Final Project

Forensic Automatic Speaker Recognition

Analyzing codecs for calibration and their impact on system performance

Authors:
V.M. Njegovec

Supervisors:
L.J. Spreeuwers
D. Meuwly
D. van der Vloed
F.W. Hahn

February, 2025

Department of Computer Science
Faculty of Electrical Engineering
Mathematics and Computer Science

Netherlands Forensic Institute
Ministry of Justice and Security

UNIVERSITY OF TWENTE.

Abstract

In this study the impact of audio codecs on the calibration performance of forensic automatic speaker recognition is analyzed, addressing challenges posed by mismatched conditions. Using the NFI-FRIDA (Netherlands Forensic Institute - Forensically Realistic Inter-Device Audio) database, a collection of speech recordings captured simultaneously with multiple recording devices relevant to forensic analysis, high quality audio samples are processed through various codecs to simulate real telephone speech and compared to actual telephone intercepts. The study uses an x-vector based automatic speaker recognition system, VOCALISE (Voice Comparison and Analysis of the Likelihood of Speech Evidence) for all experiments and system performance is measured in terms of calibration loss and cost of log likelihood ratio. The study reveals a significant performance loss due to codec mismatches and emphasizes the complexity of simulating telephone speech and replicating real world telephony conditions. Additionally, the study highlight the potential of cross-processing datasets with mismatched codecs to lower the calibration loss.

Keywords: *Forensic speaker recognition, audio codecs, calibration, automatic speaker recognition, calibration loss*

Acknowledgements

I would like to express my gratitude to the Netherlands Forensic Institute for providing me with the opportunity to carry out my thesis project as an intern. This study would not have been possible without their resources and support.

I am especially thankful to my primary supervisors, Didier Meuwly and David van der Vloed, for their invaluable guidance, constructive feedback, and continuous support throughout the project. I also extend my thanks to Luuk Spreeuwers and Florian Hahn, who served as my supervisors and provided administrative support during this journey.

Contents

List of Figures

List of Tables

Acronyms

AMR-NB - Adaptive Multi-Rate Narrowband
AMR-WB - Adaptive Multi-Rate Wideband
ASR - Automatic Speaker Recognition
Cllr - Cost of Log Likelihood Ratio
Closs - Calibration Loss
DET - Detection Error Trade-off
EER - Equal Error Rate
EVS - Enhanced Voice Services
FAR - False Acceptance Rate
FASR - Forensic Automatic Speaker Recognition
FRR - False Rejection Rate
FSR - Forensic Speaker Recognition
LLR - Log Likelihood Ratio
LR - Likelihood Ratio
MFCCs - Mel-Frequency Cepstral Coefficients
NFI - Netherlands Forensic Institute
PLDA - Probabilistic Linear Discriminant Analysis
RMEP - Rate of Misleading Evidence in favor of the Prosecution
RMED - Rate of Misleading Evidence in favor of the Defence
VAD - Voice Activity Detection

Chapter 1

Introduction

This chapter introduces the significance of forensic speaker recognition (FSR) in forensic investigations, outlining its challenges and discussing the critical role of representative data, the impact of mismatched conditions, and the importance of calibration in automatic speaker recognition (ASR) systems. The research objective is detailed to provide a comprehensive context for the study.

1.1 Problem Statement

The Netherlands Forensic Institute (NFI) is a world leading forensic laboratory, working with a variety of forensic analysis [1]. One of their divisions focus on digital and biometrical traces, playing an important role in modern forensic investigations and examinations given the increase of digital devices and biometric technologies. One area of casework and research is on speaker recognition, which entails examining if a speaker is the source of a questioned recording.

Speaker recognition has long been an interest in forensic practice. In cases where the primary evidence consists of a recorded speech utterance, the goal is to compare this recording with that of a potential suspect to determine the likelihood that both recordings originate from the same speaker. Investigations that rely heavily on speaker recognition often do so due to a lack of other substantial evidence, which further increases the importance of the accuracy of these methods. Reliable speaker recognition is crucial to inform the court correctly and avoid errors that could unfairly influence the court's sentencing of a suspect. However, FSR is a difficult task [11].

Analyzing speech as a biometric trace is inherently challenging because the sound and characteristics of a person's voice can vary. Voice changes based on the state of the speaker, someone who is tired will not sound the same as when they are energetic, emotional state such as happiness, anger or sadness will affect the voice, as well as different speaking conditions such as whispering or shouting and even intentional altering of the voice. Additionally, external factors can influence the quality of the capture, transmission and recording of the speech. Amongst others, background noise, overlapping speech, low-quality recordings and distortions must all be considered.

In many applications of speaker recognition, recordings can be made in controlled environments to ensure high-quality speech samples. However, this is not possible for forensic traces. Forensically realistic speech recordings come from uncontrolled environments, such as intercepted phone calls or collected voice notes. This results in challenging and frequently mismatched conditions between samples, creating a unique problem when applying speaker recognition in forensic cases [4].

FSR can be performed in a variety of ways. Either by trained forensic practitioners listening to the recordings and comparing the speech characteristics, or by using machine learning models in an automatic approach. At the NFI, a combination of auditory-acoustic analysis and automatic analysis is used [28]. While humans are good at distinguishing between speakers, ASR allows for the efficient comparison of large amounts of data while maintaining a scientific and consistent approach. However, improvements and further research is required for ASR to be more broadly applicable. At the NFI, ASR is used in almost a third of their forensic speaker comparison cases. The biggest limiting factor in expanding ASR use is the lack of representative data - models need to be trained with data that closely match the conditions of the case recordings [28].

Without representative data that corresponds to the specific case conditions, system performance becomes less reliable. The model will compare the recordings and generate a score, but if the model is trained on data produced under different circumstances, the score might be influenced by those differences rather than solely reflecting the actual speaker similarity [26].

The forensic practitioner is responsible for deciding what should be used as representative data, but answering this question is not straightforward. There is a multitude of various conditions to be considered, and it is not always clear which have a significant impact on system performance. The choice of representative data becomes a subjective choice made by the practitioner [26] and to ensure that the results generated by ASR systems are scientifically reliable it is important to conduct proper research on the effects of mismatched data and explore methods to mitigate any potential negative impact.

Lastly, this raises the question of what we can do if we don't have enough representative data for a particular case. Do some speaker conditions have minimal impact on system performance and can thus be used as representative data despite not being a perfect match to the case audio? Is it possible to augment the data to better represent the case audio, or to process it using specific method to mitigate the negative effects of mismatched data? It is important to recognize that a score produced by an ASR system does not necessarily mean much on its own, and care has to be taken into calibrating the models to the specific case.

1.2 Research Objective

The objective of this research is to investigate the impact of mismatched data and possible compensation strategies to enhance the performance of forensic automatic speaker recognition (FASR) systems, focusing particularly on the interchangeability of calibration. Given that forensic practitioners must make subjective judgments regarding representative data and how to interpret the score, this research aims to support these

decisions with empirical evidence. This study can contribute to a better understanding of the significance of various recording conditions and how to compensate for them, potentially enabling the NFI to extend its use of ASR. Due to the large variety of recording conditions, this research will focus on audio codec degradation.

To guide this research, the following research questions have been formulated:

1. What impact do different audio codecs in the calibration data have on the accuracy of an ASR system?
2. Can some audio codecs be used interchangeably in the calibration data of an ASR system?
3. Can we simulate real telephone speech by applying audio codecs to high-quality audio without significant performance loss?
4. Can we compensate for any potential performance loss using data augmentation?

Chapter 2

Background

The following chapter introduces forensic speaker recognition and the process involved in an automatic system, to gain a foundational understanding of its implementation and general challenges. Essential concepts such as calibration, likelihood ratio and validation are explored together with metrics that will be used in the study. Additionally, the challenges posed by a lack of representative data are assessed, along with a discussion on interchangeability of calibration sets as well as audio codecs and their potential effect on system performance.

2.1 Forensic Speaker Recognition

Speaker recognition is a type of biometric technique that utilizes speech to distinguish one person from another [11]. Biometrics refer to the science of distinguishing individuals based on their physiological or behavioural traits [13], i.e., based on who they are rather than objects or information they have access to. Physiological traits include physical attributes such as fingerprints or facial features, whereas behavioural traits encompass patterns in behaviour and interactions with the environment, such as body movements or how one writes their signature. Speech is a type of biometric trait that combines both physiological and behavioural characteristics and can be influenced by many variabilities. Not only is the sound of your voice dependent on the physical features of your vocal tract, but it is also influenced by behaviours and mental states such as emotions, health and dialect.

Forensic science involves applying scientific principles and techniques to investigate crimes and gather evidence that can be presented in a court of law. Due to the widespread use of audio recording devices, speaker recognition has become an important tool within forensic science. During a case where the primary evidence is an audio recording, and there is a suspected person, FSR is typically applied to determine if the questioned speech recording originates from that suspected person. FSR is performed by comparing the questioned recording with a known recording of the suspect and analyzing the characteristics between the two speech samples [11]. When carrying out this process, both the behavioural aspects that can alter the speech as well as external factors affecting the recordings, must be considered, making FSR a particularly challenging task.

When presenting evidence in court, it is not the forensic practitioner's role to determine the suspect's innocence or guilt. Therefore, they should refrain from providing a definitive answer regarding whether the suspect made the incriminating recording. Instead, they should provide an objective statement about the strength of evidence at the source level, indicating how much more likely the observed evidence is under one hypothesis compared to an alternative hypothesis.

When using an ASR system, the forensic practitioner inputs both the questioned speech audio and the known speech audio into the software. The system compares the samples and generates a comparison score. This score alone does not indicate the strength of evidence. The practitioner evaluates the score in light of the two hypotheses, the prosecutor's hypotheses H_p and the defence's hypotheses H_d [22], typically stated as:

H_p : The two speech samples originate from the same speaker
H_d : The two speech samples originate from different speakers

By calculating the relative likelihoods of observing the score given these two hypotheses, the practitioner can derive the likelihood ratio (LR), which represents the strength of evidence. The LR can be presented in court, where the judge can consider it alongside other relevant information to determine the outcome of the case.

Further details on how the ASR system generates the comparison score and derives the LR will be provided in the subsequent sections.

