Modeling Audio Fingerprints:

Structure, Distortion, Capacity

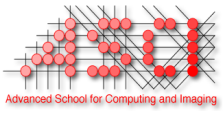
Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof.ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op woensdag 20 oktober 2010 om 12:30 uur
door Peter Jan Otto DOETS
elektrotechnisch ingenieur
geboren te Amsterdam.

Dit proefschrift is goedgekeurd door de promotor:
Prof.dr.ir. R.L. Lagendijk

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof.dr.ir. R.L. Lagendijk,	Technische Universiteit Delft, promotor
Prof.dr.ir. A.-J. van der Veen,	Technische Universiteit Delft
Prof.dr. F.M. Dekking,	Technische Universiteit Delft
Prof.dr.ir. A.W.M. Smeulders,	Universiteit van Amsterdam
Dr.ir. R.N.J. Veldhuis,	Universiteit Twente
Prof.dr. E.J. Delp,	Purdue University, Indiana, U.S.A.
Prof.dr. S. Voloshynovskiy,	University of Geneva, Geneva, Switserland



This work was carried out in the graduate school ASCI.
ASCI dissertation series number 217.

The production of this thesis has been financially supported by TNO.

ISBN: 978-90-9025704-4

To Rosanne and Douwe

Summary

An audio fingerprint is a compact low-level representation of an audio signal [28]. An audio fingerprint can be used to identify audio files or fragments thereof in a reliable way. The use of audio fingerprints for identification consists of two phases. In the enrollment phase known content is fingerprinted, and ingested into a database, together with all relevant metadata. In the identification phase, unknown audio content is fingerprinted, and the fingerprint is used as a query. The query fingerprint is compared to the fingerprints in the database. If a similar fingerprint is found in the database, the relevant metadata corresponding to the fingerprint is returned.

In this thesis we develop models for audio fingerprints. The emphasis here is on fingerprint extraction and the properties of the fingerprint, not on matching the query fingerprint to the fingerprints in the database, and the actual identification. Neither do we develop new practical fingerprinting algorithms.

There is a wide variety of applications for audio fingerprinting, including broadcast monitoring, audience measurement, forensic applications, blacklisting of unauthorized content, 'name that tune' services and linking of special offers to television or radio commercials.

Content which uses the same recorded source material, but which is in different representation, or is distorted in different ways, will generate similar audio fingerprints. This distinguishes audio fingerprints from hashes and content-based retrieval. The hash of an audio file changes even when one sample changes. Two perceptually equal audio items can have completely different hash values, but will generate similar fingerprints. Content-based retrieval looks for audio items which apply to a similar concept, like the same genre, artist or style, while fingerprinting looks for the reuse of the recorded content.

Of course, the exact requirements for a fingerprinting system strongly depend on the application. Relevant aspects for the topics discussed in this thesis are the robustness, uniqueness, accuracy (notably the False Acceptance Rate and False Rejection Rate), granularity and the size of the fingerprints.

In this thesis we make three contributions in the form of models. First, we model the structure of a particular type of audio fingerprint, the Philips Robust Hash (PRH) [44]. The PRH fingerprint extracts a series of spectral energy related features from the audio signal, which are represented efficiently but coarsely as a binary time-series.

The time-series captures the temporal and spectral dynamics of the audio signal, and has a very particular structure mainly depending on a limited number of parameters in the fingerprint extraction.

The model describes the structure of the PRH as a function of a number of parameters [35]. It can be used for better understanding and potentially optimization of the fingerprinting system. We experimentally verify the model on synthetic data in which the samples are independent identically distributed (iid) and Gaussian, and conclude that the model captures the structure of the PRH fingerprint well. This analysis was reformulated and extended by Balado, Hurley, McCarthy and Silvestre [15, 52].

Second, we observe that distortions in the audio are reflected in changes in the corresponding fingerprint. This kind of distortion affects the quality of the audio signal and changes the resulting fingerprint. The idea is to estimate the amount of distortion on the audio signal by comparing the corresponding fingerprint to a reference fingerprint extracted from a high quality copy of the same audio [38]. In this way one could extend the functionality of a fingerprinting system. We implement and compare the behavior of a number of algorithms from literature, and observe similar behavior of the distance between corresponding fingerprints due to compression.

We model the effect of particular distortions in the audio due to compression or additive white noise on the difference introduced in the PRH fingerprints. The main result of our modeling effort is a closed form relation between Signal-to-Noise Ratio (SNR) and average fingerprint distance for PRH audio fingerprints of iid signals [36, 38]. We also experimentally verify the developed models. The model fits perfectly for synthetic signals, and captures the behavior observed in a wider variety of fingerprinting algorithms on actual music.

Third, we consider an information theoretical framework developed by Westover and O'Sullivan (WOS) [104]. The main question is 'how many signals can be identified by a fingerprinting system, under certain conditions'. The conditions relate to characteristics of the fingerprint (size of the fingerprint, and representation of the fingerprint), and characteristics of the environment in which the system operates (representation and statistical characteristics of the signals that need to be identified, how much distortion is allowed). We use the results of the model developed for the PRH fingerprint to estimate up to how many signals can be identified with a binary fingerprint like the PRH. Finally, we check whether the changes in the fingerprints we observe in practice due to distortions in the audio signals, and which have been modeled in this thesis, fit in the information theoretical framework of the WOS model. We compare the WOS-model to practical implementations and outline the differences.

We finish with a list of recommendations on extending the models to jointly consider distortion and uniqueness characteristics; to take more distortion types into account, and to extend to images and video; to develop an evaluation framework for audio fingerprinting; to integrate psycho-acoustics; and to develop a theoretical framework for comparing specific algorithms to the capacity bound.

Table of Contents

Chapter 1

Introduction

1.1 Background

The last decade has shown a tremendous increase in the exchange of audio-visual information over the Internet. The popularity of content sharing portals like YouTube [13] and the various BitTorrent networks are good examples. Not all content shared on these networks are free of copyright restrictions. Rights holders often consider sharing of their content as an illegitimate act. Sometimes, they are even taking legal action against alleged violators infringing on their rights, or the platforms and networks that facilitate the sharing of their content [12]. These platforms, however, can only prevent unauthorized sharing of content if they can automatically determine whether content offered on the platform is authorized. This calls for identification of unlabeled content. A technique which can be used for this is audiovisual fingerprinting. An audiovisual fingerprint is a compact low-level representation of a multimedia signal [28]. Such fingerprints can be derived from, e.g., audio, video and images. In the content sharing scenario rights holders can derive a fingerprint from their content and put them on a black list. Content sharing platforms can then derive a fingerprint from the incoming content to be shared on the platform, and compare it the to fingerprints on the blacklist. If there is a match with one of the fingerprints on the blacklist, the associated content is blocked from the platform.

As the name fingerprinting suggests, there is some resemblance with the use of human fingerprints for identification. A person can be identified using his fingerprints, since his fingerprints are unique. That is, the probability that the fingerprints of two individuals are similar is very small. Law enforcement agencies use this fact when they find fingerprints at the scene of a crime. If the fingerprint is present in police records (in other words: the police has seen these fingerprints before), the police can identify the individual to whom the fingerprints belong, or can link crimes where the same fingerprints were present. Human fingerprints and other biometrical characteristics are widely used for identification and authentication, e.g., to gain access to a building or device.

For the identification of audio-visual content using fingerprints a collection of ref-

erence fingerprints is needed, together with the associated information (i.e., metadata). So, there are two phases as illustrated in Figure 1.1. In the *enrollment phase*, a database is filled with the fingerprints and the associated metadata of a (large) number of songs. In the *identification phase*, shown in Figure 1.1, the fingerprint of an unknown song(fragment) is extracted and compared with the items in the database. If the fingerprint of the song is present in the database, it will be found and hence identified. The song-fragment is likely to be a distorted version of the song that was used to extract the fingerprint in the database, due to compression and regular audio processing. These distortions in the audio signal result in differences in the fingerprints. Therefore, classical database lookup procedures fail, and approximate database matching procedures are needed.

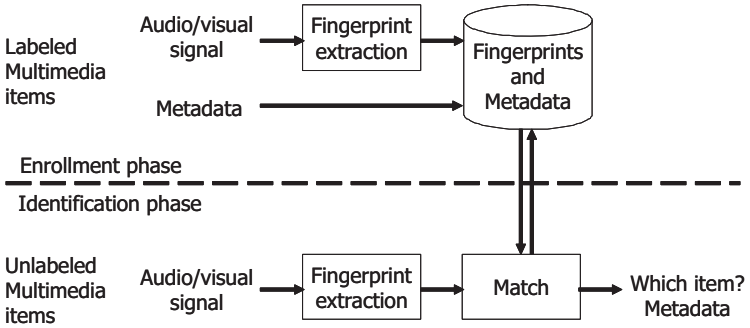


Figure 1.1: Using fingerprints for music identification: the extracted audio fingerprint (query) is matched against a database with pre-computed fingerprints and metadata.

Such a fingerprint can be used to check whether two pieces of audio content are 'the same', or to identify an unlabeled audio file or audio being played, e.g., on the radio. Although we loosely stated that by using fingerprinting you can determine whether two songs are 'the same', it is not straightforward to specify what is 'the same' or 'similar' content. Actually, in case of audio fingerprinting you are usually determining whether two audio pieces are derived from the same recording. Most fingerprinting systems are really sensitive: the same artist performing the same song twice, will generate two distinct audio fingerprints. Of course, the songs are 'the same', but on a semantic level. It also excludes the use of different modalities for query and reference fingerprint, e.g., query-by-humming [24].

In this thesis we limit ourselves to audio fingerprints, and do not consider image or video fingerprints. Audio fingerprints have been used extensively for content-based identification of unlabeled audio [101, 81, 94, 100, 48, 70, 44, 79, 27, 2]. Audio fingerprints are typically based on features computed from the audio signal. The audio fingerprint extraction is different from, e.g., video fingerprinting, since the characteristics, and thus the type of features, for audio are different from video. However, given a representation of those features, the distance metrics, search methods and database structures can be similar, see, e.g., [80].

Another typical audio fingerprinting application is the organization of music col-

lections on a storage device, such as a CD-ROM or an iPod. People digitally store huge audio visual collections on their hard disks, instead of building collections of vinyl, tapes and CDs. The combination with increasing bandwidth of Internet connections and computer processing power has led to different distribution mechanisms for music and video. During the years the emphasis has shifted from collecting entire albums to a set of individual songs and now, due to increased bandwidths available, slowly back to entire albums again. One used to know exactly what is on a compact disc: it is always the same and it goes with a booklet and title page. Even if you insert a CD into your computer, the player will consult the Compact Disc DataBase (CDDB) on the Internet to check which CD is inserted [2]. On a hard disk, however, you loose the intuitive connection between the content, the carrier and 'the things that go with it' (information or metadata). Since most people compress their music using their computer, there is a wide variety of representations of the same content: in different formats, using different bit rates, at different qualities. So, people are confronted with increasingly varying representations of increasingly diverse content. Here, fingerprints can be used to identify and index the songs on the hard disk. What then is needed is a large collection similar to the CDDB database, but now containing the reference fingerprints.

1.2 Scope and contributions

In this thesis we focus on developing models for audio fingerprints and fingerprinting systems. We do not develop new fingerprinting systems, nor do we consider video or image fingerprints. Instead, we build and use three models for audio fingerprints. The emphasis here is on fingerprint extraction and the properties of the fingerprint, not on fingerprint identification and matching. These models can be used to understand and quantify the effect of signal and system parameters on the fingerprint structure, robustness to certain distortions, and capacity.

First, we model the structure of a particular type of audio fingerprint, the Philips Robust Hash (PRH) [44]. This model describes the structure of the PRH as a function of a number of parameters [35]. It can be used for better understanding and potentially for optimization of the fingerprinting system. Furthermore, we experimentally verify the model. This analysis was reformulated and extended by Balado, Hurley, McCarthy and Silvestre [72, 73, 15].

Second, we observe that distortions in the audio are reflected in changes in the corresponding fingerprint. We model the effect of particular distortions in the audio due to compression or white noise on the distortions introduced in the fingerprints. This kind of distortion affects both the quality of the audio signal and the fingerprint. The idea is to estimate the amount of distortion on the audio signal by comparing the corresponding fingerprint to a reference fingerprint extracted from a high quality copy of the same audio [38]. In this way one could extend the functionality of a fingerprinting system. In the Music2Share paper, the authors propose a system for music distribution using Peer-to-Peer networks [56]. In this way, one could buy a copy from another person instead of buying a copy in an online music store. But then the customer would not only like to check the identity of the song for sale, but also the

quality at which the song is offered. In this way, prices can be differentiated according to the offered quality. In other words, pay less for a low quality version. The main result of our modeling effort is a closed form relation between Signal-to-Noise Ratio (SNR) and average fingerprint distance for PRH audio fingerprints of independent identically distributed (iid) signals [36, 38]. Unless stated otherwise, in the models we assume time alignment between the undistorted and distorted signal or fingerprint. In this thesis we also experimentally verify the developed models.

Third, we consider an information theoretical framework developed by Westover and O'Sullivan [104]. The main question is 'how many signals can be reliably identified by a fingerprinting system, under certain conditions'. The conditions relate to characteristics of the fingerprint (size of the fingerprint, and representation of the fingerprint), and characteristics of the environment in which the system operates (what kind of signals need to be identified, how much distortion is allowed). We use the model developed for the probability of an erroneous PRH fingerprint bit due to additive noise, and integrate this model into the framework developed by Westover and O'Sullivan (WOS). In this way we estimate up to how many signals can be identified with a binary fingerprint like the PRH. We compare our bounds with the WOS bounds in [104]. Finally, we check whether the changes in the fingerprints we observe in practice due to distortions in the audio signals, and which have been modeled in this thesis, fit in the information theoretical framework.

1.3 Organization of this thesis

In the following five chapters we present a survey of the state-of-the-art (Chapter 2); build stochastic models for a particular audio fingerprinting algorithm (Chapter 3); use this model to estimate the distortion in compressed audio (Chapter 4); estimate the number of songs that can be identified by fingerprinting systems in general (Chapter 5); and discuss the results of Chapters 2-5 (Chapter 6). In more detail:

Chapter 2: Audio fingerprinting: state-of-the-art
In this chapter we analyze the functionality and working principles of audio fingerprinting. The main building blocks that in general make up an algorithm are discussed. Furthermore, fingerprinting is compared to alternative technologies, differences and similarities are identified. In this chapter we present a thorough overview of the current state-of-the art in audio fingerprinting and compare the algorithms found in literature on several characteristics.

Chapter 3: Models for PRH generated fingerprints of iid signals
This chapter discusses one of the more successfully applied audio fingerprinting algorithms, the Philips Robust Hash (PRH). The fingerprints generated by this algorithm on audio have a particular structure. We show that the structure is predominantly determined by algorithmic choices, and to a lesser extent by the input signal characteristics. To be more precise, the structure of a fingerprint obtained from an iid signal closely resembles the structure of a fingerprint extracted from music. Following this observation, we relate the statistical properties of this fingerprint structure to a partic-

ular fingerprint extraction parameter: the relative frame overlap. Some of the results have been published in:

- P.J.O. Doets and R.L. Lagendijk, "Theoretical Modeling Of A Robust Audio Fingerprinting System", in *Proceedings of the 4th IEEE Benelux Signal Processing Symposium*, pages 101-104, April 2004.
- P.J.O. Doets and R.L. Lagendijk, "Stochastic Model of a Robust Audio Fingerprinting System", in *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR)*, pages 349-352, October 2004.

The second part of this chapter is related to the robustness characteristics of the fingerprint. We consider the effect of additive white Gaussian noise on the fingerprint. We derive a closed form expression for the probability of an erroneous fingerprint bit as a function of the SNR, and experimentally validate the result. Some of the results have been published in:

- P.J.O. Doets and R.L. Lagendijk, "Extracting Quality Parameters for Compressed Audio from Fingerprints", in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, pages 498-503, September 2005.
- P.J.O. Doets and R.L. Lagendijk, "Distortion Estimation in Compressed Music Using Only Audio Fingerprints", in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pages 302-317, February 2008.

Chapter 4: Distortion estimation in compressed music using only audio fingerprints
A fingerprinting system should be robust to many distortions, such as noise and compression. The fingerprint, however, changes slightly due to the distortion. We use the model developed in chapter 3 to estimate the distortion due to compression. The model was developed for the PRH algorithm only, but the other fingerprinting algorithms show a comparable behavior. We experimentally compare the observed behavior due to additive noise and compression for three different audio fingerprinting schemes. This chapter mainly consists of integral copies of sections from our paper:

- P.J.O. Doets and R.L. Lagendijk, "Distortion Estimation in Compressed Music Using Only Audio Fingerprints", in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pages 302-317, February 2008.

Some of the results have also been published in:

- P.J.O. Doets and R.L. Lagendijk, "Extracting Quality Parameters for Compressed Audio from Fingerprints", in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, pages 498-503, September 2005.
- P.J.O. Doets, M. Menor Gisbert and R.L. Lagendijk, "On the Comparison of Audio Fingerprints for Extracting Quality Parameters of Compressed Audio", in *Proceedings of Security, steganography, and watermarking of multimedia contents VII*, January 2006.

- P.J.O. Doets and R.L. Lagendijk, "Extension of a Stochastic Model for the Philips Audio Fingerprint", in *Proceedings of the 27th Symposium on Information Theory in the Benelux*, June 2006.

Chapter 5: Information theoretical models for fingerprinting
Fingerprinting systems are usually evaluated on their statistical properties. But how many songs can be identified, and which factors determine this number? To answer this question, we use an information theoretical model developed by Westover and O'Sullivan. We use the model developed in Chapter 3 to analyze how far binary fingerprints like the PRH are away from the maximum number of songs that can be identified under certain conditions. Finally, we analyze whether the behavior observed in the experimental comparison between algorithms fits in the information theoretical framework. The results in this chapter are recent and unpublished.

Chapter 6: Results and recommendations
In the last chapter we highlight our results and draw conclusions. Finally, we reflect on the work presented in this thesis and outline directions for future research.

Chapter 2

Audio fingerprinting: state-of-the-art

This chapter provides a high-level overview of audio fingerprinting. Section 2.1 deals with applications of fingerprinting. Section 2.2 relates fingerprinting to other content-based retrieval and identification techniques which are commonly used in the same context. In Section 2.4 we break down the fingerprint extraction and identification procedure in smaller building blocks, and discuss each of these blocks in more detail. Section 2.5 presents the main characteristics of state-of-the-art fingerprinting algorithms found in literature. Finally, in Section 2.6 we discuss one state-of-the-art algorithm, the Philips Robust Hash (PRH), in more detail. This algorithm plays a central role in the remainder of this thesis.

2.1 Applications

Establishing the identity of content is a key component in many Digital Rights Management (DRM) applications. DRM refers to technologies that support the legal distribution of digital media while protecting appropriate property rights [25]. So DRM can be seen as the whole collection of commercial, legal, and technical measures to enable trading of digital items on electronic infrastructures [54]. A typical DRM system uses a number of key components: encryption, watermarking, fingerprinting, key management, and a rights expression language [67, 41]. One of the main design philosophies of a DRM system is the separation of content from the rights [93]. This allows the content to be distributed or downloaded freely. However, it cannot be consumed without a valid license, which specifies the permissions for the various ways the associated content can be used.

Although DRM is an important application context, fingerprinting is used for a wide variety of other applications, including [28]:

- *Broadcast monitoring*
  Advertisers spend money to have their commercials aired according to a contract. However, it is very time consuming to manually check whether the commercials are actually aired according to the agreed terms. Companies like CIvolution [7] and Nielsen Broadcast Data Systems [10] offer a service called broadcast monitoring. They automatically monitor a number of radio and television channels looking for specific content, e.g., advertisements, and register when, where, how long etc. the content is aired.

- *Audience measurement*
  Fingerprinting can be used for audience measurement: to identify which programs a selected panel is watching or listening to. Similarly, statistics can be generated on what content is available on the internet. Statistics can also be generated based on relations in the metadata collected using fingerprinting.

- *Forensic applications*
  Special police teams are looking for video material of child abuse. When they raid a house, they usually get hold of a vast collection of audiovisual material. The police is interested in questions like: does the material contain child abuse scenes? What is material that we haven't seen before? Fingerprinting can assist in finding copies that have been analyzed before [8].

- *Locating unauthorized content and blacklisting*
  Rights owners are often interested in where there content is (mis)used, e.g., on content sharing platforms. A combination of web crawlers and fingerprint can locate content on various platforms [9, 7]. This can also be combined with a black list: content on the blacklist is blocked when being uploaded or distributed [11].

- *Name that tune*
  Another example is the 'name that tune' service: if you are wondering what song your listening to, e.g., on the radio, you can collect and send a few seconds of music using a cell phone. The service computes and matches the fingerprint, and returns a text message containing metadata like artist, song name, album etc. [90, 99]

- *Metadata collection*
  People collect enormous amounts of music, through different channels like CDs and downloads. Once stored on, e.g., a hard disk the metadata often is unavailable, making organization of the content very hard.

- *Find duplicates*
  A straightforward application is to find duplicates in large multimedia archives, and to reduce the amount of storage needed.

Similarity metric	The files have the same bits	The files contain the same source material	The files contain the same concept
Tool or Algorithm	Hash (MAC)	Fingerprint or AMAC	Sematic retrieval (genre detection etc.)

Table 2.1: Content based retrieval techniques and associated similarity concepts.

- *Added value services*
  Once the identity of a song or audio file is established, service can be offered based on this information. Examples include an offer to buy the song you identified using 'name that tune', targeted advertisement in social networks based on musical interests, offering of related information like biographies, lyrics, news items etc. A recent example is to capture the audio of an advertisement on the TV using a 'name that tune' service. The user is provided with a link on his cell phone to a special offer related to the advertisement [91].

2.2 Related identification technology

2.2.1 Alternative content-based identification technology

Fingerprinting can be used for content-based copy detection. Whether fingerprinting is the right technique to use depends on the notion of similarity: when are two pieces of content considered 'similar'. Usually, the aim of fingerprinting is to check whether the content originates from the same source material. The following content based identification technologies can be used to establish an increasingly wider notion of similarity, also listed in Table 2.1. Comparing the bits or cryptographic hashes (also known as Message Authentication Codes (MACs)) is relevant when one is interested in establishing whether the digital representations of two songs are bit-wise identical. Comparing waveforms, fingerprints or Approximate Message Authentication Codes (AMACs) to some degree refers to (perceptual) similarity of the waveform. Finally, semantic retrieval and classification approaches use similarity on a more conceptual level, e.g., to retrieve music from the same genre, having a similar style, or from the same artist.

We will now consider these technologies in more detail:

- *Cryptographic hash*
  A cryptographic hash is also called a message digest, or a Message Authentication Code (MAC). Well known examples are the MD5 and SHA family. A MAC is fixed-length (usually 128 or 160 bits), independent of message length. For security reasons a secret key is input for the computation of the MAC, together with the input message. A hash is bit-sensitive, i.e. changing only one bit in the message changes the entire hash. Two other important characteristics are the pre-image resistance - the inability to find a second message which results in the same hash - and the collision resistance - the probability that two arbitrary messages result in the same hash.

- *Approximate Message Authentication Codes (AMACs)*
  An AMAC is also a binary digest of fixed length [42, 109, 106]. Contrary to traditional cryptographic hashes and MACs, AMACs are distance-preserving: the probability that a bit in the AMAC changes varies monotonically as function of the number of bit changes in the message. Therefore, AMACs can be used for a probabilistic estimation of the degree of bitwise similarity of two digital messages. Usually AMACs work on binary messages, but recently AMACs have been developed that work on N-ary alphabets [42]. Such an AMAC provides control of the sensitivity to a given distortion. An AMAC can be made distance preserving, i.e., the distance between two authentication tags reflects the distance between two messages.

  Reported drawbacks of AMACs are the large variations around expected value of AMAC differences and the inability to locate the changes in the content [109] (which often is possible using audio fingerprinting). Some approaches create an AMAC of a feature vector [106], others of an entire signal [42, 109]

- *Comparison of waveforms*
  A straightforward method for identifying songs would be to compare its waveform to a series of waveforms of known songs. Besides the question what is 'the same', there are several drawbacks to such an approach. First, storing a large number of waveforms requires a lot of storage space. Second, songs that sound similar can have a large variation in their waveform representations (interclass variation). Thirdly, the computational complexity of waveform comparisons is relatively large; although feature based comparisons can be seen as a two-tier approach of the waveform comparison.

- *Semantic, content based retrieval (SCBR)*
  The aim of content based retrieval is also to find similar multimedia content items, but the similarity between content items is evaluated on a higher, semantic level. An example of SCBR is query-by-example: given (an) example song(s), find all songs that have been performed by the same artist. When comparing (A)MACs, fingerprints or waveforms the similarity is always low-level: either (approximate) bit-wise or waveform similarity. Here the notion of similarity relates more to a concept.

2.2.2 Alternative identification technology: watermarking

Watermarking is an identification technology alternative to fingerprinting. Watermarking can be defined as the 'imperceptible insertion of information into multimedia data through slight modification of the data' [30]. Literature surveys can be found in [30, 63, 77]. It can be used for similar applications, like broadcast monitoring. However, since the signal needs to be actively altered it cannot be used for legacy content, i.e. content that is already 'around', or for content over which the 'identifier' does not have full control and is thus not able to embed a watermark in. Furthermore, since the insertion needs to be imperceptible, it may potentially also be rendered undetectable without changing the perceptual characteristics of the content.

The embedded message is independent of the multimedia content, and thus can have any meaning beyond content identification, e.g., transaction tracking [84, 4, 6]. Watermarking thus makes it possible to distinguish perceptually identical copies. Three combinations of watermarking and fingerprinting are typically used in DRM applications. First, self-embedding is a technique in which a fingerprint is embedded as a watermark for authentication purposes [33, 32, 108, 43]. Second, the fingerprint can be used as an input to the watermark embedding procedure [46, 75, 19]; in this way the watermark becomes content dependent and can gain robustness to the so-called copy attack [61]. Third, to locate the start of a watermark message in an audio stream, the watermark contains markers. These markers, however, are easily located and may be removed, and pose a security risk [31]. When the embedding locations are known before-hand to the detector, there is no need for these markers. The embedding location could be revealed to the decoder in the form of a fingerprint [47]. Contrary to fingerprinting, watermarking theory is well-developed and there exist good theoretical models, e.g., [78, 68, 83].

2.2.3 Identification of individual humans: biometrics

Audio fingerprinting is related to biometrics in more aspects than just the name. Biometrics is a technology for establishing or verifying the identity of individuals based on their physiological or behavioral characteristics [53]. Example characteristics are the face, fingerprint, iris, but also gait and keystroke dynamics. Since the technical goals - identification and verification - are the same as for multimedia fingerprinting, the structure of biometrical identification systems show a lot of similarity with audio fingerprinting systems. Again there are two phases: enrollment and identification. The representations of the biometric features need to be compact for scalability.

However, there are some important conceptual differences. The biometric, e.g., a human fingerprint, is on the same conceptual level as a song. The similarity comparisons are usually carried out on features of the biometric, therefore putting these features on the level of the multimedia fingerprint. Furthermore, due to imperfections like sensor noise and personal behavior it is impossible to register the 'true' biometric, or the prototype. One can only measure distorted versions of the biometric. In multimedia fingerprinting, however, for many applications a registration of the content extremely close to the 'prototype' can be made, e.g., master tapes of a recording, a CD or high definition recording. Finally, security is a key issue in biometrics, but not in most fingerprinting applications, although key-based audio fingerprinting schemes have been published [75].

2.3 Requirements and trade-offs

Although in this thesis we do not develop or design a fingerprinting system, it is relevant to list some of the typical requirements (desired properties) for fingerprint extraction and the identification using a database of reference fingerprints. We follow the terminology and definitions used in [28, 55].

Requirements on the fingerprint extraction:

- *Robustness*
  The ability of the fingerprint representation to withstand the effect of signal processing operations, i.e. the changes in the fingerprint are limited.

- *Uniqueness*
  The discriminating capabilities of the fingerprints. This is related to the collision probability: the probability that two dissimilar signals result in two similar fingerprints.

- *Accuracy*
  Extent to which the identification results are correct. The level of accuracy is strongly dependent on the robustness and uniqueness properties of the fingerprint. Accuracy mainly relates to the False Acceptance Rate (FAR) and False Rejection Rate (FRR) - see Section 2.4.5. A related aspect is the time localization accuracy, i.e. the ability to precisely locate the starting and ending point of a query fragment located in a reference recording.

- *Fragility*
  The ability to control to which distortions the fingerprint is robust. For some applications it is desirable that the fingerprint is *only* robust to certain content-preserving operations.

- *Granularity*
  The minimum audio fragment length needed for a reliable identification. Based on a small fragment, an audio track can be identified. When a system is fine granular this means that a system is capable of reliable identification of small excerpts.

- *Fingerprint rate (size)*
  The fingerprint rate is the amount of bits (or: elements) extracted per second (or: song). To facilitate database and system scalability, the fingerprint size should be small. The size of the fingerprint is also directly related to the number of fingerprints that can be represented, and to the granularity: as a rule of thumb one can say that the larger the fingerprint rate, the finer the granularity.

- *Computational complexity*
  This refers to the amount and type of resources required for the extraction of the fingerprint, or the comparison of two fingerprints. This is a relevant issue for systems requiring real-time operation and for systems having limited computational resources. In some applications the computational burden can be divided over a client extracting a fingerprint, and a server maintaining the database and matching the fingerprints. Thus, either the fingerprint is computed locally (possibly low computational resources/bandwidth), or the query item is transmitted over a network and the fingerprint is computed centrally (high computational resources/bandwidth).

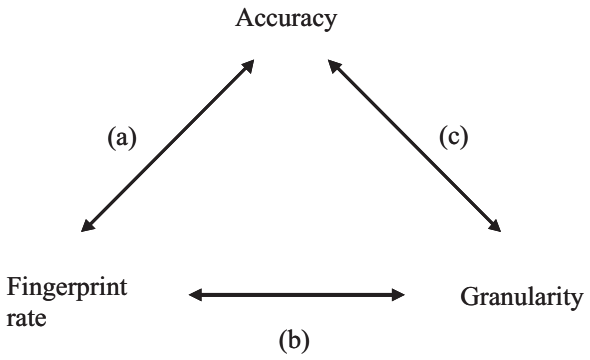


Figure 2.1: Illustration of the relation between accuracy, granularity, and fingerprint rate.

- *Security*
  For some applications it is important that the derivation of the fingerprint from the content is key dependent. One then should not be able to (mildly) change the content without changing the fingerprint. It then also should not be easy to find a different piece of content that generates the same fingerprint (collision), or to learn the key given one or more content items.

In this thesis the security and computational complexity are not considered; these are not relevant for the types of models we develop in this thesis. In the next chapter we develop a model with describes the structure of the fingerprints generated by one particular fingerprinting algorithm, the Philips Robust Hash (PRH). This structure determines the uniqueness of the fingerprints. We also develop a model that describes how fingerprints change due to additive noise, which relates to robustness. In Section 2.4.5 we discuss and illustrate detection statistics, which are coupled to granularity. The relations between the fingerprint requirements on accuracy, granularity, and compactness are related as follows, also illustrated in Figure 2.1:

(a) For a given granularity, a higher accuracy can be achieved by using a higher fingerprint rate, i.e. extract more information from the audio signal. E.g., when identifying 3 seconds of music, extracting more features from the same material and keeping other parameters constant, may provide a higher accuracy.

(b) For a given accuracy, using a higher fingerprint rate enables a finer granularity, i.e. the system is able to identify smaller audio fragments.

(c) For a given fingerprint rate, using larger minimum fragment length, i.e. a coarser granularity, can result in a larger accuracy since there is more information available in the audio fingerprint.

Although in this thesis we do not consider searching a fingerprint in a database, it is an important aspect in the use of fingerprinting systems. We briefly list some of the most important requirements for the database aspects:

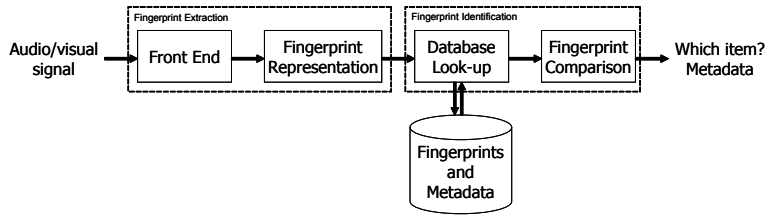


Figure 2.2: General structure of a fingerprinting system consisting of the fingerprint extraction module, fingerprint identification module and a database containing fingerprints and metadata.

- *Scalability*
  The fingerprinting system should be scalable to a large number of fingerprints. This is affected both by the database parameters (search speed, search efficiency, indexing structures), but also by the fingerprint parameters itself (how many fingerprints can be distinguished in a reliable way).
- *Search complexity*
  Search complexity refers to the complexity of the database search or evaluation of distance metric.
- *Updatability*
  It should be easy to enroll new items in the database, or remove existing items, and update the corresponding index structures.

2.4 Structure of audio fingerprinting algorithms

Most fingerprinting systems share a similar structure. In this section we analyze the structure of a fingerprinting system for identification. Most of the analysis follows the structure and terminology introduced in the survey paper by Cano *et al.* [28].

Figure 2.2 illustrates the elementary building blocks in a fingerprinting system, consisting of the fingerprint extraction and the fingerprint identification modules. The fingerprint extraction consists of two steps: extracting robust features (front end), and building the fingerprint representation based on these features (fingerprint representation).

The fingerprint identification also consists of two steps. First, the fingerprint to be identified needs to be matched against the database to retrieve potentially similar fingerprints (database matching). Fingerprinting systems need to be able to scale to large collections of fingerprints. Therefore, efficient database structures are needed. Second, the fingerprint has to be compared to each potential match (fingerprint comparison). Therefore, we investigate commonly used similarity measures and the important detection statistics.

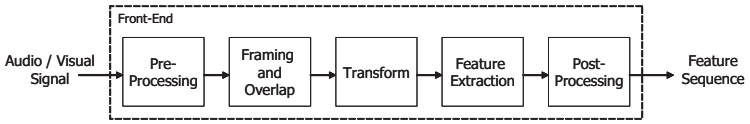


Figure 2.3: Fingerprint front end of the fingerprint extraction module.

Section 2.4.1 discusses the extraction of a feature sequences from the audio signal. The representation of the feature sequence as a fingerprint is discussed in Section 2.4.2. The following three sections deal with the fingerprint identification. Section 2.4.3 discusses the database retrieval techniques, Section 2.4.4 lists the fingerprint similarity measures, and Section 2.4.5 discusses the relevant detection statistics.

2.4.1 Front End

Figure 2.3 shows a break-up into five building blocks of the fingerprint extraction front end:

- *Pre-processing*
  The most important aspect in this step is the conversion to a standard intermediate format from which the fingerprint is computed, e.g., a mono signal at a specific sample rate. Other common operations aim a at dimension reduction, concentration on perceptually most relevant information and anticipation on specific distortions. An example is the use of bandpass filtering, and down-sampling. Down-sampling removes the high frequency information; the high frequency components usually contain less energy, are therefore more sensitive to distortion and thus are less stable.

- *Framing and overlap*
  The signal is framed in order to compute a feature sequence over time. An important parameter is the frame rate: the rate at which frames or features are extracted from the signal. Framing (or in general: discretization) introduces synchronization issues. When comparing two signals that have the same source, there is no guarantee that the frames have been put at the same location (boundary synchronization). To minimize the effect of boundary desynchronization, there commonly is a large overlap between successive frames (in the order of 50-96%). Many applications, e.g., coding, that assume (weakly) stationary signal characteristics use frames; these frames are usually in the order of 32 msec in length. The frame length used in audio fingerprinting is commonly in the order of hundreds of msec.

- *Linear transforms: Spectral estimates*
  The Human Auditory System (HAS) reacts to spectral and temporal characteristics of an audio signal. Since most music is man-made, it is intended to match the HAS characteristics. Therefore, most fingerprinting algorithms introduce windowed frames and make a time-frequency decomposition, often a FFT or

MDCT. To facilitate the computation of a spectral transform, the frames are windowed first. The time-frequency decomposition also results in decorrelation and information packing, and thus enables a more compact representation. Note that some algorithms use time domain features and thus do not compute features in the spectral domain. However, the performance reported in literature for time domain features for audio fingerprinting is worse than for frequency domain features. This may be due to the fact that some typical distortions specifically affect specific frequency regions.

- *Feature extraction*
  Feature extraction mainly aims at dimensionality reduction in the form of efficient and effective descriptions of the underlying signal. Furthermore, by using features that are based on the most robust signal elements it can increase robustness to distortions. Popular features include Mel Frequency Cepstral Coefficients (MFCC) [88, 29, 86, 107], Spectral Flatness Measure (SFM) [70, 14] and Haar features on spectral energies [44, 58].

- *Post-processing*
  This step can be used to normalize the features, to emphasize the temporal evolution of the feature sequence (derivatives) or to represent the data in an efficient form.

The order of these steps may be different, repeated, or applied on different time or frequency scales. In conclusion, we can say that each of the before-mentioned building blocks aims at one or more of the following goals:

- *Dimensionality reduction and compact representation*
  Examples include feature extraction, sample rate conversions and spectral representations, e.g., PCA, OPCA and SVD.

- *Increase robustness to distortion*
  Examples include the use of (invariant) features, coarse quantization, bandpass filtering and down-sampling.

- *Emphasize unique characteristics of the signal*
  Examples include the use of derivatives of feature time series.

- *Match perceptual characteristics*
  There are two main reasons for a fingerprinting system to consider using the perceptual characteristics and match the Human Auditory System (HAS). First, many deliberately introduced signal distortions preserve the most important perceptual characteristics. Second, some applications explicitly aim at 'perceptual similarity', or fingerprints as a perceptual digest. However, some distortions might be introduced by transmitting the signal to be identified to the fingerprint extraction engine. In the 'name that tune' scenario, for instance, the distortion introduced by the GSM channel does not necessarily preserve the perceptual characteristics.

These four goals match the requirements in Section 2.3 on compactness, robustness, uniqueness, and fragility, respectively. A good combination of feature extraction, fingerprint representation and similarity measure can increase the system's robustness to distortions.

2.4.2 Fingerprint representation

The feature stream time-series can be represented in different ways. Based on how the representation follows from the temporal evolution of the fingerprint, we distinguish three categories of fingerprint representations:

1. *Fixed size fingerprints*
   The size of the fingerprint is independent of the song length. Examples include taking the time-average over the extracted features [100] and the estimation of the probability density function (pdf) underlying the extracted feature sequence [86]. Music is non-stationary; parts with different signal and statistical characteristics are mixed in the final representation. There are three drawbacks of such a system. First, when different parts are mixed together in one model, the discriminating characteristics of such fragments are lost in the modeling procedure; when identifying shorter fragments there is only a partial match with the model derived for the entire system. Second, the timing information and the temporal order of the features is a distinguishing feature of a signal. Third, the fingerprint differences cannot be used to locate the differences between the signals. One of the advantages of loosing the temporal information is that the model potentially becomes independent of time scaling distortions.

2. *Constant rate fingerprints*
   Most fingerprinting systems extract features on regular time intervals (frames). Therefore, the fingerprint size is proportional to the song length. The main advantage is that signal characteristics that are changing over time are not mixed in the final fingerprint. Furthermore, the amount of information extracted in a certain time window can be guaranteed. Finally, when comparing the fingerprint of a distorted version to the fingerprint of the original undistorted recording, the fingerprint difference can be used to localize the changes in the distorted version.

3. *Variable rate fingerprints*
   For efficient representation the rate of fingerprint varies with acoustical events. In this way, the fingerprint only represents that salient characteristics of the underlying acoustic signal. In the Shazam fingerprint, for instance, the spectral peak locations that are most significant in both the frequency and in the temporal dimension represent the fingerprint [101]. This may result in very compact fingerprints. However, one cannot guarantee the amount of information extracted in a certain time window. Also Kurth *et al.* [60] and Lebossé *et al.* [65] use variable rate fingerprints.

For the 2nd and 3rd type we adopt the following terminology introduced by Haitsma *et al*. [45]. The part of the fingerprint that corresponds to a particular time instant is called a *sub-fingerprint*. The fingerprint of a song is thus given by a time-series of sub-fingerprints. A number of sub-fingerprints used for identification is called a *fingerprint block*.

2.4.3 Database structures

Although the emphasis in this thesis is on properties of the extracted fingerprints, the database search procedures are an essential part of any fingerprinting system for identification. Therefore, we briefly summarize some of the important aspects and characteristics.

Due to the potentially large number of items in the fingerprint database, exhaustive search is infeasible. Therefore, efficient database structures and greedy search algorithms are used. Since the fingerprint can have changes due to distortions in the audio signal, the search is usually an approximate search: the *exact* query fingerprint cannot be found in the database, but a *similar* fingerprint might be found. Therefore, it is crucial to exclude unlikely candidates without excluding matching candidates. The database search method may also affect the accuracy of the fingerprinting system: it may miss actually matching fingerprinting candidates.

There is strong relation between the fingerprint representation, the distance measure and the used database search structure. Many papers found in literature focus on the properties of the fingerprint itself, and do not consider matching strategies. Some of the common techniques include:

- *Inverted file index*
  This is a Lookup Table (LUT) of possible sub-fingerprint entries with pointers to the fingerprints in database [44, 101, 29, 60]. The applicability depends on the alphabet and the size of the sub-fingerprint. It might be infeasible to generate a list containing all possible entries (e.g., 2^{32}) and corresponding pointers. Depending on the properties of the fingerprint, the LUT might be sparsely filled. Therefore, the list might be based on another data type, e.g., cluster centers or hash table entries derived from sub-fingerprints.

  To facilitate the matching of fingerprints containing errors, either the query fingerprint might be expanded to include more possibilities, or the LUT may also contain entries corresponding to sub-fingerprint into which small errors have been introduced. One has to be careful though that assumptions that are made on the type and extent of the errors in the query fingerprint may lead to false dismissals.

- *Filtering out unlikely candidates*
  Filtering can be an efficient way to reduce the search space. Again, one should be careful not to introduce false dismissals. Several well-known techniques can be used to implement this idea. Unlikely candidates can be filtered out first using a cheap similarity measure. The remaining set is evaluated using a more complex and precise similarity measure. One could think of, e.g., computing

a similarity on a sub-sampled version of the query fingerprint. Furthermore, during the comparison process one can exclude candidates for which you know beforehand that they have a worse score that the ones considered sofar. Some tree based search methods exploit this.

- *Hierarchical search*
  The query is first compared to the fingerprints of popular items. This can include a 'most wanted' list, or a list of new releases. If there is no match in the short list, the query is matched to the entire fingerprint database.

- *Tree based search*
  Essentially, finding a similar fingerprint is a nearest neighbor search. Often trees are used for locating nearest neighbors. A good example which was designed specially for the PRH fingerprint is [76]. Here, each 5-second binary fingerprint block (8192 bits) is considered to be a point in the fingerprint space. The fingerprint block is split into 1024 8-bit patterns. The value of each consecutive 8-bit pattern determines which of the 256 possible children to descend to. A path from the root node to a leaf defines a fingerprint block. When matching a query fingerprint to the database, each 8-bit pattern is compared to the elements of the tree; at each level in the tree, the error between the query fingerprint and the best matching leaf below that node is estimated. As soon as the estimated error is higher than the best found result sofar, the search is stopped.

2.4.4 Similarity measure

Fingerprinting systems typically use either a distance measure, or a probabilistic measure to compare two fingerprints. These similarity measures result in two distinct identification rules for identifying an unknown fingerprint F_Y:

- Nearest neighbor decoding

$$\hat{w} = \arg\min_{W} d(F_W\ ,\ F_Y)$$

- Maximum confidence score

$$\hat{w} = \arg\max_{W} c(F_W\ ,\ F_Y)$$

- Maximum likelihood

$$\hat{w} = \arg\max_{W} Pr[F_W \mid F_Y]$$

where F_W denote the fingerprints stored in the database, $\hat{w}$ is the index of the most similar fingerprint, $d(\cdot,\cdot)$ represents a distance measure, $c(F_W\ ,\ F_Y)$ represents the confidence score of the match between fingerprints F_W and F_Y, and $Pr[F_W \mid F_Y]$ denotes the probability of occurrence for fingerprint F_W, given the fingerprint F_Y.

Typically, two aspects contribute to the confidence in a match between two fingerprints: the length of the matching fragment, and the (dis)similarity between the fingerprints.

To limit the number of false positives, each identification rule should be combined with a threshold:

- Nearest neighbor decoding

$$d(F_{\hat{W}}\ ,\ F_Y) \leq T_1,$$

- Maximum confidence score

$$c(F_W\ ,\ F_Y) \geq T_2.$$

- Maximum likelihood

$$Pr[F_{\hat{W}} \mid F_Y] \geq T_3.$$

Furthermore, we can distinguish *symmetric* and *asymmetric* fingerprint identification procedures. In symmetric procedures the same fingerprint model is used both for enrollment and identification. Asymmetric procedures, on the other hand, use a different fingerprint model for identification than for enrollment. Asymmetric procedures usually occur in combination with probabilistic similarity measures, e.g., [29, 86]. For instance, the fingerprint used for storage in [86] is a Gaussian Mixture Model (GMM) estimated from an extracted feature sequence. In the identification procedure, for each stored GMM, the probability is estimated that the feature sequence of the song to be identified is the result of that particular GMM.

2.4.5 Detection statistics

In this section, we discuss the most important accuracy indicators. In a typical distance based fingerprint detector, two (conditional) PDFs are of interest. Firstly, the conditional PDF of the distance between the fingerprints of similar content, i.e. the fingerprints of the distorted and the original recording. Secondly, the conditional PDF of the distance between the fingerprints of dissimilar content, i.e. between two arbitrary, unrelated fingerprints. These two PDFs are illustrated in Figure 2.4. Two probabilities are of special interest:

- *False Acceptance Rate (FAR)*
  The probability that two perceptually dissimilar objects yield similar fingerprints. This is also known as the False Positive Rate.

- *False Rejection Rate (FRR)*
  The probability that two perceptually similar objects yield dissimilar fingerprints. This is also known as the False Negative Rate.

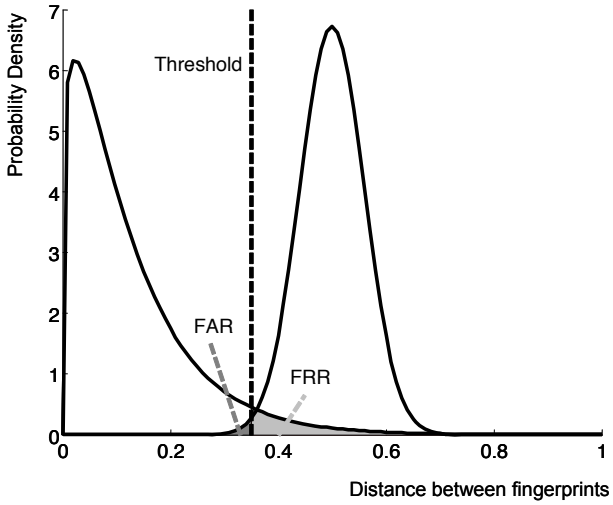


Figure 2.4: Illustration of fingerprint detector statistics, showing the conditional PDFs of the distances between similar fingerprints and between dissimilar fingerprints, the detection threshold and the resulting FAR and FRR.

Both probabilities are indicated in Figure 2.4. In practice, only the PDF of the distance between unrelated fingerprints is sufficiently known. The PDF of the distance between corresponding fingerprints is dependent on the distortion, and usually on the underlying signal characteristics. In a typical fingerprint detector design, one sets a threshold to achieve a certain FAR based on the PDF of the distance between non-corresponding fingerprints.

