many-to-One Voice Conversion Based on Eigenvoices », *in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007 ^{44, 45}
 197. Tomashenko, Natalia, Pierre Champion, Nicholas Evans, Xiaoxiao Miao, Hubert Nourtel, Massimiliano Todisco, Jean-François Bonastre, Emmanuel Vincent, Xin Wang, and Junichi Yamagishi, *The VoicePrivacy 2022 challenge evaluation plan*, 2022 ^{55, 64, 76, 78, 85, 135}
 198. Tomashenko, Natalia, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, et al., *The VoicePrivacy 2020 challenge evaluation plan*, 2020 ⁵⁶
 199. Tomashenko, Natalia, Xin Wang, et al., « The VoicePrivacy 2020 Challenge: Results and findings », *in: Computer Speech and Language*, 2022 ⁵⁸
 200. Tran, Minh and M. Soleymani, « A Speech Representation Anonymization Framework via Selective Noise Perturbation », *in: ArXiv*, 2022 ⁶⁴
 201. Trilok, Naresh P, Sung-Hyuk Cha, and Charles C Tappert, « Establishing the uniqueness of the human voice for security applications », *in: CSIS Research Day, Pace University*, 2004 ⁵⁰
 202. Turner, H.C.M., Giulio Lovisotto, Simon Eberz, and Ivan Martinovic, « I'm Hearing (Different) Voices: Anonymous Voices to Protect User Privacy », *in: ArXiv*, 2022 ⁶⁵
 203. Turner, Henry, Giulio Lovisotto, and Ivan Martinovic, « Speaker Anonymization with Distribution-Preserving X-Vector Generation for the VoicePrivacy Challenge 2020 », *in: VoicePrivacy 2020 Virtual Workshop at Odyssey*, 2020 ^{62, 80}
 204. Vaidya, Tavish and Micah Sherr, « You Talk Too Much: Limiting Privacy Exposure Via Voice Input », *in: IEEE Security and Privacy Workshops*, 2019 ⁶⁰
 205. Vaquero, Carlos, Alfonso Ortega, and EDUARDO LLEIDA SOLANO, « Partitioning of Two-Speaker Conversation Datasets », *in: Interspeech*, 2011 ³⁷

-
206. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, « Attention is All You Need », *in: Advances in Neural Information Processing Systems (NIPS)*, 2017 ¹⁹
207. Veaux, Christophe, Junichi Yamagishi, and Kirsten Macdonald, « CSTR VCTK Corpus: English Multi-speaker Corpus », *in: University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017 ⁵⁷
208. Veselý, Karel, Arnab Ghoshal, Lukáš Burget, and Daniel Povey, « Sequence-discriminative training of deep neural networks », *in: Interspeech*, 2013 ²⁹
209. Vincent, James, *Yep, human workers are listening to recordings from Google Assistant, including audio recorded by mistake*, 2019, URL: <https://www.theverge.com/2019/7/11/20690020/google-assistant-home-human-contractors-listening-recordings-vrt-nws> ^{2, 138}
210. voicebot.ai, *2022 U.S. Smart Home Consumer Adoption Report*, 2022, URL: <https://research.voicebot.ai/report-list/u-s-smart-home-consumer-adoption-report-2022/> ^{1, 137}
211. Vyas, Apoorv, Srikanth Madikeri, and Hervé Bouchard, « Comparing CTC and LFMMI for Out-of-Domain Adaptation of wav2vec 2.0 Acoustic Model », *in: Interspeech*, 2021 ¹⁰⁹
212. Waibel, Alex, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang, « Phoneme recognition using time-delay neural networks », *in: IEEE Transactions on Speech and Audio Processing*, 1989 ^{17, 29}
213. Wali, Aamir, Zareen Alamgir, Saira Karim, Ather Fawaz, Mubariz Barkat Ali, Muhammad Adan, and Malik Mujtaba, « Generative adversarial networks for speech processing: A review », *in: Comput. Speech Lang.* 2022 ²³
214. Wang, Changhan, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux, « VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation », *in: Association for Computational Linguistics*, 2021 ¹⁰⁹
215. Wang, X., S. Takaki, and J. Yamagishi, « Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis », *in: IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 2020 ⁶⁴
216. Wu, Peter, Paul Pu Liang, Jiatong Shi, Ruslan Salakhutdinov, Shinji Watanabe, and Louis-Philippe Morency, « Understanding the Tradeoffs in Client-side Privacy for Downstream Speech Tasks », *in: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2021 ^{59, 115}
217. Wu, Da-Yi and Hung-yi Lee, « One-Shot Voice Conversion by Vector Quantization », *in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020 ^{42, 45, 106}