2.2 Forensic Automatic Speaker Recognition

In a typical FASR scenario, the forensic practitioner is tasked with comparing two case recordings, the questioned speech recording (the evidence) and the known speech recording (the suspect). Before the recordings can be compared in a meaningful way, the practitioner must first train the model specifically for the case. In general, there are two datasets required for this, a calibration set and a validation set. The calibration set is used to calibrate the model to the specific case conditions and compute the final result, the LR, while the validation set is used to ensure the model's accuracy. If a sufficient amount of data is available, a set for reference normalization can also be used to compensate for minor mismatches in the data [26].

For further details on calibration and the likelihood ratio as well as validation and performance metrics, see sections 2.3 and 2.4.

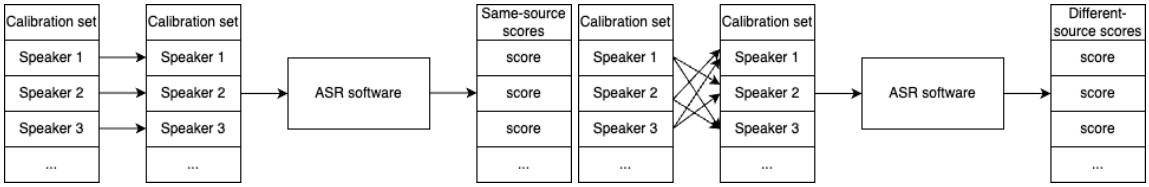
To start with, the forensic practitioner must select the calibration data. This data should be representative of the case [26], meaning that the speakers should belong to the relevant population and the audio and speaker conditions should correspond to those of the case recordings. The relevant population may include speakers of similar background such as gender, approximate age and cultural background. Representative speaker and audio conditions may include e.g., same spoken language, similar speech duration and the same type of recording device used.

Once the data has been selected, the forensic practitioner typically performs

preprocessing to prepare it for analysis. This step often involves e.g., editing audio samples to remove non-active speech segments and trimming them to match the speech duration of the case sample. Previous studies have demonstrated that a difference in duration between the calibration and evaluation datasets can significantly impact system performance [21]. After preprocessing, the data is divided into calibration, validation, and, if necessary, reference normalization datasets, each containing known speaker pairs.

Following the preparations, the forensic practitioner inputs the calibration data into the system. The system performs feature extraction to capture the speaker dependent features of the voice. One of the most commonly used features in FASR systems is Mel-Frequency Cepstral Coefficients (MFCCs) [26]. These are numerical representations of the speech signal that reflect the physical properties of the speaker's vocal tract, including the shape and size of the vocal cords and resonating cavities. The extracted features are then used to create speaker dependent models.

The generated speaker models are compared, producing a similarity score for each pair of samples. For the calibration data, sets of same-source scores and different-source scores are produced, see Fig. 2.1, serving as a ground truth used to calibrate the final case score into an LR [28].



(A) Same-speaker comparisons (B) Different-speaker comparisons

FIGURE 2.1: Same-speaker and different-speaker comparisons

The validation set is used to assess the system's reliability, ensuring that the output is suitable for use in court [20]. The validation set should contain known speaker pairs, but this information must remain hidden from the system during testing. Comparisons are carried out as previously described, and the resulting scores from the validation set are analyzed together with the scores from the calibration set to produce a set of LRs. To determine the LRs for the validation set, the same-source and different-source scores from the calibration set are plotted as separate distributions. Each test score from the validation set is then evaluated relative to these distributions, and the LR is determined based on its position within the graph, see Fig. 2.2.

This results in a set of same-speaker LRs and a set of different-speaker LRs. The system performance can then be assessed by comparing the values of the LRs with the ground truth of the speaker pairs, small LR values for same-speaker pairs and large LR values for different-speaker pairs indicate good system performance [20]. If the system performance is satisfactory, the case samples can be processed using the same method as the validation data, arriving at a final LR.

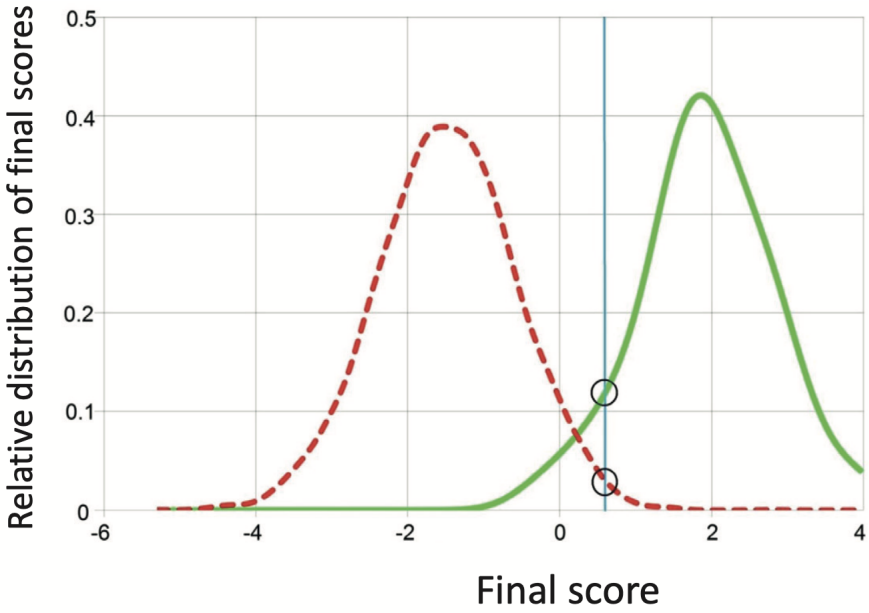


FIGURE 2.2: Score distributions and LR calculation [28]. The calibration scores determine the same-speaker curve (green) and the different-speaker curve (red), the LR is computed using the probability densities of the two distributions at the test score's location.

2.3 Calibration and Likelihood Ratio

The comparison score computed by an ASR system indicates the degree of similarity between the two speech samples, but does not provide information on how likely it is that the samples originate from the same speaker. As discussed in the previous section, calibration is used to convert these scores into LRs through a score-to-LLR function, using the calibration score distributions shown in Fig. 2.2. By interpreting the score using an LR method, we can assess the strength of the evidence in light of the two hypotheses (H_p) and (H_d). The calibration set used must contain data of similar conditions to the case data, if not, the LR may end up being too high or too low [26].

This approach enables the forensic practitioner to assess the strength of the evidence, comparing how likely it is to observe the score if the samples were made by the same speaker against how likely it is to observe the score if the samples were not made by the same speaker. Consequently, the forensic practitioner can make an objective statement regarding the evidence and express it probabilistically. Thus, the LR helps answer the question of to what extent the evidence favours one hypothesis over the other. The LR can be described by the following formula:

$$LR = \frac{P(E|H_p, I)}{P(E|H_d, I)} \tag{2.1}$$

Here, the probability of hypothesis (H_p) against hypothesis (H_d) is determined, where $P(E|H_p)$ and $P(E|H_d)$ represents the probability density functions of the same-source

and different-source score distributions respectively, evaluated at the observed evidence E. I in this context represents other information relevant to the case [6].

The LR is based on a Bayesian framework, which allows for updating the belief in a hypothesis as new information is introduced. The judge can combine prior background knowledge with new data to derive a posterior probability or belief of the hypothesis [6].

The prior odds is given by:

$$\frac{P(H_p|I)}{P(H_d|I)} \tag{2.2}$$

That is, the probability of (H_p) being true over the probability of (H_d) being true, given the background information of the case.

Using the prior odds and the LR, the posterior odds can be derived as [6]:

$$\frac{P(H_p|E,I)}{P(H_d|E,I)} = \frac{P(E|H_p,I)}{P(E|H_d,I)} \times \frac{P(H_p|I)}{P(H_d|I)} \tag{2.3}$$
$$\text{posterior odds} = \text{LR} \times \text{prior odds}$$

2.4 Validation and Performance Metrics

Validation in FASR aims to answer the question of whether a system is performing well enough to use its output in court. When developing a FASR method, validation can be split into two distinct phases, technology validation and application validation [23].

The aim is to test the system under conditions as similar as possible to real forensic casework. For this process, a validation set is used. Like the calibration set, this dataset should contain same-speaker pairs and different-speaker pairs. It is however crucial that the system being validated does not know the status of each pair. Additionally, the validation set should not contain any of the same speakers used in training the model [20].

Technology validation is used to assess the systems discriminative power, i.e., how well the model separates between the competing propositions [23]. To assess the discriminative power Equal Error Rate (EER) is used together with Detection Error Trade-off (DET) plots for graphical representation. EER measures how well the system discriminates between same-speaker and different-speaker pairs (i.e., does the system recognize same-speaker pairs better than different-speaker pairs?). There are two types of errors relevant to this context, false acceptance and false rejection. A false acceptance error happens when a different-speaker pair is misclassified as a same-speaker pair, whereas a false rejection error occurs when a same-speaker pair is misclassified as a different-speaker pair [17]. EER indicates the threshold where the false acceptance rate (FAR) meets the false rejection rate (FRR). The EER can be visualized using a DET plot, which plots FAR as a function of FRR, showing the trade-off between the two errors. The intersection of the DET-curve marks the EER [19].

Application validation [23] is the second step, where LRs are produced and evaluated for how well calibrated they are. To assess the performance of the LR system, errors are measured in terms of the rate of misleading evidence in favor of the prosecution (RMEP) and the rate of misleading evidence in favor of the defence (RMED). The Cost of Log Likelihood ratio (C_{llr}) is used [20]. This metric assess how well the LRs reflect the ground truth of the speaker pairs (same-source or different-source), and how often the system might produce misleading evidence.

C_{llr} is defined by the following formula:

$$C_{llr} = \frac{1}{2} \cdot \left(\left[\frac{1}{N_{Hp}} \sum_{i}^{N_{Hp}} log_2 \left(1 + \frac{1}{LR_{Hp_i}} \right) \right] + \left[\frac{1}{N_{Hd}} \sum_{j}^{N_{Hd}} log_2 \left(1 + LR_{Hd_j} \right) \right] \right) \tag{2.4}$$

Here, N_{Hp} represents the number of samples where (H_p) is true, and N_{Hd} represents the number of samples where (H_d) is true. Likewise, LR_{Hp} represents the LR values of samples where (H_p) is true and LR_{Hd} represents the LR values of samples where (H_d) is true. A C_{llr} value closer to 0 indicates a better performing system [30].