Items that are close in some perceptual space should also be close in the feature space; this is known in pattern recognition as the compactness hypothesis. Therefore, the fingerprint space should have the same connectivity as the original space [39]. This is illustrated in Figure 2.5. In practice, for identification one makes the assumption the other way around: fingerprints that are close imply that the audio excerpts they represent are also similar. In other words: similar features imply similar audio. The symbols in the figure represent audio items and their corresponding fingerprints. The compactness hypothesis only holds here for the symbols $\times$ and $\Diamond$. The fingerprints $\triangle$ and $\bigcirc$ are close, although the underlying audio is dissimilar, potentially giving rise to wrong identification. The opposite is true for the pair $\triangle$ and $\square$, potentially leading to a missed identification ($\square$ is isolated). The compactness hypothesis is not relevant for all content-based identification techniques listed in Section 2.2. In cryptographic hashing, for instance, the objective is the opposite of this assumption: items that are close but not identical in their signal representation should be randomly distributed in the hash space.

Some papers estimate the FAR and FRR for a specific algorithm, e.g., [44]. However, these figures correspond to one single detection. The system's performance is determined by the FRR and FAR in combination with the number of items in the

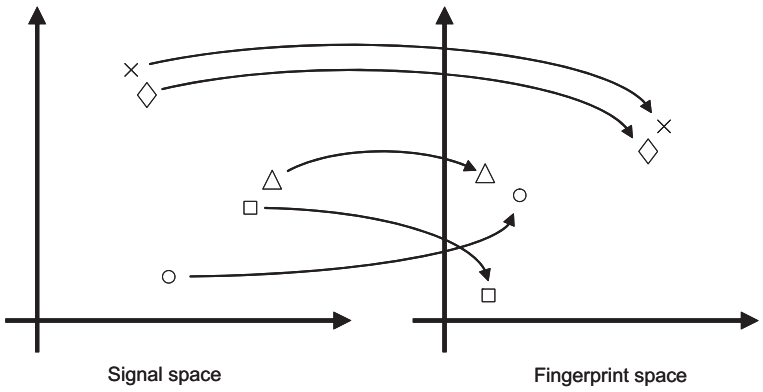


Figure 2.5: Geometric illustration of spatial connectivity in the signal and fingerprint space.

database. Now consider a database containing N items. The result of the identification process is a result set, which contains one or more results, each of which may be correct or not. The result set is obtained by computing the detection statistic and applying a threshold. For instance, consider the case of nearest neighbor decoding without applying the minimum operator but only applying a threshold on the detection statistic. In case the result set contains multiple items including the correct one, the final result after applying the minimum operator may or may not be correct. The final result is correct if the matching item is below the threshold and has the minimum distance.

In case the unknown item actually is in the database, assume there is only one item actually corresponding to the query, and $N - 1$ non-matching items. In this way, we can consider four different outcomes:

1. *No identification (1 false negative, no false positives)*
   The result set is empty; it thus does not contain the right item from the database, neither does it contain any wrong items;

2. *Wrong identification (1 false negative, at least one false positive)*
   The result set is non-empty, but does not contain the right answer;

3. *Correct identification (1 true positive, no false positives)*
   The result set is non-empty, and only contains the right answer;

4. *Mix (1 true positive, at least one false positive)*
   The result set is non-empty and contains multiple results, including the right answer. This result may be avoided by filtering the result set and keeping only the result in which the system has the highest confidence (applying the minimum or maximum operator). Then the result is either the correct or not.

Table 2.2: Illustration of fingerprint identification statistics in case the unknown item actually (a) is in the database - one matching item, $N-1$ non-matching items; (b) is not in the database - N non-matching items. The result sets contains a number of true positives (#TP), and a number of false positives (#TN)

(a)

Description of identification result	Result set contains		Probability
	#TP	#FP	
No identification	0	0	$FRR \times (1-FAR)^{N-1}$
Wrong identification	0	≥ 1	$FRR \times (1-(1-FAR)^{N-1})$
Correct identification	1	0	$(1-FRR) \times (1-FAR)^{N-1}$
Mix	1	≥ 1	$(1-FRR) \times (1-(1-FAR)^{N-1})$

(b)

Description of identification result	Result set contains		Probability
	#TP	#FP	
No identification	0	0	$(1-FAR)^N$
Wrong identification	0	≥ 1	$1-(1-FAR)^N$

In case the database actually does not contain the unknown item, we can consider two different outcomes:

1. *No identification (no false positives)*
   The result set is empty; it thus does not contain the right item from the database, but also does not contain any wrong items;

2. *Wrong identification (at least one false positive)*
   The result set is non-empty; it contains wrong items only.

Assume that the FAR and FRR values are known. Table 2.2 analyzes the overall identification statistics as a function of N for two cases: the database *contains* the item to be identified (Table 2.2(a)), or *does not contain* the item to be identified (Table 2.2(b)), respectively.

When inspecting the equations in the last column, it stands out that all results are dependent on the expression $(1-FAR)^N$. Even for situations in which the FRR is unknown, this expression is important to assess the expected performance for various database sizes. Note that the expressions are valid only for an 'exhaustive search', and does not consider the negative effect on the accuracy of, e.g., greedy search or database pruning algorithms.

Since typically $FAR \ll 1$, we can use the Taylor series expansion for the natural logarithm to get some feeling for the values of $(1-FAR)^N$:

$$(1-FAR)^N \quad = \quad e^{N \ln(1-FAR)} \approx e^{-N\ FAR}.$$

So for different values of N, we get:

$$\begin{aligned} FAR \ll \tfrac{1}{N} \qquad & (1-FAR)^N \approx 1 - N\ FAR \\ FAR = \tfrac{1}{N} \qquad & (1-FAR)^N \approx \frac{1}{e} \\ FAR \gg \tfrac{1}{N} \qquad & (1-FAR)^N \approx 0 \end{aligned}$$

In conclusion we can say that it is a necessary condition for a fingerprinting system that $N\ FAR \ll 1$.

Here, an item does not necessarily need to be an entire song, but it the basic unit for which the detection statistics are computed. So, it can be a fragment for which its corresponding fingerprint block is 'sufficiently independent' from other fingerprint blocks. This could, for instance, refer to 3s song fragments. Furthermore, two subsequent 3s fragment fingerprints from the same song having a 2.5s overlap already may generate a sufficiently different fingerprint, depending on the fingerprint extraction procedure. N can thus be very large; in the before mentioned example a 3 minute song would correspond to 355 overlapping 3 second items.

2.5 State-of-the-art algorithms

In this section we present a comprehensive overview of algorithms found in literature. Table 2.3 contains the overview; the algorithms are listed in arbitrary order. Where possible, we present the following characteristics of each algorithm:

- *Affiliation and reference*
  The institute or company where the algorithm was developed, and the reference(s) to the relevant publication(s).
- *Fingerprint type*
  Indicates whether the fingerprint extraction is constant rate (CR), variable rate (VR), or constant length (CL).
- *Symmetric*
  Indicates whether the fingerprint representation stored in the database is the same (yes) or different (no) from the representation of the fingerprint to be identified.
- *Fingerprint representation*
  The way the fingerprint is represented: time-series (TS), vector, or a specific model. In case the algorithm is asymmetric both the representation in the database and the representation of the fingerprint to be identified are listed. The models found in literature are: Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) and Vector Quantization Codebook (VQC).
- *Metric*
  The type of (dis)similarity measure: nearest neighbor (NN), confidence score (C), or maximum likelihood (ML).

- *Distance measure or confidence score*
  The measure that is used to compute the (dis)similarity of two fingerprints. Distances used: Bit error rate (BER), root mean square (RMS), mean square error (MSE), Manhattan distance, Exponential Pseudo Norm (EPN) [75]. Confidence scores used: recursive score function (SF), number of alignment spectral peaks (#co-peaks), number of similar bytes (#co-bytes).

- *Feature*
  Compact description of the features on which the fingerprint is based.

- *Fingerprint size or rate*
  In case of a CL fingerprint, the fingerprint size is listed in Bytes. Otherwise the fingerprint rate is listed in Bytes/sec. For a VR fingerprint this is the average rate.

- *Typical granularity*
  The typical fragment length used for identification as reported in the corresponding reference.

2.6 Example Audio Fingerprinting System: Philips Robust Hash

The PRH algorithm developed by Haitsma *et al*. [44] has been reported to have good performance and a simple and efficient structure. It was developed at Philips Research, and sold to Gracenote, Inc. [2] Today, it is also used by Civolution [7], a Philips spin-off.

The PRH fingerprint is derived in a number of steps from a time domain signal, $x(i)$, shown in Figure 2.6. In the PRH algorithm, the steps identified in Section 2.4.1 are easily recognizable. First, the signal is converted to mono and downsampled to 5,512.5 Hz. Then the signal is divided into frames of 371 ms (2048 samples). The frames have 96% (31/32) overlap. This strong overlap is used to prevent temporal misalignment of the frames used in the query and reference fingerprint (see Section 3.4.2). The frames used for the fingerprint are shifted 11.6 ms. Therefore, the maximum level of misalignment between the frames is 5.8 ms.

Each frame is windowed and the periodogram is estimated. The spectrum is divided into 33 logarithmically spaced frequency bands in the range 300-2000 Hz. In this way, each musical note has its own frequency band. The musical note 'A' is defined to be at 440 Hz. Twelve notes fit into an octave, making the frequency of each note being a factor $\alpha = \sqrt[12]{2} \approx 1.06$ higher than the previous one. The fingerprint is based on the lower part of the spectrum since it contains most energy, which is usually preserved in case of distortions. For instance, one use case for the algorithm is the 'Name that tune' service, where a user transmits a couple of seconds to a server using a mobile phone connection. A typical telephone bandwidth is 300-3400 Hz.

Within each band the energy is estimated. Let us denote the energy in frequency band m of frame n by $E^b(n, m)$, where $m = 0, \ldots, 32$ and $n = 0, 1, \ldots$. Differences

Table 2.3: Overview of audio fingerprinting algorithms

Affiliation	Reference	Type	Symm.	Repr.	Metric	Distance	Feature	Size	Typical granularity
Fraunhofer	[14, 57]	CR	No	VQC & TS	NN	RMS	SFM (MPEG 7 scalable)	2-800 B/sec	4-20 sec
Microsoft	[27]	CR	Yes	TS	NN	MSE	OPCA on spectral energy	64@4B	6 sec
Microsoft	[75]	VR	Yes	TS	NN	EPN [a]	Spectrogram statistics	n.a.	15 sec
Philips	[44]	CR	Yes	TS	NN	BER	Haar on spectrogram	344.8 B/sec	3-10 sec
Cantmetrix	[100]	CL	Yes	Vector	NN	RMS	Statistics of spectral energy	450 bits	First 15 sec
Shazam	[102, 101]	VR	Yes	TS	C	# co-peaks	Location of spectral peaks	20-40 B/sec	10-15 sec
MusicIP[b]	[50]	CL	Yes	Vector	NN	RMS	SVD on spectrogram	n.a.	n.a.
Audible Magic	[107]	CR	Yes	TS	NN	RMS	Time-averaged MFCCs	40 B/sec	n.a.
Dolby Labs	[85]	CR	Yes	TS	NN	BER	Spectrogram projections	1.7 kbps	n.a.
Google	[17, 18]	CR	Yes	TS	C	# co-bytes	Hashed sign of Haar wavelets	n.a.	10-30 sec
Orange	[66, 64]	VR	Yes	TS	C	recursive SF	Intervals between temporal maxima	38.3 B/sec	n.a.
Relatable[c]	[103]	CL	Yes	Vector	NN	Manhattan	Mean time and freq. features	16 B	entire songs
Tuneprint	[88]	CR	Yes	TS	NN	RMS	OPCA on Bark energies	344.1 B/sec	5 sec
Bogazici	[82]	CR	Yes	TS	NN	RMS	SVD on MFCCs	312 B/sec	3 sec
CEFRIEL	[70, 62]	CR	Yes	TS	NN	MSE	Spectral energy, SFM, SCF	48 B/sec	2 sec
KAIST	[89]	CR	Yes	TS	NN	MSE	Spectral Moments	345.8 B/sec	10 sec
Barcelona	[29]	VR	No	HMM & TS	ML	-	PCA of MFCC	500 B/sec	n.a.
Carnegie	[58]	CR	Yes	TS	ML	-	Optimized Haar on spectrogram	344.8 B/sec	3-10 sec
Budapest	[87]	CL	Yes	Vector	NN	MSE	Quantized Bark energies	n.a.	n.a.
Washington	[94]	CR	Yes	TS	NN	RMS	Centroid of modulation scale	486.4 B/sec	15 sec
Ryerson	[86]	CL	No	GMM & TS	ML	-	MFCC	stream:4828 B/sec, model:1856 Bytes	5 sec

[a] Exponential Pseudo Norm
[b] used by MusicBrainz since March 2006
[c] used by MusicBrainz till March 2006

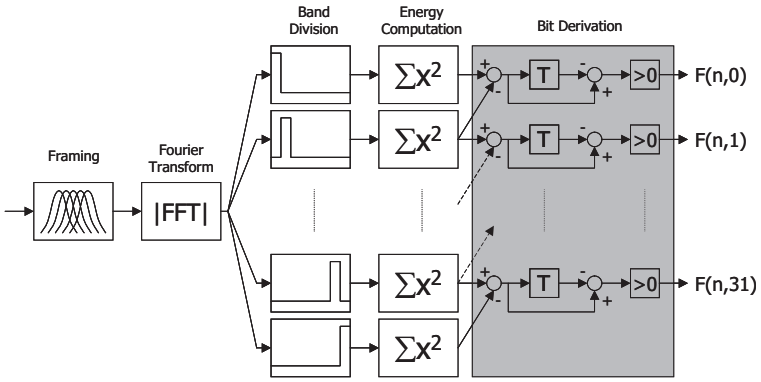


Figure 2.6: Fingerprint extraction stage of the Philips Robust Hash [44].

of these energies are taken in time *and* frequency:

$$\begin{aligned} ED(n,m) &= E^b(n,m) - E^b(n,m+1) \\ &\quad -(E^b(n-1,m) - E^b(n-1,m+1)) \end{aligned} \tag{2.1}$$

These energy differences now vary around 0. The bits of the sub-fingerprint (see Section 2.4.2) are derived by

$$F(n,m) = \begin{cases} 1 & ED(n,m) > 0 \\ 0 & ED(n,m) \leq 0, \end{cases} \tag{2.2}$$

where $F(n,m)$ denotes the m^{th} bit of the sub-fingerprint of frame n. In this way, the fingerprint bits are invariant to scaling of the signal. Each sub-fingerprint now consists of 32 bits, which can be efficiently stored as 4-Byte words.

Now consider the fingerprint block $\mathbf{F}^{N,M}(p,q)$ containing bits corresponding to M frequency indices and N sub-fingerprints, with lowest sub-fingerprint index being p and lowest frequency band index being q. This $N \times M$ fingerprint block is thus defined as the $\{0,1\}^{N \times M}$ matrix:

$$\mathbf{F}^{N,M}(p,q) \triangleq \begin{bmatrix} F(p, q) & \cdots & F(p, q+M-1) \\ \vdots & & \vdots \\ F(p+N-1, q) & \cdots & F(p+N-1, q+M-1) \end{bmatrix} \tag{2.3}$$

In this way, the nth sub-fingerprint is described by $\mathbf{F}^{1,M}(n,0)$, and a time-series of N fingerprint bits corresponding to frequency position m by $\mathbf{F}^{N,1}(0,m)$.

Figure 2.7 shows an example of the resulting fingerprint block. In the white areas the fingerprint bits are equal to one, in the black areas the bits equal to zero. Each 11.6 msec a 32 bit sub-fingerprint is computed. Due to the strong overlap there is strong correlation in the temporal dimension. This corresponds to a fingerprint rate of approximately 344.8 Bytes/sec. For identification, typically fingerprint blocks are used consisting of 256 sub-fingerprints, extracted from 3.3 seconds of music.

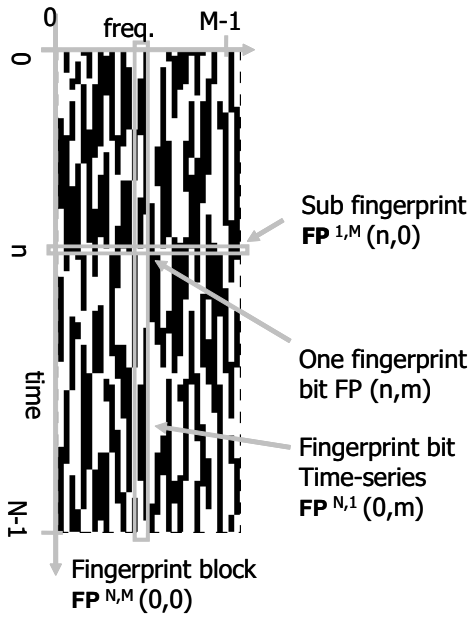


Figure 2.7: Example of a PRH fingerprint; indicated are the fingerprint block, the sub-fingerprint, and fingerprint bit time-series.

Figure 2.8 shows the search structure introduced in [44]. A look-up table (LUT) is maintained with pointers to each fingerprint position in the database containing that particular sub-fingerprint. The size of the LUT is coupled to the number of possible sub-fingerprint realizations. Assuming that some of the query sub-fingerprints are without errors, all pointers to a given sub-fingerprint can be easily retrieved. The retrieval procedure thus consists of two steps. First a matching sub-fingerprint is located based on the pointers provided in the LUT. Second, the other corresponding sub-fingerprints in the fingerprint block are directly retrieved from the database.

The assumption that at least one sub-fingerprint is without errors may be relaxed by also considering sub-fingerprints that have a certain Hamming distance, e.g., in which one or two bits may be different. This increases the search space. An alternative is to consider the energy differences on which the query sub-fingerprint is based. Assuming that the s smallest energy differences are most likely to result in erroneous bits, the search space can be expanded by considering all 2^s alternative sub-fingerprints. Instead of expanding the query, one may also include pointers to non-exact matching sub-fingerprints. In this way, an over-represented LUT is made.

Using a look-up table containing 2^{32} entries can be infeasible. Therefore, usually a hash table is used instead. The size of the stored fingerprint is thus not limited to the fingerprint itself, but also to the search structure. For each entry in the LUT a table should be maintained containing pointers to a song identifier and a position within a song. If both are represented by a 32-bit word, the effective size stored fingerprint is three times as large as the raw fingerprint itself, excluding metadata entries.

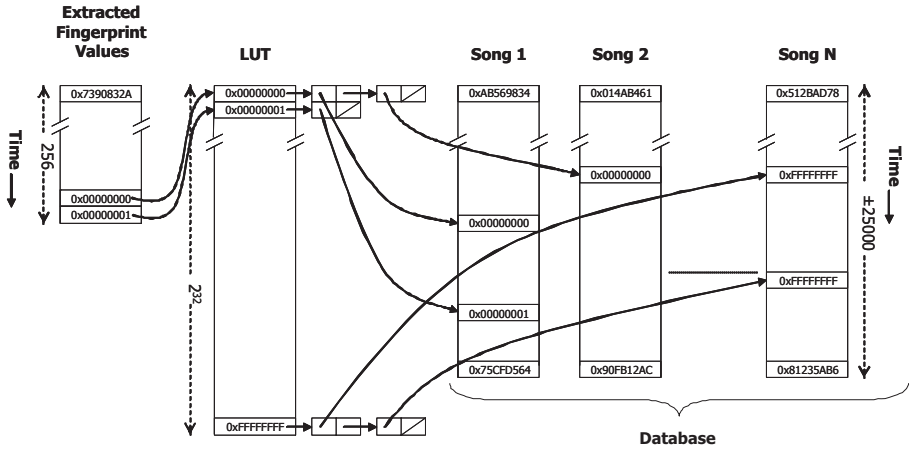


Figure 2.8: Search structure associated with the PRH [44].

An alternative search algorithm used tree-based pruning to reduce the search space [76]. Each 8-bit part of the fingerprint is considered a node in the tree; a fingerprint is represented as a path through a tree. All fingerprints in the database are then represented in one tree structure. Subtrees leading to a single child are efficiently represented. The method assumes that queries have a particular fixed length.

In the next chapter we analyze the PRH fingerprint extraction in more mathematical detail, and develop a statistical model for fingerprints extracted from i.i.d. fingerprints.

2.7 Objectives

In Table 2.3 of this chapter we have seen that there is a large number of audio fingerprinting algorithms. These algorithms typically follow the set-up in Figure 2.2, and the fingerprint extraction stage typically follows the set-up in Figure 2.3. Important characteristics for the performance of fingerprint algorithms are: the accuracy of the fingerprinting system (determined by the robustness of the fingerprint, the uniqueness of a fingerprint, and the size of the database), the granularity and the fingerprint rate, and the relation between these parameters.

There is little literature available with theoretical underpinning of the state-of-the-art fingerprinting algorithms. The objective of this thesis is to develop models for the fingerprint extraction stage, to increase the understanding of the design choices at hand, to allow for optimizations to be made, and to better understand the behavior observed in experiments. We develop these models for the PRH algorithm, which is described in the previous section. The goal is to identify the extraction and distortion parameters which affect the fingerprint and its recognition capabilities, and to analyze the effect of these parameters on the structure of the PRH fingerprint, the distortion in the fingerprint, and the identification capacity of the fingerprinting system.

We develop three models. Each model analyzes and quantifies a different property of the PRH fingerprint: the structure of the fingerprint, the effect of distortion of the input signal on the fingerprint, and the number of fingerprints that can be distinguished under certain conditions (capacity). The models have in common that they consider a fingerprint as the realization of a stochastic process, follow the algorithmic steps in Figures 2.2 and 2.3, assume Gaussian iid input signals, and are derived in particular for the PRH fingerprint.

In addition, we apply the distortion model to add functionality beyond identification. In this case, the goal is to estimate the quality of a distorted song (level of degradation, e.g., due to compression) following its identification using fingerprints. Here, we use the model for the distortion, and compare the predicted behavior to a number of practical fingerprinting algorithms, and to real music.

We now consider each of the three models in more detail.

Model for the PRH fingerprint structure
The objective of this model is to relate the structure of the PRH fingerprint to a number of parameters in the design of the PRH fingerprint.

The fingerprint structure is mainly represented by the temporal correlation between the fingerprint bits, illustrated by the black-and-white striped pattern in Figure 2.7. This structure is important for the uniqueness properties of the fingerprint; it affects the distance between arbitrary unrelated fingerprints indicated by the conditional PDF on the right-hand side in Figure 2.4. For a given fingerprint size, an increasing amount of correlation widens the distribution of the distance between unrelated fingerprints, but increases the robustness to temporal misalignment.

The model considers the PRH fingerprint as the output of a stochastic process; with each possible realization is associated a probability. To do so, we model the effect of each step in the fingerprint extraction process on the PRH fingerprint. We identify a number of parameters in the fingerprint extraction stage [35]: the relative frame overlap, the window type used in the Fourier transform, the number of frequency bands, and the bandwidth of the frequency bands. These parameters relate to the uniqueness property, the fingerprint rate and the granularity. The model relates the values for these parameters to the probability model for the structure of the fingerprint. The model is expected to provide more insight in the effect of design choices in the fingerprint extraction, and allow for subsequent optimization. The latter is not treated here in this thesis.

Model for the effect of signal distortion on the PRH fingerprint
Distortions in the input signal are reflected by changes in the corresponding fingerprint. The objective of our second model is to quantify the effect of additive white noise on the input signal and/or temporal misalignment on the difference introduced in the fingerprints.

Also in this model we closely model the steps in the fingerprint extraction. Other typical distortions commonly encountered in practice, in particular variations in the play-out speed of the audio signal are not taken into account in this model. Again the model assumes Gaussian iid input signals. The distortions considered in this model

affect the distribution of the distance between related (but distorted) fingerprints indicated by the conditional PDF on the left-hand side in Figure 2.4.

The model serves several goals. First, it relates the SNR of additive noise to the distance between corresponding fingerprints. This effectively quantifies the relation between distortion on the input signal and distortion on the fingerprint. It thus provides an indication how much the fingerprint changes under certain distortion conditions.

Second, it can be used to relate the observed distance between fingerprints to a coarse estimation of the quality difference of the input signals. It allows for new applications as mentioned in Section 1.2. The model is derived for the PRH fingerprint. We apply a framework for practical comparison of implementations of several fingerprint extraction algorithms (using the same operating point), to see whether the relation between SNR and fingerprint distance can be observed for other algorithms as well.

Model for the capacity of binary fingerprints like the PRH fingerprint

The main question for our third model is 'how many signals can be reliably identified by a fingerprinting system, under certain conditions'. The conditions relate to characteristics of the fingerprint (size of the fingerprint, and representation of the fingerprint), and characteristics of the environment in which the system operates (what kind of signals need to be identified, how much distortion is allowed). These characteristics relate to the previous models: the size and representation of the fingerprint relate to the models of the fingerprint structure; the effect of signal distortion is reflected in our second model.

Knowledge of the identification capacity of fingerprinting schemes is important since it provides upper bounds for the number of signals that can be identified. Furthermore, it provides an indication of the effects of design choices on this number, e.g., if the representation of the fingerprint is simplified, how much reduction in identification performance can we expect? How strong is the influence of certain conditions (e.g., expected SNR level) on the number of signals that can be identified.

For our third model, we consider an information theoretical framework developed by Westover and O'Sullivan (WOS) [104]. We use the model developed for the probability of an erroneous PRH fingerprint bit due to additive noise, and integrate this model into this WOS framework. In this way we estimate up to how many signals can be identified with a binary fingerprint like the PRH. We compare our bounds with the WOS bounds in [104]. Finally, we check whether the changes in the fingerprints we observe in practice due to distortions in the audio signals, and which have been modeled in this thesis, fit in the WOS framework.

Chapter 3

Models for PRH generated fingerprints of i.i.d. signals

Parts of this chapter are based on the following publications:

- P.J.O. Doets and R.L. Lagendijk, "Distortion Estimation in Compressed Music Using Only Audio Fingerprints", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pages 302-317, February 2008;
- P.J.O. Doets and R.L. Lagendijk, "Stochastic Model of a Robust Audio Fingerprinting System", in *Proceedings of ISMIR 2004*, pages 349-352, Oct. 2004;
- P.J.O. Doets and R.L. Lagendijk, "Extracting Quality Parameters for Compressed Audio from Fingerprints", in *Proceedings of ISMIR 2004*, pages 498-503, September 2005.

3.1 Introduction

The previous chapter introduced the PRH as an example audio fingerprinting algorithm. It is one of the best described algorithms found in literature, has good performance characteristics and is also used in commercial systems. Furthermore, it has a simple algorithmic structure.

Figure 3.1(a) shows a typical example of a fingerprint generated by the PRH algorithm. The fingerprint is computed on a 1.5 seconds of music. The fingerprint has a striking black-and-white pattern which is the result of the parameters that have been used in generating the fingerprint, and of the characteristics of the input signal. Figure 3.1 also shows the effect of MP3 compression at three different bitrates on the fingerprint of a song. Figure 3.1(a) shows the fingerprint of the original recording, while Figure 3.1(b) shows the fingerprint of the same fragment, but now compressed using MP3 at 128 kilobit per second (kbps). The fingerprints are largely the same, but there are some small differences. These differences are illustrated in Figure 3.1(c) by the black regions. Figures 3.1(d) and 3.1(e) show only the differences of the fingerprints of the fragments compressed at 64 kbps and 32 kbps, respectively, with respect to the original fingerprint.

From Figure 3.1 it is clear that the amount of fingerprint differences increase for lower bitrates. This relates to the robustness of the fingerprint. At some point the distortions in the song become dominant, and the differences between the fingerprint of the compressed version and the original recording exceed the threshold. These differences, on the other hand, might also be used to give some indication of the differences in the underlying songs. This observation is exploited in Chapter 4.

In this chapter we develop two models for the PRH algorithm. In the first model we capture the black-and-white structure of the PRH fingerprint by analytically computing the transition probabilities that a fingerprint bit $F(n, m) = 1$ is followed by a bit $F(n + 1, m) = 1$. The second model analyzes the probability that a fingerprint bit is flipped due to additive noise, temporal misalignment, or a combination of both. Additive noise has an effect on the fingerprint comparable to MP3 compression. Temporal misalignment occurs when in computation of the fingerprint the frames are put on a slightly different position in time than for the reference fingerprint.

In both models, we consider i.i.d. Gaussian input signals. Western music, however, typically consists of harmonic signals. Therefore, it would make sense to analyze the simple case of a fingerprint extracted from a single sinusoid. Figures 3.2(a) and 3.2(b) show two typical fingerprints extracted from a single sinusoid with random frequency and phase. The structure of the fingerprint is clearly different from the fingerprints extracted from real audio signals. Therefore, we consider the case of i.i.d. Gaussian signals. Figures 3.2(c) and 3.2(d) show two fingerprints extracted from Gaussian white noise. The spectrum of white noise is flat: on average all frequencies are equally strongly present in the signal. The structure of this type of fingerprints has a strong visual resemblance with the fingerprint shown in Figure 2.7.

To further motivate the use of Gaussian i.i.d. input signals in our modeling effort, consider the number of differences between the fingerprint bits of dissimilar content. When comparing two arbitrary fingerprint blocks, on average 50 % of the bits are different. Around this average value there is some variation. This is also clear from

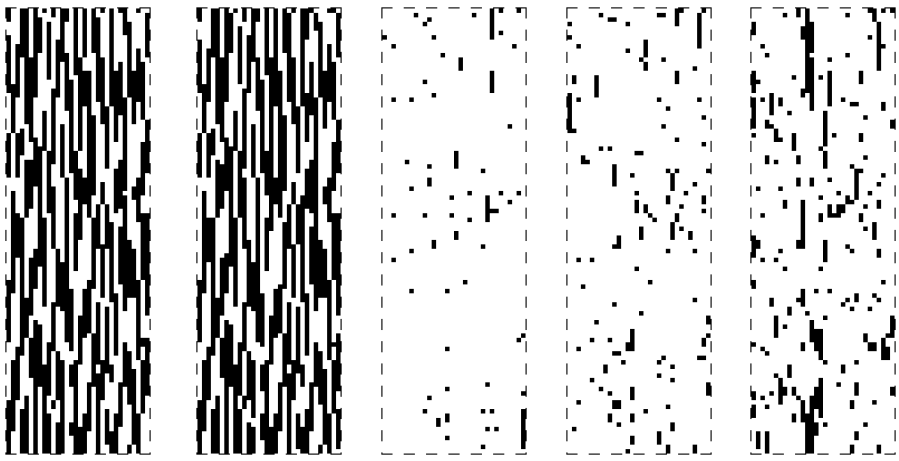


Figure 3.1: PRH fingerprints of 1.5 seconds of music (a) of the original recording; (b) of the same recording, but MP3 compressed at 128 kbps; (c) difference between fingerprints of original and of MP3@128 kbps; the black positions mark the differences with respect to the fingerprint shown in (a); (d) difference between fingerprints of original and of MP3@64 kbps; (e) difference between fingerprints of original and of MP3@32 kbps. Illustration after [44].

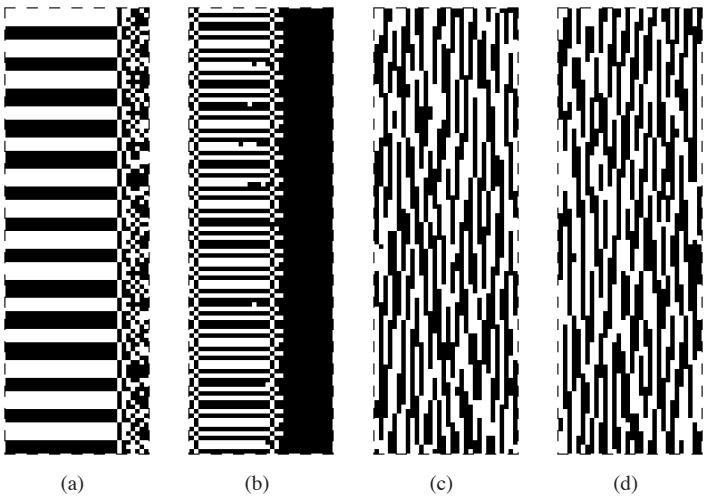


Figure 3.2: PRH fingerprints of: (a-b) Single sinusoid signals with random frequency and initial phase; (c-d) Gaussian i.i.d. fragments.

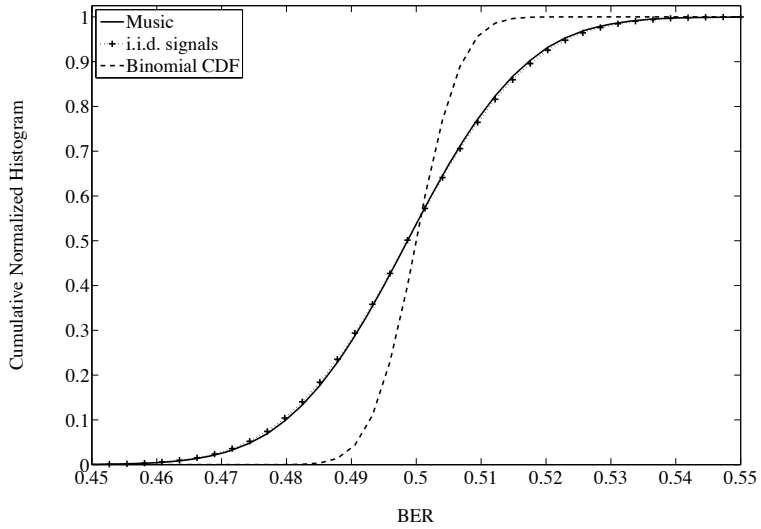


Figure 3.3: Cumulative distribution function for the BER between two arbitrary binary patterns. Three distance between three binary patterns are shown: PRH fingerprints for music ($-$); PRH fingerprints for i.i.d. Gaussian signals ($\cdot\cdot+\cdot\cdot$); normalized binomial cdf with n equal to 32×256 (bits) and $p = \frac{1}{2}$ ($- - -$).

the conditional pdf distance between the fingerprints of dissimilar content in Figure 2.4 in the previous chapter. Suppose all fingerprint bits are independent, and have equal probability of being '0' or '1'. Then also the bits in difference (XOR) between two arbitrary fingerprints have equal probability of being '0' or '1'. In this case, the number of fingerprint bits that are different between the two fingerprints follows a binomial distribution. Figure 3.3 shows the cumulative distribution of differences between arbitrary PRH fingerprints. As a reference the figure also shows the curve for a binomial distribution. The curves for music signals and i.i.d. Gaussian signals practically overlap. This supports the visual similarity between fingerprints from real music and fingerprints from i.i.d. Gaussian signals, and is a motivation to use i.i.d. Gaussian signals for our models.

This chapter is organized as follows. Section 3.2 discusses the PRH fingerprint extraction in more mathematical detail. Section 3.3 derives the first model that predicts the transition probabilities of subsequent fingerprint bits. Section 3.4 discusses the second model that predicts the probability of an erroneous fingerprint bit due to noise and misalignment. Section 3.5 looks for similarities between the PRH fingerprint and other fingerprinting algorithms for which the similar models can be made. Section 3.6 draws conclusions and discusses potential extensions of the developed models.

3.2 Philips Robust Hash: model setup

In Section 2.6 we introduced the PRH algorithm. In this section we present the algorithm in more mathematical detail. It forms the basis for the development of the two statistical models in Sections 3.3 and 3.4. Here, we only consider the fingerprint extraction stage, and do not consider the matching and actual identification.

As a first step, the time domain signal is divided into overlapping frames of length L. Each frame is shifted ΔL samples with respect to the previous frame, where $0 \leq \Delta L \leq L$. The signal in frame n is the defined as:

$$x(n,i) = x(n\Delta L + i), \qquad \begin{matrix} n{=}0,1,\ldots \\ i{=}1,\ldots,L-1 \end{matrix} \tag{3.1}$$

The Fourier transform is taken of each frame. The Fourier transform of $x(n,i)$ uses a window $w(i)$:

$$\hat{x}(n,k) = \sum_{i=0}^{L-1} w(i)x(n,i)e^{-j2\pi\frac{k}{L}i}, \qquad k = 0,\ldots,L-1 \tag{3.2}$$

The Power Spectral Density (PSD) is estimated using a periodogram estimator. Denoting the PSD of frame n at frequency bin k as $S_X(n,k)$, it is defined as:

$$S_X(n,k) = \frac{1}{L}\left|\hat{x}(n,k)\right|^2 \tag{3.3}$$

Within the PSD, $M+1$ non-overlapping frequency bands are defined. The energy of frequency band m of frame n is given by:

$$E^b(n,m) = \sum_{k\in\mathcal{K}_m} S_X(n,k) \qquad \begin{matrix} n{=}0,1,\ldots \\ m{=}0,\ldots,M, \end{matrix} \tag{3.4}$$

where $\mathcal{K}_m$ denotes the set of frequency indices which fall within frequency band m. Differences of these energies are taken in time *and* frequency:

$$\begin{aligned} ED(n,m) = {} & E^b(n,m) - E^b(n,m{+}1) \\ & -(E^b(n{-}1,m) - E^b(n{-}1,m{+}1)) \quad m{=}0,\ldots,M{-}1 \end{aligned} \tag{3.5}$$

The bits of the sub-fingerprint are derived by

$$F(n,m) = \begin{cases} 1 & ED(n,m) > 0 \\ 0 & ED(n,m) \leq 0 \end{cases}, \tag{3.6}$$

where $F(n,m)$ denotes the mth bit of the sub-fingerprint of frame n.

In order to build a stochastic model the time delay operation, T, is shifted forward yielding the equivalent arrangement shown in Figure 3.4. Here, the difference between samples in two subsequent periodogram estimates of the PSD is computed first *in time only*:

$$ED^s(n,k) \quad = \quad S_X(n,k) - S_X(n-1,k) \tag{3.7}$$

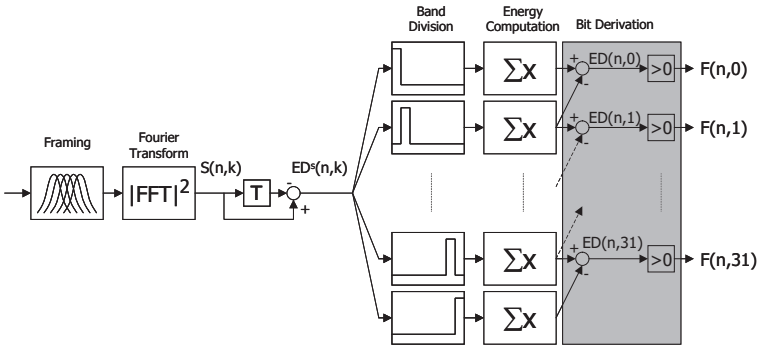


Figure 3.4: Functionally equivalent configuration of the PRH fingerprint extraction used for modeling purposes.

The *differential* energy in frequency band m is given by:

$$ED^b(n,m) \quad = \quad \sum_{k \in \mathcal{K}_m} ED^s(n,k) \tag{3.8}$$

Finally, we show that we can obtain $ED(n,m)$ as:

$$ED(n,m) \quad = \quad ED^b(n,m) - ED^b(n,m+1) \tag{3.9}$$

The statistical analysis of the rearranged configuration is easier, because the correlation between spectral samples in adjacent frames is taken into account from the beginning.

It is easy to show that Eq. (3.9) is equivalent to Eq. (3.5):

$$\begin{aligned}
ED(n,m) &\overset{(3.9)}{=} ED^b(n,m) - ED^b(n,m+1) \\
&= \sum_{k \in \mathcal{K}_m} ED^s(n,k) - \sum_{k \in \mathcal{K}_{m+1}} ED^s(n,k) \\
&= \sum_{k \in \mathcal{K}_m} (S_X(n,k) - S_X(n-1,k)) \\
&\quad - \sum_{k \in \mathcal{K}_{m+1}} (S_X(n,k) - S_X(n-1,k)) \\
&= E^b(n,m) - E^b(n,m+1) \\
&\quad -(E^b(n-1,m) - E^b(n-1,m+1)) \\
&\overset{(3.5)}{=} ED(n,m)
\end{aligned}$$

3.3 Statistics of fingerprint bits

In this section, we present a stochastic model that describes the probability that an i.i.d. Gaussian input source results in a particular binary fingerprint. The model assumes that the energy differences $ED(n,m)$ that result in the fingerprint bits, are also Gaussian.

This section is organized as follows. Section 3.3.1 outlines the model. The model is dependent on a correlation matrix with a particular structure; this is discussed in Section 3.3.2. On its turn, the correlation matrix is dependent on one specific parameter: the variance as function of the frame shift; this is discussed in Section 3.3.3. Finally, the transition probabilities for subsequent fingerprint bits are computed for rectangular windows in Section 3.3.4, and for symmetric non-rectangular windows in Section 3.3.5.

3.3.1 Notation and outline of the model

Each fingerprint bit $F(n,m)$ in the fingerprint block $\mathbf{F}^{N,M}(n,m)$ is the result of thresholding the spectral energy difference $ED(n,m)$. We denote the series of $ED(n,m)$ values resulting in the m^{th} sub-fingerprint as $\mathbf{ED}^{1,M}(n,m) = [ED(n,m), ED(n,m+1), \ldots, ED(n,m+M-1)]$, and the $ED(n,m)$ values resulting in a fingerprint block as the vector $\mathbf{ED}^{N,M}(n,m) = [\mathbf{ED}^{1,M}(n,m), \mathbf{ED}^{1,M}(n+1,m), \ldots, \mathbf{ED}^{1,M}(n+N-1,m)]$. Let us assume that $E^b(n,m)$ is stationary in n; hence $\mathbb{E}[E^b(n,m)] = \mathbb{E}[E^b(n-1,m)]$ and $\mathbb{E}[ED(n,m)] = 0$. Let us further assume that the underlying $ED(n,m)$ values are realizations of a Gaussian stochastic process. Therefore,

$$\mathbf{ED}^{N,M}(n,m) \sim \mathcal{N}(\mathbf{0}, \mathbf{C_{ED}}), \tag{3.10}$$

where $\mathbf{C_{ED}} = \mathbf{C}_{\mathbf{ED}^{N\times M}(n,m)} \in \mathbb{R}^{NM\times NM}$ represents the covariance matrix, which will be defined in Section 3.3.2. The underlying multivariate Gaussian probability density function (pdf) $f_{\mathbf{ED}^{N,M}}(\mathbf{n})$ yields the probability density that $\mathbf{ED}^{N,M}(n,m) = \mathbf{n}$, and is fully characterized by the covariance matrix $\mathbf{C_{ED}}$:

$$f_{\mathbf{ED}^{N,M}}(\mathbf{n}) = \frac{1}{\sqrt{(2\pi)^{NM}\det(\mathbf{C_{ED}})}} \exp\left[-\frac{1}{2}\mathbf{n}'\ \mathbf{C_{ED}}^{-1}\ \mathbf{n}\right] \tag{3.11}$$

where $\mathbf{n}'$ denotes the transpose of $\mathbf{n}$. The probability of a certain fingerprint block can now be computed by integrating the pdf over the area $\Psi(\mathbf{a})$ of the $\mathbf{ED}^{N,M}(n,m)$-space associated with the fingerprint block $\mathbf{F}^{N,M}(n,m) = \mathbf{a}$:

$$Pr\left[\mathbf{F}^{N,M}(n,m) = \mathbf{a}\right] \quad = \quad \int_{\Psi(\mathbf{a})} f_{\mathbf{ED}^{N,M}(n,m)}(\mathbf{n})\ d\mathbf{n} \tag{3.12}$$

For instance, assume we have a fingerprint block consisting of three times bit '1' in the temporal direction: $\mathbf{F}^{3,1}(0,0) = [F(0,0), F(1,0), F(2,0)] = [111]$, we can compute

the probability by:

$$Pr\left[\mathbf{F}^{3,1}(0,0)=\mathbf{a}\right] = \int_0^\infty \int_0^\infty \int_0^\infty f_{\mathbf{ED}^{3,1}(0,0)}(n_1,n_2,n_3)\; d(n_1,n_2,n_3) \tag{3.13}$$

In practice, the number of closed-form solutions for integration of a Gaussian pdf is limited to very specific cases, and for low dimensions only. For higher orders, the Gaussian integration can only be obtained through numerical integration, or through simplification of the model structure. One way is to describe the binary time-series as a Markov chain. A simple example shows formulation of the bits in the mth band as a first-order Markov chain.

$$Pr\left[\mathbf{F}^{N,1}(0,m)=\mathbf{a}\right] \approx Pr\left[F(0,m)=a(0)\right] \times \prod_{i=1}^{N-1} Pr\left[F(i,m)=a(i)|F(i-1,m)=a(i-1)\right] \tag{3.14}$$

where $Pr\left[F(0,m)=a(0)\right]$ refers to the a priori probability of a value for the first fingerprint bit $F(0,m)$. The conditional probabilities correspond to the transition probabilities in the Markov model, where the state is dependent on the value of the previous fingerprint bit.

3.3.2 Structure of the correlation matrix

The goal of this section is to analyze the structure of the correlation matrix $\mathbf{C}_{\mathbf{ED}^{N,M}(n,m)}$. Since the energy differences are assumed zero-mean Gaussian, the correlation matrix fully defines the pdf $f_{\mathbf{ED}^{N,M}(n,m)}(\mathbf{n})$. We separately consider the correlation in the spectral and in the temporal dimension.

First consider the correlation in the spectral dimension. The correlation between energy differences $ED(n,m)$ and $ED(n,m+1)$ is mainly introduced because they are both partly based on the same differential frequency band energy $ED^b(n,m+1)$. Assuming the neighboring bands themselves are uncorrelated

$$\text{COV}\left[ED^b(n,m), ED^b(n+l,m+p)\right]=0 \qquad p\neq 0,\ \forall\, l,n,m \tag{3.15}$$

the correlation COV$[ED(n,m), ED(n+l,m+p)]$ in neighboring spectral energy differences is dependent only on the width of the frequency bands. Within one sub-fingerprint ($l=0$), the correlation then is given by:

$$\begin{aligned}
&\text{COV}\left[ED(n,m), ED(n,m+p)\right] \\
&= \text{COV}\big[ED^b(n,m)-ED^b(n,m+1), \\
&\qquad ED^b(n,m+p)-ED^b(n,m+p+1)\big] \\
&= \begin{cases} -\text{VAR}\left[ED^b(n,m)\right] & p=-1 \\ \text{VAR}\left[ED^b(n,m)\right]+\text{VAR}\left[ED^b(n,m+1)\right] & p=0 \\ -\text{VAR}\left[ED^b(n,m+1)\right] & p=1 \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{3.16}$$

PRH uses logarithmic frequency bands, i.e. the bandwidth of band $m+1$ is $\alpha > 1$ times as large as the bandwidth of band m. These bands are designed such that each band covers exactly one harmonic tone. Twelve tones fit in an octave. Therefore, in PRH the scaling factor $\alpha = \sqrt[12]{2} \approx 1.06$. Of course, in practice the width of a frequency band is an integer. PRH with uniform frequency bands can be considered as a special case where $\alpha = 1$. For a bandwidth scaling factor α, the variance of $ED^b(n,m)$ can be approximated by:

$$\begin{aligned} \text{VAR}\left[ED^b(n,m)\right] &\approx \alpha \cdot \text{VAR}\left[ED^b(n,m-1)\right] \\ &= \alpha^m \cdot \text{VAR}\left[ED^b(n,0)\right] \end{aligned} \tag{3.17}$$

This approximation is accurate if the spectral sample autocorrelation function falls off rapidly as a function of the lag, compared to the bandwidth.

The variance of a spectral energy difference $ED(n,m)$ can thus be written as:

$$\begin{aligned} \text{VAR}\left[ED(n,m)\right] &= \text{VAR}\left[ED^b(n,m) - ED^b(n,m+1)\right] \\ &= \text{VAR}\left[ED^b(n,m)\right] + \text{VAR}\left[ED^b(n,m+1)\right] \\ &= \alpha^m \cdot \text{VAR}\left[ED^b(n,0)\right] + \alpha^m \cdot \text{VAR}\left[ED^b(n,1)\right] \\ &= \alpha^m \cdot \text{VAR}\left[ED(n,0)\right] \end{aligned} \tag{3.18}$$

Similarly, the covariance terms for $p \neq 0$ in Eq. (3.16) can be computed.