-
218. Yoo, In-Chul, Keonnyeong Lee, Seonggyun Leem, Hyunwoo Oh, Bonggu Ko, and Dong-suk Yook, « Speaker Anonymization for Personal Information Protection Using Voice Conversion Techniques », *in: IEEE Access*, 2020 ^{61, 62}
 219. Zeghidour, Neil, « Learning representations of speech from the raw waveform », *in: 2019* ¹³
 220. Zen, H., V. Dang, R. Clark, Yu Zhang, Ron J. Weiss, Y. Jia, Z. Chen, and Y. Wu, « LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech », *in: Interspeech*, 2015 ⁵⁷
 221. Zhang, Jing-Xuan, Zhen-Hua Ling, and Li-Rong Dai, « Non-Parallel Sequence-to-Sequence Voice Conversion With Disentangled Linguistic and Speaker Representations », *in: IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 2020 ^{42, 44}
 222. Zhang, Lin, Xin Wang, Erica Cooper, Nicholas Evans, and Junichi Yamagishi, « The PartialSpoof Database and Countermeasures for the Detection of Short Fake Speech Segments Embedded in an Utterance », *in: IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 2022 ⁶⁵

LIST OF ABBREVIATIONS

$D_{\leftrightarrow}^{\text{sys}}$ Linkability measure.

F_0 fundamental frequency.

AAM Additive Angular Margin.

ANN Artificial Neural Networks.

ASR Automatic Speech Recognition.

ASR-BN bottleneck from automatic speech recognition.

ASV Automatic Speaker Verification.

AWGN Additive White Gaussian noise.

Cllr log-likelihood-ratio cost function.

DCT Discrete Cosine Transform.

DNN Deep Neural Network.

DP Differential Privacy.

E2E-LF-MMI End-to-End Lattice-Free Maximum Mutual Information.

EER Equal Error Rate.

FAR False Acceptance Rate.

FRR False Rejection Rate.

GAN Generative Adversarial Network.

GDPR General Data Protection Regulation.

GMM Gaussian Mixture Models.

HMM Hidden Markov Models.

LF-MMI Lattice-Free Maximum Mutual Information.

MFCCs Mel Frequency Cepstral Coefficients.

MLP Multilayer perceptron.

MMI Maximum Mutual Information.
MOS Mean Opinion Score.
PAV Pool Adjacent Violators.
PCA Principle Component Analysis.
PII Personally Identifiable Information.
PLDA Probabilistic Linear Discriminant Analysis.
PPG Phonetic Posteriorgrams.
STFT Short-time Fourier Transform.
SVD Singular Value Decomposition.
TDNN Time-Delay Neural Network.
TDNN-F Factored Time-Delay Neural Network.
TTS Text-To-Speech.
VAE Variational Autoencoders.
VC Voice Conversion.
VPC VoicePrivacy Challenge.
VQ Vector Quantization.
WER Word Error Rate.

Title: Anonymizing Speech: Evaluating and Designing Speaker Anonymization Techniques

Keywords: Speaker anonymization, Speech recognition, Speaker verification, Privacy.

Abstract: The growing use of voice user interfaces, from telephones to remote controls, automobiles, and digital assistants, has led to a surge in the collection and storage of speech data. While data collection allows for the development of efficient tools powering most speech services, it also poses serious privacy issues for users as centralized storage makes private personal speech data vulnerable to cyber threats. Advanced speech technologies, such as voice-cloning and personal attribute recognition, can be used to access and exploit sensitive information. Voice-cloning technology allows an attacker to take a recording of a person's voice and use it to generate new speech that sounds like it is coming from that person. For example, an attacker could use voice-cloning to impersonate a person's voice to gain unauthorized access to his/her financial information over the phone. With the increasing use of voice-based digital assistants like Amazon's Alexa, Google's Assistant, and Apple's Siri, and with the increasing ease with which personal speech data can be collected and stored, the risk of malicious use of voice-cloning and speaker/gender/pathological/etc. recognition technologies have increased. Companies and organizations need to consider these risks and implement appropriate measures to protect user data in order to prevent misuse of speech technologies and comply with legal regulations (e.g., General Data Protection Regulation (GDPR)).

To address these concerns, this thesis proposes solutions for anonymizing speech and evaluating the degree of the anonymization. In this work, anonymization refers to the process of making personal speech data unlinkable to an identity, while maintaining the usefulness (utility) of the speech signal (e.g., access to the linguistic content). The goal is to protect the privacy of individuals by removing or obscuring any Personally Identifiable Information (PPI) from the acoustic of speech. PPI includes things like a person's voice, accent, and speaking style; other personal information in the speech content like, phone number, person name, etc., is out of the scope of this thesis.