In addition to C_{llr}, calibration loss (C_{loss}) can be used to quantify the relative difference between C_{llr_T}, which represents the actual C_{llr} calculated on a specific test set, and C_{llr_M}, which represents the C_{llr} produced when using a perfectly matched calibration set for that test set. This metric illustrates the extent to which the calibration performance deviates from the optimal performance achievable with an ideal calibration set.

C_{loss} is calculated using the following formula:

$$C_{loss} = \frac{(C_{llr_T} - C_{llr_M})}{C_{llr_M}} \tag{2.5}$$

2.5 Lack of Representative Data

One of the primary challenges in FASR is the lack of representative data. Calibration is a critical step in FASR, as it ensures that LRs accurately reflect the strength of the evidence under real world conditions. However, achieving reliable calibration relies heavily on the availability of representative data. For calibration data to be forensically realistic, it must accurately reflect the acoustic, environmental, and device conditions of the case audio, which can vary significantly. Additionally, forensic practitioner must consider which conditions are likely to affect the model's comparison performance. This decision involves subjective judgment, as including audio with mismatched conditions could potentially cause the system to misrepresent the strength of the evidence, resulting in misleading results, while excluding them may result in insufficient data to effectively use ASR [27].

The lack of representative calibration data has practical implications for forensic practitioners. Without calibration datasets that reflect the conditions of case data, the

LRs generated may lack reliability. This issue is heightened by the challenges of acquiring forensic data that accurately represent real cases.

2.5.1 Interchangeability

Previous studies suggest that device variability can be particularly challenging and should be considered during calibration [29]. Similarly, mismatches in speech duration and microphone distance can cause significant performance loss, whereas speakers using different but closely related languages, or speakers of different genders cause minimal to no significant performance loss [21]. Thus, if a case involves a female suspect, but the forensic practitioner only has representative data of male speakers, they may theoretically still use the male speaker data for the calibration set. This demonstrates the interchangeability of these datasets, meaning that their score-to-LLR functions produce similar LRs from the same set of scores [27].

2.6 Codec Degradation

A codec, a blendword of coder/decoder, is a component that encodes and decodes a signal to enable more efficient file transfers. When an audio signal is transmitted over a network, an audio codec compresses the signal to save space, while simultaneously trying to maintain the quality of the signal as it is decompressed at the end-point. However, some information is typically lost during this process, degrading the quality of the audio recording [32], it is unclear how this quality reduction may affect the performance of an ASR system. Fig. 2.3 and 2.4 provide a visual representation of how running an audio sample through an Adaptive Multi-Rate Narrowband (AMR-NB) codec with a bitrate of 4.75kbps can affect the audio quality. The spectogram captures the amplitude at different frequency ranges over time, with the amplitude illustrated by the color. Amplitude represents the loudness of the audio signal, while frequency represents the pitch of the audio signal. A clean, high-quality, audio sample, seen in Fig. 2.3, was passed through a codec using FFmpeg, seen in Fig. 2.4. The resulting spectogram shows that some information loss has occurred in the process, with lower and more erratic amplitude accross all frequency ranges.

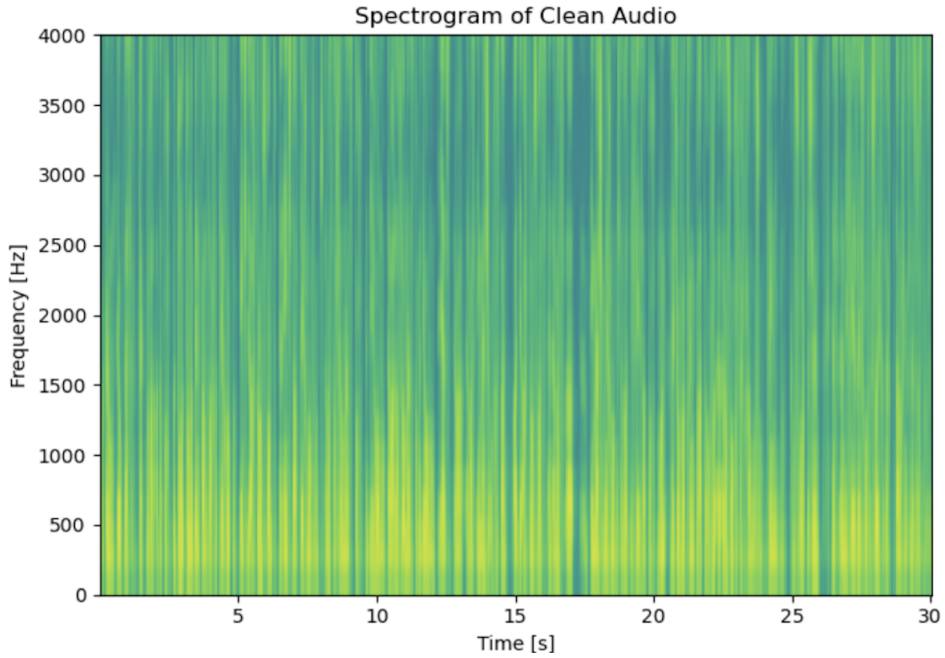


FIGURE 2.3: Spectogram of clean audio sample

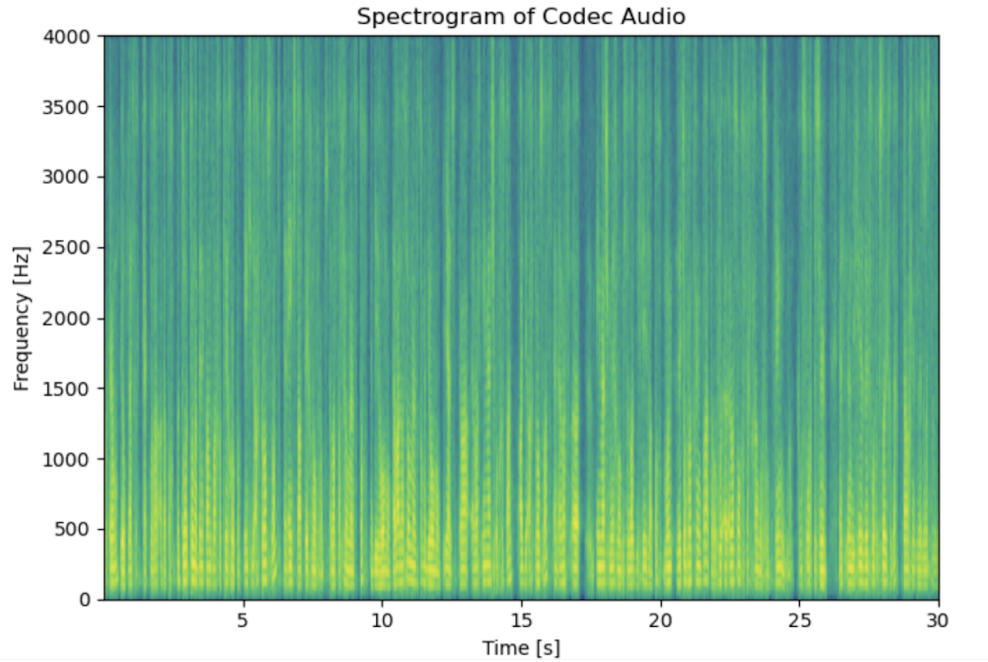


FIGURE 2.4: Spectogram of codec degraded audio sample, the same audio sample as in Fig. 2.3 was passed through an AMR-NB codec with a bitrate of 4.75kbps and a sample rate of 8000Hz

In telephony, the AMR-NB and AMR-WB (Adaptive Multi-Rate Wideband) codecs are commonly used. These codecs dynamically adjust the transmission bitrate based on the channel's capacity and current conditions. AMR-NB operates at bitrates ranging from 4.75 to 12.2 kbps, while AMR-WB ranges from 6.6 to 23.85 kbps [24]. Other widely used

telephony codecs include G.711 and Enhanced Voice Services (EVS) [16], a successor of AMR-WB.

2.7 Related Work

While previous research has examined the effects of various mismatched conditions in the calibration and evaluation data of FASR systems, studies specifically focusing on the impact of audio codecs and mismatched codecs on calibration remain limited. Some studies have investigated the impact of audio codecs on ASR systems, analyzing how codec induced distortions influence system performance when applied to training data.

[2] passed audio samples through GSM speech codecs and investigated their influence on speaker verification performance, finding that GSM coding significantly degrades both identification and verification accuracy.

[32] investigates speech recognition (distinct from speaker recognition, which identifies who is speaking rather than what is being said) using telephony data. The study applies various codecs, including G726, G722, G723.1, GSM full-rate and half-rate, Opus, and AMR-NB, to simulate real world distortions. The severity of these distortions is analyzed through spectrogram visualizations of different audio samples. The codecs are categorized into four groups, ranging from highly distorted to minimally distorted. The results show that models trained with highly distorted codec augmented data achieved better performance on real telephony data, suggesting that these codecs better reflect real world conditions.

[18] investigated the impact of codec degraded speech on a speaker recognition system based on various PLDA training scenarios. They found that including noise and reverberant speech in the PLDA model improved robustness to codec effects. Additionally, they found that the best performance was achieved when the PLDA model was trained with codec data that matched the evaluation conditions.

[14] found that a GMM-based speaker recognition system performs well for text-independent speaker identification and verification when the training and test data are encoded with the same audio codec and sample rate. Under these matched conditions, the increase in EER is minimal compared to uncoded speech. However, the system fails when there is a mismatch between the codecs of the training and test data.

These studies have explored codec induced distortions in training data for speaker and speech recognition systems, with the goal of improving recognition accuracy. In comparison, this studdy will examine the impact of codec mismatches on calibration, investigating how well the system can be calibrated under varying codec conditions to maintain reliable forensic likelihood ratios.

Chapter 3

Methodology

This chapter provides an introduction to the system and dataset used in the study, along with an overview of the data preprocessing steps undertaken. These steps include filtering the data to only include relevant samples, removing silent segments, and trimming audio samples to a consistent length. The data augmentation process is detailed, including the codecs applied and the implementation of a frequency filter to align the audio samples more closely with realistic telephony conditions.

3.1 Automatic Speaker Recognition System

For this study, VOCALISE 2019, a speaker recognition system based on the x-vector framework was used. A pre-trained model using probabilistic linear discriminant analysis (PLDA) for scoring and MFCC's for feature extraction was used [15]. This choice was made because VOCALISE 2019 is the system used locally at the NFI [28] and an x-vector framework is currently the state of the art of speaker recognition systems. PLDA outputs LRs, but due to the potential mismatches between the data that the underlying system is trained on and the case data, it is common practice to still apply calibration on the output and turn them into calibrated LRs [21]. Additionally, VOCALISE offers the option to turn on voice activity detection (VAD) to omit any potential silent parts in the audio samples, however, the datsets used were already edited and thus VAD was turned off.