Spectral correlation determines the prior probability of sub-fingerprint realizations, and is introduced by the computation of energy differences from non-overlapping frequency bands. The amount of correlation is solely determined by the width of the frequency bands through the bandwidth scaling factor α.

The correlation matrix for a single sub-fingerprint $\mathbf{C}_{\mathbf{ED}^{1,M}}$ is given by:

$$\mathbf{C}_{\mathbf{ED}^{1,M}} = \text{VAR}\left[ED(n,0)\right] \begin{bmatrix} 1 & -\frac{\alpha}{1+\alpha} & 0 & \cdots & 0 \\ -\frac{\alpha}{1+\alpha} & \alpha & -\frac{\alpha^2}{1+\alpha} & \cdots & 0 \\ 0 & -\frac{\alpha^2}{1+\alpha} & \alpha^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha^{M-1} \end{bmatrix} \tag{3.19}$$

The variable $ED^b(n,m)$ is stationary in n. To cover the temporal correlation introduced by the overlapping frames, we define the temporal correlation coefficient.

$$\rho_l \triangleq \frac{\text{COV}\left[ED^b(n,m), ED^b(n \pm l, m)\right]}{\text{VAR}\left[ED^b(n,m)\right]} \tag{3.20}$$

In combination with Eqs. (3.15) and (3.17) we can now include the temporal correlation in the expressions:

$$\begin{aligned} &\text{COV}\left[ED^b(n,m), ED^b(n \pm l, m+p)\right] \\ &= \begin{cases} \rho_l \cdot \alpha^{m-1} \cdot \text{VAR}\left[ED^b(n,1)\right] & p = 0, \quad \forall\, l, n, m \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{3.21}$$

The overall correlations defining the correlation matrix become:

$$\begin{aligned}&\text{COV}\left[ED(n,m), ED(n \pm l, m+p)\right] \\ &= \rho_l \cdot \alpha^{m-1} \cdot \text{VAR}\left[ED(n,1)\right] \cdot \begin{cases} -\frac{1}{1+\alpha} & p=-1 \\ 1 & p=0 \\ -\frac{\alpha}{1+\alpha} & p=1 \\ 0 & \text{otherwise} \end{cases}\end{aligned} \tag{3.22}$$

which yields the overall correlation matrix $\mathbf{C_{ED}}$ as a block matrix:

$$\mathbf{C_{ED}}^{N,M} = \begin{bmatrix} \mathbf{C_{ED}}^{1,M} & \rho_1 \cdot \mathbf{C_{ED}}^{1,M} & \cdots & \rho_{N-1} \cdot \mathbf{C_{ED}}^{1,M} \\ \rho_1 \cdot \mathbf{C_{ED}}^{1,M} & \mathbf{C_{ED}}^{1,M} & \cdots & \rho_{N-2} \cdot \mathbf{C_{ED}}^{1,M} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{N-1} \cdot \mathbf{C_{ED}}^{1,M} & \rho_{N-2} \cdot \mathbf{C_{ED}}^{1,M} & \cdots & \mathbf{C_{ED}}^{1,M} \end{bmatrix}. \tag{3.23}$$

The temporal correlation is dependent on the relative temporal overlap between successive frames, $1 - \frac{\Delta L}{L}$. Here, $\frac{\Delta L}{L}$ denotes the frame shift ratio. Therefore, the correlation matrix $\mathbf{C_{ED}}$ is dependent on ΔL and L.

Given the bandwidth scaling factor α and the number of frequency bands and sub-fingerprints considered, M and N respectively, the correlation matrix $\mathbf{C_{ED}}^{N,M}$ is fully parameterized by the time correlation parameters $\rho_l, l = 1, \ldots, N-1$ and the variance $\text{VAR}\left[ED(n,1)\right]$. The variance term linearly scales the covariance matrix, and has no influence on probability computations. Estimation of ρ_l can be done through the variance as a function of frame shift ΔL for a given frame length L.

As stated in the following theorem, the covariance $\text{COV}[ED(n,m), ED(n+l,m)]$ can be expressed in terms of the variance $\text{VAR}[ED^{\Delta L'}(n,m)]$, where $ED^{\Delta L'}$ denotes the variable ED generated with frame shift $\Delta L' = l\Delta L$. It is straightforward that

$$\text{VAR}\left[ED^{\Delta L'}(n,m)\right] = \begin{cases} 0 & \Delta L' = 0 \\ \text{VAR}\left[ED^{l\Delta L}(n,m)\right] & \Delta L' < L \\ \text{VAR}\left[ED^{L}(n,m)\right] & \Delta L' \geq L \end{cases}$$

Theorem 1 (Expressing the covariance of spectral energy differences in terms of variances with variable frame shift)**.**

$$\begin{aligned} &\mathit{COV}\left[ED(n,m), ED(n+l,m)\right] \\ &= -\mathit{VAR}\left[ED^{l\Delta L}(n,m)\right] \\ &\quad + \frac{1}{2}\mathit{VAR}\left[ED^{(l-1)\Delta L}(n,m)\right] + \frac{1}{2}\mathit{VAR}\left[ED^{(l+1)\Delta L}(n,m)\right] \end{aligned} \tag{3.24}$$

The proof of the theorem is given in Appendix A.2. By using this theorem, the entire correlation matrix $\mathbf{C_{ED}}$ is defined by the variance as function of the frame shift ΔL.

3.3.3 Expressing the variance as a function of the frame shift

The correlation matrix $\mathbf{C_{ED}}^{N,M}$ is fully defined by the scaling factor α, and the variance VAR$[ED^b(n,m)]$ as a function of the frame shift ΔL. This variance of $ED^b(n,m) = \sum_k ED^s(n,k)$ is given by

$$\text{VAR}[ED^b(n,m)] \quad = \quad K_m\, C_{ED^s}(0) + 2\sum_{l=1}^{K_m-1}(K_m - l)C_{ED^s}(l) \tag{3.25}$$

where $C_{ED^s}(l)$ represents the covariance function of the sample-wise spectral difference:

$$C_{ED^s}(l) \quad = \quad \text{COV}[ED^s(n,k),\ ED^s(n,k+l)] \tag{3.26}$$

where is assume that $ED^s(n,k)$ is stationary both in n and k. This covariance function is related to samples in the overlapping spectrograms:

$$\begin{aligned} C_{ED^s}(l) &= \mathbb{E}\left[ED^s(n,k)\, ED^s(n,k+l)\right] \\ &= \mathbb{E}\big[(S_X(n,k) - S_X(n-1,k)) \\ &\qquad \times (S_X(n,k+l) - S_X(n-1,k+l))\big] \\ &= 2\,\mathbb{E}\left[S_X(n,k)\, S_X(n,k+l)\right] \\ &\qquad -2\,\mathbb{E}\left[S_X(n,k)\, S_X(n-1,k+l)\right] \end{aligned} \tag{3.27}$$

In order to express Eq. (3.27) in terms of L and ΔL, we first need to consider the computation of the overlapping periodograms $S_X(n,k)$ and $S_X(n-1,k)$. Figure 3.5 illustrates the frames n and $n-1$ in time that are combined into $ED^s(n,k)$. Each of these frames can be split into overlapping and non-overlapping regions:

$$\begin{aligned} \hat{x}(n,k) &= \sum_{i=0}^{L-1} w(i)x(n,i)e^{-j2\pi\frac{k}{L}i} \\ &= \underbrace{\sum_{i=0}^{L-\Delta L-1} w(i)x(n,i)e^{-j2\pi\frac{k}{L}i}}_{\hat{a}(n,k)} + \underbrace{\sum_{i=L-\Delta L}^{L-1} w(i)x(n,i)e^{-j2\pi\frac{k}{L}i}}_{\hat{b}(n,k)} \end{aligned}$$

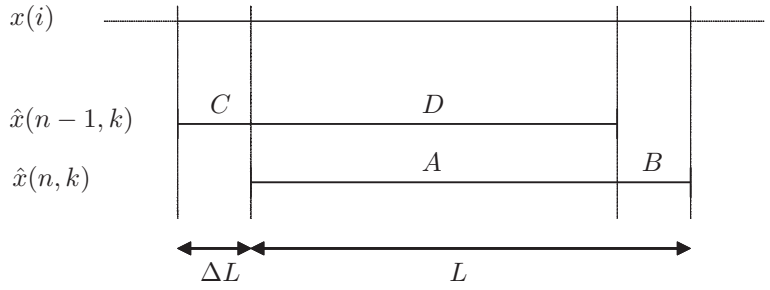


Figure 3.5: Illustration of two time frames split into overlapping and non-overlapping regions.

and:

$$\begin{aligned}\hat{x}(n-1,k) &= \sum_{i=0}^{L-1} w(i)x(n-1,i)e^{-j2\pi\frac{k}{L}i} \\ &= \underbrace{\sum_{i=0}^{\Delta L-1} w(i)x(n-1,i)e^{-j2\pi\frac{k}{L}i}}_{\hat{c}(n,k)} + \underbrace{\sum_{i=\Delta L}^{L-1} w(i)x(n-1,i)e^{-j2\pi\frac{k}{L}i}}_{\hat{d}(n,k)}\end{aligned}$$

Each sum corresponding to one of the four regions in Figure 3.5 can be split into a real and imaginary part, e.g. $\hat{a}(n,k) = R_A(n,k) + jI_A(n,k)$, where $R_A(n,k) = \mathrm{Re}\left(\hat{a}(n,k)\right)$ and $I_A(n,k) = \mathrm{Im}\left(\hat{a}(n,k)\right)$, respectively.

The correlation between overlapping frames can be expressed in terms of the real and imaginary parts of the frames:

$$\begin{aligned}&\mathbb{E}\left[S_X(n,k)\, S_X(n,k+l)\right] \\ &= \frac{1}{L^2}\mathbb{E}\left[\left(R_X^2(n,k) + I_X^2(n,k)\right)\left(R_X^2(n,k+l) + I_X^2(n,k+l)\right)\right] \\ &= \frac{2}{L^2}\left(\mathbb{E}\left[R_X^2(n,k)\right]\,\mathbb{E}\left[I_X^2(n,k)\right] + \mathbb{E}\left[R_X^2(n,k)\, R_X^2(n,k+l)\right]\right) \\ &= \frac{4}{L^2}\left(\mathbb{E}\left[R_X^2(n,k)\right]^2 + \mathbb{E}\left[R_X(n,k)\, R_X(n,k+l)\right]^2\right) \qquad (3.28)\end{aligned}$$

and overlapping regions:

$$\begin{aligned}&\mathbb{E}[(S_X(n,k)\, S_X(n-1,k+l)] \\ &= \frac{1}{L^2}\Big(2\mathbb{E}\left[R_X^2(n,k)\, R_X^2(n-1,k+l)\right] \\ &\quad + \mathbb{E}\left[R_X^2(n,k)\, I_X^2(n-1,k+l)\right] + \mathbb{E}\left[I_X^2(n,k)\, R_X^2(n-1,k+l)\right]\Big)\end{aligned}$$

$$
\begin{aligned}
&= \frac{2}{L^2}\Big(2\mathbb{E}\left[R_X^2(n,k)\right]^2 \\
&\quad +\mathbb{E}\left[R_X(n,k)\, R_X(n-1,k+l)\right]^2 + \mathbb{E}\left[I_X(n,k)\, I_X(n-1,k+l)\right]^2 \\
&\quad +\mathbb{E}\left[R_X(n,k)\, I_X(n-1,k+l)\right]^2 + \mathbb{E}\left[I_X(n,k)\, R_X(n-1,k+l)\right]^2 \Big) \\
&= \frac{2}{L^2}\Big(2\mathbb{E}\left[R_X^2(n,k)\right]^2 + \mathbb{E}\left[R_A(k)\, R_D(k+l)\right]^2 + \mathbb{E}\left[I_A(k)\, I_D(k+l)\right]^2 \\
&\quad +\mathbb{E}\left[R_A(k)\, I_D(k+l)\right]^2 + \mathbb{E}\left[I_A(k)\, R_D(k+l)\right]^2 \Big) \qquad (3.29)
\end{aligned}
$$

where we use the fact that both $R_X(n,k)$ and $I_X(n,k)$ are Gaussian, so we can use the characteristic function to derive the joint second moment:

$$\mathbb{E}[X^2\, Y^2] \;=\; \sigma_X^2\, \sigma_Y^2\, (1 + 2\rho_{XY}^2)$$

Combining the two yields the correlation function:

$$
\begin{aligned}
C_{ED^s}(l) \;&=\; 2\,\mathbb{E}\left[S_X(n,k)\, S_X(n,k+l)\right] \\
&\qquad -2\,\mathbb{E}\left[S_X(n,k)\, S_X(n-1,k+l)\right] \\
&=\; \frac{4}{L^2}\Big(2\mathbb{E}\left[R_X(n,k)\, R_X(n,k+l)\right]^2 \\
&\qquad -\Big(\mathbb{E}\left[R_A(k)\, R_D(k+l)\right]^2 + \mathbb{E}\left[I_A(k)\, I_D(k+l)\right]^2 \\
&\qquad\quad +\mathbb{E}\left[R_A(k)\, I_D(k+l)\right]^2 + \mathbb{E}\left[I_A(k)\, R_D(k+l)\right]^2 \Big)\Big) \qquad (3.30)
\end{aligned}
$$

In Sections 3.3.4 and 3.3.5 we derive expressions for Eq. (3.30) for rectangular and symmetric non-rectangular windows, respectively. Based on these expressions, we can compute the transition probabilities of subsequent fingerprint bits.

3.3.4 Transition probabilities for a rectangular window

In this section we derive an expression for the transition probabilities of subsequent fingerprint bits. The expression is given in Eq. (3.38), and is indirectly based on Eq. (3.30).

The rectangular window is simply given by $w_R(i) = 1,\ i = 0, \ldots, L-1$. The overlapping regions in two subsequent frames are related by:

$$
\begin{aligned}
\hat{d}_R(k) \;&=\; \sum_{i=\Delta L}^{L-1} x(n-1,i) e^{-j2\pi \frac{k}{L} i} \\
&=\; \sum_{i=\Delta L}^{L-1} x(i+(n-1)\Delta L) e^{-j2\pi \frac{k}{L} i} \\
&=\; e^{-j2\pi \frac{\Delta L}{L} k} \sum_{i=0}^{L-\Delta L-1} x(n,i) e^{-j2\pi \frac{k}{L} i} \\
&=\; e^{-j2\pi \frac{\Delta L}{L} k} \hat{a}_R(k) \qquad (3.31)
\end{aligned}
$$

The real part $R_{D_R}(k)$ and imaginary part $I_{D_R}(k)$ can then be formulated in terms of $R_{A_R}(k)$ and $I_{A_R}(k)$:

$$R_{D_R}(k) = \cos\left(2\pi k\tfrac{\Delta L}{L}\right) R_{A_R}(k) + \sin\left(2\pi k\tfrac{\Delta L}{L}\right) I_{A_R}(k) \quad (3.32)$$

$$I_{D_R}(k) = \cos\left(2\pi k\tfrac{\Delta L}{L}\right) I_{A_R}(k) - \sin\left(2\pi k\tfrac{\Delta L}{L}\right) R_{A_R}(k) \quad (3.33)$$

Using these expressions we can write down the correlation terms needed in Eq. (3.29):

$$\begin{aligned}
\mathbb{E}[R_{A_R}(k)R_{D_R}(k+l)] &= \cos\left(2\pi\tfrac{\Delta L}{L}(k+l)\right)\mathbb{E}[R_{A_R}(k)R_{A_R}(k+l)] \\
&\quad + \sin\left(2\pi\tfrac{\Delta L}{L}(k+l)\right)\mathbb{E}[R_{A_R}(k)I_{A_R}(k+l)] \\
\mathbb{E}[R_{A_R}(k)I_{D_R}(k+l)] &= \cos\left(2\pi\tfrac{\Delta L}{L}(k+l)\right)\mathbb{E}[R_{A_R}(k)I_{A_R}(k+l)] \\
&\quad - \sin\left(2\pi\tfrac{\Delta L}{L}(k+l)\right)\mathbb{E}[R_{A_R}(k)R_{A_R}(k+l)] \\
\mathbb{E}[I_{A_R}(k)R_{D_R}(k+l)] &= \cos\left(2\pi\tfrac{\Delta L}{L}(k+l)\right)\mathbb{E}[I_{A_R}(k)R_{A_R}(k+l)] \\
&\quad + \sin\left(2\pi\tfrac{\Delta L}{L}(k+l)\right)\mathbb{E}[I_{A_R}(k)I_{A_R}(k+l)] \\
\mathbb{E}[I_{A_R}(k)I_{D_R}(k+l)] &= \cos\left(2\pi\tfrac{\Delta L}{L}(k+l)\right)\mathbb{E}[I_{A_R}(k)I_{A_R}(k+l)] \\
&\quad - \sin\left(2\pi\tfrac{\Delta L}{L}(k+l)\right)\mathbb{E}[I_{A_R}(k)R_{A_R}(k+l)]
\end{aligned}$$

Substitution in Eq. (3.29) yields the correlation function of the overlapping spectrograms:

$$\begin{aligned}
&\mathbb{E}[(S_X(n,k)\,S_X(n-1,k+l)] \\
&= \frac{2}{L^2}\Big(\mathbb{E}[R_A(k)\,R_A(k+l)]^2 + \mathbb{E}[R_A(k)\,I_A(k+l)]^2 \\
&\qquad + \mathbb{E}[I_A(k)\,R_A(k+l)]^2 + \mathbb{E}[I_A(k)\,I_A(k+l)]^2 \\
&\qquad + 2\,\mathbb{E}[R_X^2(n,k)]^2\Big)
\end{aligned} \quad (3.34)$$

Under the assumption that the correlation terms are w.s.s., we can express the overall sample-wise correlation function:

$$\begin{aligned}
\mathbb{E}[R_A(k)\,R_A(k+l)] &= \mathbb{E}[I_A(k)\,I_A(k+l)] \\
&= C_{R_A}(l) \quad (3.35) \\
\mathbb{E}[R_A(k)\,I_A(k+l)] &= -\mathbb{E}[I_A(k)\,R_A(k+l)] \\
&= C_{R_A,I_A}(l), \quad (3.36)
\end{aligned}$$

where the correlation functions are

$$C_{R_A}(l) = \frac{1}{2}\sigma_X^2\cos(l\pi\tfrac{L-\Delta L-1}{L})\frac{\sin(l\pi\frac{L-\Delta L}{L})}{\sin(l\pi\frac{1}{L})}$$

$$C_{R_A,I_A}(l) = \frac{1}{2}\sigma_X^2\sin(l\pi\tfrac{L-\Delta L-1}{L})\frac{\sin(l\pi\frac{L-\Delta L}{L})}{\sin(l\pi\frac{1}{L})}$$

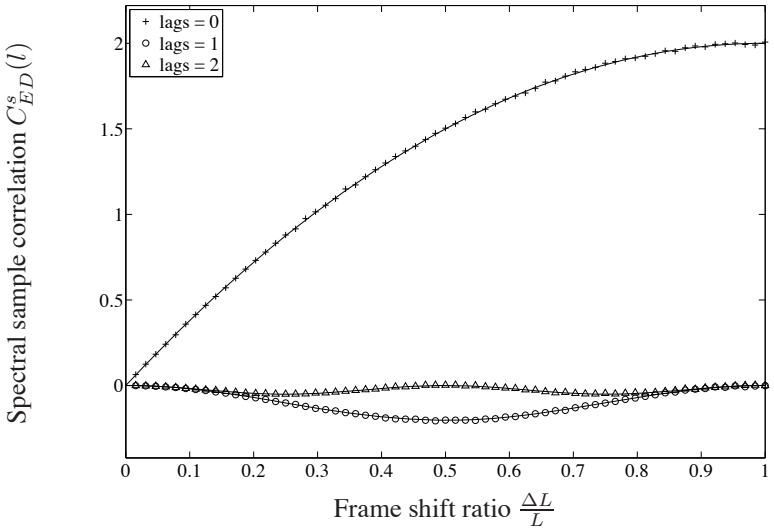


Figure 3.6: Spectral sample correlation function, $C^s_{ED}(l)$ as a function of the frame shift ratio $\frac{\Delta L}{L}$, for several lags l. The rectangular window was used to compute the spectrum.

This finally yields

$$\begin{aligned} C_{ED^s_R}(l) &= \frac{8}{L^2}\left(C^2_{R_X}(l) - \left(C^2_{R_A}(l) + C^2_{R_A,I_A}(l)\right)\right) \\ &= 2\sigma_X^4\left(\delta(l) - \frac{1}{L^2}\frac{\sin^2\left(\pi l\frac{L-\Delta L}{L}\right)}{\sin^2\left(\pi l\frac{1}{L}\right)}\right) \end{aligned} \tag{3.37}$$

Figure 3.6 shows $C_{ED^s_R}(l)$ as a function of the frame shift ratio $\frac{\Delta L}{L}$ for several lags l, in combination with experimentally estimated values. These estimations are obtained by repeatedly simulating the signal processing path for Gaussian iid input signals, and averaging the simulation results. The variance VAR$[ED(n,m)]$ can be computed using Eq. (3.25) in combination with Eq. (3.37). In the computation of VAR$[ED(n,m)]$ the values of $C_{ED^s_R}(l)$ for higher values of l play a significant role. Therefore, the variance of $ED(n,m)$ does not scale linearly with the increase in bandwidth for m. As a result, the assumptions which were used to formulate the correlation matrix $\mathbf{C_{ED}}$ are not valid: the temporal correlation varies as a function of m.

We further illustrate the characteristics of the PRH using a rectangular window with the transition probability $Pr[FP(n,m) = 1|FP(n-1,m) = 1]$ as a function of the frame shift ratio $\frac{\Delta L}{L}$. This conditional probability is based on Sheppard's formula:

$$Pr[FP(n,m) = 1|FP(n-1,m) = 1] = \frac{1}{2} + \frac{1}{\pi}\arcsin(\rho_1) \tag{3.38}$$

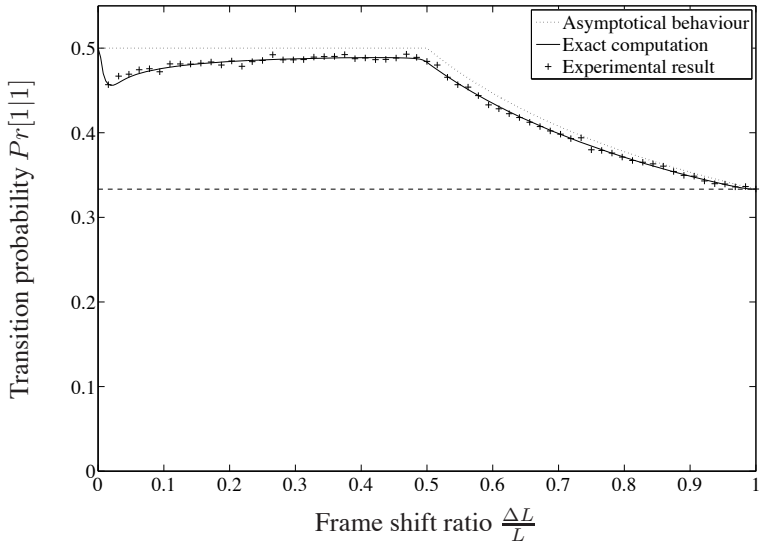


Figure 3.7: Transition probability $Pr[FP(n,m) = 1|FP(n-1,m) = 1]$ as a function of the frame shift ratio $\frac{\Delta L}{L}$ for a rectangular window.

where the temporal correlation is given by Eq. (3.24)

$$\rho_1(\Delta L) = \frac{1}{2}\frac{\text{VAR}\left[ED^{2\Delta L}(n,m)\right]}{\text{VAR}\left[ED^{\Delta L}(n,m)\right]} - 1 \tag{3.39}$$

This conditional probability is illustrated in Figure 3.7. As mentioned earlier, the transition probability for the rectangular window is dependent on m; the values presented in Figure 3.7 are averaged over the M frequency bands. Also shown is the asymptotic behavior of the transition probability as predicted by the model for extremely wide frequency bands. Here the variance VAR$[ED(n,m)]$ linearly scales with the frame shift ΔL. Therefore, we can distinguish three regions:

1. When $0 \leq \Delta L \leq \frac{1}{2}L$, the variance scales linearly with ΔL, so VAR$[ED^{2\Delta L}(n,m)] = 2VAR[ED^{\Delta L}(n,m)]$ resulting in $\rho_1 = 0$. The subsequent energy differences $ED(n-1,m)$ and $ED(n,m)$ are mutually uncorrelated, so naturally $Pr[1|1] = Pr[1] = \frac{1}{2}$

2. As stated in Eq. (3.24), for $2\Delta L > L$ the variance is constant with value VAR$[ED^{L}(n,m)]$. So, the correlation coefficient is equal to:

$$\rho_1(\Delta L) = \frac{1}{2}\frac{\text{VAR}\left[ED^{L}(n,m)\right]}{\text{VAR}\left[ED^{\Delta L}(n,m)\right]} - 1 \qquad \frac{1}{2}L < \Delta L \leq L$$

3. Similarly, $\rho_1 = -\frac{1}{2}$ for $\Delta L > L$, resulting in $Pr[1|1] = \frac{1}{3}$.

3.3.5 Transition probabilities for a non-rectangular symmetric window

In this section we extend the correlation function $C^s_{ED}(l)$ to non-rectangular symmetric windows. Using this result we compute the transition probabilities of subsequent fingerprint bits.

Since the PRH algorithm uses a discrete Fourier transform, the multiplication with the window function in the time domain $y(n,i) = w(i)x(n,i)$ results in the cyclical convolution in the spectral domain $\hat{y}(n,k) = \hat{w}(k) \odot \hat{x}(n,k)$.

The cyclical convolution $x(n) \odot y(n)$ is defined as:

$$x(n) \odot y(n) \quad \triangleq \quad \sum_{m=0}^{L-1} x((m-n) \mod L)y(m) \qquad n = 0, \ldots, L-1$$

When the window $w(i)$ is symmetrical (even), its spectral representation $\hat{w}(k)$ has no imaginary part. Then, also the real part of $\hat{x}(n,k)$ can be written as:

$$R_X(n,k) \quad = \quad \hat{w}(k) \odot R_{X_R}(n,k)$$

Using these kinds of manipulation we arrive at the expression for the sample-wise correlation function $C^s_{ED}(l)$ as stated in the following theorem. The proof is given in Appendix A.3.

Theorem 2 (Sample-wise correlation function $C^s_{ED}(l)$ for a symmetric window)**.** *The sample-wise correlation function $C^s_{ED}(l)$ for a symmetric window with spectral representation $\hat{w}(k), k = 0, \ldots, L-1$, is given by:*

$$\begin{aligned} C^s_{ED}(l) \quad = \quad & \frac{8}{L^2}(RR^2_X(l) - (RR_1(l) + RI_2(l))^2 \\ & -(RI_1(l) - RR_2(l))^2) \end{aligned} \tag{3.40}$$

where

$$\begin{aligned} RR_X(l) &= C_{R_X}(l) \odot (\hat{w}(l) \odot \hat{w}(l)) \\ RR_1(l) &= C_{R_A}(l) \odot \left(\cos(2\pi \tfrac{\Delta L}{L} l)\hat{w}(l) \odot \hat{w}(l)\right) \\ RR_2(l) &= C_{R_A}(l) \odot \left(\sin(2\pi \tfrac{\Delta L}{L} l)\hat{w}(l) \odot \hat{w}(l)\right) \\ RI_1(l) &= C_{R_A,I_A}(l) \odot \left(\cos(2\pi \tfrac{\Delta L}{L} l)\hat{w}(l) \odot \hat{w}(l)\right) \\ RI_2(l) &= C_{R_A,I_A}(l) \odot \left(\sin(2\pi \tfrac{\Delta L}{L} l)\hat{w}(l) \odot \hat{w}(l)\right) \end{aligned}$$

This formulation does allow for closed-form expressions of the sample-wise correlation function $C^s_{ED}(l)$ like Eq. (3.37) for the rectangular window case. However, the expressions rapidly become complicated. For instance, PRH uses a Hann window; its time-domain function:

$$w(i) \quad = \quad \frac{1}{2}\left(1 - \cos\left(\tfrac{2\pi}{L} i\right)\right) \qquad i = 0, \ldots, L-1$$

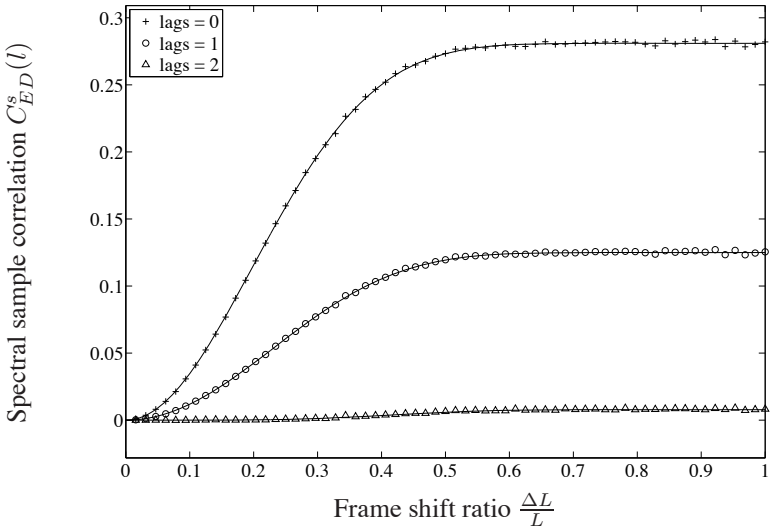


Figure 3.8: Spectral sample correlation function, $C^s_{ED}(l)$ as a function of the frame shift ratio $\frac{\Delta L}{L}$, for several lags l. The Hann window was used to compute the spectrum.

and spectral representation:

$$\begin{aligned}\hat{w}(k) &= -\frac{1}{4}\delta(k-1 \mod L) \\ &\quad +\frac{1}{2}\delta(k) - \frac{1}{4}\delta(k+1 \mod L) \qquad k = 0,\ldots,L-1\end{aligned}$$

gives the following expression for the variance $\mathrm{VAR}[ED^s] = C^s_{ED}(0)$:

$$\mathrm{VAR}[ED^s] = \frac{1}{32}\frac{\sigma_X^4}{L^2}\left((3L)^2 - \left(\left(3 - 2\sin^2\left(\pi\tfrac{\Delta L}{L}\right)\right)(L-\Delta L) + \left(3 - 2\sin^2\left(\pi\tfrac{1}{L}\right)\right)\frac{\sin\left(2\pi\frac{\Delta L}{L}\right)}{\sin\left(2\pi\frac{1}{L}\right)} \right)^2 \right)$$

Figure 3.8 shows $C^s_{ED}(l)$ for lags $l = 0, 1, 2$ together with the experimental data. Roughly speaking, the graphs consist of two regions: the frame shift ratio being smaller than, or larger than $\frac{1}{2}$, respectively. In the second region, the covariance terms are more or less constant as a function of $\frac{\Delta L}{L}$. In this region, one window has a negative slope, while the other has a positive slope. Apparently, this removes (almost) all correlation between the subsequent frames. This will also show in the transition probabilities.

Figure 3.9 shows the time-domain function $w(i)$ for several window functions:

- Hamming window:

$$w(i) = 0.54 - 0.46\cos\left(\tfrac{2\pi}{L}i\right) \qquad i = 0,\ldots,L-1$$

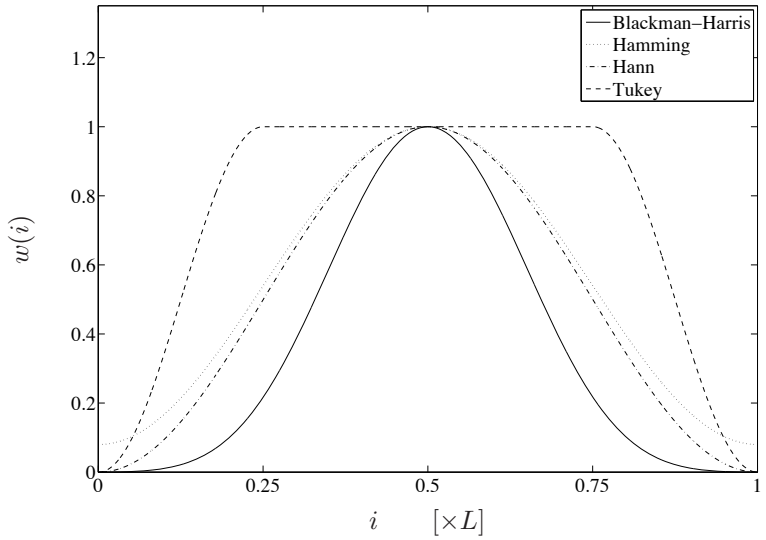


Figure 3.9: Illustration of several window functions.

- Blackman-Harris window:

$$\begin{aligned} w(i) &= 0.359 - 0.488\cos\left(\tfrac{2\pi}{L}i\right) \\ &\quad +0.141\cos\left(\tfrac{4\pi}{L}i\right) - 0.012\cos\left(\tfrac{6\pi}{L}i\right) \qquad i = 0,\ldots,L-1 \end{aligned}$$

- Tukey window, (half rectangular, half raised cosine):

$$w(i) = \begin{cases} \frac{1}{2}\left(1+\cos\left(\frac{4\pi}{L}i-\pi\right)\right) & i = 0,\ldots,\frac{1}{4}L-1 \\ 1 & i = \frac{1}{4}L,\ldots,\frac{3}{4}L-1 \\ \frac{1}{2}\left(1+\cos\left(\frac{4\pi}{L}i+\pi\right)\right) & i = \frac{3}{4}L,\ldots,\ L-1 \end{cases}$$

Using Eq. (3.40), we compute the correlation coefficient ρ_1 and the transition probability $Pr[FP(n,m)|FP(n-1,m)]$ using Eqs. (3.38) and (3.39). Figure 3.10 shows this transition probability for the window types listed above.

All curves in this figure have equal transition probability $Pr[1|1] = 1$ for $\frac{\Delta L}{L} = 0$ and tend towards $Pr[1|1] = \frac{1}{3}$ for increasing frame shift. For the first three raised cosine windows, one can say that the wider the bell-shape (the smaller the frequency smearing effect), the stronger the temporal correlation and hence the larger the transition probability. The Tukey window shows a mixed character; e.g. the effect that the transition probability is constant at is asymptotic value is prominent from $\frac{\Delta L}{L} \geq \frac{3}{4}$. This is because the slopes occupy only the first and last quarter of the window.

The relative positions of the curves for a given frame shift have a direct impact on two aspects of the performance: the uniqueness and the robustness to misalignment. This aspect is analyzed in more detail in Section 3.4.2. The larger the probability of

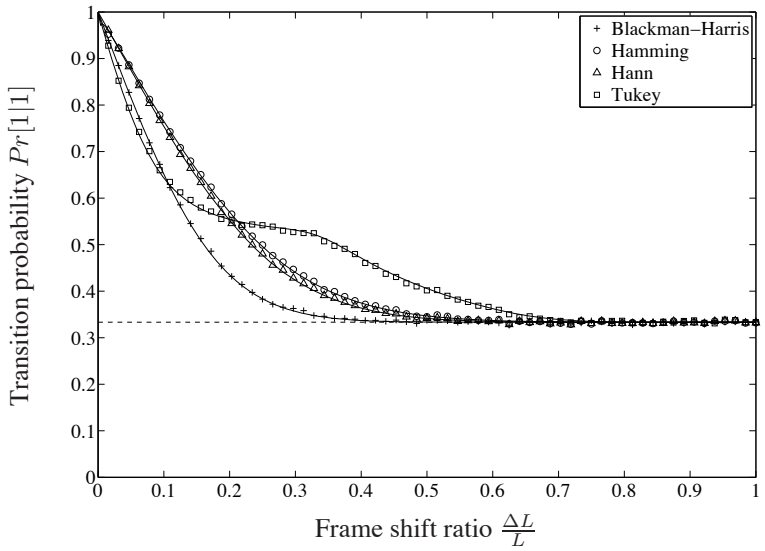


Figure 3.10: Transition probabilities for several window types as a function of the frame shift ratio $\frac{\Delta L}{L}$, taking into account the full spectrum.

the same bit value in the temporal direction, the more robust the fingerprint is to misalignment. On the other hand, more repetition in the temporal direction also implies a larger variance in the distribution of the distance between arbitrary fingerprint blocks.

3.4 Probability of an erroneous PRH fingerprint bit

In this section we consider the probability P_e that a bit in the PRH fingerprint, $F(n,m)$, is flipped due to additive white gaussian noise, temporal misalignment, or a combination of these two. Under the assumption of Gaussian i.i.d. input signals, we derive closed form expressions. In case of the additive noise, the P_e is dependent on the SNR. In case of temporal misalignment, the P_e is dependent on the amount of misalignment Δi, the frame shift ΔL and the frame length L. In the derivations of the closed-form expressions for P_e the following theorem is used.

Theorem 3 (Probability of sign change of a Gaussian random variable due to correlated Gaussian noise)**.** *Let (A,B) denote two zero-mean Gaussian random variables, drawn from a bivariate normal distribution, i.e. $(A,B) \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{AB})$, with correlation matrix $\mathbf{C}_{AB}$:*

$$\mathbf{C}_{AB} = \begin{bmatrix} \sigma_A^2 & \rho\,\sigma_A\sigma_B \\ \rho\,\sigma_A\sigma_B & \sigma_B^2 \end{bmatrix}$$

Now define $C = A + B$. The probability that the sign of C is different from the sign of A is given by:

$$\begin{aligned} P_e &= Pr\left[A \leq 0, C > 0 \quad \vee \quad A > 0, C \leq 0\right] \\ &= \frac{1}{\pi} \arctan\left(\frac{\sigma_B \sqrt{1-\rho^2}}{\sigma_A + \rho\sigma_B}\right) \end{aligned} \tag{3.41}$$

The proof of this theorem is included in Appendix A.4.

Section 3.4.1 considers additive noise, Section 3.4.2 discusses misalignment, and Section 3.4.3 discusses the combination of the two noise sources.

3.4.1 Bit-errors due to additive noise

We thus consider the following situation. Denoting the undistorted signal to be fingerprinted by $x(i)$ and the additive, normally distributed noise by $w(i)$, the distorted signal $y(i)$ is given by:

$$y(i) = x(i) + w(i) \tag{3.42}$$

We are interested in the relating the difference between the corresponding fingerprints of $x(i)$ and $y(i)$, $F_X(n,m)$ and $F_Y(n,m)$, respectively, to the statistical characteristics of $x(i)$ and $y(i)$. The probability of bit error, P_e, can be expressed in terms of the energy differences, $ED_X(n,m)$ and $ED_Y(n,m)$ (see Eq. (3.6)):

$$\begin{aligned} P_e &= Pr\left[F_X(n,m) \neq F_Y(n,m)\right] \\ &= Pr[ED_X(n,m) \leq 0, ED_Y(n,m) > 0 \\ &\qquad \vee \quad ED_X(n,m) > 0, ED_Y(n,m) \leq 0] \end{aligned} \tag{3.43}$$

We split the calculation of P_e into two parts. First, using Eq. (3.43), the following equation expresses P_e in terms of variances of $ED_X(n,m)$ and $ED_Y(n,m) - ED_X(n,m)$:

$$P_e = \frac{1}{\pi} \arctan\left(\sqrt{\frac{\text{VAR}\left[ED_Y(n,m) - ED_X(n,m)\right]}{\text{VAR}\left[ED_X(n,m)\right]}}\right) \tag{3.44}$$

This relation is based on Theorem 3 from the previous section. The additive noise $B = ED_Y(n,m) - ED_X(n,m)$ and the signal $A = ED_X(n,m)$ are uncorrelated, i.e. $\rho = 0$. Here, we assume that $ED_X(n,m)$ and $ED_Y(n,m)$ are drawn from normal distributions and have mean value zero. Appendix A.5 states a simplified version of Theorem 3 specifically for the case $\rho = 0$.

In the next step, we have to relate $\text{VAR}\left[ED_Y(n,m) - ED_X(n,m)\right]$ and $\text{VAR}\left[ED_X(n,m)\right]$ to the variances σ_X^2 and σ_W^2 of the original signal $x(i)$ and compression distortion $w(i)$, respectively. Therefore, we analyze how each of the two components $x(i)$ and $w(i)$ contribute to $ED_Y(n,m)$. To do this, we repeat the steps

in Eqs. (3.3), (3.7), (3.8) and (3.9), but now for the model in Eq. (3.42). First, the short-time Fourier transform, $\hat{y}(n,k)$, is computed for each frame n:

$$\hat{y}(n,k) = \hat{x}(n,k) + \hat{w}(n,k) \tag{3.45}$$

Second, the PSD is estimated using the periodogram:

$$\begin{aligned} S_Y(n,k) &\triangleq \frac{1}{L}|\hat{y}(n,k)|^2 \\ &= \frac{1}{L}\Big(|\hat{x}(n,k)|^2 + |\hat{w}(n,k)|^2 \\ &\qquad + 2\,\mathrm{Re}\left(\hat{x}(n,k)\overline{\hat{w}(n,k)}\right)\Big) \\ &= S_X(n,k) + S_W(n,k) + 2\mathrm{Re}\left(S_{XW}(n,k)\right) \end{aligned} \tag{3.46}$$

where $S_{XW}(n,k)$ is the (complex) cross-spectrum. Its real part is also known as the coincident spectral density or co-spectrum. Third, the difference between two spectral frames is computed:

$$\begin{aligned} ED_Y^s(n,k) &\triangleq S_Y(n,k) - S_Y(n-1,k) \\ &= S_X(n,k) + S_W(n,k) + 2\mathrm{Re}\left(S_{XW}(n,k)\right) \\ &\quad -(S_X(n-1,k) + S_W(n-1,k) \\ &\quad +2\mathrm{Re}\left(S_{XW}(n-1,k)\right)) \\ &= ED_X^s(n,k) + ED_W^s(n,k) + 2Q^s(n,k), \end{aligned} \tag{3.47}$$

where $Q^s(n,k)$ is given by:

$$Q^s(n,k) = \mathrm{Re}\left(S_{XW}(n,k) - S_{XW}(n-1,k)\right)$$

Finally, the subband energy difference, $ED_Y(n,m)$, is computed:

$$\begin{aligned} ED_Y(n,m) &\triangleq \sum_{k\in\mathcal{K}_m} ED_Y^s(n,k) - \sum_{k\in\mathcal{K}_{m+1}} ED_Y^s(n,k) \\ &= ED_X(n,m) + ED_W(n,m) + 2Q(n,m), \end{aligned} \tag{3.48}$$

where $Q(n,m)$ is defined as:

$$Q(n,m) = \sum_{k\in\mathcal{K}_m} Q^s(n,k) - \sum_{k\in\mathcal{K}_{m+1}} Q^s(n,k)$$

Using Eq. (3.48) we obtain the following expression for the numerator under the square root in Eq. (3.44):

$$ED_Y(n,m) - ED_X(n,m) = ED_W(n,m) + 2Q(n,m). \tag{3.49}$$

In Appendix A.6 we show that the variables $ED_W(n,m)$ and $Q(n,m)$ are mutually uncorrelated, yielding:

$$\begin{aligned} &\mathrm{VAR}\left[ED_Y(n,m) - ED_X(n,m)\right] \\ &= \mathrm{VAR}\left[ED_W(n,m)\right] + 4\mathrm{VAR}\left[Q(n,m)\right] \end{aligned} \tag{3.50}$$

In Appendix A.7 we show that, if we assume $x(i)$ and $w(i)$ to be normally distributed, the variances in Eq. (3.50) are proportional to VAR $[ED_X(n,m)]$:

$$\begin{aligned} &\text{VAR}\left[ED_Y(n,m) - ED_X(n,m)\right] \\ &= \left(\frac{\sigma_W^4}{\sigma_X^4} + 2\frac{\sigma_W^2}{\sigma_X^2}\right) \text{VAR}\left[ED_X(n,m)\right] \end{aligned} \tag{3.51}$$

Finally, the combination of Eqs. (3.44) and (3.51) results in:

$$P_e = \frac{1}{\pi}\arctan\left(\sqrt{\frac{\sigma_W^4}{\sigma_X^4} + 2\frac{\sigma_W^2}{\sigma_X^2}}\right) \tag{3.52}$$

Note that this expression is independent of the frame index n and the frequency band index m. The first was to be expected since the input signals are assumed stationary. In other words, since the statistical characteristics of $x(i)$ and $w(i)$ are constant over time, P_e is also constant over time. The latter is true if the subband energy difference $ED(n,m)$ satisfy the assumption that they are normally distributed. In practice this is the case if the frequency bands on which $ED(n,m)$ is based, m and $m+1$, have sufficiently large bandwidth. Eq. (3.52) was derived for Gaussian i.i.d. signals. Analyzing the assumptions necessary for the theorems to hold, it is sufficient to assume that the signal and noise are wide sense stationary (w.s.s.), zero mean, mutually uncorrelated, and have the same spectral structure, expressed in Eq. (A.25).

In the derivation of the model, the structure of the fingerprint is not taken into account. Due to the large frame overlap, the fingerprint has a slowly varying binary structure. This dependency does not have to be taken into account in the models, since we are computing the average probability of error P_e, not its variance.

Figure 3.11 shows the SNR-P_e relationship for the model of Eq. (3.52) along with experimental results on synthetic data. When the SNR is formulated as $20\log_{10}(\sigma_X/\sigma_W)$ and the P_e is plotted on a logarithmic scale, for sufficiently large SNR ($\sigma_X^2 \gg \sigma_W^2$), the SNR vs. P_e relation is a straight line. For these small distortions, the P_e as formulated in Eq. (3.52) is approximately inversely proportional to σ_X/σ_W:

$$P_e \approx \frac{1}{\pi}\arctan\left(\sqrt{2}\,\frac{\sigma_W}{\sigma_X}\right) \approx \frac{\sqrt{2}}{\pi}\,\frac{\sigma_W}{\sigma_X} \tag{3.53}$$

In practice this means that for a 20 dB increase of SNR, the P_e is expected to drop by a factor 10. The region in the curve showing the 'linear' SNR-P_e relation is of particular interest, since most audio compression algorithm operate in this region. From a quality estimation perspective, the low-SNR region is of no interest, since there the audio is degraded too severely. Furthermore, signals in the low-SNR regime generate fingerprint differences around or above the detection threshold for identification.

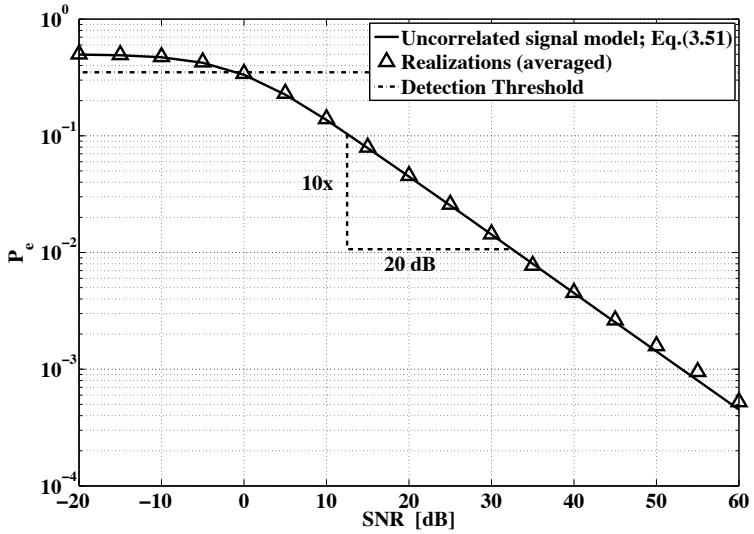


Figure 3.11: Analytical relation between SNR and P_e for the PRH. The indicated detection threshold is equal to 0.35 in [44]

3.4.2 Bit-errors due to temporal misalignment

In this section we consider bit-errors due to the misalignment of Δi samples. A single time frame of the reference signal $x(n,i)$ and the shifted signal $y(n,i)$ are given by:

$$\begin{aligned} x(n,i) &= x(i+(n-1)\Delta L) \qquad i=1,\ldots,L-1 \\ y(n,i) &= x(i+(n-1)\Delta L+\Delta i) \qquad \Delta i=-\tfrac{1}{2}\Delta L,\ldots,\tfrac{1}{2}\Delta L-1. \end{aligned}$$

The effect of misalignment in the time domain results in additive distortion in the spectral energy differences:

$$ED_Y(n,m) = ED_X(n,m) + ED_{W_{mis}}(n,m) \tag{3.54}$$

In the following derivations we will omit the time and frequency indices (n,m). In two steps we derive a closed form expression for the probability of error $P_e^{\Delta i}$ in terms of the variance VAR$[ED_X]$, based on Eq. (3.41) in Theorem 3.