Our research is built on top of existing anonymization methods based on voice conversion and existing evaluation protocols. We start by identifying and explaining several challenges that evaluation protocols need to consider to evaluate the degree of privacy protection properly. We clarify how anonymization systems need to be configured for evaluation purposes and highlight the fact that many practical deployment configurations do not permit privacy evaluation. Furthermore, we study and examine the most common voice conversion-based anonymization system and identify its weak points, before suggesting new methods to overcome some limitations. We isolate all components of the anonymization system to evaluate the degree of speaker PPI associated with each of them. Then, we propose several transformation methods for each component to reduce as much as possible speaker PPI while maintaining utility. We promote anonymization algorithms based on quantization-based transformation as an alternative to the most-used and well-known noise-based approach. Finally, we endeavor a new attack method to invert the anonymization, creating a new threat. In this thesis, we openly work on sharing anonymization systems and evaluation protocols to aid organizations in facilitating the preservation of privacy rights for individuals.

Titre : Anonymisation de la parole : Évaluation et Conception de Techniques d'Anonymisation du Locuteur

Mot clés : Anonymisation du locuteur, Reconnaissance de la parole, Vérification du locuteur, Respect vie privée.

Résumé : L'essor de l'utilisation d'assistants vocaux, présents dans les téléphones, automobiles et autres, a augmenté la quantité de données de parole collectées et stockées. Bien que cette collecte de données soit cruciale pour entraîner les modèles qui traitent la parole, cette collecte soulève également des préoccupations de protection de la vie privée.

Des technologies de pointe traitant la parole, telles que le clonage vocal et la reconnaissance d'attributs personnels (telles que l'identité, l'émotion, l'âge, le genre, etc.), peuvent être exploitées pour accéder et utiliser des informations personnelles. Par exemple, un malfaiteur pourrait utiliser le clonage vocal pour se faire passer pour une autre personne afin d'obtenir un accès non autorisé à ses informations bancaires par téléphone.

Avec l'adoption croissante des assistants vocaux tels qu'Alexa, Google Assistant et Siri, et la facilité avec laquelle les données peuvent être collectées et stockées, le risque d'utilisation abusive de technologies telles que le clonage vocal et la reconnaissance d'attributs personnels augmente. Il est donc important pour les entreprises et les organisations de prendre en compte ces risques et de mettre en place des mesures appropriées pour protéger les données des utilisateurs, en conformité avec les réglementations juridiques telles que le Règlement Général sur la Protection des Données (RGPD).

Pour répondre aux enjeux liés à la protection de la vie privée, cette thèse propose des solutions permettant d'anonymiser la parole. L'anonymisation désigne ici le processus consistant à rendre les signaux de parole non associables à une identité spécifique, tout en préservant leur utilité, c'est-à-dire ne pas modifier le contenu linguistique du message. L'objectif est de préserver la vie privée des individus en éliminant ou en rendant floues toutes les informations personnellement identifiables (PPI) contenues dans le signal acoustique, telles que l'accent ou le style de parole d'une personne. Les informations linguistiques personnelles telles que numéros de téléphone ou noms de personnes ne font pas partie du champ d'étude de cette thèse.

Notre recherche s'appuie sur les méthodes d'anonymisation existantes basées sur la conversion de la voix et sur des protocoles d'évaluation existants. Nous commençons par identifier et expliquer plusieurs défis auxquels les protocoles d'évaluation doivent faire face afin d'évaluer de manière précise le niveau de protection de la vie privée. Nous clarifions comment les systèmes d'anonymisation doivent être configurés pour être correctement évalués, en soulignant le fait que de nombreuses configurations ne permettent pas une évaluation adéquate de non-asociabilité d'un signal à une identité. Nous étudions et examinons également le système d'anonymisation basé sur la conversion de la voix le plus courant, identifions ses points faibles, et proposons de nouvelles méthodes pour en améliorer les performances. Nous avons isolé tous les composants du système d'anonymisation afin d'évaluer le niveau de PPI encodé par chaque composant. Ensuite, nous proposons plusieurs méthodes de transformation de ces composants dans le but de réduire autant que possible les PPI encodées, tout en maintenant l'utilité. Nous promovons les algorithmes d'anonymisation basés sur l'utilisation de la quantification en alternative à la méthode la plus utilisée et la plus connue basée sur le bruit. Enfin, nous proposons une nouvelle méthode d'évaluation qui vise à inverser l'anonymisation, créant ainsi une nouvelle manière d'étudier les systèmes d'anonymisation.