3.2 Database

This study made use of the NFI-FRIDA (Netherlands Forensic Institute - Forensically Realistic Inter-Device Audio) database. This is a database collected to support forensic speaker comparison casework at the NFI [29].

The database consists of speaker-sessions recorded simultaneously across multiple devices, see Table 3.1, allowing for an in-depth evaluations of how different devices affect system performance. The database includes 250 male speakers, primarily aged 18-35.

The participants come from various backgrounds, including native Dutch, Turkish-immigrant and Moroccan-immigrant populations. However, all recordings are conducted in Dutch, ensuring that the speaker-sessions belong to a similar linguistic and demographic population. This consistency facilitates a more reliable assessment of the influence of differing recording conditions on system performance.

Device	Recording device
d1	Shure WH20 HQ Headset
d2	Shure SM58 close
d3	AKG C400BL close
d4	Shure SM58 far
d5	Intercepted telephone iPhone/Nokia
d6	Video iPhone

TABLE 3.1: NFI-FRIDA Recording devices

Each speaker underwent two days of recording with 8 sessions per day, see Table 3.2, captured across 3-6 different devices. Sessions were split evenly between indoor and outdoor environments, each containing sessions in both quiet and noisy settings. One of the recording devices used was that of an intercepted phone call, alternating between the use of Nokia 1280 and an iPhone 4. The rest of the devices used include a high-quality headset, microphones at varying distances and a video recording, resulting in recordings of varying quality and levels of background noise. For a detailed breakdown of recording sessions, refer to [29].

Session	Indoors/Outdoors	Noise level	Telephone
s1	Indoors	Quiet	Nokia 1280
s2	Indoors	Quiet	iPhone
s3	Indoors	Noisy	Nokia 1280
s4	Indoors	Noisy	iPhone
s5	Outdoors	Quiet	Nokia 1280
s6	Outdoors	Quiet	iPhone
s7	Outdoors	Noisy	Nokia 1280
s8	Outdoors	Noisy	iPhone

TABLE 3.2: NFI-FRIDA Recording sessions

3.3 Data preprocessing

Data preprocessing was performed to prepare the dataset for speaker comparisons. This process involved filtering the data to retain only samples relevant to the study, eliminating any residual silent segments and standardizing the speech duration by trimming all samples to 30 seconds. These steps ensured consistency and reliability in the subsequent analyses. These preprocessing steps were automated and executed using batch processing scripts to ensure efficiency and consistency across the dataset.

Data filtering

The objective of this study is to analyze the role of audio codecs within the calibration data of forensic speaker recognition. To achieve this, the aim was to compare three distinct types of audio samples, i.e., high-quality audio recordings, real intercepted phone calls and audio samples processed through various codecs. Therefore, there were two recording devices from the FRIDA database that were of importance for the study, i.e., the high-quality Shure WH20 headset and the intercepted telephone. Recordings from other devices were excluded from the analysis as they were not relevant to the study's focus.

The study aimed to isolate the effects of audio codecs on system performance while minimizing the influence of other factors. Consequently, speaker sessions recorded in noisy environments as well as sessions made with the Nokia 1280 were excluded. The Nokia 1280 session were omitted due to the low quality of these recordings, which made them unsuitable for reliable analysis. By narrowing the dataset in this manner the study ensured that the analysis remained focused on codec related variations.

Ultimately, the data used for this research consisted of samples recorded with the Shure WH20 HQ Headset (device d1) and the intercepted iPhone (device d5) under controlled conditions, session s2 (indoors/quiet) and session s6 (outdoors/quiet). The final selection of audio samples is detailed in Table 3.3 and 3.4, marked with green. This selection resulted in a total of 710 recordings for both the high-quality headset and the intercepted telephone samples respectively, distributed across 210 speakers. Each speaker contributed up to 8 recordings.

Device	**Recording device**
d1	Shure WH20 HQ Headset
d2	Shure SM58 close
d3	AKG C400BL close
d4	Shure SM58 far
d5	Intercepted telephone iPhone/Nokia
d6	Video iPhone

TABLE 3.3: Selected devices

Session	**Indoors/Outdoors**	Noise level	Telephone
s1	Indoors	Quiet	Nokia 1280
s2	Indoors	Quiet	iPhone
s3	Indoors	Noisy	Nokia 1280
s4	Indoors	Noisy	iPhone
s5	Outdoors	Quiet	Nokia 1280
s6	Outdoors	Quiet	iPhone
s7	Outdoors	Noisy	Nokia 1280
s8	Outdoors	Noisy	iPhone

TABLE 3.4: Selected sessions

Removing silent parts

Despite prior editing of the dataset some audio samples were found to still contain silent segments. To address this issue all samples were processed using the silence filter in SoX Sound eXchange [5]. This approach was chosen over the built-in VAD filter in VOCALISE as the latter does not guarantee uniform speech duration across samples. Ensuring consistent duration was an important requirement for this study as it allowed for more reliable comparisons during speaker analysis.

Trimming samples

Since it has been shown in previous reserach that a mismatch in speech duration between calibration data and evaluation data can have a significant impact on system performance [21], all samples were trimmed to the same length of 30 seconds using the -t flag in FFmpeg [8].

3.4 Data augmentation

The next step involved augmenting the high-quality speech samples by processing them through various audio codecs. This was done to evaluate the calibration performance of different codecs and to simulate real telephone speech. The goal was to determine whether simulated telephone speech, produced by applying codecs to high-quality recordings, could achieve comparable calibration performance to real intercepted datasets, thereby enabling their interchangeable use in forensic applications. The codecs were applied to the high-quality samples using FFmpeg [9].

Codecs

The codecs selected for this study were AMR-NB, AMR-WB and G.711, as these are among the most widely used telephone codecs in modern communication systems. EVS was another codecs also considered for the study, but ultimately it was excluded from analysis due to the lack of a reliable implementation method. The G.711 codecs was pre-configured within the FFmpeg framework [9], whereas the AMR codecs required the intergration of external libraries. Specifically, OpenCORE AMR was used for AMR-NB [7], and VisualOn was used for AMR-WB [25].

For AMR-NB and G.711, audio files were encoded with a sample rate of 8000 Hz, while AMR-WB files were encoded at 16000 Hz. These sample rates were selected to align with real world telephony standards. During encoding with the AMR codecs, the bitrate settings were also configured. AMR codecs operate at variable bitrates that changes dynamically during a call based on how much traffic there is in the network. For AMR-NB, the bitrate ranges from 4.75 kbps to 12.2 kbps, and for AMR-WB, it ranges from 6.60 kbps to 23.85 kbps. To investigate the impact of these varying bitrates, a separate dataset was generated for each bitrate setting across both AMR codecs. In contrast, G.711 uses a fixed bitrate, resulting in a single dataset for this codec.

In total, 20 distinct datasets were prepared for analysis. These included one dataset containing real intercepted telephone recordings, one dataset of high-quality audio, and 18 datasets of simulated telephone speech processed through various codec configurations. All datasets comprised speech from the same recording session to ensure consistency in the analyses. An overview of the available datasets is provided in Table 3.5.

Datasets
Real telephone intercepts
High quality audio
G.711
AMR-NB (4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.2, 12.2 kbps)
AMR-WB (6.60, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05, 23.85 kbps)

TABLE 3.5: Overview of datasets

Frequency filter

Finally, a frequency filter was applied to the simulated telephone speech datasets to more closely align them with real world telephony conditions. Before this adjustment, the score distributions of the simulated datasets were visualized and compared against those of the real intercepted telephone recordings. The curves of the simulated datasets did not fully align with the real intercepts, see Fig. 3.1, indicating a discrepancy in the audio characteristics. To address this issue, a consistent frequency filter range of 180 Hz to 3600 Hz was applied across all simulated datasets. This range was selected based on the typical bandwidth characteristics of the codecs, however, it does not fully reflect the broader bandwidth of AMR-WB. Despite this limitation, the chosen filter range made for consistent processing and comparison across all datasets. In real world conditions the bandwidth of G.711 is 300–3400 Hz, AMR-NB operates within 200–3400 Hz, and AMR-WB has a significantly wider bandwidth of 50–7000 Hz.

The filter range was designed to account for the roll-off effect seen in telephone systems, where frequencies near the cutoff point are attenuated rather than entirely removed. Consequently, the selected bandwidth slightly extends beyond the actual ranges of G.711 and AMR-NB to reflect this more accurately. The frequency filter was applied using the highpass and lowpass filter of SoX Sound eXchange [10].

After visualizing the score distributions, a subset of the simulated telephone datasets were selected for further evaluation based on their impact and relevance, while the remaining datasets were discarded.