1. *Express P_e in terms of the variances of ED_X and $ED_{W_{mis}}$.*
   Since ED_Y is the result of a time-shifted version of the signal that generated ED_X, the variances of ED_X and ED_Y are identical. This leads to the following observation:

$$\begin{aligned} \text{VAR}[ED_Y] &= \text{VAR}[ED_X] + \text{VAR}[ED_{W_{mis}}] + 2\text{COV}[ED_X, ED_{W_{mis}}] \\ &= \text{VAR}[ED_X] \end{aligned}$$

   Therefore, we can express the covariance in terms of the variances:

$$\text{COV}[ED_X, ED_{W_{mis}}] = -\frac{1}{2}\text{VAR}[ED_{W_{mis}}],$$

and thus for the correlation coefficient ρ:

$$\begin{aligned}\rho &= \frac{\text{COV}[ED_X, ED_{W_{mis}}]}{\sigma_{ED_X}\,\sigma_{ED_{W_{mis}}}} \\ &= -\frac{1}{2}\frac{\sigma_{ED_{W_{mis}}}}{\sigma_{ED_X}} \end{aligned} \tag{3.55}$$

This relation can be plugged into the expression for P_e – Eq. (3.41) in Theorem 3 – with $\sigma_A = \sigma_{ED_X}$ and $\sigma_B = \sigma_{ED_{W_{mis}}}$:

$$\begin{aligned}P_e &= \frac{1}{\pi}\arctan\left(\frac{\sigma_{ED_{W_{mis}}}\sqrt{1-\rho^2}}{\sigma_{ED_X}+\rho\sigma_{ED_{W_{mis}}}}\right) \\ &= \frac{1}{\pi}\arctan\left(\sqrt{\frac{4\sigma^4_{ED_X}-(2\sigma^2_{ED_X}-\sigma^2_{ED_{W_{mis}}})^2}{(2\sigma^2_{ED_X}-\sigma^2_{ED_{W_{mis}}})^2}}\right) \\ &= \frac{1}{\pi}\arccos\left(\frac{2\sigma^2_{ED_X}-\sigma^2_{ED_{W_{mis}}}}{2\sigma^2_{ED_X}}\right) \end{aligned} \tag{3.56}$$

where we used the relation:

$$\frac{1}{\pi}\arctan\left(\sqrt{c^2-1}\right) = \frac{1}{\pi}\arccos\left(\frac{1}{c}\right)$$

2. *Express the variance of $ED_{W_{mis}}$ in terms of the variance of ED_X.*
We can rewrite the variance of $ED_{W_{mis}}$ using Eq. (3.54):

$$\begin{aligned}\text{VAR}[ED_{W_{mis}}] &= \text{VAR}[ED_Y - ED_X] \\ &= 2\text{VAR}[ED_X] - 2\text{COV}[ED_X, ED_Y] \end{aligned} \tag{3.57}$$

Like in the temporal correlation in Eq. (3.24), the covariance term can be expressed in terms of variances with different frame shift ratios:

$$\begin{aligned}&\text{COV}[ED_X, ED_Y] \\ &= -\text{VAR}[ED_X^{\Delta i}] + \frac{1}{2}\left(\text{VAR}[ED_X^{\Delta i-\Delta L}] + \text{VAR}[ED_X^{\Delta i+\Delta L}]\right) \end{aligned} \tag{3.58}$$

Plugging Eq. (3.58) back into Eq. (3.57) leads to:

$$\begin{aligned}\text{VAR}[ED_{W_{mis}}] = {} & 2\left(\text{VAR}[ED_X^{\Delta L}] + \text{VAR}[ED_X^{\Delta i}]\right) \\ & - \left(\text{VAR}[ED_X^{\Delta L-\Delta i}] + \text{VAR}[ED_X^{\Delta L+\Delta i}]\right) \end{aligned} \tag{3.59}$$

which finally yields the result:

$$\begin{aligned}P_e^{\Delta i} = {} & \frac{1}{\pi}\arccos\Bigg(-\frac{\text{VAR}[ED_X^{\Delta i}]}{\text{VAR}[ED_X^{\Delta L}]} \\ & + \frac{1}{2}\frac{\text{VAR}[ED_X^{\Delta L-\Delta i}] + \text{VAR}[ED_X^{\Delta L+\Delta i}]}{\text{VAR}[ED_X^{\Delta L}]}\Bigg) \end{aligned} \tag{3.60}$$

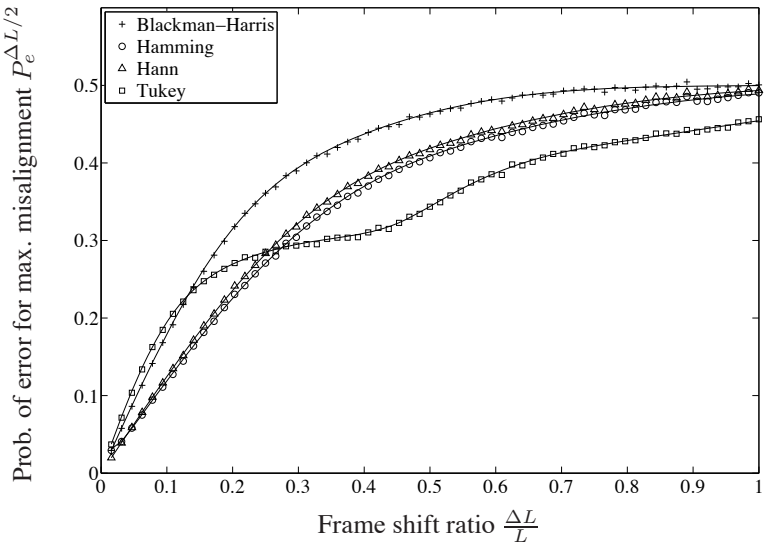


Figure 3.12: Probability of error P_e due to maximum misalignment as a function of the frame shift ratio $\frac{\Delta L}{L}$. The curves are generated for several window types.

The variances were computed in Section 3.3.4 for rectangular windows and in Section 3.3.5 for non-rectangular windows.

Figure 3.12 shows the probability of error $P_e^{\Delta L/2}$ due to maximum misalignment as a function of the frame shift ratio $\frac{\Delta L}{L}$. Figure 3.13 shows the probability of error due to misalignment as a function of the amount of misalignment Δi for a given frame shift ratio $\frac{\Delta L}{L} = \frac{1}{32}$ and frame length $L = 2048$. In practice, each amount of desynchronization is equally likely:

$$P_e \quad = \quad \frac{1}{\Delta L} \sum_{\Delta i=-\frac{1}{2}\Delta L}^{\frac{1}{2}\Delta L-1} P_e^{\Delta i}.$$

3.4.3 Bit-errors due to additive noise and temporal misalignment

In this section we derive an expression for the probability of error P_e due to a combination of additive noise and temporal misalignment. For this combination the starting point becomes:

$$ED_Y \quad = \quad ED_X + ED_W \tag{3.61}$$

where now the noise component consists of contributions from additive noise and misalignment:

$$ED_W \quad = \quad ED_{W_{mis}} + ED_{W_{add}} \tag{3.62}$$

In two steps we derive a closed form expression for the probability of error P_e in terms of the variance VAR$[ED_X]$ and the SNR, based on Eq. (3.41) in Theorem 3.

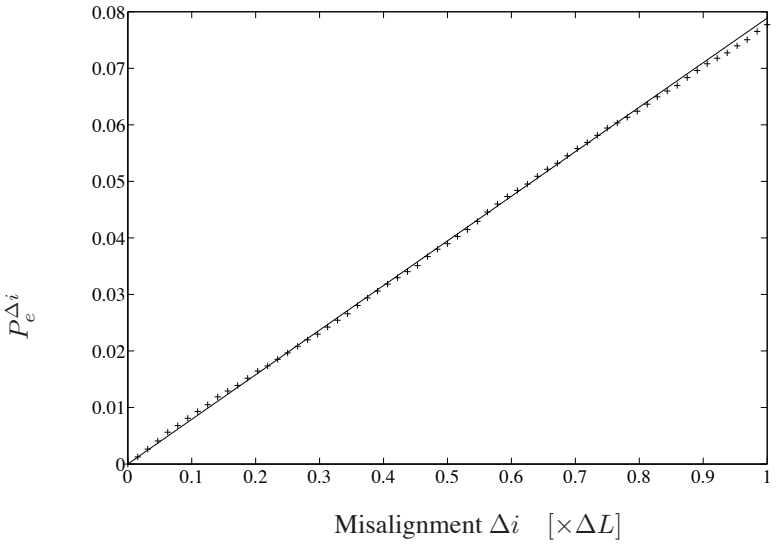


Figure 3.13: Probability of error P_e due to misalignment as a function of the amount of misalignment Δi for a given frame shift ratio $\frac{\Delta L}{L} = \frac{1}{32}$. The curve is generated for the use of a Hann window.

1. *Express P_e in terms of the variances of ED_X and ED_W.*
   An observation similar to Eq. (3.55) can be made: the variance of the spectral energy differences due to misalignment and additive noise is equal to the variance of the spectral energy differences due to additive noise alone:

$$\begin{aligned} \text{VAR}[ED_Y] &= \text{VAR}[ED_X] + \text{VAR}[ED_W] + 2\text{COV}[ED_X, ED_W] \quad (3.63) \\ &= \text{VAR}[ED_X + ED_{W_{add}}] \\ &= \text{VAR}[ED_X] + \text{VAR}[ED_{W_{add}}] \quad (3.64) \end{aligned}$$

   The additive noise is independent of the signal itself: $\text{COV}[ED_X, ED_{W_{add}}] = 0$. By comparing Equations (3.63) and (3.64) it can be seen that the covariance of the signal and the total noise is equal to:

$$\begin{aligned} \text{COV}[ED_X, ED_W] &= \frac{1}{2}\left(\text{VAR}[ED_{W_{add}}] - \text{VAR}[ED_W]\right) \\ &= -\frac{1}{2}\text{VAR}[ED_{W_{mis}}] \quad (3.65) \end{aligned}$$

   Therefore the parameters of the Gaussian PDF become:

$$\begin{aligned} \sigma^2_{ED_X} &= \text{VAR}\left[ED_X^{\Delta L}\right] \quad (3.66) \\ \sigma^2_{ED_W} &= \text{VAR}\left[ED_{W_{mis}}\right] + \text{VAR}\left[ED_{W_{add}}\right] \quad (3.67) \\ \rho &= -\frac{\text{VAR}\left[ED_{W_{mis}}\right]}{2\sigma_{ED_X}\sigma_{ED_W}} \quad (3.68) \end{aligned}$$

This relation can be plugged into the expression for P_e (Eq. 3.41) in Theorem 3 with $\sigma_A = \sigma_{ED_X}$ and $\sigma_B = \sigma_{ED_W}$:

$$P_e = \frac{1}{\pi} \arctan\left(\frac{\sigma_{ED_W}\sqrt{1-\rho^2}}{\sigma_{ED_X} + \rho\sigma_{ED_W}}\right)$$

2. *Express the variance of ED_W in terms of the variance of ED_X and the SNR.* The variances of the additive noise and misalignment noise are given in the previous sections by Eqs. (3.51) and (3.59), respectively:

$$\begin{aligned}\text{VAR}[ED_{W_{add}}] &= \left(\frac{\sigma_W^4}{\sigma_X^4} + 2\frac{\sigma_W^2}{\sigma_X^2}\right) \cdot \text{VAR}\left[ED_X^{\Delta L}\right] \\ \text{VAR}[ED_{W_{mis}}] &= 2\text{VAR}\left[ED_X^{\Delta L}\right] \\ &\quad - \left(\text{VAR}\left[ED_X^{\Delta L+\Delta i}\right] - \text{VAR}\left[ED_X^{\Delta L-\Delta i}\right]\right)\end{aligned}$$

Filling in the parameters from Eqs (3.66 - 3.68) we get for the nominator:

$$\begin{aligned}&\sigma_{ED_W}\sqrt{1-\rho^2} \\ &= \frac{\sqrt{4\sigma_{ED_X}^2\left(\sigma_{ED_{W_{add}}}^2 + \sigma_{ED_{W_{mis}}}^2\right) - \sigma_{ED_{W_{mis}}}^4}}{2\sigma_{ED_X}} \\ &= \frac{\sqrt{4\sigma_{ED_X}^2\left(\sigma_{ED_X}^2 + \sigma_{ED_{W_{add}}}^2\right) - \left(2\sigma_{ED_X}^2 - \sigma_{ED_{W_{mis}}}^2\right)^2}}{2\sigma_{ED_X}}\end{aligned} \tag{3.69}$$

and denominator:

$$\sigma_{ED_X} + \rho\sigma_{ED_W} = \frac{2\sigma_{ED_X}^2 - \sigma_{ED_{W_{mis}}}^2}{2\sigma_{ED_X}} \tag{3.70}$$

We finally end up with:

$$\begin{aligned}P_e^{\Delta i} &= \frac{1}{\pi}\arctan\left(\sqrt{\frac{4\sigma_{ED_X}^2\left(\sigma_{ED_X}^2 + \sigma_{ED_{W_{add}}}^2\right)}{\left(2\sigma_{ED_X}^2 - \sigma_{ED_{W_{mis}}}^2\right)^2} - 1}\right) \\ &= \frac{1}{\pi}\arccos\left(\frac{\text{VAR}\left[ED_X^{\Delta L+\Delta i}\right] - \text{VAR}\left[ED_X^{\Delta L-\Delta i}\right]}{2\left(1+\frac{\sigma_W^2}{\sigma_X^2}\right)\text{VAR}\left[ED_X^{\Delta L}\right]}\right)\end{aligned} \tag{3.71}$$

Figure 3.14 illustrates the probability of error P_e due to additive noise and maximum misalignment $\Delta i = \frac{1}{2}\Delta L$ in Eq. (3.71) as a function of SNR for various frame shift ratios $\frac{\Delta L}{L}$. For low SNR-values the additive component is dominant, and the curves resemble the curve for additive noise only shown in Figure 3.11. For high SNR-values the P_e converges to the values that correspond to misalignment only, shown in Figure 3.12.

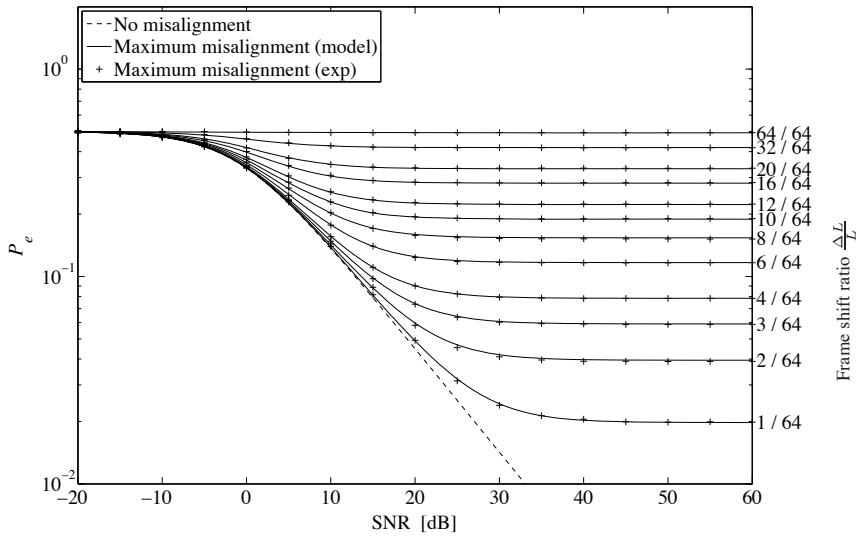


Figure 3.14: Probability of error P_e due to additive noise and maximum misalignment as a function of SNR for various frame shift ratios $\frac{\Delta L}{L}$. The curves are generated for the use of a Hann window.

3.5 Relation with other binary fingerprinting algorithms

Several other audio fingerprinting algorithms use binary energy features, e.g. [58, 65]. These algorithms might be modeled in a similar fashion. In these algorithms, including the PRH, fingerprint bits are derived by filtering the spectrogram. In the following, we use the matrix notation introduced in Section 2.6, but the time indices will be omitted. The computation of the binary features consists of three steps:

1. *Computation of coarse spectrogram.*
   Within each frame, the energy within frequency bands are computed. The result is a coarse spectrogram $\mathbf{E}^{N,P}$, consisting of N frames and P frequency band energies over time. Let $\mathbf{E}^{N,1}(m)$ denote the energy in the mth frequency band over time.

2. *Convolution with a filter.*
   The spectral energy differences are computed by filtering the spectrogram in the temporal direction:

$$\mathbf{ED}^{N,1}(m) \quad = \quad \mathbf{E}^{N,w_m}(s_m) * \mathbf{H}_m^{t_m,w_m} \qquad m = 1, \ldots, M$$

   where $\mathbf{H}_m^{t_m,w_m}$ denotes the mth filter with t_m coefficients in the temporal di-

rection, and w_m coefficients in the frequency direction:

$$\mathbf{H}_m^{t_m,w_m} = \begin{bmatrix} h(1,1) & \dots & h(1,w_m) \\ \vdots, & & \vdots \\ h(t_m,1) & \dots & h(t_m,w_m) \end{bmatrix}$$

and $\mathbf{E}^{w_m}(s_m)$ denotes a selection of w_m frequency bands from the spectrogram:

$$\mathbf{E}^{N,w_m}(s_m) = [\mathbf{E}^{N,1}(s_m), \dots, \mathbf{E}^{N,1}(s_m + w_m - 1)]$$

where s_m denotes the lowest frequency band index of the selection. These filtered time series are concatenated into the spectral energy differences block:

$$\mathbf{ED}^{N,M} = [\mathbf{ED}^{N,1}(1), \dots, \mathbf{ED}^{N,1}(M)]$$

3. *Binarization*
The fingerprint bits are the result of applying a filter-dependent threshold to the spectral energy differences:

$$FP(n,m) = \begin{cases} 0 & ED(n,m) < T_m \\ 1 & ED(n,m) \geq T_m \end{cases}$$

Usually, the threshold T_m is chosen equal to the median of $f_{ED(m)}(t)$, the distribution of the samples in time for frequency band m. PRH uses a fixed threshold $T = 0$; for a zero-mean Gaussian distribution this is equal to the median value.

PRH uses one specific Haar filter of fixed size for each time-series:

$$\mathbf{H}^{2,2} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

Let $\mathbf{1}^{n,m}$ represent a matrix of all ones of size $n \times m$. Lebosse *et al.* use a filter which combines the average of a series of neighboring frequency bands with the instantaneous value of the next frequency band:

$$\mathbf{H}^{2,m+1} = \begin{bmatrix} \frac{1}{m}\mathbf{1}^{1,m} & -1 \\ -\frac{1}{m}\mathbf{1}^{1,m} & 1 \end{bmatrix},$$

Ke *et al.* learn a set of filters from training data by employing Adaboost, a pairwise boosting algorithm known from computer vision. The consider the following

(a) (b) (c) (d) (e)

Figure 3.15: Illustration of the prototype filters employed in [58]. White regions represent filter coefficients equal to '1', black regions represent filter coefficients equal to '-1'. PRH uses the leftmost type of filter.

five prototype filters, also illustrated in Figure 3.15:

$$
\begin{aligned}
\mathbf{H}_1^{2t,2w} &= \begin{bmatrix} \mathbf{1}^{t,w} & -\mathbf{1}^{t,w} \\ -\mathbf{1}^{t,w} & \mathbf{1}^{t,w} \end{bmatrix} \\
\mathbf{H}_2^{t,2w} &= \begin{bmatrix} \mathbf{1}^{t,w} & -\mathbf{1}^{t,w} \end{bmatrix} \\
\mathbf{H}_3^{2t,w} &= \begin{bmatrix} \mathbf{1}^{t,w} \\ -\mathbf{1}^{t,w} \end{bmatrix} \\
\mathbf{H}_4^{t,3w} &= \begin{bmatrix} -\mathbf{1}^{t,w} & \mathbf{1}^{t,w} & -\mathbf{1}^{t,w} \end{bmatrix} \\
\mathbf{H}_5^{3t,w} &= \begin{bmatrix} -\mathbf{1}^{t,w} \\ \mathbf{1}^{t,w} \\ -\mathbf{1}^{t,w} \end{bmatrix}
\end{aligned}
$$

The algorithm selects M filters from a set of filters from the above types with different widths w_m, heights t_m, and frequency locations s_m.

We expect that the two models developed for PRH can also be used for the algorithms of Lebosse *et al.* and Ke *et al.*

Also Mihçak *et al.* [75] and Kim *et al.* [59] both compute binary features. However, Mihçak *et al.* compute their features on the MDCT and use error correcting codewords. Kim *et al.* use the spectral centroids which they optimize and convert to a binary representation in a way similar to Ke *et al.* We believe that our models might be extended to these algorithms as well.

3.6 Conclusion and discussion

In this chapter we have developed two statistical models for PRH when the input signal is i.i.d. Gaussian. The first model predicts the structure of the PRH fingerprint. The second predicts the probability of an erroneous bit in case of temporal misalignment and additive noise. We have experimentally verified that the models provide accurate predictions. We have also shown that the predictions from the first model apply to real-life data as well.

In a series of publications, McCarthy, Silvestre, Hurley and Balado have modeled and optimized several individual aspects of the PRH fingerprint [71, 72, 73, 15, 16, 51, 52], such as the method for converting the real-valued features into a binary representation [71], the optimal window for the smallest desynchronization effect [15], and the collision probability [52]. The first models [72, 73, 15] follow our approach to a great extent, but the elegant and compact formulation in quadratic form allows for optimizations, and extensions to Gaussian signals which are more complex than iid. The collision probability is not limited to binary fingerprints, but can also be modeled for M-ary fingerprints.

We expect that the models can be extended to other types of distortions, e.g. variations in play-back speed of the audio signal, and to other audio fingerprinting algorithms. These other algorithms are expected not to be limited to fingerprints based directly on the spectral energy. Also other types of features, e.g. MFCC or spectral centroid features may be modeled in a similar fashion. The main assumption is that the features before binarization follow a Gaussian distribution.

Chapter 4

Distortion Estimation in Compressed Music Using Only Audio Fingerprints

---

This chapter is largely based on the publication "Distortion Estimation in Compressed Music Using Only Audio Fingerprints", by P.J.O. Doets and R.L. Lagendijk in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pages 302-317, February 2008.

4.1 Introduction

One of the applications of fingerprinting is to identify music on the Internet. However, if two copies of the same song are identified as being the same music, they can still differ in quality to a large extend. Therefore, one would like to discriminate between qualities of songs identified. A consumer prefers to obtain the version with the highest quality. A platform moderator, however, might want to block high quality versions of copyrighted content, but allow a low-quality preview version to be uploaded. Therefore, it is desirable to use the same mechanism for quality discrimination.

In this chapter, we extend the functionality of fingerprinting to estimate the Signal-to-Noise Ratio (SNR) between the original recording and a compressed version. This SNR-estimation can then serve as a simple, yet coarse quality indicator, using fingerprints only. The SNR-estimation is based on the way the fingerprint reflects the changes in the audio signal introduced by compression, as will be explained next.

Figure 4.1(a) schematically shows the procedure proposed in this chapter for estimating the SNR of compressed content using fingerprinting technology. After the song on the Internet has been identified, we have two fingerprints: the fingerprint of the original high quality recording from the database, F_X, and the fingerprint of the compressed version of the same song from the internet, F_Y. Due to compression, the waveform of the compressed recording, Y, is slightly different from its original recording, X. This difference in waveform then results in a difference in the corresponding fingerprints, $d(F_X, F_Y)$. Figure 4.1(b) shows an illustration of the relationship between fingerprint differences and audio quality. In this example, we can roughly estimate the audio quality of the compressed music from the difference between F_X an F_Y, i.e. $d(F_X, F_Y)$. The accuracy of the estimation is dependent on the spread in $d(F_X, F_Y)$ - indicated here by the shaded area - for a given quality level, and vice versa.

At first sight, there are several alternatives to obtain the quality of compressed audio, e.g. the bitrate from the compressed audio file header and perceptual quality assessment algorithms [96]. The bitrate, however, like other metadata is unreliable. The bitrate is not a required parameter for decoding in every audio compression format (e.g. Ogg Vorbis [5]) and therefore not always present. Furthermore, the quality of the compressed audio content is a result of the selected compression bitrate, within the limits and settings of the specific implementation. Even compressing the same song with the same algorithm at the same bitrate but using different implementations may result in significantly different quality. The variability is even larger when comparing versions compressed with different algorithms at the same bitrate.

Another alternative that comes to mind is to use an algorithm that estimates the perceptual quality of the compressed version with respect to its original recording. A wide variety of algorithms can be found in literature [23, 22, 95, 49], some of which are used in the Perceptual Audio Quality (PEAQ) measure adopted by the ITU [96]. These algorithms use elaborate psycho-acoustic models to mimic the effects of the Human Auditory System (HAS). They need, however, the original uncompressed version as a reference. Because in our envisioned application scenario's this reference is unavailable, in our proposed technique the fingerprint of the original uncompressed recording takes the role of the reference. In this way the resulting quality indication is

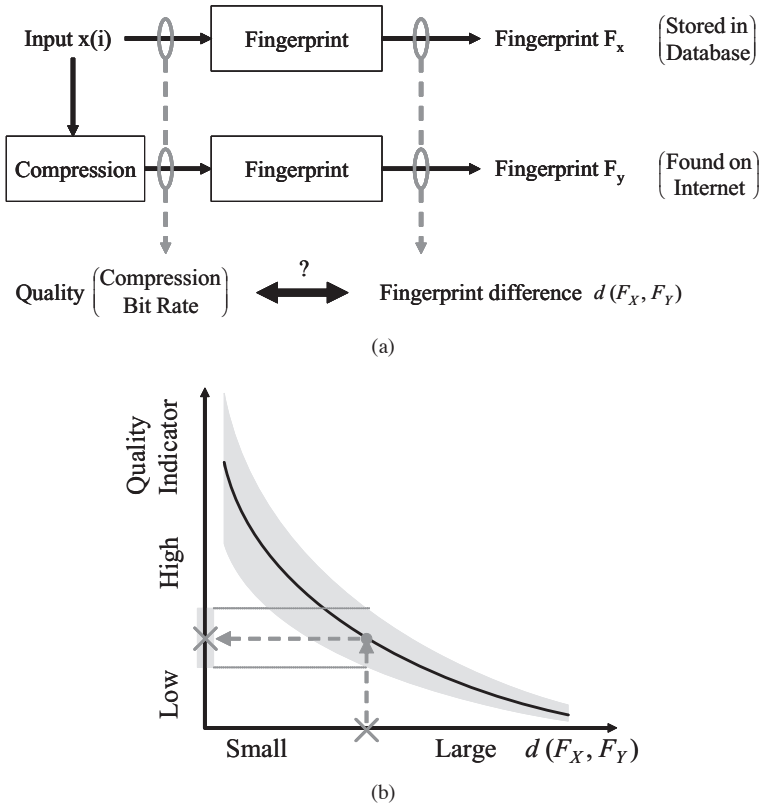


Figure 4.1: Using fingerprints for music quality assessment; (a) Relating differences in audio fingerprints of two versions of the same recording, X and Y, to differences in perceptual quality of these recordings. (b) Example relationship between fingerprint differences and a quality indicator (compression bitrate).

only indirectly based on the difference between the original and compressed version.

Our technique does not intend to predict the subjective quality or to match the capabilities of subjective quality predicting algorithms. These are much more accurate and reliable, and have a better correlation with human perception, but they need information that is not available in our scenarios. Furthermore, for our envisioned application scenarios outlined in this introduction, such accuracy also is not needed. The only common factor with perceptually motivated techniques is the use of a reference to give a content-based indication of the difference between the compressed content and its original.

This chapter is organized therefore as follows. Section 4.2 provides an overview of fingerprinting algorithms described in literature. Three algorithms which are considered representative for the field are reviewed. In Section 4.3 we model the distortion introduced by compression as additive noise and develop a model that expresses the fingerprint differences in terms of the SNR for one of the three algorithms. This model

provides the theoretical foundation for experiments in Section 4.4 that relate the bitrate used for compression, and the resulting SNR, to the distance between the fingerprints. Section 4.5 draws conclusions and outlines directions for future research.

4.2 Audio Fingerprinting Algorithms

In the last decade, several fingerprinting systems have been developed. Cano *et al*. present a good survey of fingerprinting algorithms [28]. A fingerprinting system has to meet three requirements:

- *Robustness*: The fingerprint of a distorted piece of music has to be sufficiently close to the fingerprint of the undistorted recording,
- *Collision-resistance*: The fingerprints of two different pieces of music should be sufficiently different.
- *Database search efficiency*: In order to keep the database scalable, the fingerprint representation has to allow for efficient database search.

These requirements are primarily concerned with identification. To use fingerprints for indicating the quality (SNR) of compressed music, the fingerprinting system has to meet a fourth criterion: the distance between the fingerprints of the original and compressed version should also reflect the amount of compression.

Each algorithm tries to meet these requirements in a different way. However, in their paper Cano *et al*. identify a number of steps and procedures common to the fingerprint extraction of almost all audio fingerprinting systems. Figure 4.2 shows a schematic view of these steps in the fingerprint extraction process. In the preprocessing step the audio signal is usually converted to mono, filtered using a lowpass filter and downsampled to a (lower) standard sample rate. Then, the signal is divided into (strongly) overlapping frames. The frame lengths range from 50-400 ms, the overlap varies from 50% to 98%. Each frame is multiplied by a window and converted to a spectral representation. In many algorithms the spectrum is divided into several frequency bands. Features are extracted from each frequency band in every frame. Each feature is then represented by a number of bits in the post-processing step. The compact representation of the time-frequency features of a single frame is called a sub-fingerprint. Due to the large overlap, subsequent sub-fingerprints are (strongly) correlated and vary slowly in time. The fingerprint of a song consists of a sequence of sub-fingerprints, which are stored in a database. A song-fragment is identified by matching a sequence of sub-fingerprints, called a fingerprint block, to the items in the database. A fingerprint block usually corresponds to several seconds of music.

The main differences between the algorithms found in literature are due to the (time-frequency) features that are used [28]. Based on the information used for extracting the feature sequence, we have divided fingerprinting algorithms into three categories [34]. From each category we selected one algorithm we consider to be representative for the category. Next, these three algorithms will be presented in more detail and they are used in the experiments presented in Section 4.4.

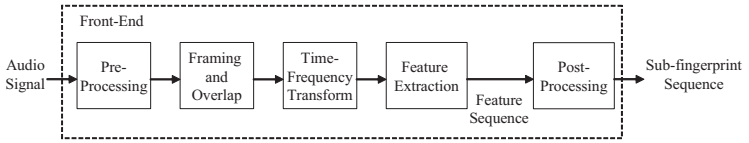


Figure 4.2: Fingerprint extraction procedure.

The three categories differ in the way they combine spectral information. The first category extracts a feature from each frequency band, the second category extracts features that are combined from multiple frequency bands and the third category extracts features that are based on the entire spectral range, while the combination is obtained through off-line training.

4.2.1 Systems that use features based on a single band

Shazam uses the locations of peaks in the spectrogram to represent the fingerprint [101]. This algorithm does not reflect the distortions related to compression, especially at medium and high bitrates. Özer *et al.* use periodicity estimators and a Singular Value Decomposition of the MFCC matrix [81]. Sukkittanon and Atlas propose frequency modulation features [94]. These papers do not address the response to compression. MusicDNA uses global mean and standard deviation of the energies within 15 subbands of 15 seconds of music, thus creating a 30 dimensional vector [100]. The effect of moderate compression is shown to be minimal. Both Fraunhofer's AudioID and the algorithm developed by Mapelli *et al.* use spectral shape descriptors to represent the fingerprint: the Spectral Flatness Measure (SFM) and Spectral Crest Factor (SCF) [48, 70]. The latter algorithm is well-defined and the response to compression is discussed in literature. Based on its reported response to compression and its full description, we have selected the latter SFM/SCF algorithm to represent this category. In the remainder of this chapter we refer to this algorithm by the abbreviation SSD (Spectral Shape Descriptors).

Figure 4.3 shows the SSD fingerprinting algorithm proposed by Mapelli *et al.* [70]. The algorithm extracts features from the periodogram estimate of the Power Spectral Density (PSD). The PSD of frame n at frequency bin k, $S(n,k)$, is estimated from the length-L windowed Fourier transform of the corresponding frame $\hat{X}(n,k)$:

$$S(n,k) = \frac{1}{L}\left|\hat{X}(n,k)\right|^2 \tag{4.1}$$

The extracted features are the mean energy (ME), the Spectral Flatness Measure (SFM) and the Spectral Crest Factor (SCF). We follow the approach in [48] to extract the features within each of several subbands per frame. The features are based on the arithmetic and geometric means of (subband) energies. Define the arithmetic mean of signal $x(i), \quad i = 1, \ldots, N$, as:

$$M_a(x) = \frac{1}{N}\sum_{i=1}^{N} x(i) \tag{4.2}$$

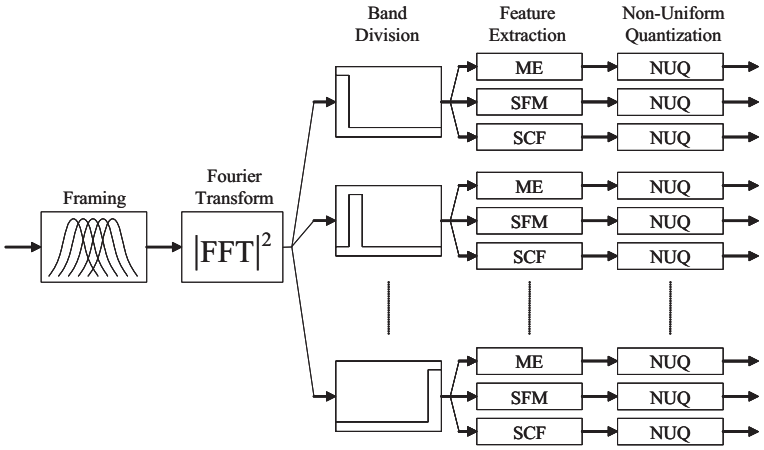


Figure 4.3: Fingerprint extraction stage of Cefriel SSD [70]

and the geometric mean as:

$$M_g(x) = \sqrt[N]{\prod_{i=1}^{N} x(i)} \tag{4.3}$$

In frame n and subband m the ME, SFM and SCF features are extracted from the periodogram $S(n,k)$ are then given as:

$$\begin{aligned} Feat(m,n,0) &= ME(n,m) \\ &= M_a\left(S(n,k)\right), \quad k \in \mathcal{K}_m \end{aligned} \tag{4.4}$$

$$\begin{aligned} Feat(m,n,1) &= SFM(n,m) \\ &= 10\log_{10}\left(\frac{M_g\left(S(n,k)\right)}{M_a\left(S(n,k)\right)}\right), \; k \in \mathcal{K}_m \end{aligned} \tag{4.5}$$

$$\begin{aligned} Feat(m,n,2) &= SCF(n,m) \\ &= 10\log_{10}\left(\frac{\max_k\left(S(n,k)\right)}{M_a\left(S(n,k)\right)}\right), \; k \in \mathcal{K}_m \end{aligned} \tag{4.6}$$

where $\mathcal{K}_m$ is the set of frequency bin indices belonging to subband m.

Within each band, each feature is quantized using a (different) 4-bit Non-Uniform Quantizer (NUQ). For more information the NUQ please refer to [70]. The fingerprint is thus defined as the quantization level index of each feature of the three features:

$$F(n,m,p) = NUQ_p\left(Feat(m,n,p)\right), \quad p = 0,1,2 \tag{4.7}$$

The distance between two fingerprint blocks is computed using the Mean Square Error (MSE):

$$MSE = \frac{1}{3MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\sum_{p=0}^{2}\left(F_X(m,n,p) - F_Y(m,n,p)\right)^2 \tag{4.8}$$

4.2.2 Systems that use features based on multiple subbands

The PRH algorithm presented in Section 2.6 uses the sign of the difference between energies in Bark scaled frequency bands [44]. While it is reported to be highly robust against distortions [44], the difference between fingerprints of original and compressed content also reflects compression artifacts [37].

The distance between two realizations $FP_X(n,m)$ and $FP_Y(n,m)$ is computed as the Bit Error Rate (BER):

$$\text{BER} = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} F_{diff}(n,m), \tag{4.9}$$

where

$$F_{diff}(n,m) = \text{XOR}(\, F_X(n,m),\, F_Y(n,m)\,). \tag{4.10}$$

4.2.3 Systems that use optimized subband- or frame-combinations

Batlle *et al*. use Hidden Markov Models (HMMs) to describe their fingerprint [20]. The HMMs are trained based on audio examples. In a second algorithm from the same authors, the states sequences of the HMMs are interpreted as 'Audio Genes' [79]. Both systems use complex distance measures, use the Viterbi algorithm for identification and implementation is far from straightforward. Microsoft Research uses dimensionality reduction techniques to extract the fingerprint in their Robust Audio Recognition Engine (RARE) [27]. The two-stage dimension reduction is based on training using examples. Compression artifacts are reflected in the distances between fingerprints of the original and the compressed content. Therefore, we select Microsoft's RARE to represent the third category of algorithms.

Figure 4.4(a) shows the fingerprint extraction of RARE, which uses the log power spectrum of the Modulated Complex Lapped Transform (MCLT) for the time-frequency representation of the data. The log power spectra are pre-processed to remove the effects of equalization and volume adjustment. A second pre-processing step removes the non-audible frequency components from the spectrum based on a simple Psycho-Acoustic Model (PAM) [69]. The entire pre-processing procedure is shown in Figure 4.4(b).

Features are extracted by means of a two-stage projection of the log power spectra. Each projection is the result of Oriented Principle Component Analysis (OPCA) which uses both undistorted and distorted data for a one-time, off-line training. OPCA projects the data onto those directions in the MCLT-frequency space that maximize the ratio of signal energy and distortion energy in the training data. These directions are the result of the eigenvalue decomposition of the covariance matrices of pre-processed log-power spectra of the training data. The first OPCA projection is based on the pre-processed log-power spectra of the training data, the second OPCA projection is based on a number of concatenated, projected spectra from the first OPCA projection. The fingerprint consists of the floating point representation of the trace of features, i.e. the trace of projected spectra. The distance between two fingerprints is computed using the Euclidean (Root Mean Square) distance.

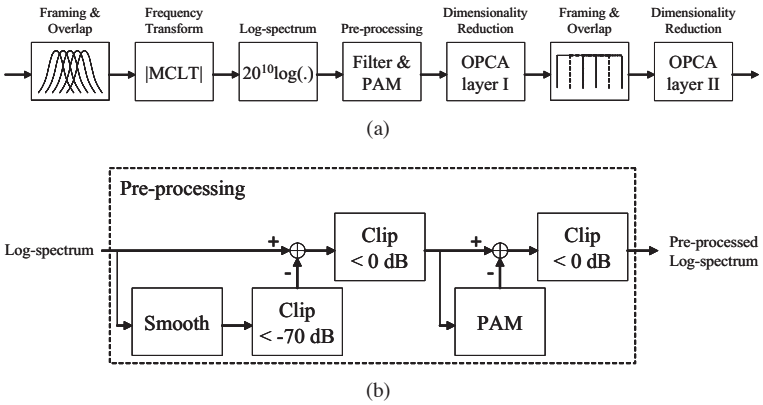


Figure 4.4: Microsoft's Robust Audio Recognition Engine (RARE) [27] (a) Fingerprint Extraction (b) Pre-processing.

4.3 Stochastic Models of the Philips Robust Hash

Each algorithm reviewed in the previous section has been developed for the *identification* of music. In the introduction we motivated that we want to use fingerprinting algorithms for *estimating the quality* of compressed music as well, as an add-on feature after the music has been identified. We base the quality estimation on the difference between the fingerprint stored in the database and the fingerprint extracted from the compressed content for identification.

In this section, we model the compression artifacts as additive white noise. We shall show that this relatively simple model for compression degradations leads to expressions that match experimental data very well. For the binary fingerprints of the PRH, we derive an expression for the probability of bit error, P_e, in terms of the SNR due to additive noise. We choose to model the PRH algorithm for three reasons. First, this algorithm is proven to be robust and used in practical applications [3, 2]. Second, it is well-documented [44] and therefore the subsequent steps in the fingerprint algorithm can be well understood. Finally, these steps can be modeled for simple signal models (uncorrelated and correlated stochastic signal models). Although the model is based on one specific algorithm (PRH) we expect the behavior to be indicative for the other algorithms as well, since the features in SSC, PRH and RARE are also based on linear combinations of components in the (log-)magnitude spectrum. In Appendix B.1 we sketch a relation between the MSE and SNR for the log-magnitude spectrum for uncorrelated signals. This relation is easily extendible to the RMS distance measure.

We thus consider the following situation. Denoting the undistorted signal to be fingerprinted by $x(i)$ and the additive, normally distributed noise by $w(i)$, the distorted signal $y(i)$ is given by:

$$y(i) = x(i) + w(i) \tag{4.11}$$

We are interested in the relating the difference between the corresponding fingerprints

of $x(i)$ and $y(i)$, $F_X(n,m)$ and $F_Y(n,m)$, respectively, to the statistical characteristics of $x(i)$ and $y(i)$. The probability of an erroneous bit in frame n at frequency band m, $P_e(n,m)$, can be expressed in terms of the energy differences, $ED_X(n,m)$ and $ED_Y(n,m)$ (see Eq. (3.6)):

$$\begin{aligned} P_e(n,m) &= Pr\left[F_X(n,m) \neq F_Y(n,m)\right] \\ &= Pr[ED_X(n,m) \leq 0, ED_Y(n,m) > 0 \\ &\quad \vee \quad ED_X(n,m) > 0, ED_Y(n,m) \leq 0] \end{aligned} \tag{4.12}$$

Section 4.3.1 extends the model from Section 3.4.1 to correlated signals $x(i)$. Section 4.3.2 uses the model from Section 4.3.1 to predict the behavior for music. Finally, section 4.3.3 addresses the problem of the large variance in $d(F_X, F_Y)$ for a given bitrate or SNR level, and proposes a modified distance measure to reduce this variance.

4.3.1 Synthetic signals

We start our analysis from the model outlined in Section 3.4.1 of the previous chapter. The model expresses the probability of an erroneous bit due to additive noise, P_e, in terms of the signal and noise variances, σ_X^2 and σ_W^2, respectively, and is given by Eq. (3.52):

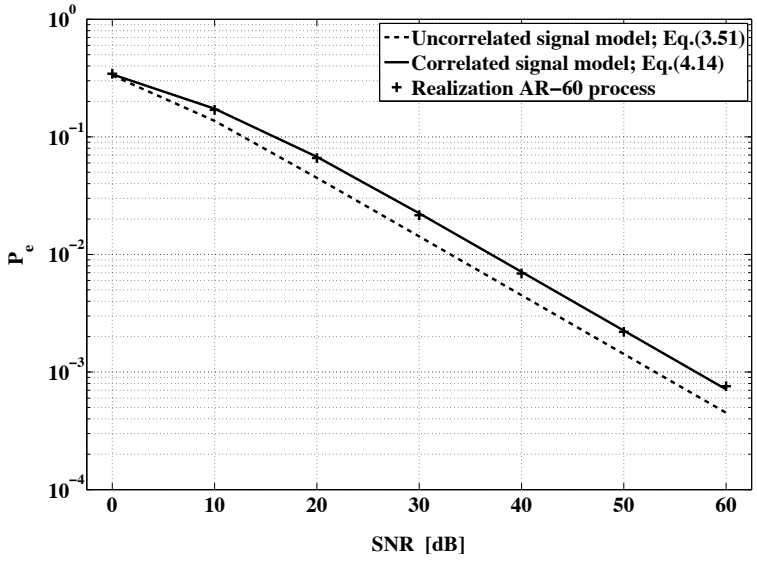
$$P_e = \frac{1}{\pi} \arctan\left(\sqrt{\frac{\sigma_W^4}{\sigma_X^4} + 2\frac{\sigma_W^2}{\sigma_X^2}}\right)$$

Since the fingerprint bits are stationary in n, and the white noise spectrum is flat, the probability $P_e(n,m) = P_e$ is independent of the indices n and m. Figure 3.11 shows P_e as a function of SNR; for high SNR, P_e drops by a factor 10 when the SNR improves with 20 dB.

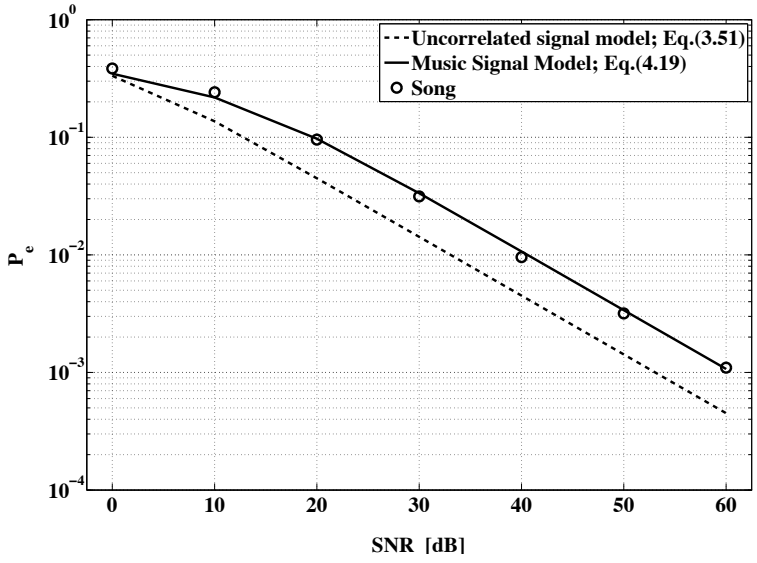
This model assumes that the signal is uncorrelated, and hence the PSD is constant. Therefore, all frequency bands have an identical robustness to additive noise and have equal probability of bit errors. When the signal $x(i)$ is correlated in time, the spectrum is not flat. Then, the bands in the periodogram having a relatively high average energy density (power/Hz) are more robust to additive white noise than those which have relatively low average energy density.

An extension to the model of Eq. (3.53), is to take the average energy and noise densities in the *individual* frequency bands into account. Let $\sigma^2_{X_m}$ denote the average energy density in frequency band m, and let $\sigma^2_{X_{m:m+1}}$ denote the average energy density in bands m and $m+1$; similar for $\sigma^2_{W_{m:m+1}}$. Then the Probability of Error of a bit in position m, which corresponds to the signal and noise in bands m and $m+1$, can be approximated by:

$$P_e(m) \approx \frac{1}{\pi} \arctan\left(\sqrt{\frac{\sigma^4_{W_{m:m+1}}}{\sigma^4_{X_{m:m+1}}} + 2\,\frac{\sigma^2_{W_{m:m+1}}}{\sigma^2_{X_{m:m+1}}}}\right) \tag{4.13}$$



(a)



(b)

Figure 4.5: SNR-BER relation for (a) an AR model of order 60 (model: '-', realization: '+') in the presence of additive noise. (b) model of a song (model: '-', realization: 'o'). As reference, the uncorrelated signal model ('- -') is also shown in (a-b).

Now assume that the noise is white, and as a consequence $\sigma^2_{W_{m:m+1}} = \sigma^2_W$. The model can then further be simplified to:

$$\begin{aligned} P_e(m) \approx & \frac{1}{\pi} \arctan\left(\sqrt{\frac{\sigma^4_W}{\sigma^4_{X_{m:m+1}}} + 2\,\frac{\sigma^2_W}{\sigma^2_{X_{m:m+1}}}} \right) \\ = & \frac{1}{\pi} \arctan\left(\sqrt{\frac{\sigma^4_W}{\sigma^4_X}\,\frac{\sigma^4_X}{\sigma^4_{X_{m:m+1}}} + 2\,\frac{\sigma^2_W}{\sigma^2_X}\,\frac{\sigma^2_X}{\sigma^2_{X_{m:m+1}}}} \right) \end{aligned} \tag{4.14}$$

It is easy to see that the ratio $\sigma^2_X/\sigma^2_{X_{m:m+1}}$ effectively scales the σ^2_W/σ^2_X argument according to the local average signal power. Of course, if band m contains $N_m = |\mathcal{K}_m|$ samples, the average power over all frequency bands, σ^2_X, is related to the average power in subband m, $\sigma^2_{X_m}$, through

$$\sigma^2_X = \sum_{m=0}^{M} N_m \sigma^2_{X_m} \Big/ \sum_{m=0}^{M} N_m. \tag{4.15}$$

In practical systems like the PRH, the subbands don't cover the entire spectral range; Eq. (4.15) assumes that the behavior in the $M+1$ subbands is representative for the behavior in the entire spectrum. This assumption is also implicitly made when using fingerprinting for identification: the fingerprint is based on part of the signal, but is assumed to be representative for the entire signal.

The overall BER can be expressed as the average of the M frequency band BERs:

$$P_e = \frac{1}{M} \sum_{m=0}^{M-1} P_e(m) \tag{4.16}$$

The model in Eq. (4.14) assumes that the PSD of the signal is flat within two subsequent bands and the model in Eq. (4.16) that the probabilities are independent over m. Eq. (4.16) again results in a - more complicated - $\arctan(\cdot)$ relation, since

$$\begin{aligned} & \arctan(a) + \arctan(b) \\ & = \arctan\left(\frac{a+b}{1-ab}\right) + \begin{cases} 0 & ab < 1 \\ \pi & ab > 1,\ a > 0 \\ -\pi & ab > 1,\ a < 0 \end{cases} \end{aligned} \tag{4.17}$$

As an illustration, Figure 4.5(a) shows the modeled and experimental SNR-BER curves for a 60^{th} order AR process. The coefficients were obtained by fitting the AR model onto a frame of real music. This example shows a perfect fit.

4.3.2 Music

Previous sections considered synthetic signal models. Here, we will extend the analysis to real audio signals. Although the model in Eqs. (4.14) and (4.16) assumes a stationary signal, it does reflect the influence of a non-flat spectrum. In music, the

spectral peaks correspond to reliable bits and the low-energy, noise-like regions correspond to unreliable bits. For music and additive noise we can extend the analysis by taking the non-stationarity into account. The errors between individual fingerprint bits reflect the SNR, *localized* both in time *and* frequency.

The expected probability of error, P_e of a fingerprint of size $N \times M$ is related to the ratio σ_X^2/σ_W^2 by:

$$P_e = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} P_e(n,m) \tag{4.18}$$

where

$$P_e(n,m) = \frac{1}{\pi} \arctan\left(\sqrt{2 \cdot \frac{\sigma_W^2}{\sigma_X^2} \cdot \frac{\sigma_X^2}{\sigma_{X_{n,m}}^2}} \right) \tag{4.19}$$

Here, $\sigma_{X_{n,m}}^2$ denotes the signal variance at position (n,m) and thus $\sigma_{X_{n,m}}^2/\sigma_W^2$ represents the SNR level corresponding to fingerprint bit $F(n,m)$. Equations (3.5) and (3.6) relate the value of this fingerprint bit to the energy in two frequency bands in two frames. The energy density of the signal reflected in the BER is assumed to be the maximum of the four energies in (3.5). This assumption is based on the observation that spectral peaks correspond to reliable fingerprint bits, but may lead to near zero subband energy differences $ED(n,m)$. Experiments show that for most music fragments the model in Eq. (4.19) fits better if the SNR is not solely based on frames $n-1$ and n, but estimated over a larger window size of $2r+1$ frames:

$$\sigma_{X_{n,m}}^2 = \max_{i,j} E^b(i,j) \quad \begin{matrix} i = n-r, \ldots, n, \ldots, n+r \\ j = m, m+1 \end{matrix} \tag{4.20}$$

In our experiments we used $r = 13$. The predicted and experimental curves for a 3 second music segment is shown in Figure 4.5(b).

4.3.3 Reducing the variance in the SNR-P_e relation for PRH

When in a song the spectral energy is concentrated in a few spectral components, the fingerprint bits corresponding to these peaks are very reliable since most processing preserves the spectral peaks. On the other hand, the spectral regions in between these spectral peaks become very unreliable. This is easily illustrated by the fact that the bandwidth of a subband in the Philips algorithm approximately a semitone. Is some classical music pieces with only one or a few instruments playing one or a few notes at a time the spectral energy within a frame is concentrated in few spectral peaks. This results in other subbands having near-zero energy and corresponding energy differences which are also around zero, and therefore generates fingerprint bits which are unreliable. This is easily illustrated by setting $\sigma_{X_{n,m}}^2 \ll \sigma_X^2$ in the model in Eq. (4.19), to represent the regions with near-zero energy differences. In this case, the relative noise level σ_W/σ_X is amplified by the small value of $\sigma_{X_{n,m}}^2$, pushing the

$\arctan(\,.\,)$-function towards its saturation level. The differences in spectral shape between different songs and the non-stationarity of music in general, result in a large variance of the P_e for a given SNR. If we like to estimate the SNR of a song using the fingerprint distance, this variance is a problem.

There are two ways to improve the estimation result. First, we can use longer song fragments, if available. However, the effect within a song is limited, due to the non-stationary character of music. Furthermore, the effect averaged over multiple songs is limited, due to the different spectral characteristics of different songs.

Second, we can use the model in the Section 4.3.2 to estimate the behavior of a specific song to additive distortions. By analyzing the spectrogram, we can estimate the probability of error for individual bits by using Eq. (4.14). This estimation can be used in two different ways: either the estimation is used to correct the SNR-estimation for a specific song. This information can either be stored in the database, or be estimated from the spectrogram of the song to be identified. The alternative is to use only those bits from the fingerprint to estimate the SNR that reflect the additive distortion level in the same way as in the case of white noise.

That is, we define a subset $\mathcal{L}$ of the fingerprint bit positions $\{n, m\}$ to compute the distance between the fingerprints, such that by considering only these fingerprint bits $F(n, m),\ \{n, m\} \in \mathcal{L}$, the (SNR, $P_{e,est}$) behaves approximately the same as the theoretical (SNR, P_e)-curve for white noise, i.e.:

$$\mathcal{L} \quad \text{s.t.} \quad P_{e,est}(\text{SNR} \mid \mathcal{L}) \approx P_e(\text{SNR}) \tag{4.21}$$

where $P_{e,est}$ denotes the average probability of bit error estimated for a specific song, obtained using the model in Eqs. (4.19) and (4.20). Also in this case, the set of usable fingerprint bits $\mathcal{L}$ can be stored additionally in the database, or be estimated from the spectrum of the (distorted) song that is (to be) identified. After identification of a song using its fingerprint, the SNR can be estimated from the BER of the bits indicated in $\mathcal{L}$:

$$\text{BER}_W = \frac{1}{|\mathcal{L}|} \sum_{\{n,m\}\in\mathcal{L}} F_{diff}(n, m), \tag{4.22}$$

where $|\cdot|$ denote the cardinality of the set.

We now focus on how to obtain the set of usable fingerprint bits $\mathcal{L}$. Using Eq. (4.14) the behavior of a small fragment of $2r + 1$ frames can be predicted from the spectrum. Let's denote the averaged behavior within a number of frames explicitly by the function

$$P_{e,est}(\text{SNR}, \mathcal{L}) = \frac{1}{\mathcal{L}} \sum_{\{n,m\}\in\mathcal{L}} P_e(n, m, \text{SNR}) \tag{4.23}$$

Now, those fingerprint bits are selected that statistically approximate the white noise fingerprint bit flip probability:

$$\mathcal{L} \quad \text{s.t.} \quad P_{e,est}(\text{SNR} \mid \mathcal{L}) \approx P_e(\text{SNR}) \tag{4.24}$$

The set $\mathcal{L}$ is obtained in the following, iterative way. Since the strongest spectral peaks generate the most reliable bits, in iteration i we select the bits corresponding to the $|\mathcal{L}_i|$ strongest spectral components. One can see that for a given SNR level, adding a spectral component which is weaker that those already selected increases $P_{e,est}(\text{SNR} \mid \mathcal{L})$, i.e.

$$P_{e,est}(\text{SNR} \mid \mathcal{L}_i) < P_{e,est}(\text{SNR} \mid \mathcal{L}_{i+1}) \quad \mathcal{L}_i \subset \mathcal{L}_{i+1} \tag{4.25}$$

In order to determine when we have to stop selecting additional spectral components, we evaluate the cost function L_i:

$$L_i = \int_{-\infty}^{\text{SNR}_{\max}} \big\{\log(P_e(\text{SNR})) - \log(P_{e,est}(\text{SNR}, \mathcal{L}_i))\big\}^2 d\text{SNR} \tag{4.26}$$

The cost function expressed the distance between the two curves $P_{e,est}(\text{SNR} \mid \mathcal{L})$ and $P_e(\text{SNR})$. Due to the increasing nature of Eq. (4.25) the cost function is convex and has a minimum for a certain iteration i. The SNR-region of interest is limited by $\text{SNR}_{\max}$ for three reasons. First, the integral does not converge for the limit $\text{SNR} \to \infty$. Second, in most practical compression systems, the SNR resulting from audio coding is not infinite. Third, due to the limited fingerprint block range, extremely small error probabilities cannot be reliably estimated from the fingerprint difference. For convergence there not necessarily needs to be a lower SNR-bound, since $\lim_{\text{SNR}\to-\infty} P_{e,est} = 0.5$.

Figure 4.6(a) shows the result of applying this strategy to music and additive noise. The variance in the BER for a given SNR level is greatly reduced.

4.4 Experiments using music

In Section 4.2 we split up the field of audio fingerprinting algorithms into three categories and presented one algorithm for each category. In Section 4.3 we presented stochastic models for the PRH algorithm. In this section, we experimentally compare the three algorithms presented in Section 4.2 with each other.

Section 4.4.1 discusses the details of the comparison process. Sections 4.4.3 and 4.4.2 compare the algorithms in a compression bitrate-vs.-$d(F_X, F_Y)$ and a SCNR-vs.-$d(F_X, F_Y)$ setting.

4.4.1 Enabling algorithmic comparison

The fingerprinting systems described in Section 4.2 not only use different features, but also have different operating conditions like sampling rates, frame length, granularity, etc. A fair comparison requires similar operating conditions. Therefore, we set the following parameters for all systems:

- Sampling rate of 5512.5 Hz
- Frequency bands between 300 and 2000 Hz for the PRH and SSD system
- Fingerprint block length of about 3.1 seconds
- Framelength of 2048 samples (371.5 ms)
- Fingerprint block size of 4096 bits

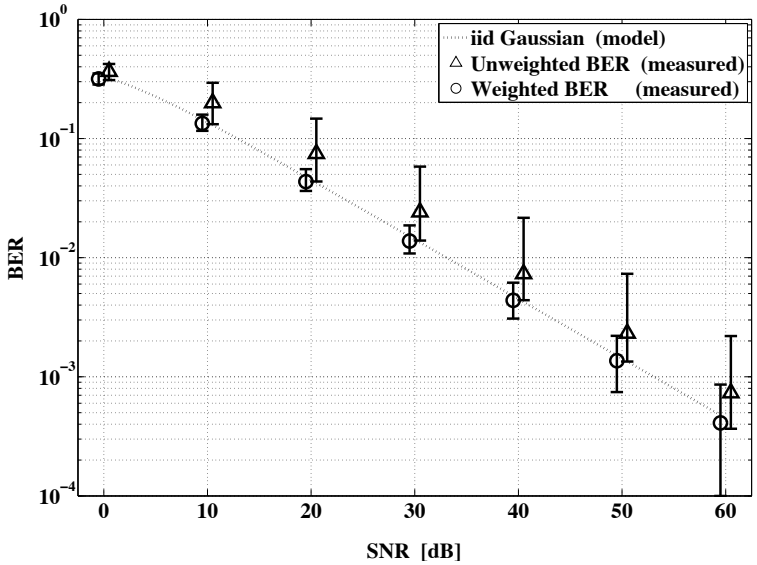
In order to achieve these settings, we can modify the frame overlap ratio, the number of frequency bands, the number of features, the number of bits to represent each feature. In addition we have changed the overlap ratio in the second OPCA layer of Microsoft's RARE system. Table 4.1 compares the settings for the different systems.

We have used 275 song fragments of 40 seconds each; 100 of these fragments have been used for training Microsoft's RARE system. This is in the same order of magnitude as the number of songs mentioned in [27]. For each of these 100 song fragments we have generated 9 distorted versions. These distortions are mainly non-linear amplitude distortions and two pitch shifts. Compression is not one of the distortions.

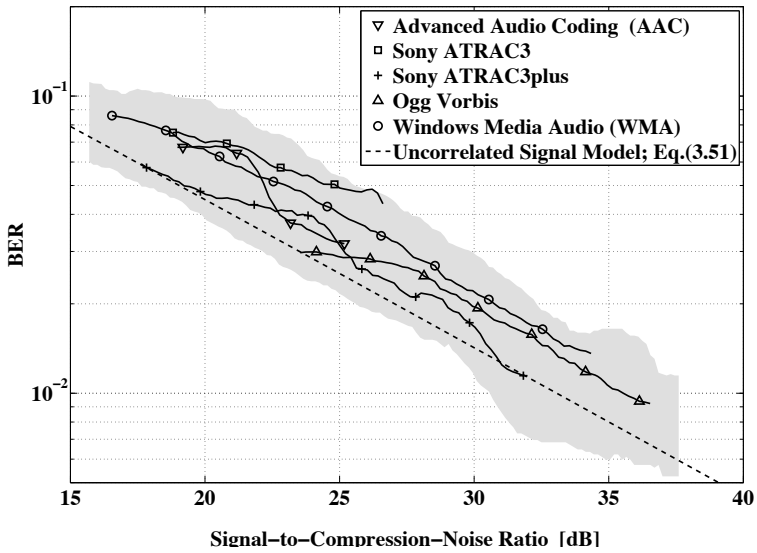
For the large scale experiments discussed later in this section, we have used MP3 compression using the LAME codec [1]. The selected bitrates for MP3 compression range from 32-256 kilobit-per-second (kbps) using constant bitrate. To test the variability over different compression algorithms, we have conducted a small-scale experiment shown in Figure 4.6(b) (for the PRH algorithm only) with a number of different, widely-used audio codecs, including Advanced Audio Coding (AAC) [26], Sony ATRAC(plus) [97, 92], Ogg Vorbis [5], and Window Media Audio (WMA) [74]. They all show a comparable behavior on the SNR-Fingerprint difference plots. This was to be expected, since our model does not model one specific coding scheme, but uses a white noise model. Furthermore, all of these audio coders are waveform coders - as opposed to parametric coders, such as sinusoidal coders - using a subband decomposition and/or a MDCT time-frequency transform. It other words, although they differ a lot in performance and implementation, they all use the same basic tools to achieve the compression.

For each system we have set a threshold for identification, such that all system operate under the same false positive rate per fingerprint block, P_{fa}. The P_{fa} is based on a Gaussian approximation of the distances between fingerprint blocks of original, undistorted fragments. We have chosen $P_{fa} = 10^{-5}$, which is quite high for a practical fingerprinting system, when compared to some of the numbers reported in literature[1]. However, $P_{fa} = 10^{-5}$ is achievable for all three systems and we are interested in the relation between compression and fingerprint distance, given a fixed false alarm rate P_{fa}.

[1]False positives reported in literature can be as low as 10^{-20} for PRH [44], but 10^{-5} to 10^{-8} for RARE (depending on the experiment) [27].



(a)



(b)

Figure 4.6: (a) SNR-BER relation for additive noise on music averaged over 11 songs: SNR-BER ('△') and SNR-BER$_W$ ('o'). The markers indicate the median. Errorbars indicate lower and upper 10% BER values for a given SNR. The curves have been shifted slightly horizontally in order not to overlap. The iid model is shown as a reference ('- -'). (b) SNR-BER relation for 9 songs, comparing the behavior of the PRH algorithm in its original form [44] for 5 different compression algorithms: AAC ('▽'), Sony ATRAC ('□'), Sony ATRAC3plus ('+'), Ogg Vorbis ('△') and WMA ('o'), and the curve for the uncorrelated signal model (Eq. (3.52)).

Table 4.1: Comparison between parameters for original and modified versions of selected systems (a) PRH and SSD (b) RARE.

(a)

	PRH		SSD	
	Original System	Modified System	Original System	Modified System
Sample rate [Hz]	5512.5	5512.5	44100	5512.5
Frequency Range [Hz]	300-2000	300-2000	300-3400	300-2000
Window length [ms]	371.5	371.5	743	371.5
Frame overlap ratio	31/32	31/32	63/64	31/32
# Bits er feature	1	1	4	4
# Frequency bands	33	17	1	4
# Features	1	1	3	1
# Frames per segment (sec.)	256 (3.1 s)	256 (3.1 s)	64 (1.5 s)	256 (3.1 s)

(b)

Microsoft	Original System	Modified System
Sample rate (Hz)	11025	5512.5
Window length (ms)	371.5	371.5
Frame overlap ratio	1/2	1/2
Overall OPCA reduction	$32 \times 2048 \rightarrow 64$	$16 \times 1024 \rightarrow 64$
Fingerprint block length (frames)	32 (6.2 s)	16 (3.1 s)
Overlap ratio in 2^{nd} OPCA layer	0	1/2

4.4.2 Experimental relation between bitrate and $d(F_X, F_Y)$

Figures 4.7 and 4.8 compare the relation between compression bitrate and fingerprint differences for the original algorithms with their modified counterparts. In general, the behavior of the modified algorithms is comparable to the algorithms using the original settings. Since the differences have been normalized such that the algorithms achieve a similar P_{fa}, the scale of the curves is related to the variance of the distribution of the fingerprints of the uncompressed songs.

If one would try to estimate the bitrate from the fingerprint differences, the spread in the curves for a given bitrate should be as small as possible. Visual inspection learns that for each curve, the standard deviation at a certain bit rate compared to the corresponding mean value is in the same order of magnitude. Therefore, we can conclude that there is not one algorithm that stands out in its potential for bitrate estimation.

4.4.3 Experimental relation between SNR and $d(F_X, F_Y)$

Audio compression introduces compression noise. In the stochastic models in the previous section, the compression noise was modeled as independent, stationary, uncorrelated noise. In practice, however, this is not the case. Audio compression algorithms apply psycho-acoustic models to shape the compression noise in the temporal and spectral domain, such that the artifacts are rendered inaudible. Figures 4.9 and 4.10 show the Signal-to-Compression-Noise for the three algorithms. Figure 4.9(b) - 4.10(a) compares the modified version with an implementation using settings described in literature.

The shading indicates the spread in fingerprint differences of the curves. After being normalized to achieve the common P_{fa}, some of the curves have been shifted for display purposes, resulting in a vertical shift in the plot, to avoid overlap. The scaling factors are indicated in the caption of Figures 4.9 and 4.10. It is quite clear that all curves have approximately the same gradient in the SNR plots. Although the SNR $-$ P_e in Eq. (3.52) was derived for an uncorrelated signal in the presence of additive, uncorrelated noise, the experimental SCNR-$d(F_X, F_Y)$ for all three algorithms follow the $\arctan(\cdot)$-regime. RARE and SSD make use of the log-magnitude spectrum. In Appendix B.1 we roughly outline the relation between MSE and the SNR for i.i.d. Gaussian data.

Due to the fact that in compression the bitrates are chosen, and the SNR levels are a result of the selected bitrate, it is not straightforward to indicate the spread in the curves. Since the points are not aligned on certain SNR levels, the shading indicates the 1/6-percentile and 5/6-percentile within an overlapping bin of SNR levels. The binning introduces the effect that the angle of the averaged curves changes slightly (becomes less steep at the end points). Curves for one single fragment show a clear relation between SNR and fingerprint difference: if the SNR is increased by 20 dB, the fingerprint difference becomes 10 times smaller.

4.5 Conclusion and discussion

4.5.1 Conclusions

A wide variety of audio fingerprinting systems has been presented in literature over the last couple of years. The main difference between the systems is the features that are used. We have shown that although the features and projections that are used in the three systems that have been compared are very different, the fingerprint differences behave in a comparable fashion as a function of SNR or compression bitrate. This behavior matches the behavior predicted by the models presented in Section 4.3. For these distortions, the actual detection performance for identification is mainly dependent on the distribution of the differences between arbitrary fingerprints. This determines the threshold for identification.

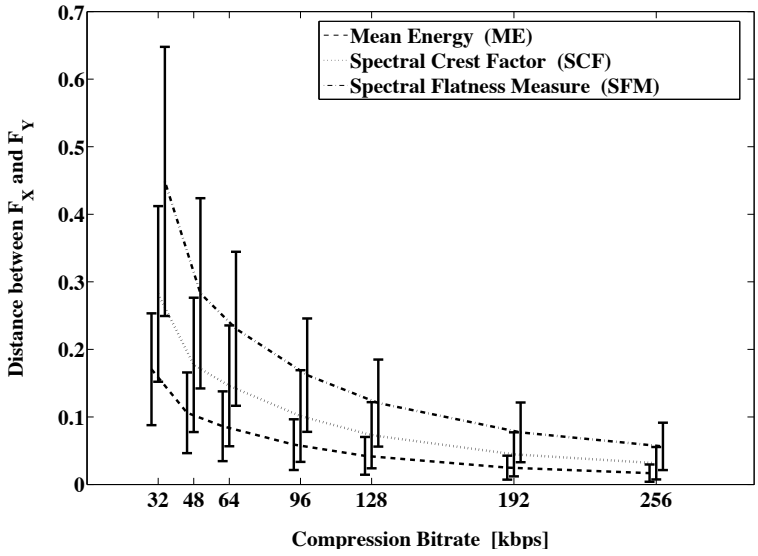
The differences between fingerprints reflect the difference between an original recording and a compressed version and can be used to roughly estimate the quality of compressed content. The main obstacle for doing this is the large variance of the fingerprint difference for a given compression bit rate. All algorithms in our study suffer from a variance which is relatively large. This limits the classification possibilities to 3 or maybe 5 classes of different SNR level, which should be enough for our intended use. We have shown that for the PRH this variance can be reduced by discarding certain unreliable bits in computing the distance between two fingerprints. For the other two algorithms, the variance reduction still is an open issue.

4.5.2 Extension to perceptually motivated distortion measures

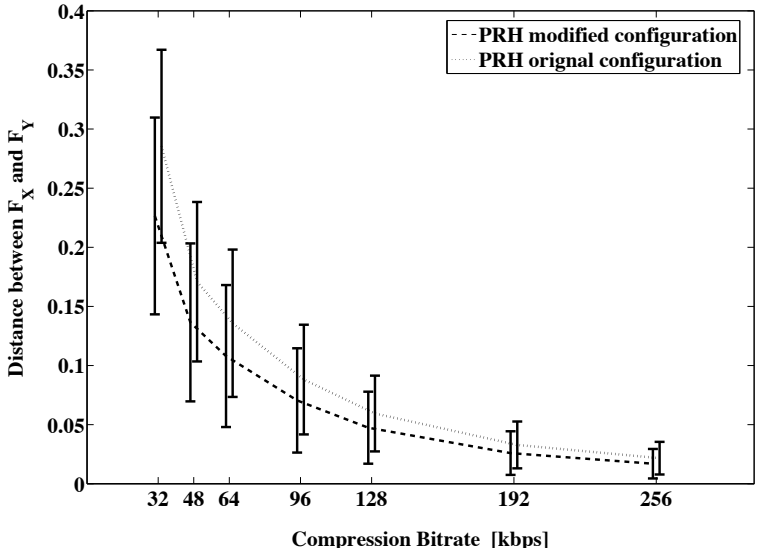
Our current approach relates the fingerprint differences to SNR. Although SNR is suitable for our envisioned application scenarios, we foresee two options to alter the current setup to relate the fingerprint differences to more perceptually motivated distortion measures.

In coding applications and in systems that predict the subjective quality in given audio signal with respect to the reference, psycho-acoustical models are used to estimate the so-called masking threshold. The masking threshold models the fact that some components in the audio signal, can mask - make less audible - other components which are close-by in time and frequency. The estimation procedure of the masking threshold models the way the Human Auditory System (HAS) reacts to sounds. Spectral components that fall below this masking threshold are not audible and are therefore considered irrelevant.

To match fingerprint differences to a distance measure involving psycho-acoustics, we can distinguish between two different approaches: altering the fingerprinting scheme and altering the fingerprint distance measure. In both cases the masking threshold can be estimated from the spectrum, even on a subband basis.

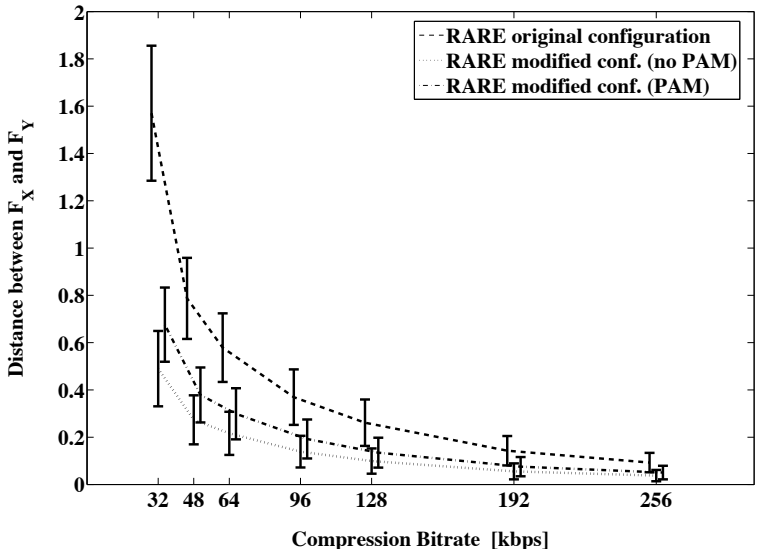


(a)

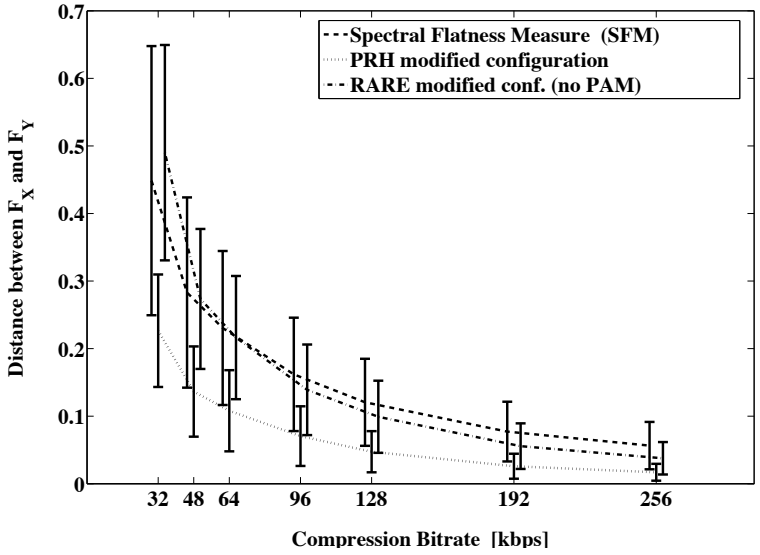


(b)

Figure 4.7: Compression bitrate vs. fingerprint differences. The curves have been shifted such that there is no overlap. (a) The features in the SSD algorithm: From top to bottom: Energy (−−), SCF (· · ·), SFM (−.), (b) PRH: Modified (−−), Original (· · ·).

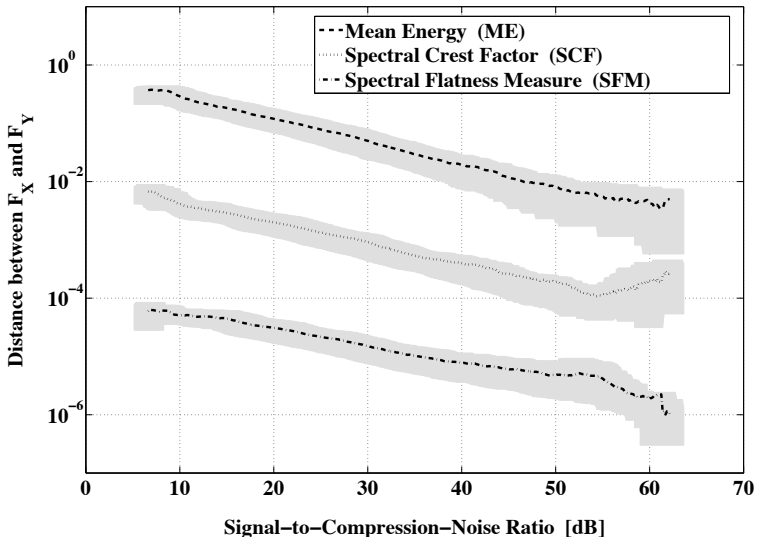


(a)

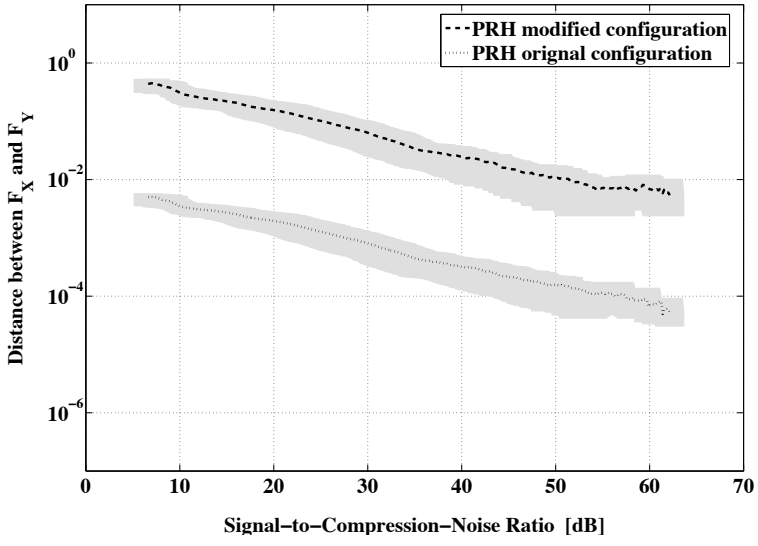


(b)

Figure 4.8: Compression bitrate vs. fingerprint differences. The curves have been shifted such that there is no overlap. (a) RARE: Original (−−), Modified, no Psycho-Acoustic Model (· · ·), Modified, using a Psycho-Acoustic Model (−.) (b) Comparison between the modified versions of SFM (−−), PRH (· · ·), RARE (−.).

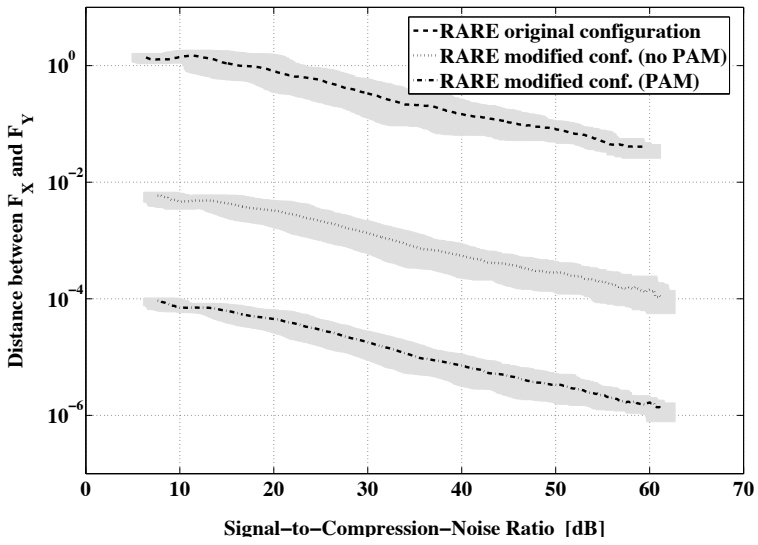


(a)

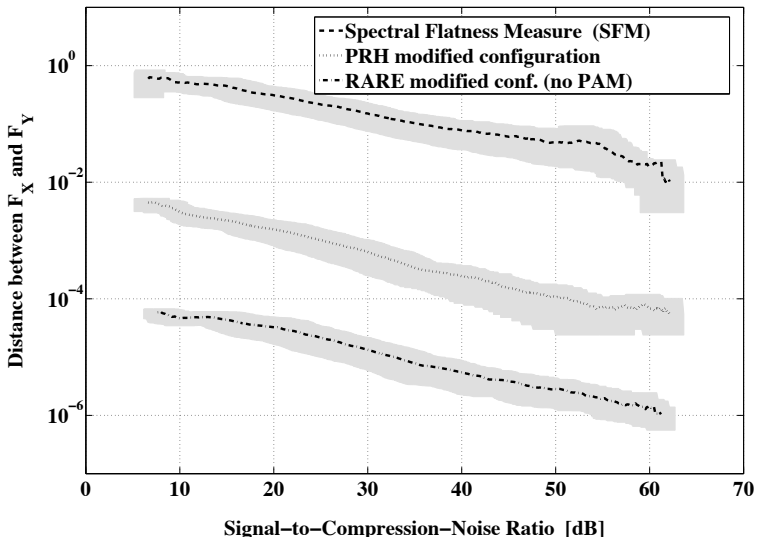


(b)

Figure 4.9: Compression SCNR vs. fingerprint distances. The lines mark the average behavior, the shaded areas indicate the spread. The curves have been scaled such that there is no overlap. (a) The features in the SSD algorithm: From top to bottom: Energy ($--$, not scaled), SCF ($\cdots$, scaled by factor 10^{-2}), SFM ($-.$, scaled by factor 10^{-4}), (b) PRH: Modified ($--$, not scaled), Original ($\cdots$, scaled by factor 10^{-2}).



(a)



(b)

Figure 4.10: Compression SCNR vs. fingerprint distances. The lines mark the average behavior, the shaded areas indicate the spread. The curves have been scaled such that there is no overlap. (a) RARE: Original ($--$, not scaled), Modified, no Psycho-Acoustic Model ($\cdots$, scaled by factor 10^{-2}), Modified, using a Psycho-Acoustic Model ($-.$, scaled by factor 10^{-4}) (b) Comparison between the modified versions of SFM ($--$, not scaled), PRH ($\cdots$, scaled by factor 10^{-}), RARE ($-.$, scaled by factor 10^{-4}).

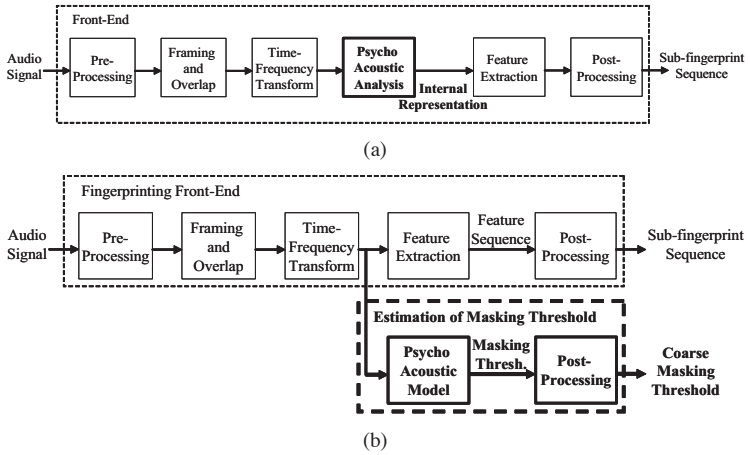


Figure 4.11: Towards perceptually motivated fingerprint distances: including Psycho-acoustical models (a) in the audio fingerprint extraction stage, (b) parallel to the fingerprint extraction stage.

In the first approach, the fingerprint extraction procedure outline in Figure 4.2 is changed to estimate the sound representation inside the human ear using the masking threshold, shown in Figure 4.11(a). Spectral components that exceed the masking threshold are scaled by it; components that fall below the masking threshold can be considered inaudible and can therefore be removed from the spectrum. The fingerprint features can then be extracted from the estimated internal representation instead of from the raw spectrum.

In the other approach, shown in Figure 4.11(b), the masking threshold is computed in parallel with the fingerprint, but not included in the derivation of the fingerprint itself. Together with the reference fingerprint, a rough approximation of the - e.g. average masking per critical band which has a bandwidth equal to that of multiple fingerprint subbands - can be efficiently stored in the database. This masking threshold can be used to estimate the Noise-to-Mask Ratio (NMR), a feature used for psycho-acoustic analysis [21]. The main idea is to combine a local estimation of SNR and a local estimation of Signal-to-Mask Ratio (SMR) in the following way:

$$\mathrm{NMR} = \mathrm{SMR} - \mathrm{SNR} \qquad [\mathrm{dB}]$$

The SNR is estimated using the techniques described in this chapter. To estimate the SMR we need an estimation of the signal variance and the masking threshold. Each can be estimated from the query signal, or de derived from components in the database. The first approach is less reliable since the masking threshold should be estimated from the reference signal. The second approach needs either the masking threshold or the SMR to be stored in the database in parallel with the fingerprint used for identification. Due to the strong frame-overlap both masking threshold and SMR are expected to slowly develop in time enabling efficient storage.

Whatever psycho-acoustical measure is introduced, the results will never compete

with the subjective quality predicting algorithms like PEAQ, nor should they. To illustrate the limitations of such models in fingerprinting scenario's we refer to the fact that the frame lengths used in algorithms like PEAQ are very small compared to those used in fingerprinting.

4.5.3 Further development of fingerprint models

The model we developed for the behavior of the PRH is confirmed by experiments, both on simple stochastic signals, and on real music. Here, the model was used to predict how the SNR relates to the P_e. In a previous modeling approach, we developed a model describing the structure of the PRH fingerprint itself (so $F_X(n,m)$ instead of $d(F_X(n,m), F_Y(n,m))$) [35]. This triggered another modeling approach by McCarthy *et al*. [15]. These models describing the behavior of fingerprinting systems can also be used to predict and improve the performance of these systems.

The fact that the systems behave more or less the same - the relation between compression bitrate and fingerprint differences and between noise and fingerprint differences have comparable shapes - leads us to believe that there is more to fingerprinting than just extraction of robust features. There seems to be more common ground to behavior of the algorithms than the steps preceding the feature extraction. Therefore, it makes sense to analyze fingerprinting on a more abstract level, and to analyze the relation between compression and audio fingerprinting in general without considering specific implementations or systems.

Chapter 5

Information Theoretical Models for Fingerprinting

5.1 Introduction

An audio fingerprint is a compact representation which can be used for identification, in a way which is robust to audio degradations. From this short description it is intuitive that these two aspects – audio degradations and compactness of the fingerprint – influence the identification performance of the fingerprinting system. If the audio signals to be identified – and therefore the fingerprints – are not distorted, the size and representation of the fingerprint determine the number of signals that can be represented. For instance, if each song is mapped onto a 3-dimensional binary vector, $2^3 = 8$ songs can be represented. This analysis is complicated by the fact that the signal to be identified usually is a distorted version of the reference signal. Hence, the fingerprint of the distorted signal is a distorted version of the reference fingerprint. If only 1 bit in the 3-bit fingerprint is allowed to be different, then a maximum of only 2 fingerprints can be discriminated. From this example, it is intuitive that the maximum allowed distortion ('when are two signals the same'), the size of the fingerprint ('how many bits are extracted per second or per sample') and representation of the fingerprint ('what does the fingerprint look like') determine the maximum number of identifiable signals.

In this chapter, the two research questions are: *how many songs can be reliably identified for a given degradation level and a certain fingerprint size?* and *is the fingerprint degradation behavior we observed in the experiments in the previous chapters characteristic for audio fingerprinting?* The starting point for the analysis is a recent paper by Westover and O'Sullivan [104]. In their paper, Westover and O'Sullivan (WOS) generalize the concept of a pattern recognition system [104]. The goal of such a system is to reliably recognize distorted versions of signals that have been learnt by the system. Figure 5.1 shows the setup of the WOS model. Like in fingerprinting, two phases are distinguished: a training phase and an identification phase. In the training phase, a number of signals M_c are presented to the system. Each signal in the training

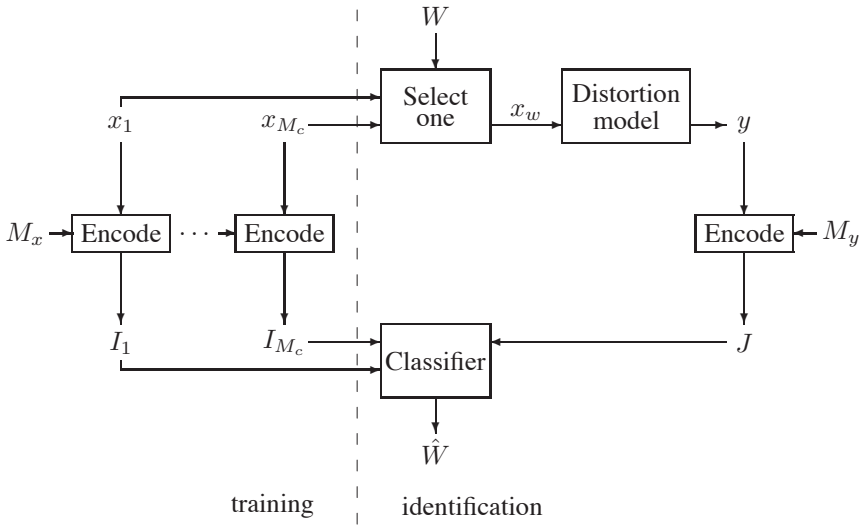


Figure 5.1: Functional diagram of the WOS framework. Each training signal is encoded in the training stage. In the identification stage a randomly selected training signal is distorted, encoded, and compared to the encoded training signals.

phase $x_1, \ldots, x_{M_c}$ is represented in a compact form $I_1, \ldots, I_{M_c}$ and stored in memory. In the identification phase, one of the signals from the training phase is selected x_W, distorted (resulting in signal y) and presented to the system. The task of the system now is to determine which of the signals from the training phase is presented to the system. This is done by comparing a compact representation of the distorted signal, J, to the representations $I_1, \ldots, I_{M_c}$ stored in memory.

The paper by Westover and O'Sullivan presents and discusses a model to theoretically determine how many signals can maximally be reliably identified by pattern recognition systems. In the WOS model, the design choices are the compressed representations of the signals in the training phase and in the identification phase, and the method how to identify the signal; the model for the distortion and the representation (alphabet) of the signals are the constraints. The WOS model does not tell how to make these design choices. The model considers the number of signals that can be distinguished in both compressed representations and the number of signals that can be identified, given the two constraints. Given the alphabet of the training signals, and a probabilistic model for the distortion, it derives a bound for the maximum number of signals that can be reliably identified as a function of the bitrate employed in representing the fingerprints. In many aspects the constraints and design choices identified in the WOS model closely match the practice of audio fingerprinting.

This chapter is organized as follows. Section 5.2 presents the details of the WOS model. The WOS model is not the only model found in literature with similarities to fingerprinting (e.g. cf. [105, 98]), but it is the one which most closely matches the practice of fingerprinting and the experiments carried out in previous chapters. Sections 5.3 and 5.4 use the WOS model in two different ways. Section 5.3 considers

the PRH fingerprint and the distortion model derived in Section 3.4.1 – Eq. (3.52) –, and analyzes the PRH from a capacity perspective. In other words, how many i.i.d. signals can be distinguished by the PRH system, when we allow a signal to be distorted by additive noise up to a certain SNR level. This relates to the first research question posed in the beginning of this chapter. Section 5.4 takes the WOS model, and analyzes it from a distortion perspective; this relates to the second research question. Finally, Section 5.5 draws conclusions and discusses the relation between the models and experiments in this chapter, and real-life audio fingerprinting.

5.2 The WOS model

This chapter presents the WOS model in more detail, and is organized as follows. In Section 5.2.1 we present the setup and notation of the WOS model. In their paper, Westover and O'Sullivan derive general expressions for the bounds on the number of signals, M_c, that can be reliably identified. General here means that they are formulated such that they are not restricted to a particular representation (alphabet) of the signals or a particular distortion model. The main question is: 'how large can M_c be, under certain conditions?'. The conditions refer to the distortion model, the size and representation of the fingerprint. We present these results in Section 5.2.2. Westover and O'Sullivan derive closed form solutions of the bounds for two specific source signals: binary signals and Gaussian signals. In this chapter we omit the results for the binary case; we present the results for the Gaussian signals in Section 5.2.3.

5.2.1 Model setup and definitions

Figure 5.2 shows the set-up of the WOS model. In the training phase, a total number of M_c signals x are fingerprinted and stored in memory. Each signal consists of n samples. In the identification phase, one of these known signals is selected, distorted, and presented to the system to be identified. The index w denotes which of the M_c training signals has been selected. The recognition is based on a fingerprint derived from the distorted signal y. The fingerprinting in the training and identification phases are performed by two different mappings ϕ_x and ϕ_y, respectively. The result of representing the signal x by the mapping ϕ_x is an index I. The number of different indices that are possible outcomes of the mapping ϕ_x is denoted by M_x. Hence, the possible values for I can be enumerated, e.g. $I \in \{1, \ldots, M_x\}$. Similar arguments apply for the mapping ϕ_y which is applied to the distorted signal y in the identification phase, resulting in a number M_y of possible indices J. Hence, the possible values for J can be enumerated, e.g. $J \in \{1, \ldots, M_y\}$. The distortion follows a conditional probability function $Pr(y|x)$, but the exact distortion is unknown. An identification is made by a classifier g which compares the training indices $I_1, \ldots, I_{M_c}$ to the received index J. The result of the identification is an estimate $\hat{w}$ of which training signal was selected and distorted. An error is made if the identification is incorrect, i.e. $\hat{w} \neq w$.