3.5 Analysis

Once the data had been gathered, selected and processed, it was analyzed using the VOCALISE system, where speaker comparisons were made within each dataset. Each audio sample in a dataset was compared to every other sample within the same dataset,

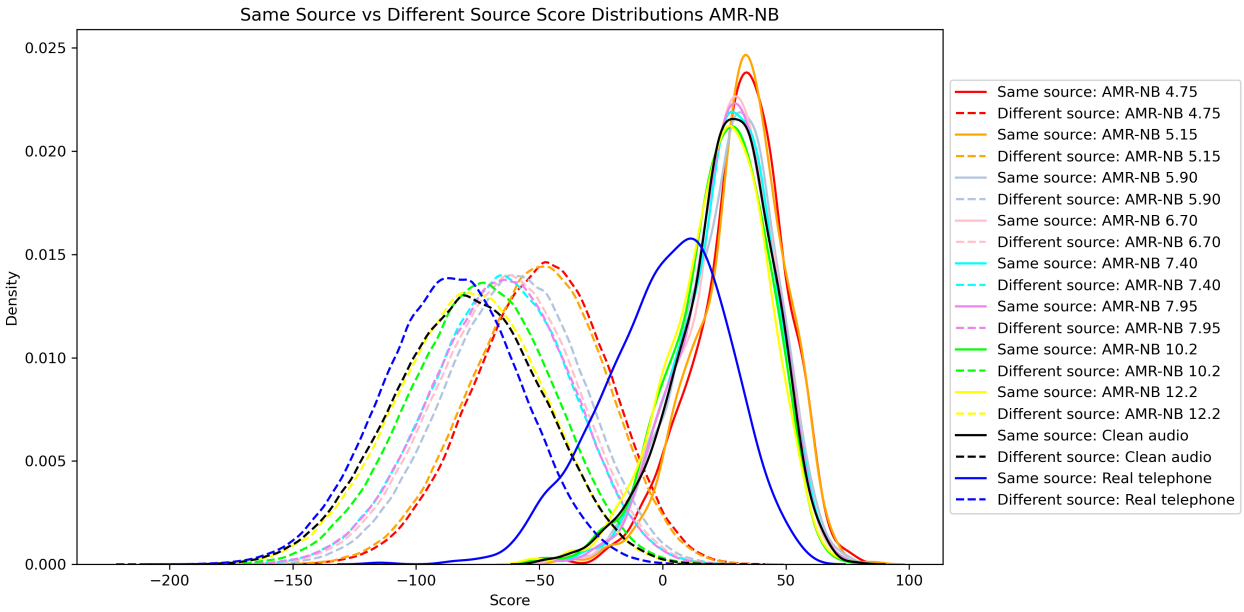


FIGURE 3.1: Score distributions of AMR-NB codecs compared to real intercepts and high-quality audio, there were similar results for the simulated AMR-WB and G.711 audio samples

generating a set of same-source and different-source scores for each dataset. These scores provided the foundation for further analysis through various metrics and visualizations.

To better understand the system's performance, several types of plots were utilized to visualize the generated scores. Score distributions were plotted to provide insights into the separation between same-source and different-source scores, offering a direct comparison of the simulated telephone datasets against the real telephone intercepts and high-quality datasets. Score-to-LLR functions were plotted to evaluate the datasets as calibration sets, these plots illustrate the relationship between raw similarity scores and their corresponding likelihood ratios, highlighting how different calibration sets influence the interpretation of the same similarity scores. For the calibration process, linear logistic regression [3] was applied using the NFI's LiR library [12], with ELUB boundaries [31] to limit extreme log likelihood ratio (LLR) values which might skew the interpretability of the results.

Lastly, C_{llr} and C_{loss} values were calculated and visualized using heatmaps to evaluate calibration performance across mismatched conditions. Each dataset was split into two parts, one for calibration and one for testing. Each test set was calibrated using all available calibration sets, including one matched set. This process was then repeated, swapping the roles of calibration and test sets. The resulting C_{llr} values were pooled to calculate the corresponding C_{loss} values.

In a final experiment, datasets processed with one codec were subjected to additional processing with another codec to create dual-processed datasets. This approach aimed to investigate the effects of initially mismatched conditions, where the calibration set is processed using codec A and the evaluation set is processed using Codec B, and to explore whether performance could be improved by subsequently matching these

conditions. Specifically, the experiment involved processing datasets with Codec A through Codec B and vice versa. The purpose was to evaluate whether aligning the codec conditions, even in varying sequences, could mitigate the impact of the initial mismatch on system performance. C_{llr} and C_{loss} values were calculated for these dual-processed datasets and visualized on heatmaps, offering insights into the extent to which codec matching can enhance calibration reliability and overall system performance.

Chapter 4

Results

This chapter presents the research findings, detailing the evaluation and analysis of calibration performance for speaker recognition systems under various audio codec conditions. Key metrics include score distributions, score-to-LLR functions, C_{llr} and C_{loss}.

4.1 Score distributions

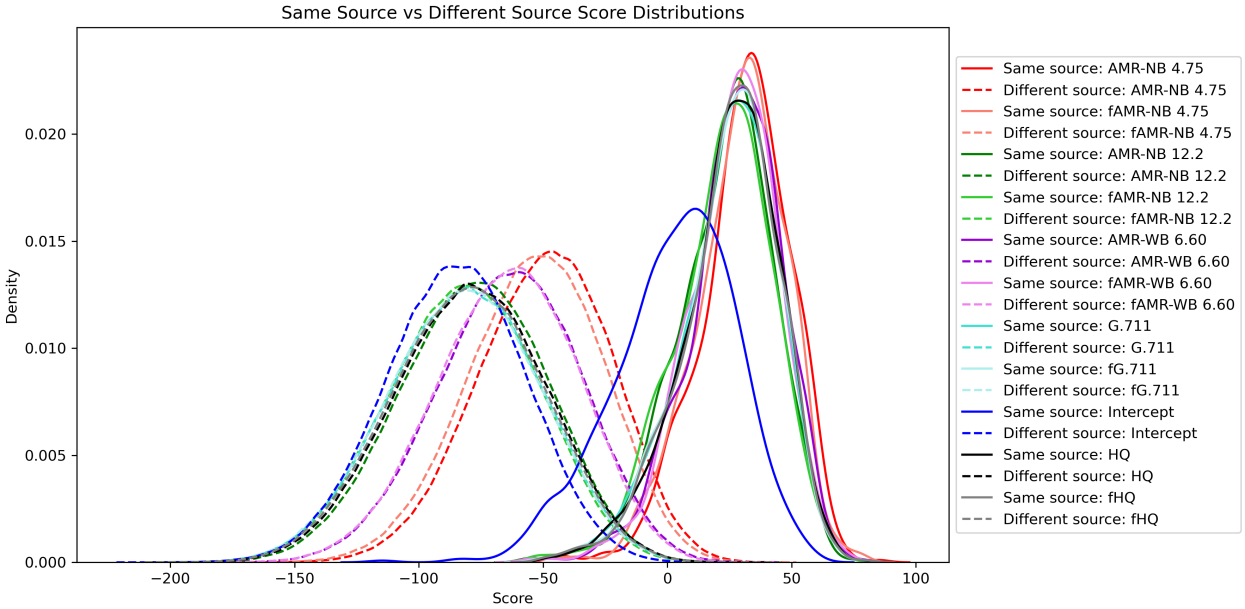


FIGURE 4.1: Score distributions of selected codecs compared to real intercepts and high-quality audio. Sets labeled with a lowercase f in the beginning of the name, such as fAMR-NB 4.75 represent a dataset processed through a frequency filter

Fig. 4.1 shows the score distributions for same-source and different-source comparisons across various datasets, including high-quality, real telephone intercepts and simulated telephone speech processed with different codecs. These distributions provide insights

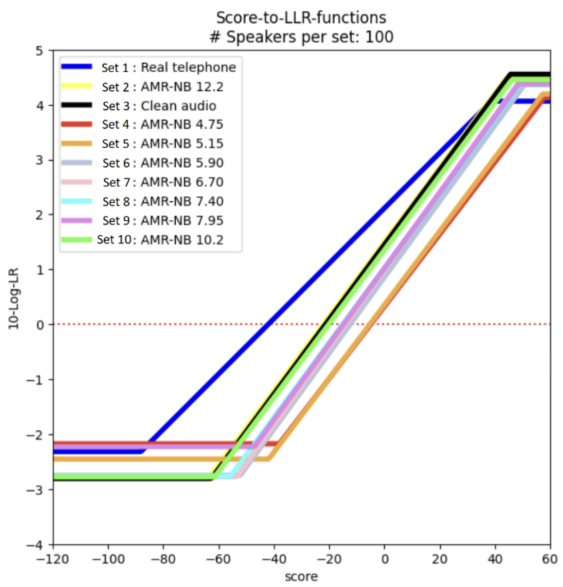
into how closely the simulated datasets align with the real intercept data and how the choice of codec impacts the system's ability to differentiate between same-source and different-source audio samples.

The high-quality dataset demonstrates a more pronounced separation between same-source and different-source scores, as indicated by the clear gap between the respective curves, in comparison to the real telephone intercepts which displays some overlap between same-source and different-source scores. This reflects the challenges posed by real world telephony conditions.

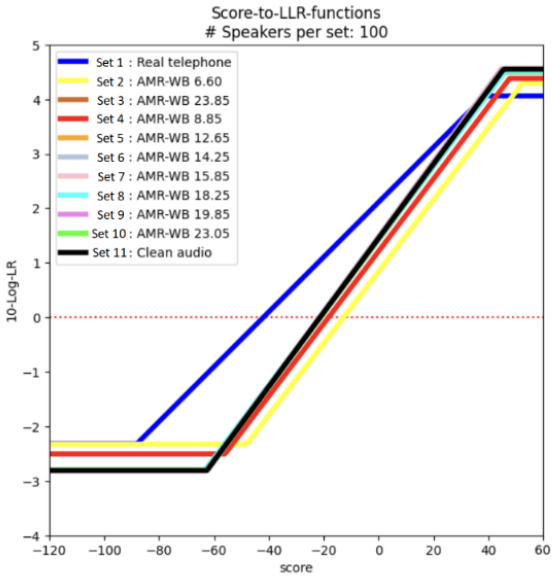
The simulated datasets, processed through various codecs, show distributions that do not fully match the curves of the real intercept dataset. The same-source and different-source score distributions indicate a closer alignment with high-quality speech rather than with real intercepts. This suggests that additional processing may be required to bring the simulated datasets closer to realistic conditions.

A frequency filter was applied to the simulated datasets to better reflect the bandwidth limitations inherent in telephone systems. However, while the datasets processed through both a codec and a frequency filter show a slightly closer alignment with the real intercept dataset, they still align closer to the high-quality dataset.

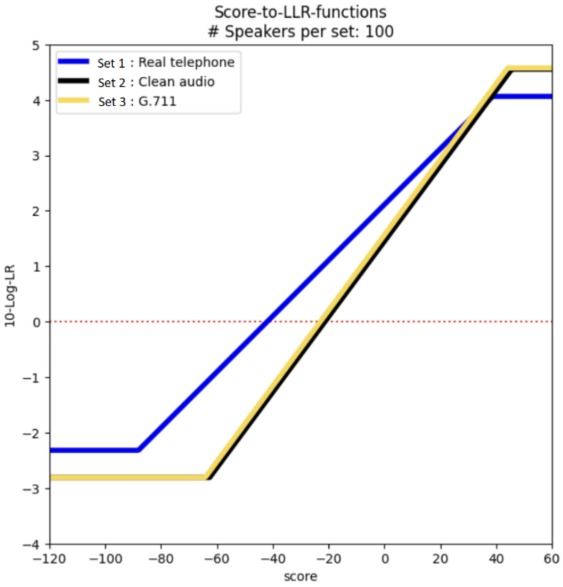
4.2 Score-to-LLR



(A) AMR-NB comparison



(B) AMR-WB comparison



(C) G.711 comparison

FIGURE 4.2: Score-to-LLR functions

Fig. 4.2 presents the score-to-LLR functions for datasets processed with AMR-NB, AMR-WB, and G.711 codecs, alongside real telephone and clean high-quality audio datasets. These plots visualize the relationship between similarity scores and their corresponding LR values, providing insights into the calibration performance of the system under different conditions. Two sets are interchangeable if their respective score-to-LLR functions generate similar sets of LRs given the same scores [27], meaning that they could be used interchangeably with each other as calibration sets.