Since both the signals x and y consist of n samples, rates can be associated with

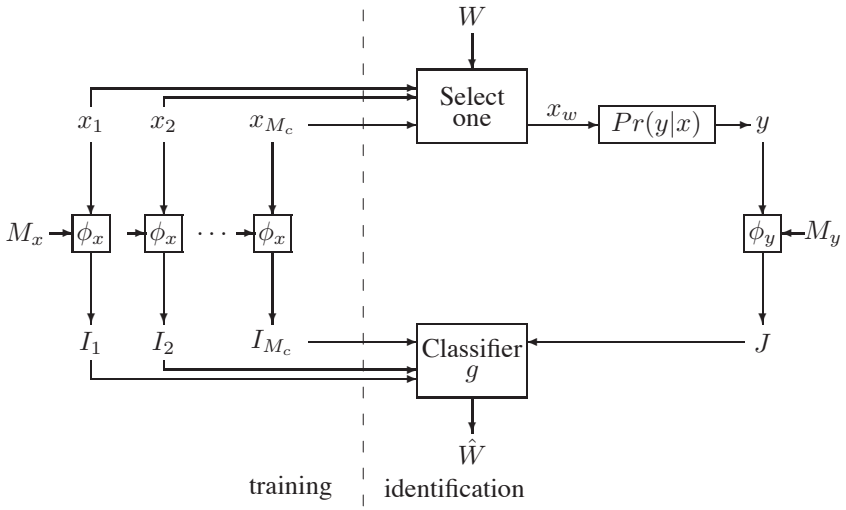


Figure 5.2: Functional diagram of the WOS framework. Each training signal is encoded using mapping ϕ_x in the training stage. In the identification stage a randomly selected training signal is distorted according to a model $Pr(y|x)$, encoded using mapping ϕ_y, and compared to the encoded training signals using classifier g.

the number the messages M_x and M_y, respectively:

$$R_x = \frac{1}{n}\log(M_x) \tag{5.1}$$

$$R_y = \frac{1}{n}\log(M_y) \tag{5.2}$$

Similarly, the number of signals in the training phase, M_c, can be normalized by the number of samples in the form of a rate:

$$R_c = \frac{1}{n}\log(M_c) \tag{5.3}$$

Note that the rates R_x and R_y indicate how many signals can be represented by the mappings ϕ_x and ϕ_y, while R_c indicates how many signals can be distinguished by the classifier g.

The model makes two assumptions on the type of signals and distortions. First, the signals x are iid: $p(x) = \prod_{i=1}^{n} p(x(i))$. Second, the distortion is memoryless: $p(y|x) = \prod_{i=1}^{n} p(y(i)|x(i))$. Furthermore, the assumption is made that each training signal has equal probability of being selected for identification. This leads to an expression for the average probability of error: $P_e = \frac{1}{M_c}\sum_{w=1}^{M_c} Pr[\hat{w} \neq w]$.

The WOS model attempts to indicate which rate combinations (R_x, R_y, R_c) are 'achievable'. When a rate combination is achievable, this means as much as that we can control the error rate: the probability of error P_e can be made lower than an arbitrary level $\epsilon > 0$ by increasing the signal length n (i.e. P_e tends to zero for increasingly large n). In practice, this is seen in the distributions of the detection statistic. As

shown in Figure 2.4, two conditional distributions of the distance between two fingerprints play a role: 1) the distribution given that the audio signals are similar, and 2) the distribution given that the audio signals are dissimilar. By increasing the number of seconds of music used to identify the song fragments, these distribution become less wide: the variance gets smaller relative to the mean values of these distributions. This makes the distributions better separable, and the probability error decreases.

The mappings ϕ_x and ϕ_y, and the classifier g are considered channels. The output of the mapping ϕ_x is an index, I, which can simply be enumerated; alternatively, the output of the mapping may be represented as a signal, U. As an example, lets assume the mapping ϕ_x is realized as a vector quantizer. In this comparison the index I corresponds to the index of the Voronoi region, and U corresponds to the centroid of the Voronoi region. A similar approach may be taken for the output of ϕ_y. Hence, to derive the achievability bounds two auxiliary random variables are introduced: U to represent the result of the mapping ϕ_x, and V to represent the result of the mapping ϕ_y. A joint probability distribution $P_{U,X,Y,V}(u, x, y, v)$ characterizes the relations between the variables[1].

The mappings ϕ_x and ϕ_y need to be able to 'cover' all the messages that can be 'transmitted' over this 'channel' between X and U, and Y and V, respectively. The associated minimum rates R_x and R_y and the maximum rate R_c are dependent on the probability distribution $P_{U,X,Y,V}(u, x, y, v)$.

The aim is to divide the space of all rate combinations (R_x, R_y, R_c) into two regions: those rates that are achievable, and those that are not achievable. This is done by maximizing R_c over all probability distributions that give rise to a certain minimum rate $(R_x, R_y) = (r_x, r_y)$.

However, such a clear division into achievable and non-achievable rates cannot derived by Westover and O'Sullivan. They consider two approximations of the achievable rate region: the inner bound region, and the outer bound region. All rate combinations within the inner bound region are achievable. Non of the rate combinations outside the outer bound region is achievable. Some rate combinations are inside the outer bound region, but not inside the inner bound region. For these rates it is unclear if they are achievable or not. Figure 5.3 shows an example of the rate region and the bounds as a function of the bitrate used to represent the fingerprints, for a given distortion model. For simplification, we consider a setup in which $R_x = R_y$.

The figure shows the rate R_c associated with the number of signals that can be discriminated by a fingerprinting system (M_c) as a function of the rate R_x associated with the mapping ϕ_x. Suppose $M_x = 256$ for signals of length $n = 4$, then the associated rate $R_x = -\frac{1}{n}\log_2(M_x) = 2$ bits per sample. Increasing R_x means that the mapping ϕ_x is capable of representing more different signals X. Larger values for R_c imply that the system can discriminate more different signals (U, V). Let $r_{in}(r_x)$ and $r_{out}(r_x)$ denote the inner and outer bound as a function of the rate $R_x = r_x$, respectively. In Figure 5.3, the inner bound region is marked 'Achievable rates'; for a given fingerprint rate $R_x = r_x$, all rates $R_c \leq r_{in}$ are achievable. The area outside the outer bound region is marked 'Non-achievable rates'; for a given fingerprint rate $R_x = r_x$, all rates $R_c > r_{out}$ are not achievable. The figure clearly shows the gap

[1] in case of continuous signals: the joint probability density function $f_{U,X,Y,V}(u, x, y, v)$

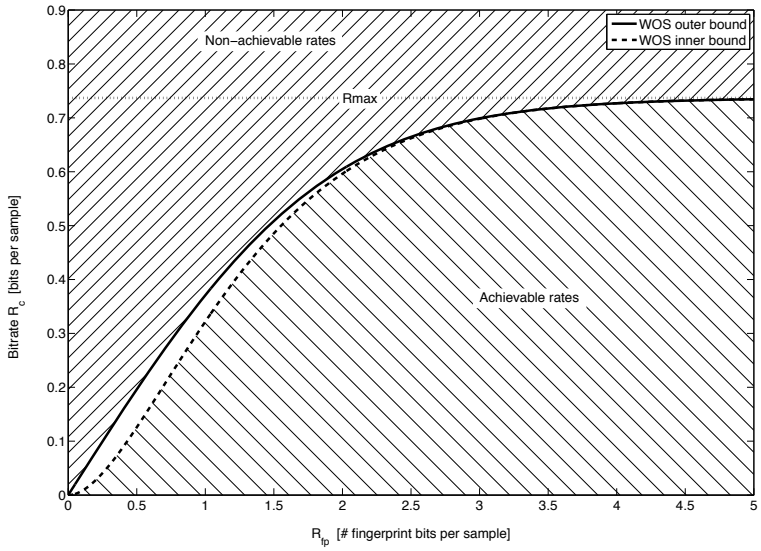


Figure 5.3: Example of the achievable rate regions (WOS model bounds for Gaussian signals, corresponding to distortion correlation coefficient $\rho_{xy} = 0.8$. The curve saturates at $R_{max} = -\frac{1}{2}\log(1 - 0.8^2) = 0.74$ bits per sample; illustrated by the dotted line.)

between the two bounds; for these rate combinations it is unclear if they are achievable or not. The gap is most prominent for low values of R_x. For large rates R_x, the R_c curve saturates and tends towards the value R_{max}. This value is dependent on the maximum allowed distortion in the $X - Y$ channel.

5.2.2 Bounds on the achievable rates

Some probability distributions give rise to rate combinations that are achievable, some to rate combinations that are not achievable. By optimizing over all probability distributions that give rise to achievable rates, the bound can be found. However, it is difficult to say beforehand which probability distributions give rise to achievable rate. Furthermore, it is impossible to optimize over all probability distributions. Westover and O'Sullivan show that all probability distributions satisfying a certain structure are achievable. In this way, they limit the number of probability distributions over which one needs to maximize.

Two sets of joint probability distributions are formed. The first set $\mathcal{P}_{out}$ contains all joint distributions $P_{U,X,Y,V}(u, x, y, v)$ which satisfy the following two Markov chain relations between the variables: $U - X - Y$ and $X - Y - V$. The second set $\mathcal{P}_{in}$ contains all joint distributions $P_{U,X,Y,V}$ which satisfy the following three Markov chain relations between the variables: $U - X - Y$, $X - Y - V$, and $U - (X, Y) - V$. The Markov chain relation $U - X - Y$ represents the statement 'U and Y are conditionally independent given X,' i.e., $p(u, y|x) = p(u|x)p(y|x)$. This set of probability

distributions $\mathcal{P}_{in}$ is thus a subset of $\mathcal{P}_{out}$.

The achievable rate region is characterized by its surfaces $r_{in}(r_x, r_y)$ and $r_{out}(r_x, r_y)$ as a function of $(R_x, R_y) = (r_x, r_y)$. Here $r_x = I(U;X)$, where $I(U;X)$ denotes the mutual information for discrete variables X and U defined as:

$$I(U;X) = \sum_{u\in U}\sum_{x\in X} p_{U,X}(u,x)\log\left(\frac{p_{U,X}(u,x)}{p_U(u)p_X(x)}\right) \tag{5.4}$$

in case of continuous variables X and U, the mutual information is defined as:

$$I(U;X) = \int_{u\in U}\int_{x\in X} f_{U,X}(u,x)\log\left(\frac{f_{U,X}(u,x)}{f_U(u)f_X(x)}\right)d(u,x) \tag{5.5}$$

The surfaces for the achievable rate region can be expressed by:

$$r_* = \max_{(U,V)\in\mathcal{C}_*(r_x,r_y)} I(U;V) - I(U;V|XY) \tag{5.6}$$

$$= r_x + r_y - \min_{(U,V)\in\mathcal{C}_*(r_x,r_y)} I(XY;UV) \tag{5.7}$$

where $*$ can be replaced by either *in* or *out* for the inner and outer bound, respectively, and

$$\mathcal{C}_{in}(r_x,r_y) = \{(U,V)\in\mathcal{P}_{in} : r_x = I(U;X), r_y = I(V;Y)\} \tag{5.8}$$

$$\mathcal{C}_{out}(r_x,r_y) = \{(U,V)\in\mathcal{P}_{out} : r_x = I(U;X), r_y = I(V;Y)\}, \tag{5.9}$$

where $\mathcal{C}_{in}$ and $\mathcal{C}_{out}$ represent the variables (U,V) from a distribution in $\mathcal{P}_{in}$ or $\mathcal{P}_{out}$, respectively, and result in a rates r_x and r_y.

Optimizing the expressions over $\mathcal{C}_{out}$ yields the outer achievability bound; optimizing over $\mathcal{C}_{in}$ yields the inner bound. Due to the construction of $\mathcal{P}_{in}$, for the inner bound the term $I(U;V|XY) = 0$, yielding $r_{in} = I(U;V)$. In other words, the inner bound is given by the mutual information in the $U-V$ channel only.

5.2.3 Gaussian signals

The Gaussian version[2] of the problem considers originals signals X and distorted signals Y that are jointly Gaussian with correlation coefficient ρ_{xy}. Since this perfectly matches the setup used in Chapter 3 with the fingerprints of iid signals in the presence of additive Gaussian noise, we consider this specific case in more detail. For Gaussian signals, the WOS paper derives closed form expressions for the surfaces of the inner and outer bounds expressed as optimizations in Eqs. (5.6) and (5.7).

[2]In deriving the expressions in the WOS model, assumptions are made that are only valid for discrete alphabets. However, the model is also applied to signals with continuous magnitudes like Gaussian signals. Although the framework and the expressions seem to be correct, no arguments are presented why the model derived for discrete alphabets should also be valid for continuous sources. The Gaussian framework seems in line with related model in [105].

Since the mappings ϕ_x and ϕ_y are modeled as Gaussian channels, the rates r_x and r_y can be expressed in terms of the correlation coefficients ρ_{ux} and ρ_{yv}, respectively:

$$r_x = -\frac{1}{2}\log(1-\rho_{ux}^2) \tag{5.10}$$

$$r_y = -\frac{1}{2}\log(1-\rho_{yv}^2) \tag{5.11}$$

The correlation coefficients ρ_{ux} models the coding ('compression') of X into U through the mapping ϕ_x. Similar arguments apply to ρ_{yv}.

The authors assume that the joint distributions $f_{U,X,Y,V}(u,x,y,v)$ that maximize the two bounds are both jointly Gaussian. Let us define the vectors $\mathbf{A} = [X,Y]$ and $\mathbf{B} = [U,V]$. Since the distribution is jointly Gaussian, the mutual information $I(XY;UV)$ is fully determined by the correlation matrix $\mathbf{C}_{XYUV}$[3]:

$$\begin{aligned}\mathbf{C}_{XYUV} &= \left[\begin{array}{cc|cc} \sigma_X^2 & \rho_{xy}\sigma_X\sigma_Y & \rho_{ux}\sigma_U\sigma_X & \rho_{xv}\sigma_X\sigma_V \\ \rho_{xy}\sigma_X\sigma_Y & \sigma_Y^2 & \rho_{uy}\sigma_U\sigma_Y & \rho_{yv}\sigma_Y\sigma_V \\ \hline \rho_{ux}\sigma_U\sigma_X & \rho_{uy}\sigma_U\sigma_Y & \sigma_U^2 & \rho_{uv}\sigma_U\sigma_V \\ \rho_{xv}\sigma_X\sigma_V & \rho_{yv}\sigma_Y\sigma_V & \rho_{uv}\sigma_U\sigma_V & \sigma_V^2 \end{array}\right] \\ &= \left[\begin{array}{c|c} \mathbf{C_A} & \mathbf{C_{AB}} \\ \hline \mathbf{C'_{AB}} & \mathbf{C_B} \end{array}\right], \end{aligned} \tag{5.12}$$

where ρ_{xy} denotes the correlation coefficient for the variables X and Y, etc.

The differential entropy for variable $\mathbf{A}$ is equal to:

$$h(\mathbf{A}) = \frac{1}{2}\log((2\pi e)^2\det(\mathbf{C_A})). \tag{5.13}$$

Similarly, the differential entropy for $\mathbf{A}|\mathbf{B}$ is equal to:

$$h(\mathbf{A}|\mathbf{B}) = \frac{1}{2}\log((2\pi e)^2\det(\mathbf{C_{A|B}})), \tag{5.14}$$

where

$$\mathbf{C_{A|B}} = \mathbf{C_A} - \mathbf{C_{AB}}\mathbf{C_B}\mathbf{C'_{AB}}. \tag{5.15}$$

Now, we obtain the mutual information $I(XY;UV)$:

$$\begin{aligned} I(XY;UV) &= I(\mathbf{A};\mathbf{B}) \\ &= h(\mathbf{A}) - h(\mathbf{A}|\mathbf{B}) \\ &= -\frac{1}{2}\log(\det(\mathbf{I} - \mathbf{C_A}^{-1}\mathbf{C_{AB}}\mathbf{C_B}^{-1}\mathbf{C'_{AB}})) \\ &= \frac{1}{2}\log\left(1 + \frac{2\rho_{uv}\gamma - \beta}{1-\rho_{uv}^2}\right), \end{aligned} \tag{5.16}$$

[3]The relations $\rho_{uy} = \rho_{ux}\,\rho_{xy}$ and $\rho_{xv} = \rho_{xy}\,\rho_{yv}$ apply because of the Markov chain conditions $U - X - Y$ and $X - Y - V$.

where β and γ are defined as:

$$\begin{aligned}\beta &= \rho_{ux}^2 + \rho_{yv}^2 - \left(1-\rho_{xy}^2\right)\rho_{yv}^2\rho_{ux}^2 \\ &= 1-(1-\rho_{ux}^2)(1-\rho_{yv}^2)+\gamma^2 \qquad (5.17)\\ \gamma &= \rho_{ux}\rho_{xy}\rho_{yv}. \qquad (5.18)\end{aligned}$$

The achievable rate region is expressed in Eqs. (5.6) and (5.7) in terms of its surfaces $r_{in}(r_x, r_y)$ and $r_{out}(r_x, r_y)$. Using this expression for $I(XY;UV)$, the surface $r_*(r_x, r_y)$ can be expressed in terms of r_x and r_y through ρ_{ux} and ρ_{yv}, respectively.

$$\begin{aligned}r_* &= r_x + r_y + \frac{1}{2}\log\left(1+\frac{2\rho_{uv,*}\gamma-\beta}{1-\rho_{uv,*}^2}\right)\\ &= -\frac{1}{2}\log\left(\frac{(1-\rho_{uv,*}^2)(1-\rho_{ux}^2)(1-\rho_{yv}^2)}{(1-\rho_{ux}^2)(1-\rho_{yv}^2)-(\gamma-\rho_{uv,*})^2}\right), \qquad (5.19)\end{aligned}$$

where '*' can be replaced by in or out. The surfaces $r_{in}(r_x, r_y)$ and $r_{out}(r_x, r_y)$ are obtained by substitution of $\rho_{uv,in}$ and $\rho_{uv,out}$, respectively, in Eq. (5.19). $\rho_{uv,*}$ is the result of maximizing r_* through setting $\frac{\partial}{\partial\rho_{uv}}\ r_* = 0$. $\rho_{uv,in}$ is the result of maximizing of variables for which the Markov constraints in $\mathcal{P}_{in}$ apply. Similarly, $\rho_{uv,out}$ is the result of maximizing of variables for which the Markov constraints in $\mathcal{P}_{out}$ apply.

The constraints on the dependencies between the random variables in the two sets $\mathcal{P}_{in}$ and $\mathcal{P}_{out}$ for the Gaussian case take the form of relations between the elements in correlation matrix. Applying the Markov conditions for the inner bound, the correlation coefficient $\rho_{uv,in}$ can be written as $\rho_{uv,in} = \gamma$. Therefore, the inner bound becomes:

$$\begin{aligned}r_{in}(r_x, r_y) &= I(U;V)\\ &= -\frac{1}{2}\log\left(1-\gamma^2\right) \qquad (5.20)\end{aligned}$$

The outer bound is obtained by minimizing the mutual information $I(XY;UV)$ over ρ_{uv} through setting $\frac{\partial}{\partial\rho_{uv}}\ I(XY;UV) = 0$. The solution yields the optimal value:

$$\rho_{uv,out} = \frac{\beta}{2\gamma} - \sqrt{\left(\frac{\beta}{2\gamma}\right)^2 - 1} \qquad (5.21)$$

The outer bound is now given by:

$$\begin{aligned}r_{out}(r_x, r_y) &= r_x + r_y + \frac{1}{2}\log\left(1+\frac{2\rho_{uv,out}\gamma-\beta}{1-\rho_{uv,out}^2}\right)\\ &= -\frac{1}{2}\log\left(\frac{(1-\rho_{uv,out}^2)(1-\rho_{ux}^2)(1-\rho_{yv}^2)}{(1-\rho_{ux}^2)(1-\rho_{yv}^2)-(\gamma-\rho_{uv,out})^2}\right), \qquad (5.22)\end{aligned}$$

which now contains $\rho_{uv,out}$.

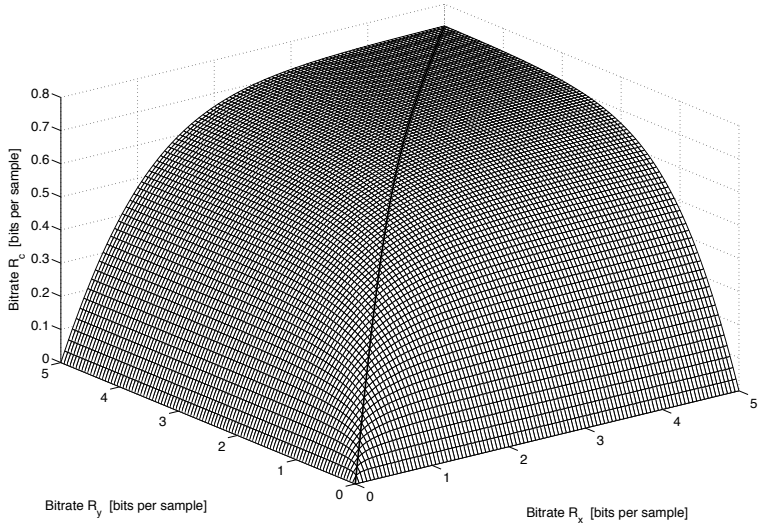


Figure 5.4: WOS model inner bound for Gaussian signals, and a Gaussian distortion corresponding to a correlation coefficient $\rho_{xy} = 0.8$.

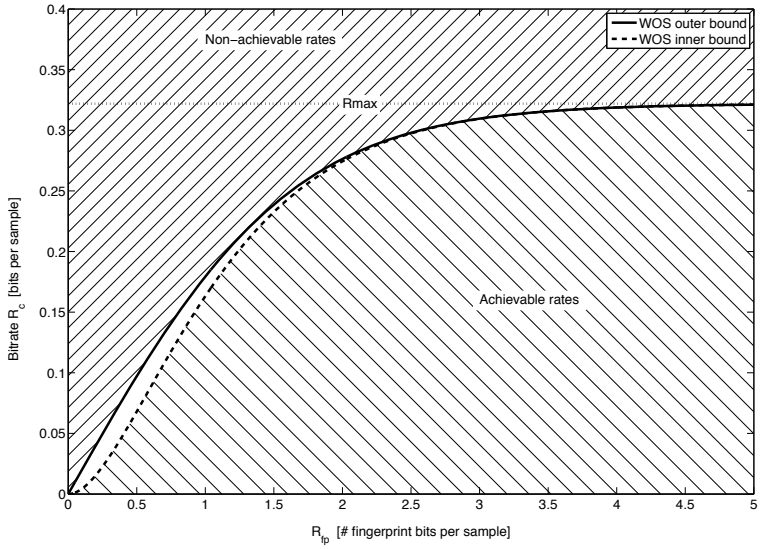
Figure 5.4 shows the computed inner bound surface $r_{in}(r_x, r_y)$ for the case when the distortion $f_{Y|X}(x, y)$ is characterized by the correlation coefficient $\rho_{xy} = 0.8$. The example in Figure 5.3 used the same value for ρ_{xy}, and shows the inner and outer bounds $r_{in}(r_x)$ and $r_{out}(r_x)$, respectively, with $r_x = r_y$. Figure 5.5 shows the inner and outer bounds with $r_x = r_y$ for two other distortion levels ρ_{xy}.

From these equations and examples the following observations can be made on the bounds in some limiting cases. First, when either one of the correlation coefficients ρ_{ux} or ρ_{yv} tends to zero, so do r_{in} and r_{out}. This situation corresponds to $R_x = 0$ or $R_y = 0$, respectively. This matches the intuitive notion that when the different signals x (or y) cannot be distinguished based on their compressed representations (zero bits per sample), no signals can be identified.

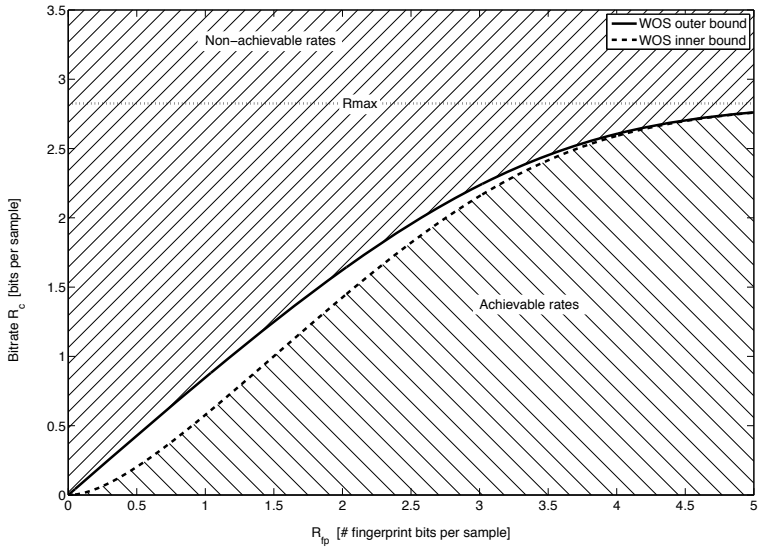
Second, when the rate r_x used in the mapping ϕ_x is increased, more signals can be represented in memory. In this case the correlation coefficient ρ_{ux} tends towards one. As a result, the outer bound collapses onto the inner bound. Lets assume $\rho_{ux} = 1$ (no limitation on the number of signals that can be represented by ϕ_x). In this case $\beta = 1 + \rho_{yv}^2 \rho_{xy}^2 = 1 + \gamma^2$, and therefore $\rho_{uv,out} = \gamma$, making the outer bound equal to the inner bound. Similar results apply when $\rho_{yv} = 1$.

Third, when both ρ_{ux} and ρ_{yv} tend towards one, the maximum rate R_c in the outer bound is limited solely by the capacity of the distortion channel, $r_c \leq -\frac{1}{2}\log(1-\rho_{xy}^2)$, as indicated by the dotted horizontal lines in Figures 5.3 and 5.5. In other words, at high bitrates $r_x = r_y$ the maximum value of $r_{out}(r_x, r_y) \approx r_{in}(r_x, r_y)$ is determined by ρ_{xy}.

In their paper, Westover and O'Sullivan exclusively use the WOS model to compute the recognition capacity as a function of the bitrates r_x and r_y, for a *given* statisti-



(a)



(b)

Figure 5.5: WOS model bounds for Gaussian signals, corresponding to distortion correlation coefficients (a) $\rho_{xy} = 0.6$; $R_{max} = 0.32$; (b) $\rho_{xy} = 0.99$; $R_{max} = 2.83$.

cal distortion model $Pr[y|x]$ (or in the continuous case: $f_{Y|X}(x,y)$). In our further analysis we will assume Additive White Gaussian Noise, i.e. $Y = X + W$, where $W \sim \mathcal{N}(0, \sigma_W^2)$. For our analysis it is interesting to fix the rates $r_x = r_y$, and to vary the level of the additive distortion σ_W^2. We thus consider the capacity bounds as a function of the SNR, for a given bitrate $r_x = r_y$. This corresponds to symmetric fingerprinting systems.

The WOS bounds as a function of SNR are obtained by converting the SNR for additive noise into ρ_{xy} through

$$\text{SNR} = 10\log_{10}\left(\frac{\sigma_X^2}{\sigma_W^2}\right) = 10\log_{10}\left(\frac{\rho_{xy}^2}{1-\rho_{xy}^2}\right), \tag{5.23}$$

where we use the relation of the correlation coefficient ρ_{xy} to σ_X and σ_W:

$$\rho_{xy} = \sqrt{\frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}}. \tag{5.24}$$

Further, in Eqs. (5.20) and (5.22), $R_x = R_y$ is converted into $\rho_{ux} = \rho_{yv}$ using Eqs. (5.10) and (5.11).

Figure 5.6 shows an example of the bounds as function of SNR, for a given bitrate of $R_{fp} = R_x = R_y = 1$ bit per sample. For increasing SNR, the inner and outer bounds increase as well, and saturate at levels dependent on the used bitrate R_{fp}. As in the (R_{fp}, R_c)-plots, there is a clear gap between the inner and outer bound. For large SNR the outer bound converges to the bitrate $r_x = r_y$. This is line with the intuitive notion that the system cannot recognize more signals than it can uniquely represent. To proof this, we need to show that:

$$\lim_{\rho_{xy}\uparrow 1} r_{out}(\rho_{xy}) = r_x \tag{5.25}$$

Starting point is the substitution of the expression for $\rho_{uv,out}$ in Eq. (5.21) into Eq. (5.22):

$$\begin{aligned} r_{out}(r_x) &= 2r_x + \frac{1}{2}\log\left(1 + \frac{2\rho_{uv,out}\gamma - \beta}{1-\rho_{uv,out}^2}\right) \\ &= 2r_x + \frac{1}{2}\log\left(1 - \frac{4\gamma^2\sqrt{\beta^2-4\gamma^2}}{2(4\gamma^2-\beta^2)+2\beta\sqrt{\beta^2-4\gamma^2}}\right) \\ &= 2r_x + \frac{1}{2}\log\left(1 - \frac{2\gamma^2}{-\sqrt{\beta^2-4\gamma^2}+\beta}\right) \end{aligned} \tag{5.26}$$

Using $\rho_{ux} = \rho_{yv}$, computation of the limit for β and γ yields:

$$\lim_{\rho_{xy}\uparrow 1} \beta = 2\rho_{ux}^2 - (1-\rho_{xy}^2)\rho_{ux}^4 = 2\rho_{ux}^2 \tag{5.27}$$

$$\lim_{\rho_{xy}\uparrow 1} \gamma = \rho_{ux}^2\rho_{xy} = \rho_{ux}^2 \tag{5.28}$$

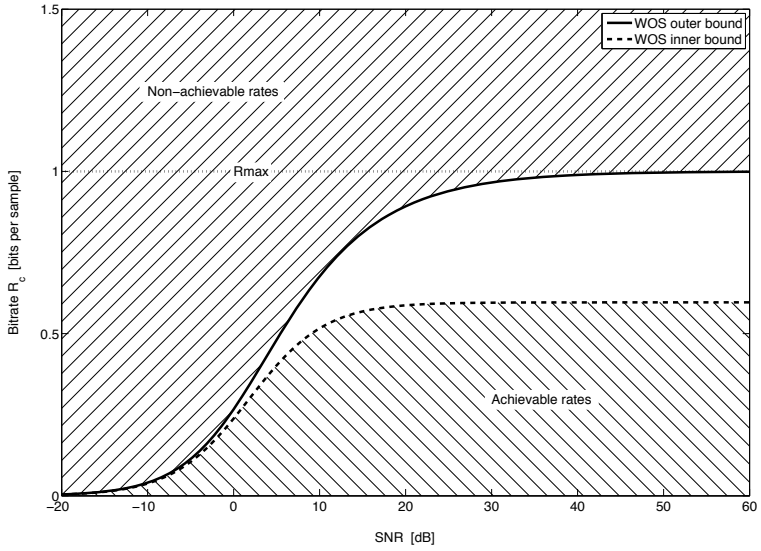


Figure 5.6: WOS model SNR vs. R_c for a given bitrate $R_{fp} = R_x = R_y = 1$ bits per sample.

Combining these results yields:

$$\begin{aligned}\lim_{\rho_{xy}\uparrow 1} r_{out}(\rho_{xy}) &= 2r_x + \frac{1}{2}\log\left(1 - \frac{2\gamma^2}{\beta}\right) \\ &= 2r_x + \frac{1}{2}\log\left(1-\rho_{ux}^2\right) \\ &= -\frac{1}{2}\log\left(1-\rho_{ux}^2\right) \\ &= r_x \end{aligned} \tag{5.29}$$

Similarly, for large SNR the inner bound converges to:

$$\lim_{\rho_{xy}\uparrow 1} r_{in}(\rho_{xy}) = -\frac{1}{2}\log\left(1-\rho_{ux}^4\right) \tag{5.30}$$

We will further use this SNR vs. capacity perspective in the following section.

5.3 The PRH model from a capacity perspective

In this section we analyze the PRH algorithm from a capacity perspective. In analogy to the WOS model we try to answer the question: how many signals (messages) can maximally be identified by binary fingerprints like the PRH fingerprint. We compare the bounds derived for the PRH with the bounds from the WOS model.

Figure 5.7 shows the fingerprinting setup in analogy to the WOS model. A number of signals, $x_1, \ldots, x_{M_c}$, is fingerprinted using the PRH fingerprinting algorithm;

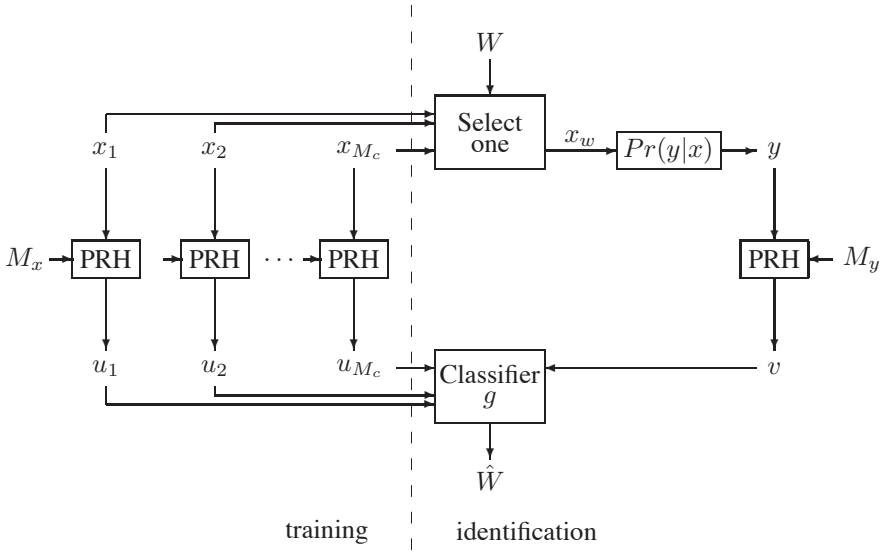


Figure 5.7: PRH fingerprint set-up for capacity analysis based on the WOS- model.

these fingerprints $u_1, \ldots, u_{M_c}$ are stored in a database. The input signals x are Gaussian i.i.d. signals. An arbitrary reference signal $x_w, w = 1, \ldots, M_c$ is chosen, and distorted by additive white Gaussian noise. The system tries to identify the resulting noisy signal y. This signal is also fingerprinted using the PRH algorithm. The query fingerprint v is compared to the reference fingerprints in the database $u_1, \ldots, u_{M_c}$, and an identification is made. Both branches of the PRH fingerprints use the same mapping (including same rate, i.e. $R_x = R_y$). Now we ask ourselves: given a binary fingerprinting algorithm operating at rate $R_x = R_y$, behaving according to the differential PRH-model in Section 3.4.1, and given an SNR-level, how many signals (M_c) can be identified with arbitrarily small error?

In practice signals of finite duration are fingerprinted and identified. The WOS model, however, considers signals of infinite length. To be able to compare to the WOS model, we also assume that the signal length n – and therefore the fingerprint length – tends to infinity.

For the capacity analysis of PRH we use the results from Section 3.4.1. For a given Signal-to-Noise level SNR [dB] $= 20 \log_{10}(\sigma_X/\sigma_W)$, the average fraction of erroneous bits in a fingerprint is given by the differential PRH-model in Eq. (3.52):

$$\alpha = \frac{1}{\pi} \arctan\left(\sqrt{\frac{\sigma_W^4}{\sigma_X^4} + 2\,\frac{\sigma_W^2}{\sigma_X^2}}\right), \tag{5.31}$$

where

$$\frac{\sigma_W}{\sigma_X} = 10^{-\text{SNR}/20}. \tag{5.32}$$

We use this relation to derive bounds on the capacity for binary fingerprint systems like the PRH system in two different ways. In Section 5.3.1 we model the erroneous fingerprint bits due to signal distortions as a binary symmetric channel. In Section 5.3.2 we draw parallels between fingerprinting and error correcting coding (ECC), and adapt well-known ECC bounds to the fingerprinting framework.

5.3.1 PRH bound based on binary symmetric channel capacity

Consider the following practical implementation of a fingerprinting system. Again the signals x are Gaussian, and the distortion model $f_{Y|X}(x, y)$ is additive Gaussian. Consider a quantizer for both mappings, e.g. a one bit quantizer based on the sign of the input value. Thus $U = Q(X)$ and $V = Q(Y)$ are discrete variables based on continuous inputs.

Since U is a deterministic mapping of X, there is no uncertainty about U given the value X. The same argument holds for Y and V. The implication is that $I(U, V|X, Y) = 0$. Therefore, the achievable rate is bounded by $R_c \leq I(U; V)$. Actually, this turns the (U, V)-channel into a binary symmetrical channel (BSC). Its capacity is determined by the cross-over probability α, which is a function of the SNR in the (X, Y)-channel. Hence,

$$\begin{aligned} R_{bsc} &\leq I(U; V) \\ &= 1 - H(\alpha). \end{aligned} \tag{5.33}$$

where $H(\alpha)$ denotes the binary entropy function:

$$H(\alpha) = -\alpha \log_2(\alpha) - (1 - \alpha) \log_2(1 - \alpha) \tag{5.34}$$

For an m-bit codeword transmitted over a BSC-channel, the bound on the number of messages that can be distinguished is $M_{bsc} = 2^{mR_{bsc}}$. A PRH fingerprint generated at bitrate R_x extracts R_x bits for every sample of the input signal x. The resulting fingerprint has size $m = nR_x$ bits. Therefore, the overall bound on the rate is

$$\begin{aligned} R_{prh} &= R_x\, R_{bsc} \\ &\leq R_x \left(1 + \alpha \log_2(\alpha) + (1 - \alpha) \log_2(1 - \alpha)\right) \end{aligned} \tag{5.35}$$

Since the mappings ϕ_x and ϕ_y are now deterministic, there is no distinction between the inner and outer bound; they are the same. This is due to the fact that $I(U; V|X, Y) = 0$ in Eq. 5.6

Figure 5.8 compares the inner- and outer bound for the WOS model with the binary symmetrical channel model of the PRH for bitrate $R_x = R_y = 1$. Due to its structure involving the quantization, the PRH BSC model corresponds to the WOS inner bound. For high SNR, however, the BSC bound converges to the WOS outer bound. This

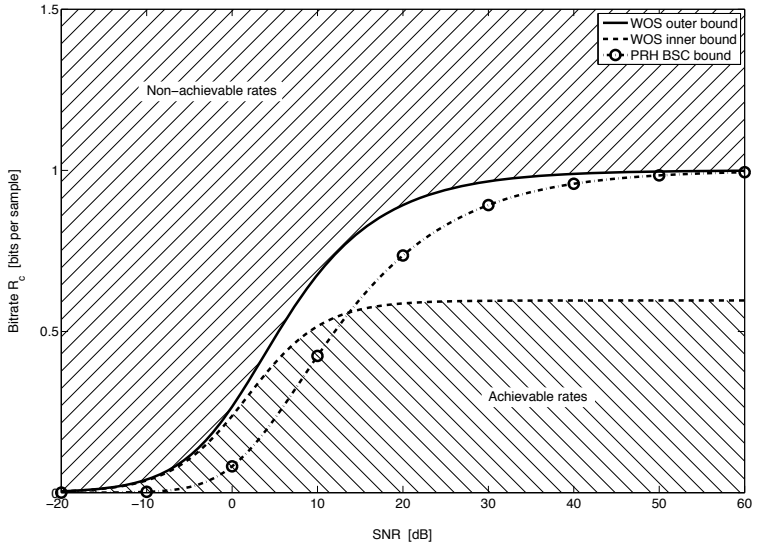


Figure 5.8: WOS inner and outer bound for Gaussian variables as a function of the SNR, at a bitrate $R_x = R_y = 1$, in case of additive white Gaussian noise. Also shown here is the PRH BSC bound for the same (quantizer) bitrate.

indeed shows that certain rates higher than the WOS inner bound, but lower than the WOS outer bound, are achievable.

The rates inside the gap between the WOS outer and inner bound might be achievable, but from the WOS-model this is not certain. The PRH bound is an indication that tighter bounds are possible. For SNR values higher than approximately 15 dB, the PRH bound lies in between the WOS inner and outer bound. Hence, the PRH (inner) bound indicates that certain rates are achievable; rates for which the WOS-model does not make a clear statement. This is a useful observation, since it provides a hint on how to close the gap between the WOS-bounds. However, there might be a mismatch between the two models which causes the difference between the bounds. For SNR values lower than approximately 15 dB, rates that are achievable according to the WOS inner bound, are not achievable according to the PRH (upper) bound. This is not a conflict between the models. The WOS model indicates that certain rates are achievable; the inner bound itself is achieved when U and V are Gaussian. In the PRH bound, however, the signals U and V are binary. Thus, the comparison between the models yield that the rates that may be achieved using a binary representation are lower than the rates when using a Gaussian representation. Indeed, in their paper Westover and O'Sullivan assume that the rate R_c is maximized when the representations U and V are Gaussian.

For a given value of R_{bsc}, the rate R_{prh} scales linearly with the bitrate employed in the mapping R_x. A drawback of this analysis is that there is no 'natural' bound when the bitrate R_x increases. This is an important difference with the WOS-model. In the WOS-model, there is a natural upper limit for R_c, R_{max}, dependent on the

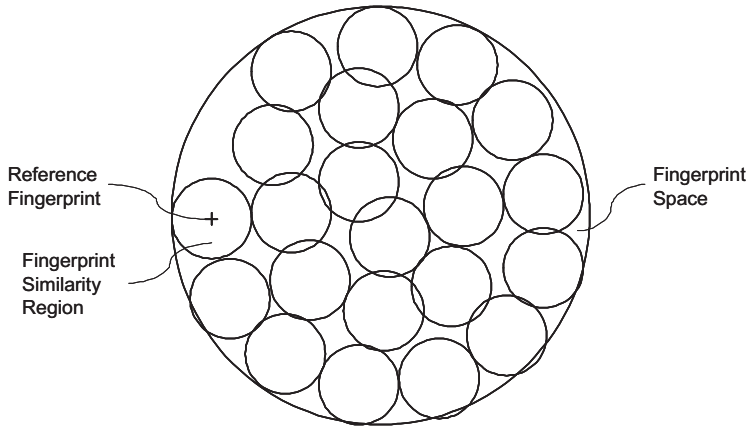


Figure 5.9: Sphere packing example.

distortion in the $X - Y$ channel, and the employed bit rates in the coding branches, i.e. $R_x = R_y$. For increasing R_x the recognition rate R_c in the WOS model converges to the channel capacity in the (X, Y)-channel; R_c is bounded by $-\frac{1}{2}\log(1 - \rho_{xy}^2)$ for the Gaussian channel.

5.3.2 PRH bound based of error correcting codes

When the signal to be identified is distorted by additive Gaussian noise, a fraction of α bits is flipped in the binary PRH fingerprint. Here, α depends on the SNR. For an m-bit fingerprint on average αm bits are flipped. The $m = nR_x$-bit fingerprint is the result of fingerprinting the length-n signal x at bitrate R_x. Of course, the fraction of bits that is actually flipped varies around this average value α, but since the fingerprint length m tends to infinity along with the signal length n, the amount of variation relative to αm can be made arbitrarily small.

The total number of possible m-bit fingerprints that differ up to αm bits from a reference fingerprint is given by the following binomial summation:

$$S(m, \alpha) = \binom{m}{0} + \binom{m}{1} + \ldots + \binom{m}{\alpha m} \tag{5.36}$$

Figure 5.9 illustrates a thought model which leads to an upper bound on the number of signals that can be identified using an m-bit binary fingerprint. Such a fingerprint can represent 2^m different signals in total. This is called the fingerprint space illustrated by the big circle. One of the reference fingerprints is marked '+'. The small circle around it marks all $S(m, \alpha)$ fingerprints with a hamming distance smaller than, or equal to, αm.

To identify each reference signal without errors, the small circles may not overlap for this distortion level α. Therefore, an upper bound is obtained by dividing the

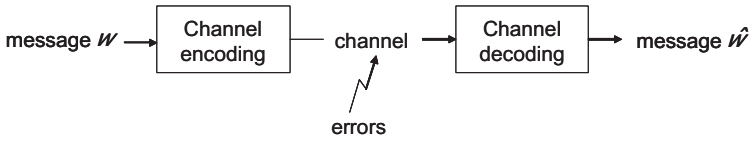


Figure 5.10: Channel coding is used to increase the robustness of messages sent over an error-prone channel, such that the received message $\hat{w}$ is equal to the sent message w.

volume of the big circle by the volume of a small circle:

$$M_c \leq \frac{2^m}{S(m,\alpha)} \tag{5.37}$$

This bound is called the sphere packing bound. Due to the dependence of the summation $S(m,\alpha)$ on m it is difficult to express the bound in the form of a rate independent of m, which is needed for the comparison to the WOS bounds. Therefore, we turn to asymptotic bounds found in literature on error correcting codes (ECCs) [40].

ECCs are channel codes. Their aim is to protect messages transmitted over an error-prone channel, as shown in Figure 5.10. This is done by choosing the codewords such that they have a minimum distance with respect to each other. As a result the received codewords are not misinterpreted when a limited number of errors are introduced in the signal transmitted. Figure 5.9 can also be interpreted in an ECC context. Then the '+' marks a channel codeword. Each such codeword represents a message. The small circle contains all received words that contain a limited number of errors, while the large circle contains the set of all possible received signals. Again, the small circles may not overlap. The ratio of the volumes - the sphere packing bound - provides an upper bound on the number of messages that an be distinguished by the receiver.

To be robust against up to $\frac{d}{2}$ bit errors, two arbitrary binary code words of length m can be distinguished at the receiving end when they differ at least d bits at the sender. For the sphere packing bound, this implies that $d = \lceil 2\alpha m \rceil$. In the context of the WOS model we consider asymptotic bounds when the fingerprint length goes to infinity. For large m, the fraction $\frac{d}{m}$ tends to a constant δ, where $0 \leq \delta \leq 1$.

Several asymptotic bounds are presented in reference [40]. These are bounds on the rate $R_c = \frac{1}{m}\log_2(M_c)$ - and thus on the number of messages M_c - as a function of δ. The distance δ scales the relative volume of the small circles in Figure 5.9. Increasing values of δ result in increasing volume of the small circles and thus in a decreasing number of messages with can be discriminated with arbitrarily small error.

The following four expressions are well-known outer bounds for binary codes [40]:

- Sphere packing bound

$$R_{SP}(\delta) = 1 - H(\delta/2), \qquad 0 \leq \delta \leq 1; \tag{5.38}$$

- Singleton bound

$$R_S(\delta) = 1 - \delta, \qquad 0 \leq \delta \leq 1; \tag{5.39}$$

- Bassalygo-Elias bound

$$R_{BE}(\delta) = \begin{cases} 1 - H\left(\frac{1}{2} - \frac{1}{2}\sqrt{1-2\delta}\right), & 0 \leq \delta \leq \frac{1}{2}, \\ 0, & \frac{1}{2} < \delta \leq 1; \end{cases} \tag{5.40}$$

- McEliece-Rodemich-Rumsey-Welch bound

$$\begin{aligned} &R_{MRRW}(\delta) \\ &= \begin{cases} \min\limits_{0 \leq u \leq 1-2\delta} 1 + g(u^2) - g(u^2 + 2\delta u + 2\delta), & 0 \leq \delta \leq \frac{1}{2}; \\ 0, & \frac{1}{2} < \delta \leq 1; \end{cases} \end{aligned} \tag{5.41}$$

where

$$g(x) = H\left(\frac{1 - \sqrt{1-x}}{2}\right). \tag{5.42}$$

These outer bounds say that for a given δ the rate of a code cannot exceed the bound $R(\delta)$. It is unknown whether the rates indicated by the bounds are actually achievable.

Inner bounds exist as well for ECCs [40]. The Gilbert-Varshamov bound expresses which rates are in theory achievable for a given δ.

$$R_{GV}(\delta) = \begin{cases} 1 - H(\delta) & 0 \leq \delta \leq \frac{1}{2} \\ 0 & \frac{1}{2} < \delta \leq 1 \end{cases} \tag{5.43}$$

The Gilbert-Varshamov does not tell that rates exceeding the bound are not achievable, but that for a given distance δ the rates lower than the bound are achievable.

Figure 5.11 illustrates these rate-bounds as a function of δ. It can be seen that in this figure indeed the sphere packing bound is a relative loose bound. The McEliece-Rodemich-Rumsey-Welch bound is the tightest outer bound shown here.