To generate the score-to-LLR functions, the same-source and different-source scores from all datasets respectively were used to train separate calibration functions. A set of all the scores were used to calculate LRs, once using each calibration function [27].

Looking at the three functions in Fig. 4.2, the simulated datasets demonstrate a closer alignment with the high-quality audio dataset, failing to fully replicate the calibration characteristics of the real intercepts dataset. These results are consistent with those obtained from the score distribution plots, further highlighting the challenges in simulating realistic telephony conditions.

4.3 C_{loss} and C_{llr}

To evaluate the calibration performance of the speaker recognition system both C_{loss} and C_{llr} are presented. The C_{llr} value is as an absolute measure of calibration quality, representing how effectively the system transforms similarity scores into meaningful LRs. A lower C_{llr} indicates better calibration, with a value of 0 representing a perfectly calibrated system.

While C_{llr} provides a detailed and absolute evaluation, it does not show the extent of degradation relative to perfect calibration. This is why C_{loss} has been calculated alongside C_{llr}. C_{loss} expresses the relative calibration loss as a percentage, offering an intuitive understanding of how far the system's performance deviates from the ideal, and is calculated directly from the C_{llr} values.

C_{loss} alone does not provide sufficient information, as it lacks the absolute calibration context. A high C_{loss} value would suggest that the system's calibration is noticeably worse compared to an ideal calibration set, however, this does not necessarily mean that the system performs poorly. It means that its performance is significantly degraded compared to a perfectly calibrated system. Therefore, both metrics are presented together to provide a comprehensive analysis.

4.3.1 One codec

Calibration performance was first evaluated for datasets processed through a single codec. This section shows the effect that each type of codec has on calibration performance. Fig. 4.3 presents the $C_{loss}(\%)$ values for the real intercepts dataset, high-quality dataset and simulated telephone speech datasets. On the x-axis of Fig. 4.3 are the test sets whose underlying scores are converted into LRs using the calibration sets, the y-axis represents the calibration sets used for this conversion. Each test set has a corresponding perfectly matched calibration set, these values are 0 by definition. Some C_{loss} values are negative, meaning that the corresponding calibration set achieved better calibration performance than the perfectly matched calibration set.

FIGURE 4.3: $C_{loss}(\%)$ values of single codecs, real intercepts and high-quality audio. Datasets labeled with a lowercase 'f' at the beginning of their names, such as fAMR-NB 4.75, represent those processed with the frequency filter

All of the calibration sets, except for the matched set, performed poorly on the test set of real intercepted telephone calls, with high C_{loss} values accross all sets. Among the simulated telephone sets, the datasets processed with the G.711 codec outperformed the high-quality dataset when used to calibrate the intercepted telephone calls set, while all other simulated datasets underperformed in relation to the high-quality set. The datasets processed through the AMR-NB 4.75 codec displayed the highest C_{loss} value for the intercepted telephone calls test set, indiciating that these were the worst performing calibration sets for that test set.

Datasets with the frequency filter applied generally show improved calibration performance compared to their unfiltered counterparts, with the exception of the G.711 dataset which showed a slight degradation in performance when filtered. Analyzing the C_{loss} values for the column and row corresponding to the intercepted telephone calls as both the test set and the calibration set reveals that the simulated datasets fail to replicate the calibration performance of real telephone speech. Additionally, the distinction between the performance among the different codecs highlights the impact of codec selection on calibration performance.

Fig. 4.4 shows the underlying C_{llr} values, used to calculate the C_{loss} values of Fig. 4.3.

FIGURE 4.4: C_{llr} values of single codecs, real intercepts and high-quality audio. Datasets labeled with a lowercase 'f' at the beginning of their names, such as fAMR-NB 4.75, represent those processed with the frequency filter

4.3.2 Multiple codecs

Calibration performance was then evaluated for datasets processed through two codecs. The high quality dataset was first processed through one codec (referred to here as Codec A) followed by a second codec (referred to here as Codec B), and then processed in the reverse order (Codec B followed by Codec A). Fig. 4.5 shows the C_{loss} values for the real intercepts dataset, the high-quality dataset and the simulated telephone speech datasets processed through a combination of two codecs each.

As in the previous heatmap, the x-axis displays the test sets and the y-axis displays the calibration sets and each test set has a corresponding matched calibration set where the C_{loss} value is 0.

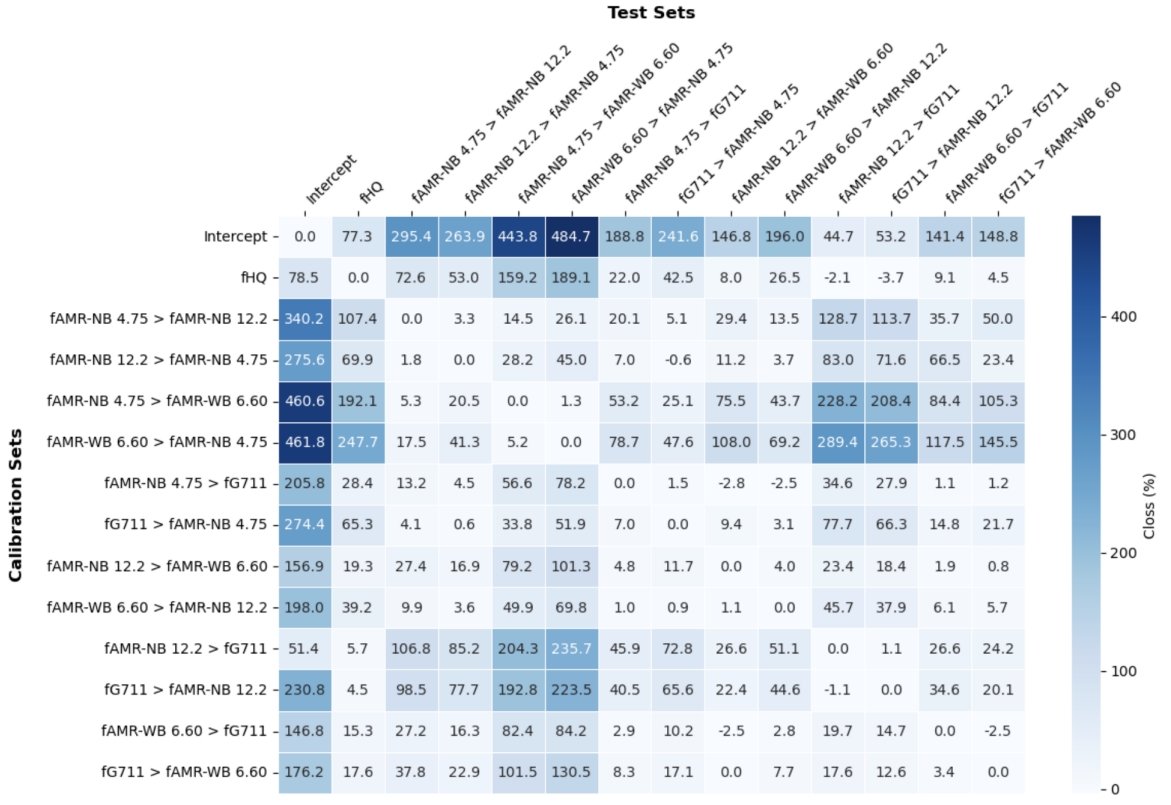


FIGURE 4.5: $C_{loss}(\%)$ values of dual codecs, real intercepts and high-quality audio. Simulated telephone speech datasets are processed through two codecs with a frequency filter each

The results presented in Fig. 4.5 and 4.3 together demonstrate that calibration performance improves when datasets encoded with different codecs are processed through each other's codecs prior to calibration.

The main comparison lies between two scenarios, i.e., (1) the calibration performance when a test set encoded with Codec A is directly calibrated using a calibration set encoded with Codec B, and (2) the calibration performance when the test set encoded with Codec A is processed through Codec B, and the calibration set encoded with Codec B is processed through Codec A. This approach of cross-processing the datasets minimizes the mismatch between the test and calibration sets, thereby enhancing calibration performance.

Calibrating the test set processed through an AMR-NB 4.75 codec (with the frequency filter applied) directly with the calibration set processed through a G.711 codec (with the frequency filter applied) results in significant calibration loss, producing C_{loss} values of 52.2% and 69.4% when reversed (fG.711 as the test set and fAMR-NB 4.75 as the calibration set), as shown in Fig. 4.3. However, when the datasets are processed through each other's codecs (e.g., fG.711 processed through the AMR-NB 4.75 codec with a frequency filter and vice versa), the calibration loss decreases drastically to 7.0% and 1.5%, respectively, as shown in Fig. 4.5.

A similar improvement is observed with fAMR-NB 4.75 and fAMR-NB 12.2. Direct

calibration between these datasets results in calibration losses of 61.5% and 59.9%, respectively, as shown in Fig. 4.3. However, processing both datasets through each other's codecs reduces the calibration loss significantly, giving C_{loss} values of 3.3% and 1.8%, as shown in Fig. 4.5.

Fig. **??** shows the underlying C_{llr} values, used to calculate the C_{loss} values of Fig. 4.5.