We derive bounds for PRH based on the ECC bounds, and compare these with the WOS model in the (SNR, R_c) view presented in Figure 5.6. The ECC bounds can be related to fingerprints and the WOS bounds as follows. The ECC bounds R_{ecc} are expressed as function of δ, where R_{ecc} corresponds to a number of messages equal to $M_{ecc} = 2^{mR_{ecc}}$. A PRH fingerprint generated at bitrate R_x extracts R_x bits for every sample of the input signal x. The resulting fingerprint has size $m = nR_x$ bits. In this way, a bound for the PRH corresponds to $M_{prh} = 2^{n(R_x R_{ecc})} = 2^{nR_{prh}}$. We use the PRH differential model in Eq. (5.31) to relate the SNR level to α, and through the relation $\delta = 2\alpha$ relate the SNR to δ. This relates the SNR of the Gaussian signals y to the asymptotic ECC bounds. In this way, we obtain (SNR, R_{prh})-pairs.

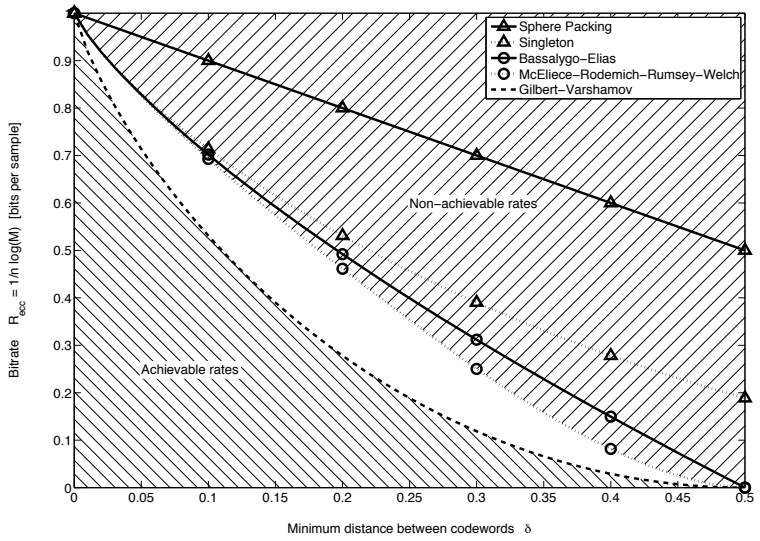


Figure 5.11: Bounds on the achievable coding rates of ECCs [40].

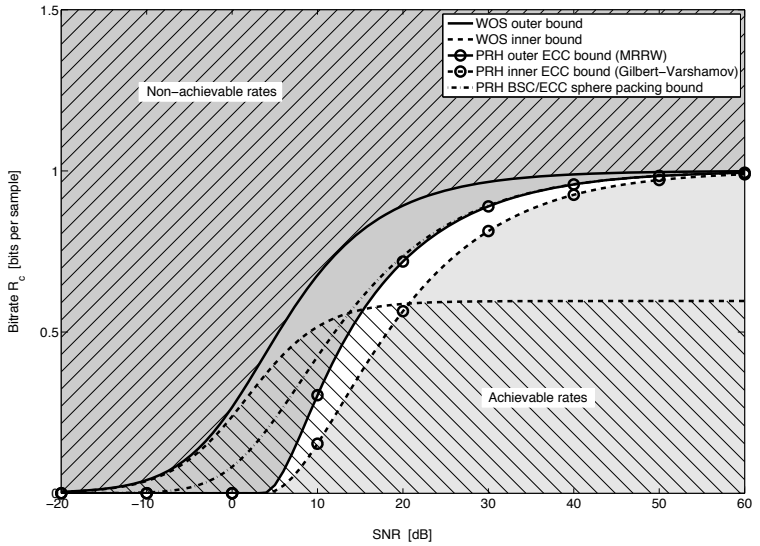


Figure 5.12: Gaussian WOS achievability bounds, combined with the binary PRH achievability bounds, at a bitrate $R_x = R_y = 1$. The rate regions the are achievable and non-achievable according to the WOS-model are indicated by the striped regions; similar regions according to the PRH ECC bounds are indicated by shading. For readability of the figure, the regions according to the PRH BSC bound are not marked since there is no gap between the inner and outer bound.

For the PRH bounds constructed as described above, we only consider the Gilbert-Varshamov inner bound, and the tightest outer bound (McEliece-Rodemich-Rumsey-Welch bound). The sphere packing bound coincides with the BSC capacity bound. Figure 5.12 illustrates both the WOS and PRH achievability bounds, as function of SNR for a given rate. The rates used are $R_{fp} = R_x = R_y = 1$ bits per sample. This corresponds to generating 32-bit sub-fingerprints every 5.8 msec from a signal which is sampled at 5,5025 Hz. The PRH algorithm downsamples to this frequency and extracts fingerprints at half this bitrate.

The ECC bounds are so tight that they reach zero capacity for $\delta = \frac{1}{2}$. This corresponds to $\alpha = \frac{1}{4}$. Then the argument in Eq. (5.31) is equal to:

$$\sqrt{\left(\frac{\sigma_W^2}{\sigma_X^2} + 1\right)^2 - 1} = \tan\frac{1}{4}\pi \tag{5.44}$$

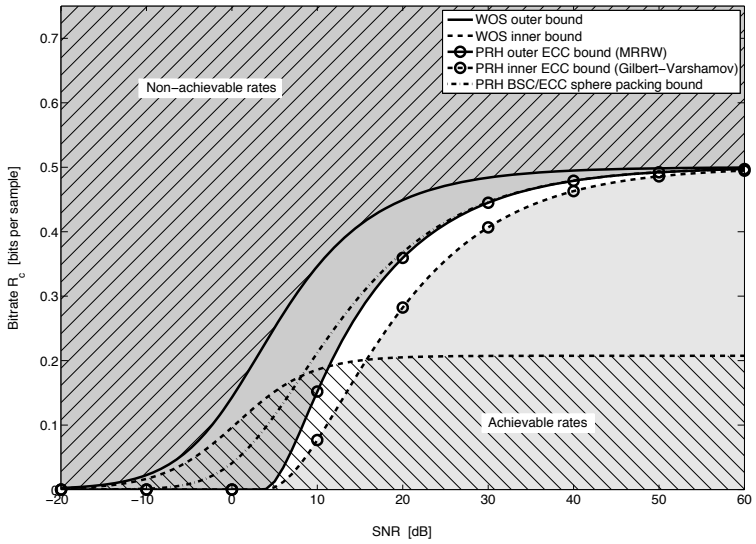
Further derivation yields that the SNR value at which the bounds reach zero capacity is given by:

$$\begin{aligned} \text{SNR}_{\text{zero cap}} &= 10\log\left(\frac{1}{\sqrt{2}-1}\right) \\ &= 3.83 \text{ dB} \end{aligned} \tag{5.45}$$

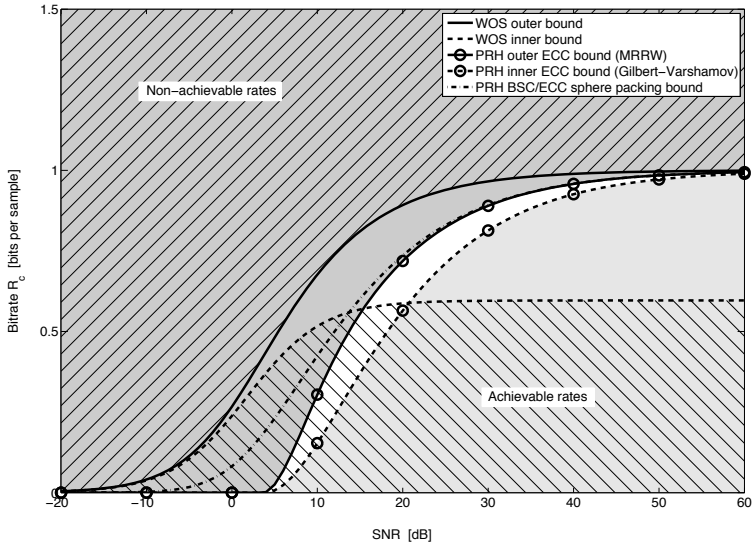
The bounds used here are the MRRW-bound and the Gilbert-Varshamov bound. These are derived by Ericson [40] for constant weight codes and may not fully apply to the PRH fingerprint. The important curve is the Gilbert-Varshamov bound since rates lower than this bound are actually achievable.

Figures 5.13 and 5.14 again show the SNR vs. Rate curves, but now for various fingerprint bit rates: $R_{fp} = 0.5$ (Figure 5.13(a)), $R_{fp} = 1$ (Figure 5.13(b)), $R_{fp} = 2$ (Figure 5.14(a)), and $R_{fp} = 6$ (Figure 5.14(b)). From Figure 5.14(b) it can be seen that for high bitrates the PRH models intersect and exceed the WOS outer bound. Here, the PRH model contradicts the WOS model: rates that are non-achievable according to the WOS-model are achievable in the PRH models. This contradiction occurs for high fingerprint bitrates. However, the rates in the WOS model are assumed to be maximized by the Gaussian representation of U and V; other signal representations (like the binary PRH fingerprints) should yield lower or equal rates when compared to the Gaussian case. This might be due to the unconstrained scaling of R_c with the fingerprint bitrate $R_{fp} = R_x$.

Figures 5.15 and 5.16 show the WOS-bounds as a function of fingerprint bit rate for a given SNR level, but now also indicating the PRH bounds. As mentioned at the end of Section 5.3.1 the PRH bounds are proportional to the bit rate R_{fp}. While the WOS-bounds are naturally limited by the capacity of the (X, Y) channel $-\log_2(1 - \rho_{xy}^2)$, the PRH models are unbound. For high SNR (Figure 5.16(b)), both the PRH bounds and the WOS outer bound are proportional to R_x.

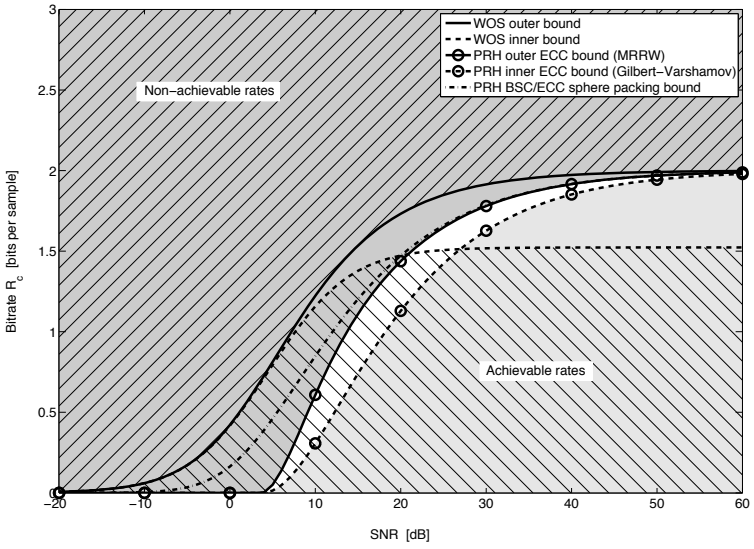


(a)

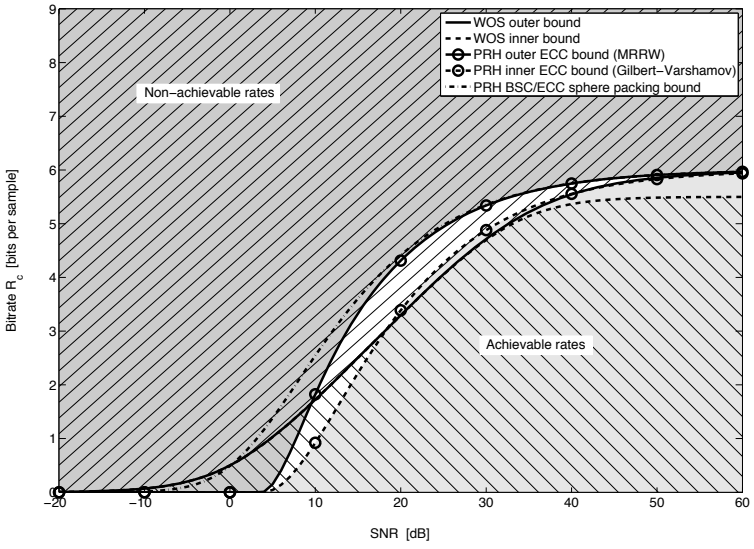


(b)

Figure 5.13: Gaussian WOS achievability bounds, combined with the binary PRH achievability bounds, at bitrates $R_{fp} =$ (a) 0.50, (b) 1 bit per sample. The marking of the achievable rate regions follows the conventions outlined for Figure 5.12.

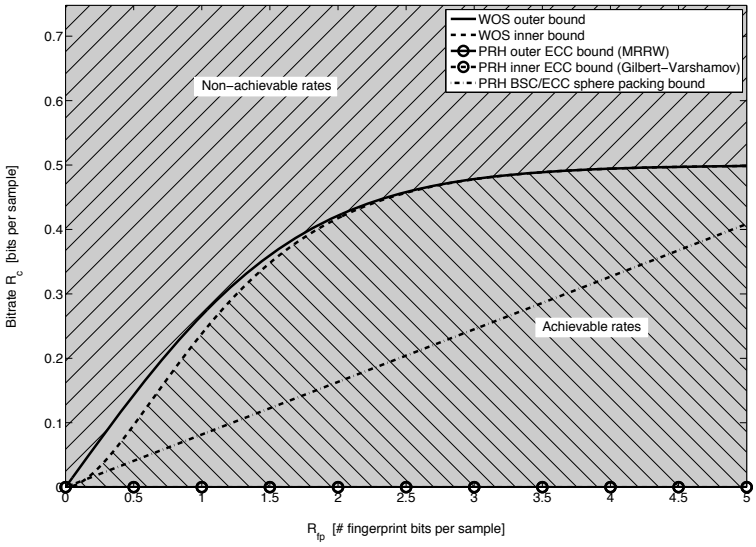


(a)

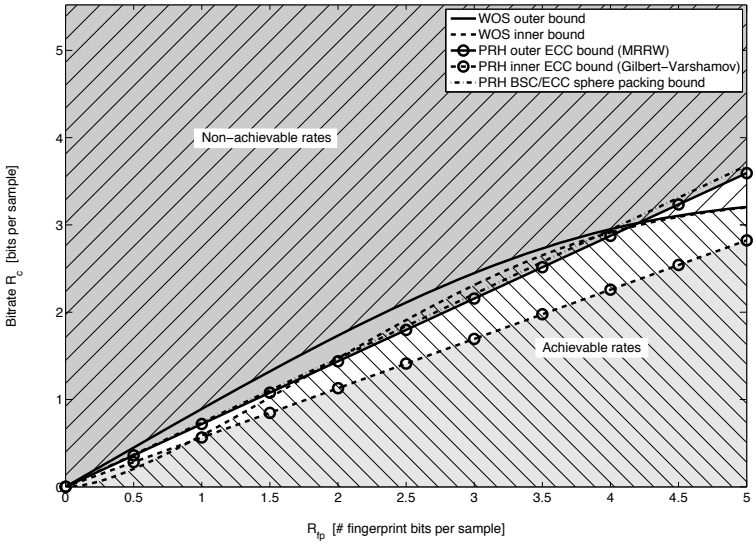


(b)

Figure 5.14: Gaussian WOS achievability bounds, combined with the binary PRH achievability bounds, at bitrates $R_{fp} =$ (a) 2, (b) 6 bit per sample. The marking of the achievable rate regions follows the conventions outlined for Figure 5.12.

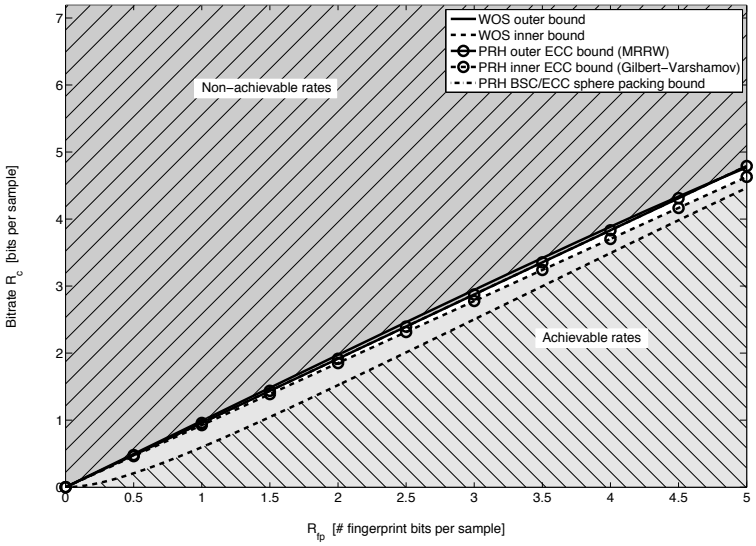


(a)

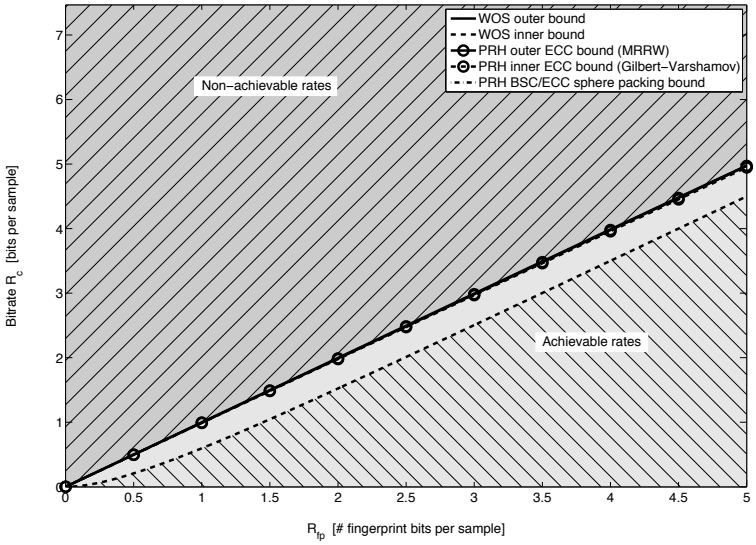


(b)

Figure 5.15: Gaussian WOS achievability bounds, combined with the binary PRH achievability bounds, at SNR = (a) 0 dB, (b) 20 dB. The marking of the achievable rate regions follows the conventions outlined for Figure 5.12.



(a)



(b)

Figure 5.16: Gaussian WOS achievability bounds, combined with the binary PRH achievability bounds, at SNR = (a) 40 dB, (b) 60 dB. The marking of the achievable rate regions follows the conventions outlined for Figure 5.12.

5.3.3 Conclusions

In this section we have derived two types of bounds for binary fingerprints in the context of the WOS model. Although there are some potential mismatches between the PRH bounds and the WOS model, e.g. to the conversion from continuous Gaussian variables to discrete binary values, the bounds fit in well. For large rates $R_x = R_y$, however, the PRH bounds exceed the WOS-bounds. In the WOS-model, for large rates R_x and/or R_y the distortion in the signal path $X - Y$ is the dominating limitation in the calculation of the rate R_c. One cannot distinguish more signals based on the compressed representations U and V than on the original signal representations X and Y. This limiting factor is not taken into account in the PRH bounds.

From literature it is not known how to close the gap between the inner and outer bounds in the WOS model, i.e. whether the inner bound is too tight or the outer bound is too loose. In the (SNR, R_c) plane the PRH bounds exceed the inner bound in the WOS model, and clip to the outer bound. Therefore, assuming the model assumptions between the PRH bounds and the WOS model align, the PRH bounds give rise to the notion that the WOS inner bound is too tight; especially the bound based on the Gilbert-Varshamov bound, since it is a lower bound.

5.4 The WOS model from a distortion perspective

In the previous chapters we have observed that the distortion introduced in the signal is reflected in the distance between the corresponding fingerprints. More specifically, a closed form equation was derived for the distance between PRH fingerprints when Gaussian iid signals are distorted by additive white Gaussian noise with a certain SNR. For high SNR values the log fingerprint distance is linearly related to the SNR on a decibel scale. The model was derived for Gaussian iid signals in the presence of additive noise for the PRH algorithm, but the linear relationship is also observed in practice for this and other algorithms when music is distorted by e.g. compression. In this section we explore the differences observed by comparing the practical observations with the setup of the WOS model for Gaussian signals.

Assuming the distortion on X is additive white Gaussian noise, the (positive) correlation coefficient ρ_{xy} between X and Y is related to the variance of the distortion σ_W^2 by Eq. (5.24). In the WOS model for Gaussian signals, the distributions that maximize the rate are also assumed to be Gaussian. Therefore, the variables U and V are assumed Gaussian. In this case, U can be regarded a version of X distorted by additive white Gaussian noise. Similarly, we consider V a distorted version of Y. To distinguish between the different distortions, the distortion in the $X - Y$ channel is denoted by W_{xy}, and its variance by $\sigma^2_{W_{xy}}$. In this slightly adapted notation Eq. (5.24) now reads:

$$\rho_{xy} = \sqrt{\frac{\sigma_X^2}{\sigma_X^2 + \sigma^2_{W_{xy}}}} \tag{5.46}$$

Similarly, the correlation coefficients ρ_{ux} and ρ_{yv} can be related to $\sigma^2_{W_{ux}}$ and $\sigma^2_{W_{yv}}$, respectively.

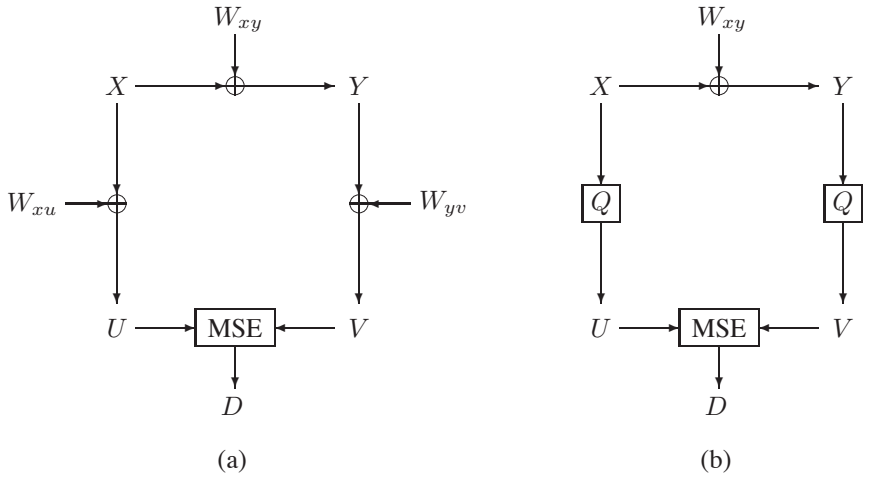


Figure 5.17: Experimental set-up (a) distortions in the WOS model, (b) distortions in the PRH model

Assume the system operates on a point on the inner bound. Figure 5.17(a) shows the set-up including the different variables. Now let us consider how the relative distortion $\frac{\sigma^2_{W_{xy}}}{\sigma^2_X}$ has its effect on the relative distortion of the (Gaussian) fingerprint, $\frac{\sigma^2_{W_{uv}}}{\sigma^2_U}$. For the inner bound correlation between U and V is expressed through

$$\rho^2_{uv,in} = \gamma^2 = \rho^2_{xy}\rho^2_{ux}\rho^2_{yv}. \tag{5.47}$$

Similar to Eq. (5.24) we can relate $\rho_{uv,in}$ to the distortion in the fingerprint channel (u, v):

$$\rho_{uv,in} = \sqrt{\frac{\sigma^2_U}{\sigma^2_U + \sigma^2_{W_{uv}}}} \tag{5.48}$$

Therefore, we can derive:

$$\frac{\sigma^2_{W_{uv}}}{\sigma^2_U} = \frac{\sigma^2_{W_{xy}}}{\sigma^2_X} \underbrace{\left(\frac{1}{\rho^2_{ux}\rho^2_{yv}}\right)}_{\zeta(R_x)} + \underbrace{\left(\frac{1}{\rho^2_{ux}\rho^2_{yv}} - 1\right)}_{\xi(R_x)} \tag{5.49}$$

From this equation it is clear that according to the model two effects contribute to the relative distortion in the fingerprint: the relative distortion of the signal scaled by ζ, and a constant term ξ. Both terms are dependent on the fingerprint bit rate. Due to the

constant term, there still is a minimum distance between the fingerprints, also in case of zero distortion ($\sigma^2_{W_{xy}} = 0$). This term $\xi(R_x)$ is there because the two mappings ϕ_x and ϕ_y are modeled as independent. In practice, however, in most symmetric fingerprinting systems the mappings are identical. Therefore, in practice when the signal distortion approaches zero, the distance between the fingerprints also approaches zero.

To illustrate the practical behavior of the PRH fingerprint for Gaussian inputs at various bitrates, consider the following experiment. The set-up is shown in Figure 5.17(b). As before the Gaussian signal X is distorted by additive white Gaussian noise resulting in a certain SNR-level on signal Y. The fingerprints of both X and Y – U and V, respectively – are derived by quantizing the continuous signals with the *same* non-uniform quantizer, i.e. $U = Q[X]$ and $V = Q[Y]$. The quantizer is designed such that each quantizer bin is selected with equal probability when quantizing x. Due to the distortion W_{xy} a different quantizer bin might be selected. The representation levels are the quantizer bin centroids. The distortion D is computed using the MSE on the quantizer outputs, i.e. $D = \frac{1}{N}\sum_{i=1}^{N}(Q[x(i]) - Q[y(i)])^2$.

Figure 5.18 illustrates the MSE between fingerprints as a function of the SNR; for the WOS model in Figure 5.18(a) and for the quantization in Figure 5.18(b). The curves are based directly on the models, not on numerical simulation. In both figures, the dashes line illustrates the case where no quantization or coding has been applied on X and Y, i.e. $U = X$ and $V = Y$.

For the quantizer experiment in Figure 5.18(b), the MSE between the fingerprint signals U and V typically shows three regions as function of the SNR:

1. For low SNR the noise component is dominant. The fingerprinting curve is saturated (constant) as function of SNR, due to the quantization operation. For each original fingerprint sample in U its distorted counterpart in V is a random selected quantizer output.

2. For medium SNR the MSE fingerprint distance is proportional to $\frac{\sigma^2_{W_{xy}}}{\sigma^2_X}$.

3. For high SNR the MSE fingerprint distance is proportional to $\frac{\sigma_{W_{xy}}}{\sigma_X}$.

For the WOS model, however, Figure 5.18(b) shows the following distortion pattern:

1. For low SNR the fingerprinting distance is proportional to $\frac{\sigma^2_{W_{xy}}}{\sigma^2_X}$, corresponding to the term $\zeta(R_x)$.

2. For high SNR the MSE fingerprint distance is constant, corresponding to the term $\xi(R_x)$.

The figures show a completely different behavior in the two models. For high SNR, the quantized fingerprints are nearly identical, while the distance between the WOS fingerprints to saturates at $\xi(R_x)$. For low SNR, distance between the fingerprints saturates at a level dependent on the quantizer representation levels; the distances between WOS fingerprints increases along with the noise level.

We conclude that the WOS model does not reflect the distortion characteristics seen in practice on fingerprints, e.g. in Chapter 4. In the recognition stage the WOS model contains a classifier g and does not make any further assumptions on the characteristics of the classifier. In our experiments, we compare the fingerprints using MSE and BER distance measures. The WOS model is designed with a different purpose: to analyze the capacity. In practice, many systems assume that related fingerprints are close in the fingerprint using a distance measure like MSE or BER (nearest neighbor classification, with a threshold on the maximum distance between the fingerprints). Moreover, in practice the actual procedures for fingerprint extraction in the enrollment and identification phase are typically related or even the same. To better match the WOS model to these practical fingerprinting characteristics and to allow for distortion analysis within the information theoretical framework, we recommend the WOS model to be reformulated to specifically contain the coupling between the fingerprint extraction stages and the distance measures in the classification process.

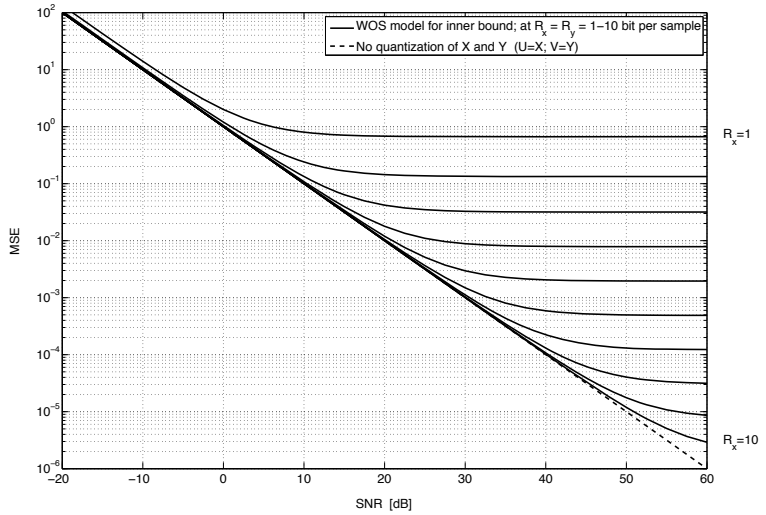
5.5 Discussion

In this chapter we have compared capacity bounds derived for PRH fingerprints to the bounds in the WOS model. As concluded in Section 5.3.3, there are some conflicts between the bounds from the different models, but in general the PRH bounds fit in well with the WOS model bounds. Further, in this chapter we have compared one operating point in the WOS configuration (system operating on the inner bound for Gaussian input signals) to a practical implementation using simple quantizers. As concluded in Section 5.4, the WOS model from a distortion perspective shows different behavior than observed in practice and in these experiments.

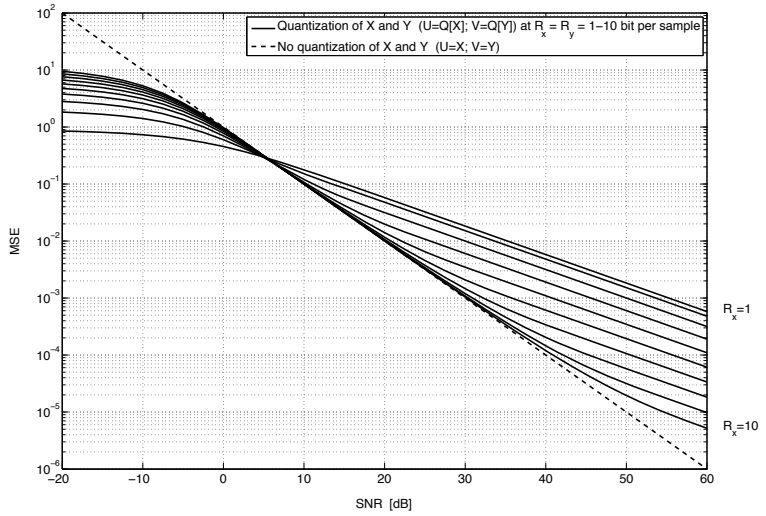
The following observations can be made when comparing the WOS model to practical implementations of fingerprinting systems:

1. *In practice, the mappings ϕ_x and ϕ_y are dependent*
   The mappings ϕ_x and ϕ_y are considered to be independent. In practice, often the same fingerprinting algorithm is used in both the training and identification phase. Then, the mappings are dependent. One of the practical consequences is that if no distortion is applied to the signals the fingerprints in both phases are equal. In the WOS model, however, even when the same rates are used, i.e. $R_x = R_y$, the fingerprints are not identical. Of course, in practice this can also happen when the signals in the training and identification phase are not exactly aligned when the signal is framed and the features are extracted. This temporal misalignment effect was modeled for the PRH in Section 3.4.2 for Gaussian i.i.d. signals.

2. *In practice, the rate R_x is smaller than, or equal to, the rate R_y*
   In our discussions we have limited ourselves fingerprinting systems that compute the same type of fingerprint in the training and identification phase (coined 'symmetric systems' in Section 2.5). This corresponds to $R_x = R_y$ in the WOS model. However, often the fingerprint in the training phase is based on the same feature representation as used in the identification phase, but is compressed



(a)



(b)

Figure 5.18: Comparison of models from a distortion perspective; SNR vs. MSE between fingerprints at various fingerprint bit rates for (a) WOS model; (b) quantization.

even further by generating a parametric model based on the feature representation (coined 'asymmetric systems' in Section 2.5). Examples of such parametric models are Gaussian Mixture Models (GMMs), Hidden Markov Models or PDFs. This corresponds to $R_x < R_y$, although it is difficult to say something about the relative size of R_x w.r.t. R_y, that is about the ratio $\frac{R_x}{R_y}$.

3. *Practical systems use finite length signals*
   The WOS model derives asymptotic models for i.i.d. signals. It assumes that n can be made very large to bring down the error probability. However, many practical applications rely on the fact that an entire song can be identified based on a small fragment of a few seconds.

4. *In practice, the structure of the fingerprints is highly correlated*
   Practical fingerprinting algorithms like the PRH often are based on features with strong correlation. This is done to make the fingerprint robust to desynchronisation between the fingerprint stored in the database and the query fingerprint. However, from a capacity point-of-view this is not optimal, and makes the bound not achievable for this type of fingerprint.

Given these observations and the results in this chapter, we conclude that there is a gap between the analysis provided by the WOS model, and behavior observed of practical algorithms. We therefore recommend to reformulate the WOS model to take into account the dependency between the mappings observed in practice.

Furthermore, it is difficult to compare practical systems to these theoretical bounds. We can evaluate practical algorithms which compute fingerprints on audio signals of finite length using Receiver Operating Characteristic (ROC) curves. We might also experiment with the signal length used in the experiments and see how the performance in the ROC curves changes as a function of the signal length. However, from such experiments is impossible to see what is the upper limit on the number of signals that can be identified. On the other hand, in the WOS model it is difficult to compare practical implementations to the theoretical bounds. Therefore, we recommend to develop 'something in the middle' which allows to evaluate how large the capacity gap is between practical systems and the bounds from the WOS or PRH models.

Chapter 6

Results and Recommendations

6.1 Results

In this thesis we have developed several models for audio fingerprints and fingerprinting systems, with an emphasis on the fingerprint extraction and the properties of the fingerprint. Our modeling approach was outlined in Section 1.2 'Scope and contributions' and 2.7 'Objectives'. We now outline the main results and draw conclusions for each of the three models.

First, we have developed a model relating the statistical structure of the PRH fingerprint bits to a number of system parameters. The model applies to input signals which are Gaussian iid sources. The system parameters (e.g. relative frame overlap, window type) determine the correlation matrix underlying the fingerprint structure. A PRH fingerprint can thus be seen as the realization of a stochastic process, which may be approximated by a Markov chain. Experimental verification shows that the model captures the fingerprint structure well. The model can be used to optimize the fingerprint structure, as shown by Balado *et al.* who base their optimization on a reformulation of the model described in this thesis.

Second, we have developed a model that relates the SNR for additive white Gaussian noise to the probability of an erroneous bit. The model is extended to include temporal desynchronization. The model is verified experimentally, and for Gaussian iid sources fits well to the simulation results. We expect that both models of the PRH structure and PRH fingerprint distortion can be extended to other fingerprint algorithms - with slight modifications, of course -, especially those extracting in binary fingerprints from the spectrogram.

The models have been successfully extended to synthetic and real audio signals. Given some basic knowledge on the spectrogram, model predicts the BER for a given SNR. However, the variance in the predicted BER is relatively large. Therefore, for a measured BER only a coarse indication of SNR can be provided. The variations are the result of the fact that in real signals like music some feature realizations are more

stable than others; this is directly related to the spectral characteristics of the audio signal. It is shown in this thesis that a more accurate relation with SNR can be made if these spectral characteristics are taken into account to some extend in computing the fingerprint distance.

The model for PRH predicts that for Gaussian iid input signals a 20dB increase in SNR results in a drop of the fingerprint distance with a factor 10. The model is extended to other input signal models. The predicted behavior is not only typical for PRH, but can also be observed in practice for two other audio fingerprinting. We thus conclude that this behavior is typical for a wider class of audio fingerprinting algorithms, on a wider class of input signals.

Third, the WOS model provides a framework to answer the following question: 'how many signals can be reliably identified by a fingerprinting system, under certain conditions'. The conditions relate to characteristics of the fingerprint (size of the fingerprint, and representation of the fingerprint), and characteristics of the environment in which the system operates (what kind of signals need to be identified, how much distortion is allowed). The WOS model applies to iid signals, and has closed form bounds for Gaussian signals and distortions.

Within the WOS framework, we have derived bounds for the binary PRH fingerprints. To do so, we use our second model that relates the SNR to the probability of an erroneous fingerprint bit. We have applied two parallel strategies: consider the effect of distortion on fingerprints as a binary symmetric channel, and apply bounds from error correcting coding.

The model fits well into the framework, but there is some friction between our model and the results from the WOS-model. The WOS model is more general in the sense that it assumes that the Gaussian representation of the fingerprints maximizes the number of (Gaussian) signals that can be identified. Application of the PRH models is more specific in the sense that it applies to a specific representation of fingerprints (binary) that behave in a particular way to additive Gaussian distortion in the signal space. Therefore, we expect that the number of signals that can be distinguished at a certain operating point (fingerprint bitrate and SNR) for the PRH model is smaller or equal to the Gaussian WOS model. However, for high bit rates the PRH model indicates that certain identification rates are achievable that according to the WOS model are not achievable. We believe the friction originates from the fact that in our formulation using the PRH model the number of signals that can be distinguished based on the fingerprint, is not limited by the number of the signals that can be distinguished in the signal space itself. Further, the PRH bounds indicate that the gap between the inner and outer achievability bound might be closed by extending the inner bound.

Using a simple experiment we have shown that the effects of additive distortion in quantization based fingerprints differs significantly from the effects seen in the WOS-model for Gaussian iid input signals.

In this thesis we have presented several models for audio fingerprints. The impact beyond the models themselves is that it shows that it is feasible to derive models for fingerprint extraction, which can be input to optimization procedures. One of the

challenges for future work is to translate the modeling results into practical design rules, i.e. a 'design recipe' which provides guides for a step-by-step fingerprint design process including trade-offs.

Further, this thesis introduces the notion that a signal distortion parameter can be related to the actual performance, not just in a qualitative fashion, but also quantatively. This allows both to relate observed fingerprint distances to signal distortion or quality, but also to perform sensitivity analysis: what if the distortion is slightly increased, or decreased, what is the effect on the fingerprint distance, and thus on the identification performance.

6.2 Recommendations

Based on our results we recommend the following for future research and development.

Develop a model framework for joint optimization of the robustness and collision probability
Our current PRH models capture the structure of the PRH fingerprint, and the average robustness to additive noise and temporal desynchronization. We recommend to model the actual conditional distributions discussion in Section 2.4.5 and illustrated in Figure 2.4, i.e. the conditional distribution of the distance between the fingerprints of similar content, and of the distance between the fingerprints of dissimilar content. In this way the optimal threshold can be determined for a given distortion model $f_{Y|X}(x, y)$, fingerprint size $N \times M$, and fingerprint parameters L and ΔL.

In a series of publications, Balado *et al.* [71, 72, 73, 15, 16, 51, 52] have modeled and optimized several individual aspects of the PRH fingerprint, such as the method for converting the real-valued features into a binary representation [71], the optimal window for the smallest desynchronization effect [15], and the collision probability [52]. The first models [72, 73, 15] follow our approach to a great extend, but the compact formulation in quadratic form allows for extensions to Gaussian signals which are more complex than iid, and for optimizations such as the window size needed for minimal desynchronization. Their model for the collision probability is not limited to binary fingerprints.

Both our models and the models and optimizations by Balado *et al.* consider a limited number of aspects in isolation, e.g. the fingerprint structure or the probability of an erroneous bit. We recommend to consider the optimization problem over the entire detection chain, to derive the optimal operating point for a given application scenario. Ideally, such models are not just mathematically correct, but also elegantly formulated such that the provide insight in the trade-offs at hand.

Furthermore, it is interesting to compare such modeling results with the results of data driven optimization approaches (e.g. AdaBoost), such as used by Ke *et al.* The joint insight might lead to improvements in the design of practical algorithms.

Extend the PRH models to other audio fingerprints, other modalities and other distortion types.
The PRH models have been developed for a specific audio fingerprinting system, for two distortion types: additive noise and temporal desynchronization. The PRH models might be extended to other fingerprint representations and fingerprint distance measures, as well as other distortion types such as variations in play-out speed.

In addition, many of the underlying modeling steps are not audio specific. It would therefore be interesting to see to what extend the models also apply to image and video fingerprinting.

Set-up an evaluation framework for audio fingerprinting
Further, we recommend to set up a framework for the evaluation of audio fingerprinting schemes. Currently there are, to our knowledge, two evaluation practices.

First, there are the ad-hoc comparisons found in papers where the authors compare their algorithm with an implementation of other algorithms from literature. Because each author sets up his evaluation in a different way, it is difficult to compare the performance of different algorithms reported in different papers.

Second, the TRECVID evaluation framework organized by the National Institute for Science and Technology (NIST) in 2008 and 2009 contained a Content-Based Copy Detection (CBCD) task, mainly aiming at the evaluation of video fingerprinting algorithms. In TRECVID, the contestants are given a dataset for training (with ground truth) and a dataset for evaluation (without ground truth). The submissions are only evaluated on their actual detection performance, and to a limited extend on the time localizations of the algorithms. Since the contestants train and optimize their own algorithms, the performance is usually indicative of the algorithm itself, and is not degraded by the potentially sub-optimal implementation by a third party.

The TRECVID framework, however, is not capable of comparing sub-aspects within the algorithms, or the processing speed of the different algorithms. For the ad-hoc comparisons of algorithms it is not easy to compare results between papers. For true optimization of algorithms, a more detailed framework overcoming these shortcomings is necessary.

Integrate psycho-acoustics in fingerprinting schemes to allow a better perceptual comparison between audio fragments
Our current approach relates the fingerprint differences to SNR. Although SNR is suitable for our envisioned application scenarios, we foresee two options to alter the current setup to relate the fingerprint differences to more perceptually motivated distortion measures: altering the fingerprinting scheme and altering the fingerprint distance measure. In both cases the masking threshold can be estimated from the spectrum, even on a subband basis.

A theoretical framework for comparing a specific algorithm, in stead of a class of algorithms, to the capacity bounds.
In Section 5.3 we compared two types of bounds derived for binary fingerprints to the WOS-bounds. The bounds apply for the *group* of binary fingerprints for which the probability of an erroneous bit due to additive distortion is given by Eq. (3.52). It would be helpful to develop a framework in which it is possible to see what is the gap between a *specific implementation* and a theoretical upper bound like the WOS-bounds.

Development of a practical fingerprinting framework and algorithm, which can be easily tuned to the application domain (design recipe).
Each application has its own constraints and operating conditions for the fingerprinting system. It is desirable for a fingerprint to be optimized for specific use, starting from a generic structure. Therefore, we recommend to develop a fingerprinting scheme with associated design rules which can be used to tune the fingerprint to the application scenario.

Appendix A

Background for Chapter 3

A.1 Statistical properties of the Fourier transform of white noise

In this section, we summarize the cross-correlation of a real and imaginary parts of the Fourier transform of a white noise sequence $x(i)$, i.e. $x(i) \sim \mathcal{N}(0, \sigma_X^2)$. Section A.1.1 discusses the full length Fourier transform: the length of the transform, L, is equal to the number of samples used. Section A.1.2 considers the Fourier transform of a zero-padded signal. Here, the length of the signal is $\Delta L \geq 0$ samples shorter than the length of the Fourier transform.