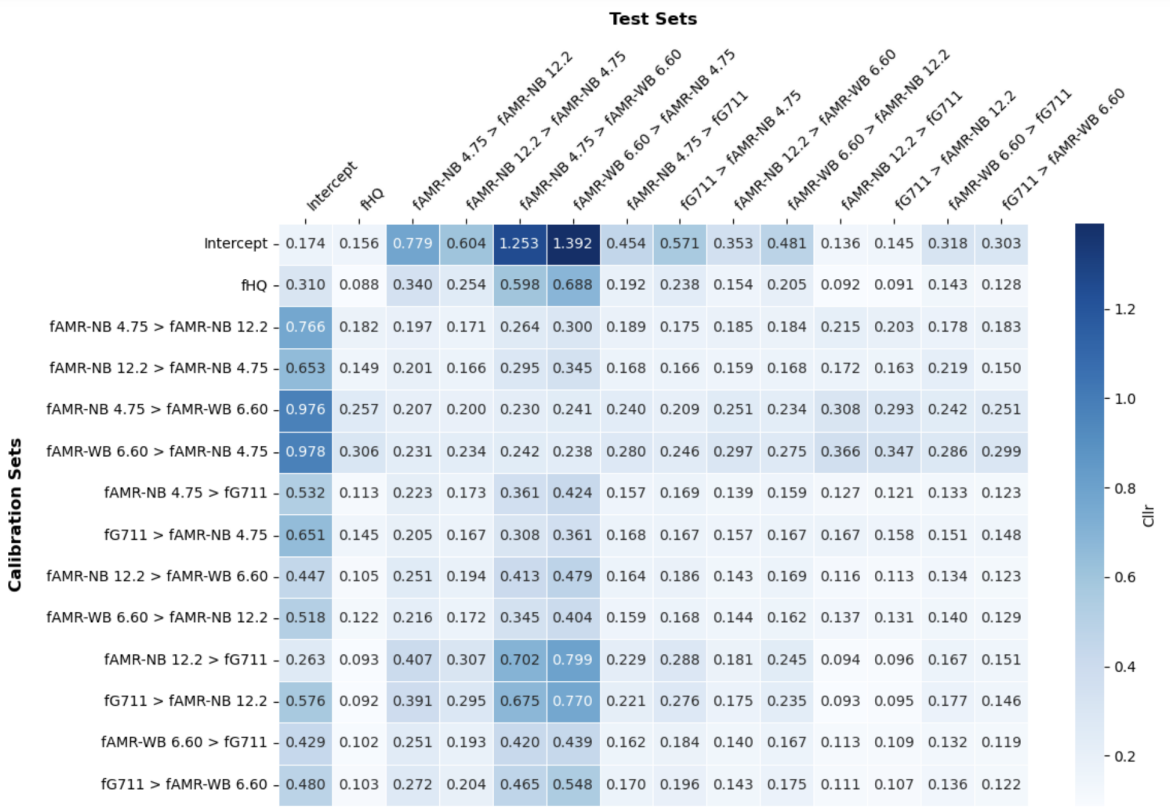


FIGURE 4.6: C_{llr} values of dual codecs, real intercepts and high-quality audio. Simulated telephone speech datasets are processed through two codecs with a frequency filter each

To further analyze the impact of cross-processing on calibration performance, C_{llr} values were evaluated for the described scenarios. Table 4.1 summarizes the C_{llr} values for direct calibration and cross-processed calibration, corresponding to the C_{loss} results shown in Fig. 4.3 and 4.5.

Calibration Type	Test Set	Calibration Set	Cllr
Direct	fAMR-NB 4.75	fG.711	0.258
Direct	fG.711	fAMR-NB 4.75	0.147
Cross-Processed	fAMR-NB 4.75 >fG.711	fG.711 >fAMR-NB 4.75	0.168
Cross-Processed	fG.711 >fAMR-NB 4.75	fAMR-NB 4.75 >fG.711	0.169
Direct	fAMR-NB 4.75	fAMR-NB 12.2	0.271
Direct	fAMR-NB 12.2	fAMR-NB 4.75	0.155
Cross-Processed	fAMR-NB 4.75 >fAMR-NB 12.2	fAMR-NB 12.2 >fAMR-NB 4.75	0.201
Cross-Processed	fAMR-NB 12.2 >fAMR-NB 4.75	fAMR-NB 4.75 >fAMR-NB 12.2	0.171

TABLE 4.1: Comparison of C_{llr} values between direct and cross-processed calibration

For the fAMR-NB 4.75 test set calibrated with the fG.711 calibration set, direct calibration resulted in C_{llr} values of 0.258 and 0.147, depending on the direction. Cross-processing reduced these values to 0.168 and 0.169, reflecting an overall improvement in calibration performance. Similarly, for the fAMR-NB 4.75 and fAMR-NB 12.2 datasets, direct calibration produced C_{llr} values of 0.271 and 0.155, while cross-processing overall reduced these values to 0.201 and 0.171.

These reductions align with the observed decreases in C_{loss}, as shown in Fig. 4.3 and 4.5. This indicates that cross-processing mitigates calibration errors caused by codec mismatches, improving both absolute calibration quality and relative performance.

The C_{llr} values from cross-processed calibration sets (e.g., 0.168 and 0.169 for fAMR-NB 4.75 and fG.711) are lower than the worst case direct calibration values (0.258) but slightly higher than the best case direct calibration values (0.147). Thus, while cross-processing may not always achieve the optimal performance of a specific direct calibration pairing, it results in more consistent C_{llr} values accross both directions and reduces the overall mismatch. Even if it slightly increases the C_{llr} in the "better" direction, the trade-off is that it significantly reduces the worst case calibration error, resulting in more balanced and robust performance.

Chapter 5

Discussion

This chapter discusses the implications of the findings presented in Chapter 4, addressing the impact of codecs on system performance, the limitations of simulated telephone speech, the potential benefits of minimizing mismatch between test and calibration sets by cross-processing and finally directions for future research.

5.1 Codecs impact on system performance

To address the first research question, codecs do have a clear impact on system performance, with different codecs influencing calibration performance to varying degrees. The heatmap of C_{lossS} values for datasets processed through a single codec in Fig. 4.3 illustrates this. When the high-quality dataset is used as the test set, calibrations sets processed through AMR-WB 6.60 and AMR-NB 4.75 codecs have a significant calibration loss, while the other simulated telephone speech datasets have an insignificant calibration loss. This suggest two key points, first, codecs can degrade calibration performance likely due to their destructive nature, altering the underlying acoustic features. Second, not all codecs have the same impact and some codecs cause more significant degradation than others.

The results also highlight a clear mismatch between not only the different codecs, but also between the varying bitrate settings within the same codec. As an example, AMR-NB 4.75 performs poorly as a calibration set on all other simulated telephone speech datasets, but the calibration loss is significantly higher when used with AMR-NB 12.2 and G.711 test sets compared to the AMR-WB 6.60 test set. This emphasizes the complexity of how different codecs and variations in bitrate settings affect the preservation of speaker specific features and may reflect how each codec encodes and compresses the audio signal differently, e.g., lower bitrate settings, such as AMR-NB 4.75, are designed for efficiency in telecommunication environments but sacrifice fine grained acoustic detail which is crucial for reliable speaker recognition. These findings demonstrate the need for careful consideration of both codec type and bitrate configuration when analyzing forensic speaker recognition systems.

5.2 Interchangeability of calibration sets processed through different audio codecs

The alignment of the score-to-LLR functions shown in Fig. 4.2 indicates that the calibration sets produce similar transformations of scores into LRs. This suggests that, in theory, the simulated datasets and the high-quality dataset could be used interchangeably for calibration without significant differences in the LRs produced from the same scores. Essentially, the mapping of scores to LRs for these datasets is comparable. However, the interchangeability of calibration sets should not be treated as a universal conclusion but rather as a context dependent judgment that must account for the specific contexts of the case.

Despite the general alignment of the score-to-LLR functions, there are some deviations to be considered. The AMR-NB codecs, particularly at lower bitrates, and the AMR-WB 6.60 codec deviate more noticeably from the score-to-LLR function of the high-quality dataset, as opposed to the G.711 codec which is more closely aligned with the high-quality dataset. This is also consistent with the results from the score distributions plot in Fig. 4.1 and the C_{loss} values presented in Fig. 4.3 where the calibration loss from using AMR-NB 4.75 and AMR-WB 6.60 to calibrate the high-quality dataset is higher compared to AMR-NB 12.2 and G.711, although all simulated datasets show relatively low C_{loss} values when compared to the higher calibration loss observed when the real telephone intercepts are used either as the test set or as the calibration set with the simulated datasets. These results also consistently indicate that the simulated telephone speech datasets and the high-quality dataset are not interchangeable with the real telephone intercepts dataset. The score-to-LLR functions for the real intercepts deviate significantly from those of the simulated datasets and the high-quality dataset, as shown in Fig. 4.2.

5.3 Limitations of simulated telephone speech

One of the key findings of this research is the inability to fully simulate real telephone speech by applying codecs and a frequency filter to high quality audio. While these steps mimic some aspects of telephony conditions, real telephone speech is influenced by a multitude of additional factors that are not accounted for in this experiment. This highlights the inherent limitations of relying on simulations when working with forensic audio. Real telephone audio is subject to complex processes and environmental influences that extend beyond the application of codecs.

Real telephone calls are subject to dynamic codec changes, since, e.g., the AMR-NB and AMR-WB codecs are not static, but they instead adapt dynamically to network conditions. This can cause variations in bitrate and compression artifacts within a single call, during periods of high network traffic codecs may lower the bitrate resulting in additional audio degradation that is not captured in a fixed simulation like in this experiment. Another factor that is not captured in this experiment is that of base station transitions, audio signals often pass over multiple base stations during a call, when, e.g., the user physically moves around. Transitioning between base stations can cause disturbances such as packet loss, jitter or temporary drop in signal quality. A real

telephone call also include environmental factors such as background noise and reverberations which are absent in the simulated datasets of this study, these are factors that can further obscure speaker specific features that are important for the speaker recognition system to make reliable distinctions. Lastly, real world telephone communication involves simultaneous processing of incoming and outgoing audio streams, and all of these mentioned factors come from each side of the phone call, potentially causing cumulative distortions.

The more pronounced gap between same-source and different-source score distributions in the high-quality dataset compared to the real intercepts dataset seen in Fig. 4.1 illustrates these challenges. High-quality audio is recorded using high-fidelity equipment, ensuring a cleaner and more detailed signal. This allows the ASR system to extract speaker specific features with greater precision, resulting in more distinct and separated scores for same-source and different-source comparisons. As seen in the results of the comparisons made with the simulated telephone speech datasets, processing through codecs and applying a frequency filter was insufficient to fully replicate the technical factors of a real telephone call and there is more to it that affect the ASR system performance. While simulations can provide valuable insights under controlled conditions, they fall short of capturing the full complexity of real world telephony.

5.4 Reducing performance loss with dual codecs

Another finding of this research is the potential to reduce mismatched codec calibration performance loss by processing datasets through both codecs.

Imagine a forensic scenario where calibration data and test data contain different telephony conditions, such as different codecs, e.g., if the evaluation data is encoded using Codec A but the available calibration data is encoded using Codec B. Using dual codec processing could prove valuable in helping to mitigate the mismatch and produce more reliable LRs. Additionally, the concept of cross-processing raises a broader question, can the strategy of mixing mismatched data to create matched conditions be applied to other mismatched scenarios beyond codecs? For instance in cases of mismatched environmental noise conditions, it may be possible to "normalize" datasets by introducing comparable noise characteristics to both calibration and test data. Similarly, mismatches in recording equipment, such as microphone types, could potentially be mitigated by processing audio through filters or transformations that emulate the characteristics of the other device. These ideas could be explored further to determine whether the principle of cross-processing extends to other mismatched conditions impacting calibration performance.

Lastly, this finding raises a critical question, should practitioners deliberately reduce audio quality by, e.g., introducing an additional codec to improve calibration performance? On one hand, this approach enables more reliable LRs which can be highly beneficial when forensic practitioners face mismatched calibration and test data. On the other hand, deliberately degrading audio quality compromises the integrity of the evidence by reducing its original detail.

The decision to cross-process mismatched conditions likely depends on the context of the

case. In scenarios where reliable calibration is paramount and codec mismatches are unavoidable, cross-processing offers a pragmatic solution. However, the trade-off in audio quality must be carefully considered, particularly when the integrity of the evidence is critical.

5.5 Future research

Future research in this area holds significant potential for advancing the field of forensic speaker recognition and aiding practitioners in making more informed subjective judgments when selecting representative data for calibration. The multitude of conditions that can affect calibration performance makes this an ongoing process with no clear endpoint. This highlights the importance of continually expanding the scope of research to investigate how different factors influence system calibration and reliability. While this study focused on the effects of audio codecs, future work could explore other conditions that may impact performance. These conditions could, e.g., include environmental factors such as varying levels of background noise or reverberation, or new factors that have not been commonly considered such as the impact of speaker emotion or speech styles. Another promising direction for future research is the exploration of combinations of conditions and how they interact with each other. Given the endless possibilities of conditions and their potential interactions future studies are crucial in broadening the understanding of how FASR systems perform under diverse conditions.

Researching cross-processing with other conditions presents another intriguing opportunity for future work. While this study demonstrated the benefits of cross-processing datasets through mismatched codecs to reduce calibration loss, this concept could be extended to other mismatched conditions. By extending the principle of cross-processing to a wider range of mismatched conditions, future research could provide a more generalizable framework for improving calibration performance in diverse forensic scenarios.

Additionally, the search for a well simulated telephone speech dataset remains an interesting area for future research. Current simulations are limited in their ability to fully replicate realistic telephony conditions. Future work could build on the understanding of how various codecs impact performance by incorporating more complex variables, for instance, dynamic bitrate changes within single audio samples could be simulated to reflect how real codecs adapt to fluctuating network conditions. Similarly, adding background noise and reverberations could make the datasets more environmentally realistic and network specific factors, such as jitter in phone calls or packet loss, could also be included to simulate inconsistent networks and their impact on audio quality. Ultimately, these efforts would aim to bridge the gap between simulated and real telephone speech, providing forensic practitioners with datasets that better reflect the challenges of real world conditions.

Chapter 6

Conclusion

This study explored the impact of audio codecs on the calibration performance of FASR systems, focusing on mismatched conditions and methods to mitigate their effects. The findings demonstrate that audio codecs influence calibration performance with varying effects depending on the codec and the bitrate setting. While some simulated datasets and the high-quality dataset exhibited close alignment in their calibration functions, suggesting potential interchangeability, the real telephone intercept dataset remained distinct.

The research showed that while processing high quality datasets with codecs and frequency filters approximated some aspects of real telephone conditions, these simulations were insufficient to fully replicate the complexity of real world telephony. This limitation highlights the influence of additional factors such as, e.g., dynamic bitrate changes, network transitions, and environmental noise on the performance of FASR systems.

Additionally, the introduction of dual codec processing was shown to mitigate calibration loss caused by mismatched datasets, providing a promising strategy for improving performance. While effective in reducing calibration loss, this approach introduces questions regarding integrity of the data.

In summary, this study contributes to a deeper understanding of the role of audio codecs in FASR system calibration and potential mitigation strategies of mismatched conditions.

Bibliography

[1] Netherlands forensic institute. https://www.forensicinstitute.nl/, 2024. Accessed: 2024-07-17.

[2] L. Besacier, S. Grassi, A. Dufaux, M. Ansorge, and F. Pellandini. Gsm speech coding and speaker recognition. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 2, pages II1085–II1088. IEEE, June 2000.

[3] Niko Brümmer and Johan Du Preez. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2-3):230–275, 2006.

[4] Joseph P. Campbell, Wade Shen, William M. Campbell, Reva Schwartz, Jean-François Bonastre, and Driss Matrouf. Forensic speaker recognition. *IEEE Signal Processing Magazine*, 26(2):95–103, 2009.

[5] Digital Cardboard. The sox of silence. https://digitalcardboard.com/blog/2009/08/25/the-sox-of-silence/comment-page-2/, August 2009. Accessed: 2025-01-14.

[6] Christophe Champod and Didier Meuwly. The inference of identity in forensic speaker recognition. *Speech Communication*, 31(2-3):193–203, 2000.

[7] Belledonne Communications. Opencore-amr library. https://github.com/BelledonneCommunications/opencore-amr/tree/master, n.d. Accessed: 2025-01-14.

[8] FFmpeg Developers. *ffmpeg - Multimedia Framework*, n.d. Accessed: 2025-01-14.

[9] FFmpeg Developers. *FFmpeg Codecs Documentation*, n.d. Accessed: 2025-01-14.

[10] SoX Developers. *SoX - Sound Exchange*, n.d. Accessed: 2025-01-14.

[11] John H.L. Hansen and Taufiq Hasan. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, 32(6):74–99, 2015.

[12] Netherlands Forensic Institute. Lir library: Likelihood ratio computation tools. https://github.com/NetherlandsForensicInstitute/lir, n.d. Accessed: 2025-01-20.

[13] Anil K. Jain, Ruud Bolle, and Sharath Pankanti. *Introduction to Biometrics*. Springer US, 1996.

[14] T. Jiang, B. Gao, and J. Han. Speaker identification and verification from audio coded speech in matched and mismatched conditions. In *2009 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 2199–2204. IEEE, December 2009.

[15] F. Kelly, O. Forth, S. Kent, L. Gerlach, and A. Alexander. Deep neural network based forensic automatic speaker recognition in vocalise using x-vectors. In *Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics*. Audio Engineering Society, June 2019.

[16] Manfred Lutzky and Markus Schnell. Enhanced voice service (evs) codec. Technical report, Fraunhofer.

[17] Miranti Indar Mandasari. Speaker recognition system in forensic conditions: The calibration and evaluation of the likelihood ratio. 2018.

[18] M. McLaren, V. Abrash, M. Graciarena, Y. Lei, and J. Pesán. Improving robustness to compressed speech in speaker recognition. In *INTERSPEECH*, pages 3698–3702, August 2013.

[19] Didier Meuwly, Daniel Ramos, and Rudolf Haraksim. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International*, 276:142–153, 2017.

[20] Geoffrey Stewart Morrison, Ewald Enzinger, Vincent Hughes, Michael Jessen, Didier Meuwly, Cedric Neumann, S. Planting, William C. Thompson, David van der Vloed, Rolf J.F. Ypma, Cuiling Zhang, A. Anonymous, and B. Anonymous. Consensus on validation of forensic voice comparison. *Science & Justice*, 61(3):299–309, 2021.

[21] Mahesh Kumar Nandwana, Luciana Ferrer, Mitchell McLaren, Diego Castan, and Aaron Lawson. Analysis of critical metadata factors for the calibration of speaker recognition systems. In *INTERSPEECH*, 2019.

[22] Amy Neustein and Hemant A. Patil. *Forensic Speaker Recognition*. Springer, 2012.

[23] Daniel Ramos Rolf Ypma and Didier Meuwly. Ypma, rolf jf, daniel ramos, and didier meuwly. "ai-based forensic evaluation in court: The desirability of explanation and the necessity of validation. *Artificial Intelligence (AI) in Forensic Sciences*, 2, 2023.

[24] Johan Sjoberg, Ari Lakaniemi, Magnus Westerlund, and Qiaobing Xie. RTP Payload Format and File Storage Format for the Adaptive Multi-Rate (AMR) and Adaptive Multi-Rate Wideband (AMR-WB) Audio Codecs. RFC 4867, April 2007. URL: https://www.rfc-editor.org/info/rfc4867, doi:10.17487/RFC4867.

[25] Mårten Storsjö. Vo-amrwbenc library. https://github.com/mstorsjo/vo-amrwbenc, n.d. Accessed: 2025-01-14.

[26] David van der Vloed. Data strategies in forensic automatic speaker comparison. *Forensic Science International*, 350, 2023.

[27] David van der Vloed. Interchangeability of calibration audio datasets for forensic automatic speaker recognition. In *2024 12th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, April 2024.

[28] David van der Vloed and Tina Cambier-Langeveld. How we use automatic speaker comparison in forensic practice. *International Journal of Speech, Language & the Law*, 29(2), 2022. doi:10.1558/ijsll.23955.

[29] David van der Vloed, Finnian Kelly, and Anil Alexander. Exploring the effects of device variability on forensic speaker comparison using vocalise and nfi-frida, a forensically realistic database. In *Odyssey*, 2020.

[30] Stijn van Lierop, Daniel Ramos, Marjan Sjerps, and Rolf Ypma. An overview of log likelihood ratio cost in forensic science – where is it used and what values can we expect? *Forensic Science International: Synergy*, 8, 2024.

[31] Peter Vergeer, Anouk van Es, Arjen de Jongh, Ivo Alberink, and Reinoud Stoel. Numerical likelihood ratios outputted by lr systems are often based on extrapolation: When to stop extrapolating? *Science & Justice*, 56(6):482–491, 2016.

[32] Thi-Ly Vu, Zhiping Zeng, Haihua Xu, and Eng-Siong Chng. Audio codec simulation based data augmentation for telephony speech recognition. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019.