A.1.1 Full length Fourier Transform

The Fourier transform of a realization, both of length L, is defined as:

$$\begin{aligned} \hat{x}_R(k) &= \sum_{i=0}^{L-1} x(i) e^{-j\frac{2\pi ik}{L}} \\ &= R_{X_R}(k) - j\, I_{X_R}(k), \qquad\qquad k = 0, \ldots, L-1 \end{aligned}$$

where j is the imaginary unit and:

$$\begin{aligned} R_{X_R}(k) &= \sum_{i=0}^{L-1} x(i) \cos\left(\frac{2\pi ik}{L}\right) \\ I_{X_R}(k) &= \sum_{i=0}^{L-1} x(i) \sin\left(\frac{2\pi ik}{L}\right) \end{aligned}$$

In line with the main text in this thesis, the additional subscript $_R$ denotes the use of a rectangular window. The autocorrelation of the real part is given by:

$$\begin{aligned} E[R_{X_R}(p)\, R_{X_R}(q)] &= \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} E[x(i)x(j)] \cos\left(\frac{2\pi p}{L}\, i\right) \cos\left(\frac{2\pi q}{L}\, j\right) \\ &= \sigma_X^2 \sum_{i=0}^{L-1} \cos\left(\frac{2\pi p}{L}\, i\right) \cos\left(\frac{2\pi q}{L}\, i\right) \\ &= \frac{\sigma_X^2}{2} \sum_{i=0}^{L-1} \cos\left(\frac{2\pi (p-q)}{L}\, i\right) + \frac{\sigma_X^2}{2} \sum_{i=0}^{L-1} \cos\left(\frac{2\pi (p+q)}{L}\, i\right) \end{aligned}$$

Similarly, the correlation between the real and imaginary parts can be shown to be:

$$\begin{aligned} E[R_{X_R}(p)\, I_{X_R}(q)] &= -\frac{\sigma_X^2}{2} \sum_{i=0}^{L-1} \sin\left(\frac{2\pi (p-q)}{L}\, i\right) + \frac{\sigma_X^2}{2} \sum_{i=0}^{L-1} \sin\left(\frac{2\pi (p+q)}{L}\, i\right) \\ E[I_{X_R}(p)\, R_{X_R}(q)] &= \frac{\sigma_X^2}{2} \sum_{i=0}^{L-1} \sin\left(\frac{2\pi (p-q)}{L}\, i\right) + \frac{\sigma_X^2}{2} \sum_{i=0}^{L-1} \sin\left(\frac{2\pi (p+q)}{L}\, i\right) \\ E[I_{X_R}(p)\, I_{X_R}(q)] &= \frac{\sigma_X^2}{2} \sum_{i=0}^{L-1} \cos\left(\frac{2\pi (p-q)}{L}\, i\right) - \frac{\sigma_X^2}{2} \sum_{i=0}^{L-1} \cos\left(\frac{2\pi (p+q)}{L}\, i\right) \end{aligned}$$

In the derivation we used the fact that $E[x(i)x(j)] = \sigma_X^2 \delta(i-j)$ since the noise is white. For $p, q = 0, \ldots, \frac{L}{2} - 1$:

$$\sum_{i=0}^{L-1} \cos\left(\frac{2\pi (p-q)}{L} i\right) = L\, \delta(p-q) \tag{A.1}$$

$$\sum_{i=0}^{L-1} \cos\left(\frac{2\pi (p+q)}{L} i\right) = L\, \delta(p)\delta(q) \tag{A.2}$$

$$\sum_{i=0}^{L-1} \sin\left(\frac{2\pi (p-q)}{L} i\right) = 0 \tag{A.3}$$

$$\sum_{i=0}^{L-1} \sin\left(\frac{2\pi (p+q)}{L} i\right) = 0 \tag{A.4}$$

The zero-outcomes in equations (A.1) and (A.2) are due to the summation over a multiple of the period of the (co)sinusoids in the expressions. The overall results for

$p, q = 0, \ldots, \frac{L}{2} - 1$ are given by:

$$\begin{aligned}
E[R_{X_R}(p)\, R_{X_R}(q)] &= \begin{cases} \frac{1}{2} L\sigma_X^2 \delta(p-q), & p, q \neq 0 \\ L\sigma_X^2 \delta(p-q), & p, q = 0 \end{cases} \\
E[R_{X_R}(p)\, I_{X_R}(q)] &= 0, \qquad \forall\, p,\, q \\
E[I_{X_R}(p)\, R_{X_R}(q)] &= 0, \qquad \forall\, p,\, q \\
E[I_{X_R}(p)\, I_{X_R}(q)] &= \begin{cases} \frac{1}{2} L\sigma_X^2 \delta(p-q), & p, q \neq 0 \\ 0, & p, q = 0 \end{cases}
\end{aligned}$$

A.1.2 Zero-padded Fourier transform

In this section we review the covariance of real and imaginary parts in a zero-padded Fourier transform, i.e.

$$\begin{aligned}
\hat{A}_R(k) &= \sum_{i=0}^{L-\Delta L-1} x(i) e^{-j2\pi \frac{k}{L} i} \\
&= R_{A_R}(k) - j\, I_{A_R}(k), \qquad\qquad k = 0, \ldots, L-1
\end{aligned}$$

Define functions $SS_A(k, \alpha, L - \Delta L)$ and $SC_A(k, \alpha, L - \Delta L)$:

$$\begin{aligned}
SC_A(k, \alpha) &= \sum_{i=0}^{L-\Delta L-1} \cos(2\pi\alpha \frac{k}{L} i) \\
&= \frac{\cos\left(\alpha\pi \frac{L-\Delta L-1}{L}\right) \sin\left(\alpha\pi \frac{L-\Delta L}{L}\right)}{\sin\left(\alpha\pi \frac{1}{L}\right)} \\
SS_A(k, \alpha) &= \sum_{i=0}^{L-\Delta L-1} \sin(2\pi\alpha \frac{k}{L} i) \\
&= \frac{\sin\left(\alpha\pi \frac{L-\Delta L-1}{L}\right) \sin\left(\alpha\pi \frac{L-\Delta L}{L}\right)}{\sin\left(\alpha\pi \frac{1}{L}\right)}
\end{aligned}$$

Now the correlations for the real and imaginary terms, $R_A(k)$ and $I_A(k)$, respectively, can be written as:

$$\begin{aligned}
E[R_A(p)\, R_A(q)] &= \frac{\sigma_X^2}{2} SC_A\left(k, p-q, L-\Delta L\right) + \frac{\sigma_X^2}{2} SC_A\left(k, p+q, L-\Delta L\right) \\
E[R_A(p)\, I_A(q)] &= -\frac{\sigma_X^2}{2} SS_A\left(k, p-q, L-\Delta L\right) + \frac{\sigma_X^2}{2} SS_A\left(k, p+q, L-\Delta L\right) \\
E[I_A(p)\, R_A(q)] &= \frac{\sigma_X^2}{2} SS_A\left(k, p-q, L-\Delta L\right) + \frac{\sigma_X^2}{2} SS_A\left(k, p+q, L-\Delta L\right) \\
E[I_A(p)\, I_A(q)] &= \frac{\sigma_X^2}{2} SC_A\left(k, p-q, L-\Delta L\right) - \frac{\sigma_X^2}{2} SC_A\left(k, p+q, L-\Delta L\right)
\end{aligned}$$

A.2 Expressing the covariance of spectral energy differences in terms of variances with variable frame shift

Theorem 1 (Expressing the covariance of spectral energy differences in terms of variances with variable frame shift)**.**

$$\begin{aligned}
&\mathit{COV}\left[ED(n,m), ED(n+l,m)\right] \\
&= -\mathit{VAR}\left[ED^{l\Delta L}(n,m)\right] \\
&\quad +\frac{1}{2}\mathit{VAR}\left[ED^{(l-1)\Delta L}(n,m)\right] \\
&\quad +\frac{1}{2}\mathit{VAR}\left[ED^{(l+1)\Delta L}(n,m)\right]
\end{aligned} \tag{A.5}$$

Proof. From Eq. (3.22) we know that:

$$\begin{aligned}
&\mathrm{COV}\left[ED(n,m), ED(n+l,m)\right] \\
&= \mathrm{COV}\left[ED^b(n,m), ED^b(n+l,m)\right] \\
&\quad +\mathrm{COV}\left[ED^b(n,m+1), ED^b(n+l,m+1)\right]
\end{aligned} \tag{A.6}$$

Each of the covariance terms can be expressed in terms of the energies in the individual frequency bands:

$$\begin{aligned}
&\mathrm{COV}\left[ED^{b,\Delta L}(n,m), ED^{b,\Delta L}(n+l,m)\right] \\
&= -2\mathrm{COV}\left[E^{b,\Delta L}(n,m), E^{b,\Delta L}(n+l,m)\right] \\
&\quad +\mathrm{COV}\left[E^{b,\Delta L}(n,m), E^{b,\Delta L}(n+l+1,m)\right] \\
&\quad +\mathrm{COV}\left[E^{b,\Delta L}(n,m), E^{b,\Delta L}(n+l-1,m)\right].
\end{aligned} \tag{A.7}$$

On the other hand, the variance of a spectral energy difference involving a single frequency band, but with a frame shift of $l\Delta L$ also involves these covariance terms:

$$\begin{aligned}
&\mathrm{VAR}\left[ED^{b,l\Delta L}(n,m)\right] \\
&= 2\,\mathrm{VAR}\left[E^{b,\Delta L}(n,m)\right] - 2\,\mathrm{COV}\left[E^{b,\Delta L}(n,m), E^{b,\Delta L}(n-l,m)\right]
\end{aligned}$$

Rearranging the terms:

$$\begin{aligned}
&\mathrm{COV}\left[E^{b,\Delta L}(n,m), E^{b,\Delta L}(n-l,m)\right] \\
&= \mathrm{VAR}\left[E^{b,\Delta L}(n,m)\right] - \frac{1}{2}\mathrm{VAR}\left[ED^{b,l\Delta L}(n,m)\right],
\end{aligned} \tag{A.8}$$

and plugging this back into Eq. (A.7) yields:

$$\begin{aligned}
&\mathrm{COV}\left[ED^b(n,m), ED^b(n+l,m)\right] \\
&= -\mathrm{VAR}\left[ED^{b,l\Delta L}(n,m)\right] \\
&\quad +\frac{1}{2}\mathrm{VAR}\left[ED^{b,(l+1)\Delta L}(n,m)\right] \\
&\quad +\frac{1}{2}\mathrm{VAR}\left[ED^{b,(l-1)\Delta L}(n,m)\right]
\end{aligned}$$

In combination with Eq. (A.6) this results in the desired expression. □

A.3 Sample-wise correlation function $C^s_{ED}(l)$ for a symmetric window

In Section 3.3.5 the following theorem was stated without proof.
Theorem 2 (Sample-wise correlation function $C^s_{ED}(l)$ for a symmetric window)**.** *The sample-wise correlation function $C^s_{ED}(l)$ for a symmetric window with spectral representation $\hat{w}(k), k = 0, \ldots, L-1$, is given by:*

$$\begin{aligned} C^s_{ED}(l) \;=\; \frac{8}{L^2}(RR^2_X(l) - (RR_1(l) + RI_2(l))^2 \\ -(RI_1(l) - RR_2(l))^2) \end{aligned} \tag{A.9}$$

where

$$\begin{aligned} RR_X &= C_{R_X}(l) \odot (\hat{w}(l) \odot \hat{w}(l)) \\ RR_1(l) &= C_{R_A}(l) \odot \big(\cos(2\pi\tfrac{\Delta L}{L}\, l)\hat{w}(l) \odot \hat{w}(l)\big) \\ RR_2(l) &= C_{R_A}(l) \odot \big(\sin(2\pi\tfrac{\Delta L}{L}\, l)\hat{w}(l) \odot \hat{w}(l)\big) \\ RI_1(l) &= C_{R_A,I_A}(l) \odot \big(\cos(2\pi\tfrac{\Delta L}{L}\, l)\hat{w}(l) \odot \hat{w}(l)\big) \\ RI_2(l) &= C_{R_A,I_A}(l) \odot \big(\sin(2\pi\tfrac{\Delta L}{L}\, l)\hat{w}(l) \odot \hat{w}(l)\big) \end{aligned}$$

Proof. The sample-wise correlation function C^s_{ED} is given in Eq. (3.30):

$$\begin{aligned} C^s_{ED}(l) \;=\; \frac{4}{L^2}\Big(2E\left[R_X(n,k)\, R_X(n,k+l)\right]^2 \\ -\Big(E[R_A(k)R_D(k+l)]^2 + E[I_A(k)R_D(k+l)]^2 \\ +E[R_A(k)I_D(k+l)]^2 + E[I_A(k)I_D(k+l)]^2\Big)\Big) \end{aligned}$$

In the following we use the short-hand notation $\alpha = 2\pi\frac{\Delta L}{L}$. For the correlation between $R_A(k)$ and $R_D(k)$ we can thus write

$$\begin{aligned} &E[R_A(k)R_D(k+l)] \\ &= E[(\hat{w}(k) \odot R_{A_R}(k))(\hat{w}(k+l) \odot R_{D_R}(k+l)] \\ &= E[(\hat{w}(k) \odot R_{A_R}(k)) \\ &\quad (\hat{w}(k+l) \odot (\cos(\alpha(k+l))R_{A_R}(k+l) + \sin(\alpha(k+l))I_{A_R}(k+l))] \\ &= E[(\hat{w}(k) \odot R_{A_R}(k))(\hat{w}(k+l) \odot \cos(\alpha(k+l))R_{A_R}(k+l))] \\ &\quad +E[(\hat{w}(k) \odot R_{A_R}(k))(\hat{w}(k+l) \odot \sin(\alpha(k+l))I_{A_R}(k+l))] \end{aligned} \tag{A.10}$$

For the first part of Eq. (A.10) can be written as:

$$
\begin{aligned}
&E[(\hat{w}(k) \odot R_{A_R}(k))(\hat{w}(k+l) \odot R_{D_R}(k+l)] \\
&= E\Bigg[\left(\sum_{\kappa} \hat{w}(\kappa) R_A(k-\kappa \mod L) \right) \\
&\qquad \left(\sum_{\lambda} \hat{w}(\lambda) \cos(\alpha(k+l-\lambda \mod L)) R_A(k+l-\lambda \mod L) \right) \Bigg] \\
&= \sum_{\kappa} \sum_{\lambda} \cos(\alpha(k+l-\lambda) \mod L) \\
&\qquad \hat{w}(\kappa)\hat{w}(\lambda) E\left[R_A(k-\kappa \mod L) R_A(k+l-\lambda \mod L) \right] \\
&= \sum_{\kappa} \sum_{\lambda} \left(\cos(\alpha(k+l)) \cos(\alpha\lambda) + \sin(\alpha(k+l)) \sin(\alpha\lambda) \right) \\
&\qquad \hat{w}(\kappa)\hat{w}(\lambda) E\left[R_A(k-\kappa \mod L) R_A(k+l-\lambda \mod L) \right] \\
&= \cos(\alpha(k+l)) \sum_{\kappa} \sum_{\lambda} \cos(\alpha\lambda) \hat{w}(\kappa)\hat{w}(\lambda) C_{R_A}(l+\kappa-\lambda \mod L) \\
&\quad + \sin(\alpha(k+l)) \sum_{\kappa} \sum_{\lambda} \sin(\alpha\lambda) \hat{w}(\kappa)\hat{w}(\lambda) C_{R_A}(l+\kappa-\lambda \mod L) \\
&\overset{a}{=} \cos(\alpha(k+l)) \sum_{\kappa} \sum_{\lambda} \cos(\alpha\lambda) \hat{w}(-\kappa)\hat{w}(\lambda) C_{R_A}(l+\kappa-\lambda \mod L) \\
&\quad + \sin(\alpha(k+l)) \sum_{\kappa} \sum_{\lambda} \sin(\alpha\lambda) \hat{w}(-\kappa)\hat{w}(\lambda) C_{R_A}(l+\kappa-\lambda \mod L) \\
&\overset{b}{=} \cos(\alpha(k+l)) \sum_{m} C_{R_A}(l-m \mod L) \sum_{\lambda} \cos(\alpha\lambda) \hat{w}(\lambda)\hat{w}(m-\lambda \mod L) \\
&\quad + \sin(\alpha(k+l)) \sum_{m} C_{R_A}(l-m \mod L) \sum_{\lambda} \sin(\alpha\lambda) \hat{w}(\lambda)\hat{w}(m-\lambda \mod L) \\
&= \cos(\alpha(k+l)) \sum_{m} C_{R_A}(l-m \mod L) \left(\cos(\alpha m) \hat{w}(m) \odot \hat{w}(m) \right) \\
&\quad + \sin(\alpha(k+l)) \sum_{m} C_{R_A}(l-m \mod L) \left(\sin(\alpha m) \hat{w}(m) \odot \hat{w}(m) \right) \\
&= \cos(\alpha(k+l)) \left(C_{R_A}(l) \odot (\cos(\alpha l) \hat{w}(l) \odot \hat{w}(l)) \right) \\
&\quad + \sin(\alpha(k+l)) \left(C_{R_A}(l) \odot (\sin(\alpha l) \hat{w}(l) \odot \hat{w}(l)) \right)
\end{aligned}
\tag{A.11}
$$

Where we used:

(a) Since the window is symmetric, imaginary part is zero. Therefore, $\hat{w}(k) = \hat{w}(-k)$.

(b) Substitution of $m = l + \kappa$

Through similar manipulations we get:

$$
\begin{aligned}
&E[(\hat{w}(k) \odot R_{A_R}(k))(\hat{w}(k+l) \odot R_{D_R}(k+l)] \\
&\quad = \cos(\alpha(k+l))(RR_1(l) + RI_2(l)) + \sin(\alpha(k+l))(RI_1(l) - RR_2(l)) \\
&E[(\hat{w}(k) \odot R_{A_R}(k))(\hat{w}(k+l) \odot I_{D_R}(k+l)] \\
&\quad = \cos(\alpha(k+l))(IR_1(l) + II_2(l)) + \sin(\alpha(k+l))(II_1(l) - IR_2(l)) \\
&E[(\hat{w}(k) \odot I_{A_R}(k))(\hat{w}(k+l) \odot R_{D_R}(k+l)] \\
&\quad = \cos(\alpha(k+l))(RI_1(l) - RR_2(l)) - \sin(\alpha(k+l))(RR_1(l) + RI_2(l)) \\
&E[(\hat{w}(k) \odot I_{A_R}(k))(\hat{w}(k+l) \odot I_{D_R}(k+l)] \\
&\quad = \cos(\alpha(k+l))(II_1(l) - IR_2(l)) - \sin(\alpha(k+l))(IR_1(l) + II_2(l))
\end{aligned}
$$

where

$$
\begin{aligned}
RR_1(l) &= C_{R_A}(l) \odot (\cos(\alpha l)\hat{w}(l) \odot \hat{w}(l)) \\
RR_2(l) &= C_{R_A}(l) \odot (\sin(\alpha l)\hat{w}(l) \odot \hat{w}(l)) \\
RI_1(l) &= C_{R_A,I_A}(l) \odot (\cos(\alpha l)\hat{w}(l) \odot \hat{w}(l)) \\
RI_2(l) &= C_{R_A,I_A}(l) \odot (\sin(\alpha l)\hat{w}(l) \odot \hat{w}(l)) \\
IR_1(l) &= C_{I_A,R_A}(l) \odot (\cos(\alpha l)\hat{w}(l) \odot \hat{w}(l)) \\
IR_2(l) &= C_{I_A,R_A}(l) \odot (\sin(\alpha l)\hat{w}(l) \odot \hat{w}(l)) \\
II_1(l) &= C_{I_A}(l) \odot (\cos(\alpha l)\hat{w}(l) \odot \hat{w}(l)) \\
II_2(l) &= C_{I_A}(l) \odot (\sin(\alpha l)\hat{w}(l) \odot \hat{w}(l))
\end{aligned}
$$

Since $C_{I_A}(l) = C_{R_A}(l)$ and $C_{I_A,R_A}(l) = -C_{R_A,I_A}(l)$ we get $II_x(l) = RR_x(l)$ and $IR_x(l) = -RI_x(l)$ for $x = 1, 2$

Combining the above finally yields:

$$
\begin{aligned}
C^s_{ED}(l) &= \frac{4}{L^2}\Big(2E\left[R_X(n,k)\, R_X(n,k+l)\right]^2 \\
&\qquad -\Big(E[R_A(k)R_D(k+l)]^2 + E[I_A(k)R_D(k+l)]^2 \\
&\qquad\quad +E[R_A(k)I_D(k+l)]^2 + E[I_A(k)I_D(k+l)]^2\Big)\Big) \\
&= \frac{8}{L^2}(RR_X^2(l) - (RR_1(l) + RI_2(l))^2 \\
&\qquad -(RI_1(l) - RR_2(l))^2)
\end{aligned} \tag{A.12}
$$

where

$$
RR_X = C_{R_X}(l) \odot (\hat{w}(l) \odot \hat{w}(l)) \tag{A.13}
$$

□

A.4 Probability of a sign change of a Gaussian variable due to correlated Gaussian distortion

In Section 3.4 the following theorem was stated without proof.

Theorem 3 (Probability of sign change of a Gaussian random variable due to correlated Gaussian noise)**.** *Let (A, B) denote two zero-mean Gaussian random variables, drawn from a bivariate normal distribution, i.e. $(A, B) \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{AB})$, with correlation matrix $\mathbf{C}_{AB}$:*

$$\mathbf{C}_{AB} = \begin{bmatrix} \sigma_A^2 & \rho\,\sigma_A\sigma_B \\ \rho\,\sigma_A\sigma_B & \sigma_B^2 \end{bmatrix}$$

Now define $C = A + B$. The probability that the sign of C is different from the sign of A is given by:

$$\begin{aligned} P_e &= Pr\left[A \leq 0, C > 0 \quad \vee \quad A > 0, C \leq 0\right] \\ &= \frac{1}{\pi} \arctan\left(\frac{\sigma_B\sqrt{1-\rho^2}}{\sigma_A + \rho\sigma_B}\right) \end{aligned} \tag{A.14}$$

Proof. The derivation is made in four steps.

1. The Gaussian approximation of the PDF, $f_{A,B}(a, b)$, is fully defined by the variances σ_A^2 and σ_B^2 and the correlation coefficient ρ.

   The P_e is obtained by integrating the PDF:

   $$\begin{aligned} P_e &= Pr\left[A \leq 0, C > 0 \quad \vee \quad A > 0, C \leq 0\right] \\ &= 2\,Pr\left[A > 0, C \leq 0\right] \\ &= 2\,Pr\left[A > 0, B \leq -A\right] \\ &= 2\int_0^\infty \int_{-\infty}^{-a} f_{A,B}(a, b)\,db\,da \end{aligned}$$

   The PDF, $f_{A,B}(a, b)$, is shown in Figure A.1(a) and the integration area is indicated by the shaded area.

2. The vertical axis is scaled such that its variance is equal to the signal variance: $B' = \frac{\sigma_A}{\sigma_B}B$. The line $B = -A$, shown in Figure A.1(b), now has an angle α with the vertical axis:

   $$\alpha = \arctan\left(\frac{\sigma_B}{\sigma_A}\right)$$

   Since the variances along both axes now are equal to each other, the main diagonal of the PDF has an angle $\theta = -\frac{1}{4}\pi$ with the horizontal axis.

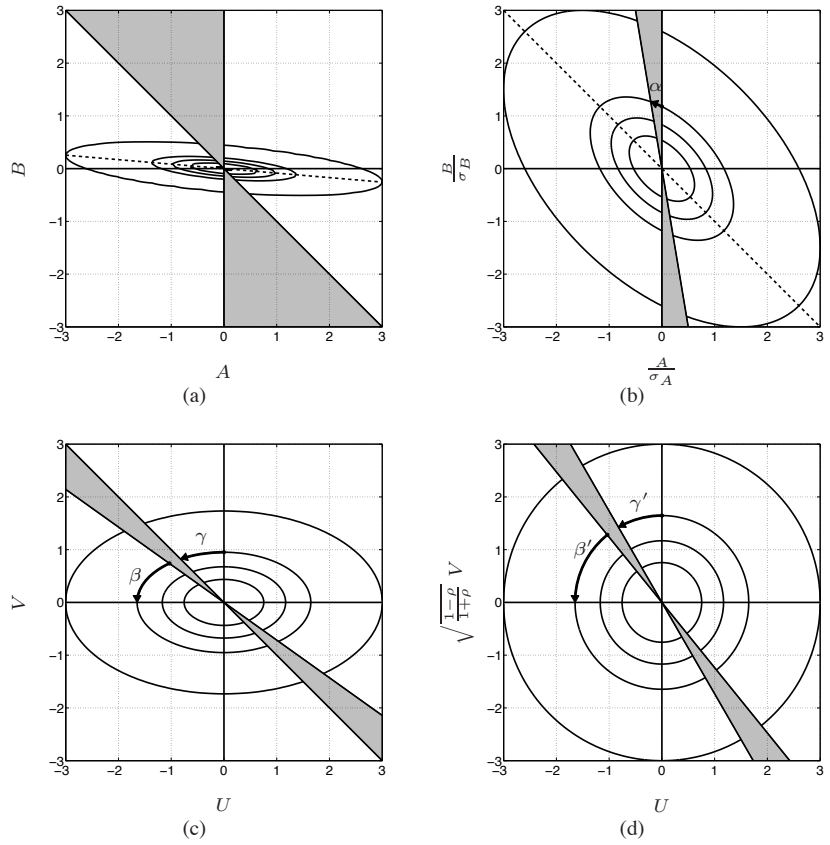


Figure A.1: Illustration of the four steps in the derivation of Eq. A.14: (a) Gaussian approximation of the PDF, $f_{A,B}(a,b)$; the shaded area illustrates the integration area; (b) normalization of the variances; (c) decorrelation, and (d) normalization of the variances.

3. After rotation (decorrelation) the new variables, U and V, have variance $\sigma_U^2 = (1-\rho)\sigma_A^2$ and $\sigma_V^2 = (1+\rho)\sigma_A^2$, respectively. Since there has been no scaling, the angle with the main diagonal of the PDF is unaltered. Figure A.1(c) shows this PDF, including the angles β and γ. For the next step, the $\tan(\beta)$ and $\tan(\gamma)$ need to be known. We will use the following relations:

$$\begin{aligned} \tan(\alpha) &= \frac{\sigma_B}{\sigma_A} \\ \tan\left(\tfrac{1}{4}\pi + \alpha\right) &= \frac{1+\tan(\alpha)}{1-\tan(\alpha)} \end{aligned}$$

This yields the expressions for $\tan(\beta)$ and $\tan(\gamma)$:

$$\begin{aligned}\tan(\gamma) &= 1\\ \tan(\beta) &= \frac{1}{\tan\left(\frac{1}{4}\pi+\alpha\right)}\\ &= \frac{\sigma_A-\sigma_B}{\sigma_A+\sigma_B}\end{aligned}$$

4. Finally, the vertical axis is scaled such that both axis have equal variance, yielding a rotation symmetric PDF. The vertical scaling factor is equal to $\sqrt{\frac{1-\rho}{1+\rho}}$. So now the tangens values are changed into:

$$\begin{aligned}b = \tan(\beta') &= \sqrt{\frac{1-\rho}{1+\rho}}\tan(\beta)\\ c = \tan(\gamma') &= \sqrt{\frac{1+\rho}{1-\rho}}\tan(\gamma)\end{aligned}$$

Using the relationship

$$\arctan(b)+\arctan(c)=\arctan\left(\frac{b+c}{1-bc}\right)$$

The value of α' can be computed:

$$\alpha' = \frac{1}{2}\pi-\arctan\left(\frac{b+c}{1-bc}\right)$$

Filling in the appropriate values for b and c, we obtain:

$$\begin{aligned}b+c &= \frac{(1-\rho)\tan(\beta)+(1+\rho)\tan(\gamma)}{\sqrt{1-\rho^2}}\\ &= \frac{\tan(\beta)+\tan(\gamma)+\rho(\tan(\gamma)-\tan(\beta))}{\sqrt{1-\rho^2}}\\ &= \frac{2}{\sigma_A+\sigma_B}\cdot\frac{\sigma_A+\rho\sigma_B}{\sqrt{1-\rho^2}}\\ 1-bc &= 1-\tan(\beta)\tan(\gamma)\\ &= 1-\frac{\sigma_A-\sigma_B}{\sigma_A+\sigma_B}\\ &= \frac{2}{\sigma_A+\sigma_B}\cdot\sigma_B\end{aligned}$$

Using α' it is straightforward to compute P_e:

$$\begin{aligned} P_e &= 2\int_0^{\infty}\int_{-\infty}^{-a} f_{A,B}(a,b)\,db\,da \\ &= 2\int_0^{\infty}\int_{-u\tan(\alpha'+\beta')}^{-u\tan(\beta')} f_{U,V'}(u,v)\,dv\,du \\ &= \frac{\alpha'}{\pi} \\ &= \frac{1}{2} - \frac{1}{\pi}\arctan\left(\frac{\sigma_A+\rho\sigma_B}{\sigma_B\sqrt{1-\rho^2}}\right) \\ &= \frac{1}{\pi}\arctan\left(\frac{\sigma_B\sqrt{1-\rho^2}}{\sigma_A+\rho\sigma_B}\right) \end{aligned}$$

□

A.5 Relation between $ED_X(n,m)$, $ED_Y(n,m)$ and P_e

Eq. (3.44) relates the energy differences $ED_X(n,m)$ and $ED_Y(n,m)$ to the probability of error P_e. This relation is based on the following theorem, stated here in terms of two Gaussian distributions, A and C. Using this theorem and substituting $A = ED_X(n,m)$ and $C = ED_Y(n,m)$, we immediately obtain Eq. (3.44).

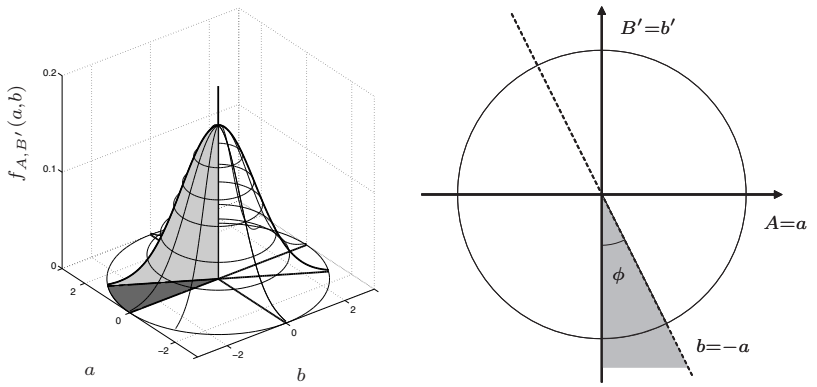
Theorem 4. *Let $A \in \mathcal{N}(0,\sigma_A^2)$ and $B \in \mathcal{N}(0,\sigma_B^2)$ denote two zero-mean, mutually independent, normally distributed random variables. Now define $C = A + B$. The probability that the sign of C is different from the sign of A is given by:*

$$\begin{aligned} P_e &= Pr\left[A \le 0, C > 0 \quad \vee \quad A > 0, C \le 0\right] \\ &= \frac{1}{\pi}\arctan\left(\frac{\sigma_B}{\sigma_A}\right) \\ &= \frac{1}{\pi}\arctan\left(\sqrt{\frac{VAR\left[C-A\right]}{VAR\left[A\right]}}\right) \end{aligned} \tag{A.15}$$

Proof. Due to symmetry $Pr\left[C > 0 | A \le 0\right] = Pr\left[C \le 0 | A > 0\right]$ and $Pr\left[A \le 0\right] = Pr\left[A > 0\right] = \frac{1}{2}$. Therefore,

$$\begin{aligned} P_e &= Pr\left[A \le 0, C > 0 \quad \vee \quad A > 0, C \le 0\right] \\ &= Pr\left[C > 0 | A \le 0\right] Pr\left[A \le 0\right] \\ &\quad + Pr\left[C \le 0 | A > 0\right] Pr\left[A > 0\right] \\ &= Pr\left[C \le 0 | A > 0\right] \\ &= Pr\left[B \le -A | A > 0\right] \end{aligned} \tag{A.16}$$

Define $\alpha = \frac{\sigma_B}{\sigma_A}$ and introduce the normalized version of B, viz. $B' = \frac{\sigma_A}{\sigma_B} \cdot B = \frac{1}{\alpha} \cdot B, \quad B' \in \mathcal{N}(0,\sigma_A^2)$. Due to the scaling factor α, the joint-PDF $f_{A,B'}(a,b)$ is



(a) (b)

Figure A.2: Probability Density Function $f_{A,B'}(a,b)$, (a) 3D visualization (b) Projection onto the ground plane (contour line)

rotation symmetric with respect to the origin, as illustrated in Figure A.2(a). P_e is related to $f_{A,B'}(a,b)$ by:

$$\begin{aligned} P_e &= Pr\left[B \leq -A | A > 0\right] \\ &= \int_0^{\infty} \int_{-\infty}^{-a} f_{A,B}(a,b)\, db\, da \\ &= \int_0^{\infty} \int_{-\infty}^{-\alpha a} f_{A,B'}(a,b)\, db\, da \end{aligned} \tag{A.17}$$

The angle between the vertical axis and the integration boundary is denoted by the angle ϕ, where $\phi = \arctan(\alpha)$, as illustrated in Figure A.2(b). If $\alpha = 1$, i.e. $\sigma_A^2 = \sigma_B^2$, we have $\phi = \pi/2$. Due to the rotational symmetry around the origin[1], the probability P_e is proportional to the $0 \leq \phi < \pi$. We can now express P_e in terms of ϕ as follows:

$$P_e = \int_0^{\infty} \int_{-\infty}^{-\alpha a} f_{A,B'}(a,b)\, db\, da = \frac{\phi}{\pi} = \frac{1}{\pi} \arctan\left(\frac{\sigma_B}{\sigma_A}\right)$$

□

A.6 Correlation between $ED_W(n,m)$ and $Q(n,m)$

The fact that the variables $ED_W(n,m)$ and $Q(n,m)$ are mutually uncorrelated is used in Section 3.4.1 to derive Eq. (3.50).

[1]The result in Eq. (A.18) holds for any rotation-symmetric PDF $f_{A,B'}(a,b)$. If the PDF is not symmetric, the analysis procedure stays the same as long as the analysis can be done using a projection onto the (A, B')-plane. The resulting expression might be different.

Theorem 5. *The variables $ED_W(n,m)$ and $Q(n,m)$ are mutually uncorrelated, and as a result:*

$$\begin{aligned}&\mathit{VAR}\left[ED_Y(n,m) - ED_X(n,m)\right]\\ &= \mathit{VAR}\left[ED_W(n,m) + 2Q(n,m)\right]\\ &= \mathit{VAR}\left[ED_W(n,m)\right] + 4\mathit{VAR}\left[Q(n,m)\right]\end{aligned} \tag{A.18}$$

Proof. Because $ED_W(n,m)$ and $Q(n,m)$ are based on summations of terms $ED_W^s(n,k)$ and $Q^s(n,k)$, respectively, it is sufficient to show that $\mathrm{COV}\left[ED_W^s(n,k), Q^s(n,k+l)\right] = 0$. Using the short-hand notation $R_X(n,k) = \mathrm{Re}\left(\hat{x}(n,k)\right)$ and $I_X(n,k) = \mathrm{Im}\left(\hat{x}(n,k)\right)$ we can express $Q^s(n,k)$ in terms of the two input components $\hat{x}(n,k)$ and $\hat{w}(n,k)$:

$$\begin{aligned}&Q^s(n,k)\\ &= \frac{1}{L}\,\mathrm{Re}\left(\hat{x}(n,k)\overline{\hat{w}(n,k)} - \hat{x}(n{-}1,k)\overline{\hat{w}(n{-}1,k)}\right)\\ &= \frac{1}{L}\left(R_X(n,k)R_W(n,k) - R_X(n{-}1,k)R_W(n{-}1,k)\right)\\ &\quad + \frac{1}{L}\left(I_X(n,k)I_W(n,k) - I_X(n{-}1,k)I_W(n{-}1,k)\right)\end{aligned} \tag{A.19}$$

The covariance can now be computed:

$$\begin{aligned}&\mathrm{COV}\Big[ED_W^s(n,k), Q^s(n,k{+}l)\Big]\\ &= \frac{1}{L}\Big(\mathrm{COV}\Big[ED_W^s(n,k), R_X(n,k{+}l)R_W(n,k{+}l)\Big]\\ &\qquad + \mathrm{COV}\Big[ED_W^s(n,k), I_X(n,k{+}l)I_W(n,k{+}l)\Big]\\ &\qquad - \mathrm{COV}\Big[ED_W^s(n,k), R_X(n{-}1,k{+}l)R_W(n{-}1,k{+}l)\Big]\\ &\qquad - \mathrm{COV}\Big[ED_W^s(n,k), I_X(n{-}1,k{+}l)I_W(n{-}1,k{+}l)\Big]\Big)\\ &= \frac{1}{L}\Big(E\Big[ED_W^s(n,k)R_W(n,k{+}l)\Big]E\Big[R_X(n,k{+}l)\Big]\\ &\qquad + E\Big[ED_W^s(n,k)I_W(n,k{+}l)\Big]E\Big[I_X(n,k{+}l)\Big]\\ &\qquad - E\Big[ED_W^s(n,k)R_W(n{-}1,k{+}l)\Big]E\Big[R_X(n{-}1,k{+}l)\Big]\\ &\qquad - E\Big[ED_W^s(n,k)I_W(n{-}1,k{+}l)\Big]E\Big[I_X(n{-}1,k{+}l)\Big]\Big)\\ &= 0\end{aligned} \tag{A.20}$$

□

A.7 Relation between VAR $[ED_W(n,m)]$, VAR $[Q(n,m)]$ and VAR $[ED_X(n,m)]$

Equation (3.51) is Section 3.4.1 relates the variance VAR $[ED_Y(n,m) - ED_X(n,m)]$ to the variance VAR $[ED_X(n,m)]$.

Theorem 6. *The variance VAR* $[ED_Y(n,m) - ED_X(n,m)]$ *is proportional to VAR* $[ED_X(n,m)]$ *and is equal to:*

$$\begin{aligned} &\textit{VAR}\,[ED_Y(n,m) - ED_X(n,m)] \\ &= \left(\frac{\sigma_W^4}{\sigma_X^4} + 2\frac{\sigma_W^2}{\sigma_X^2}\right) \textit{VAR}\,[ED_X(n,m)] \end{aligned} \tag{A.21}$$

Proof. Theorem 5 expressed the variance on the left-hand side of the equation as:

$$\begin{aligned} &\text{VAR}\,[ED_Y(n,m) - ED_X(n,m)] \\ &= \text{VAR}\,[ED_W(n,m)] + 4\text{VAR}\,[Q(n,m)] \end{aligned} \tag{A.22}$$

Since $ED_X(n,m)$, $ED_W(n,m)$ and $Q(n,m)$ are based on summations of $ED_X^s(n,k)$, $ED_W^s(n,k)$ and $Q^s(n,k)$, respectively, over index k, it is sufficient to relate COV$\Big[ED_W^s(n,k), ED_W^s(n,k+l)\Big]$ and COV$\Big[Q^s(n,k), Q^s(n,k+l)\Big]$ to COV$\Big[ED_X^s(n,k), ED_X^s(n,k+l)\Big]$.

In the following we only consider these covariances. We first express the covariance COV $[ED_X^s(n,k), ED_X^s(n,k+l)]$ in terms of $R_X(n,k)$ and $I_X(n,k)$:

$$
\begin{aligned}
&\mathrm{COV}[ED_X^s(n,k), ED_X^s(n,k+l)] \\
&= \mathrm{COV}[S_X(n,k) - S_X(n-1,k), \\
&\qquad S_X(n,k+l) - S_X(n-1,k+l)] \\
&= 2\big(\mathrm{COV}[S_X(n,k), S_X(n,k+l)] \\
&\quad -\mathrm{COV}[S_X(n,k), S_X(n+1,k+l)]\big) \\
&= \frac{2}{L^2}\mathrm{COV}\big[R_X^2(n,k) + I_X^2(n,k), \\
&\qquad R_X^2(n,k+l) + I_X^2(n,k+l)\big] \\
&\quad -\frac{2}{L^2}\mathrm{COV}\big[R_X^2(n,k) + I_X^2(n,k), \\
&\qquad R_X^2(n+1,k+l) + I_X^2(n+1,k+l)\big] \\
&= \frac{4}{L^2}\Big(\mathrm{COV}\left[R_X^2(n,k), R_X^2(n,k+l)\right] \\
&\qquad -\mathrm{COV}\big[R_X^2(n,k), R_X^2(n+1,k+l)\big]\Big) \\
&= \frac{8}{L^2}\Big(\mathrm{COV}\big[R_X(n,k), R_X(n,k+l)\big]^2 \\
&\qquad -\mathrm{COV}\big[R_X(n,k), R_X(n+1,k+l)\big]^2\Big)
\end{aligned}
\tag{A.23}
$$

Here we used two properties of the Fourier transform of an uncorrelated signal: first, the real part $R_X(n,k)$ and imaginary part $I_X(n,k)$ are mutually uncorrelated; second, the fact that the autocorrelation function of the imaginary part is equal to the autocorrelation function of the real part. Furthermore, we used the following relation for two zero-mean, normally distributed random variables X_1 and X_2:

$$
\mathrm{COV}\left[X_1^2, X_2^2\right] = 2\,\mathrm{COV}\left[X_1, X_2\right]^2 \tag{A.24}
$$

Since the autocorrelation functions of $R_X(n,k)$ and $R_W(n,k)$ are proportional to the variances σ_X^2 and σ_W^2, respectively, it is straightforward to relate these to each other:

$$
\begin{aligned}
&\mathrm{COV}\left[R_W(n,k), R_W(n+p,k+l)\right] \\
&= \frac{\sigma_W^2}{\sigma_X^2} \cdot \mathrm{COV}\left[R_X(n,k), R_X(n+p,k+l)\right].
\end{aligned}
\tag{A.25}
$$

Hence, we can express $\mathrm{COV}\left[ED_W^s(n,k), ED_W^s(n,k+l)\right]$ as:

$$
\begin{aligned}
&\mathrm{COV}\left[ED_W^s(n,k), ED_W^s(n,k+l)\right] \\
&= \frac{\sigma_W^4}{\sigma_X^4} \cdot \mathrm{COV}\left[ED_X^s(n,k), ED_X^s(n,k+l)\right].
\end{aligned}
\tag{A.26}
$$

We can now relate COV $[Q^s(n,k), Q^s(n,k+l)]$ to COV $[ED_X^s(n,k), ED_X^s(n,k+l)]$:

$$\begin{aligned}
&\text{COV}\,[Q^s(n,k), Q^s(n,k+l)] \\
&= \frac{4}{L^2}\big(\text{COV}\,[R_X(n,k), R_X(n,k+l)]\cdot \\
&\qquad \text{COV}\,[R_W(n,k), R_W(n,k+l)] \\
&\quad -\text{COV}\,[R_X(n,k), R_X(n+1,k+l)]\cdot \\
&\qquad \text{COV}\,[R_W(n,k), R_W(n+1,k+l)]\,\big) \\
&= \frac{4}{L^2}\frac{\sigma_W^2}{\sigma_X^2}\big(\text{COV}\,[R_X(n,k), R_X(n,k+l)]^2 \\
&\qquad -\text{COV}\,[R_X(n,k), R_X(n+1,k+l)]^2\,\big) \\
&= \frac{1}{2}\frac{\sigma_W^2}{\sigma_X^2}\text{COV}\,[ED_X^s(n,k), ED_X^s(n,k+l)] \qquad \text{(A.27)}
\end{aligned}$$

Combining Eqs. (3.48), (A.26) and (A.27) results in:

$$\begin{aligned}
&\text{VAR}\,[ED_Y(n,m) - ED_X(n,m)] \\
&= \left(\frac{\sigma_W^4}{\sigma_X^4} + 2\frac{\sigma_W^2}{\sigma_X^2}\right)\text{VAR}\,[ED_X(n,m)] \qquad \text{(A.28)}
\end{aligned}$$

□

Appendix B

Background for Chapter 4

B.1 Relating SNR to MSE for log-spectra and Gaussian iid data

Both the SSD and RARE algorithms use features that are extracted from the log-spectrum, in conjunction with a MSE or RMS distortion measure. In our implementation of RARE we used RMS as the fingerprint-distance measure. For SSD we used the MSE. Since the RMS value is just the square root of the MSE value, in the following we relate the MSE between two unquantized fingerprints (cf. RARE) to the distortion in the fingerprint. The different choices for the distortion measure follow from the difference in quantization of the features used in the fingerprint. In our RARE implementation, the features are represented using 32-bit single precision floats. In SSD the features are quantized into 4-bit characters. There, SNR is directly related to the MSE on feature-level, but the actually observed SNR-MSE relation originates from the quantization procedure.

Consider a log-spectral sample from the original and the distorted version; the distribution of the fingerprint distance would be related to:

$$\begin{aligned} E[MSE] &\propto E[(FP_Y - FP_x)^2] \\ &\propto E[(\log(S_Y(n,k)) - \log(S_X(n,k)))^2] \\ &= E[Z^2], \end{aligned}$$

where $Z = \log\left(\frac{S_Y(n,k)}{S_X(n,k)}\right)$.

In the following we derive the pdf for Z, $f_Z(z)$, and it first and second moment, $E[Z]$ and $E[Z^2]$: Denoting the real and imaginary parts of $x(n,k)$ by random variables x_1 and x_2, respectively, the spectrogram $S_X(n,k)$ can be written as:

$$S_X = x_1^2 + x_2^2$$

The same way we write $S_Y = y_1^2 + y_2^2$. The joint-PDF for $f_{X_1,X_2,Y_1,Y_2}(x_1,x_2,y_1,y_2) = f_{X_1,Y_1}(x_1,y_1)f_{X_2,Y_2}(x_2,y_2)$ consists of the product two zero-mean normal distributions $f_{X_i,Y_i}(x_i,y_i), i = 1,2$ with covariance matrix:

$$C = \frac{1}{2}\begin{bmatrix} \sigma_X^2 & \sigma_X^2 \\ \sigma_X^2 & \sigma_X^2 + \sigma_W^2 \end{bmatrix}$$

Converting both (x_1, x_2) and (y_1, y_2) to polar coordinates ($u = \sqrt{x_1^2 + x_1^2}, v = \sqrt{y_1^2 + y_1^2}, \phi = \arctan(x_2/x_1), \theta = \arctan(y_2/y_1)$) and integrating out the phase components ϕ and θ yields a PDF $f_{U,V}(u, v)$:

$$\begin{aligned} f_{U,V}(u,v) &= \int_0^{2\pi}\int_0^{2\pi} f_{U,V,\Phi,\Theta}(u,v,\phi,\theta)\, d\phi\, d\theta \\ &= \frac{4uv}{\sigma_X^2\sigma_W^2}\exp\left(-\frac{u^2}{\sigma_X^2} - \frac{u^2+v^2}{\sigma_W^2}\right)\sum_{l=0}^{\infty}\frac{1}{(l!)^2}\left(\frac{uv}{\sigma_W^2}\right)^{2l} \end{aligned}$$

Making a conversion to variable $r = v^2/u^2 = S_Y(n,k)/S_X(n,k)$, we obtain the PDF $f_R(r)$:

$$f_R(r) = \frac{\sigma_W^2}{\sigma_X^2}\sum_{l=0}^{\infty}\frac{(2l+1)!}{(l!)^2}\frac{r^l}{\left(r+1+\frac{\sigma_W^2}{\sigma_X^2}\right)^{2l+2}}$$

Since $z = \ln(r)$, the PDF we are looking, $f_Z(z)$, is given by:

$$f_Z(z) = \frac{\sigma_W^2}{\sigma_X^2}\sum_{l=0}^{\infty}\frac{(2l+1)!}{(l!)^2}\left(\frac{\exp(z)}{\left(\exp(z)+1+\frac{\sigma_W^2}{\sigma_X^2}\right)^2}\right)^{l+1}$$

The p^{th} moment of Z can be obtained through integration:

$$E[Z^p] = \int_{-\infty}^{\infty} z^p\, f_Z(z)\, dz$$

Its mean is given by:

$$E[Z] = \ln\left(1+\frac{\sigma_W^2}{\sigma_X^2}\right)$$

and its second moment is given by:

$$E[Z^2] = \ln\left(1+\frac{\sigma_W^2}{\sigma_X^2}\right)^2 - 2Li_2\left(1-\frac{\sigma_W^2}{\sigma_X^2+\sigma_W^2}\right) + \frac{1}{3}\pi^2$$

where $Li_2(\cdot)$ is the polylogarithm function with $n = 2$:

$$Li_n(x) = \sum_{k=1}^{\infty}\frac{x^k}{k^n} \qquad |x| \le 1$$

The first term in $E[Z^2]$ is much smaller than the other terms and can thus be ignored. For large SNR, $\sigma_X^2 + \sigma_W^2 \approx \sigma_X^2$. Converting to SNR on a decibel scale, we obtain:

$$E[Z^2] \approx \frac{1}{3}\pi^2 - 2Li_2\left(1 - 10^{-\text{SNR}/10}\right)$$

Using the relation:

$$\frac{\pi^2}{6} - Li_2(1-x) = Li_2(x) + \log_2(x)\log_2(1-x)$$

and using $Li_2(0) = 0$ and elementary properties of the $\ln(\cdot)$-function, we obtain:

$$E[Z^2] \approx \frac{\ln(10)}{5\,\ln(2)^2} \cdot \text{SNR} \cdot 10^{-\text{SNR}/10}$$

On a log-scale this works out into:

$$\log_{10}(E[Z^2]) \approx \log_{10}(\text{SNR}) - \frac{\text{SNR}}{10} + \text{const}$$

For large SNR, the linear term is dominant, and thus the MSE between the fingerprints is expected to drop by a factor 10 for and increase in SNR with 10 dB. Using the RMS measure - like we did in RARE - the fingerprint distance reduces by a factor 10, for an SNR increase of 20 dB, like we experimentally observed.

Samenvatting

Een *audio fingerprint* (letterlijk: vingerafdruk van audio) is een compacte representatie van een audio signaal. Een audio fingerprint kan worden gebruikt om een audio bestand of een audio fragment automatisch te herkennen. Het identificeren van een audio fragment gebeurt in twee stappen. In de eerste stap, de *enrollment*, worden de audio fingerprints van een collectie bekend audio materiaal berekend en in een database gestopt, voorzien van relevante metadata zoals de naam van het liedje en de artiest. Het doel van de tweede stap is het herkennen van een audio fragment. Daartoe wordt van dit fragment de audio fingerprint berekend, en vergeleken met de fingerprints in de database. Als de database een vergelijkbare fingerprint bevat wordt het fragment herkend. Het systeem kan dan de juiste metadata uit de database presenteren.

In dit proefschrift ontwikelen we een drietal modellen voor audio fingerprints. De nadruk ligt hier op de wijze waarop de fingerprint berekend wordt en de eigenschappen van de fingerprints, en niet zozeer op het daadwerkelijke vergelijken van de onbekende fingerprint met de fingerprints in de database en de resulterende herkenning. Ook zijn er geen nieuwe fingerprinting algoritmes ontwikkeld.

Er zijn vele toepassingen van audio fingerprinting bekend of denkbaar, waaronder het herkennen en vastleggen wat er op radio of TV is uitgezonden, hoeveel mensen daarnaar luisterden, forensische toepassingen, het herkennen van ongeautoriseerde uploads van bestanden, het herkennen van muziek op de radio, en het automatisch doorschakelen naar aanbiedingen op Internet op basis van het geluid van een reclame op radio of TV.

Wanneer een audio bestand in een ander digitaal formaat wordt opgeslagen of wanneer een audio fragment wordt verstoord, blijft de fingerprint vergelijkbaar met de oorspronkelijke fingerprint. Hiermee onderscheidt audio fingerprinting zich van twee andere technieken: hashing en content based retrieval. De hash van een audio bestand is eveneens een compacte representatie, maar de hash-waarde verandert onherkenbaar als er ook maar een enkele bit in het audio bestand gewijzigd wordt. Twee audio bestanden die identiek klinken, maar een verschillende digitale representatie hebben, zullen twee verschillende hash-waardes opleveren, maar hebben een vergelijkbare audio fingerprint. Content-based retrieval wordt gebruikt om die bestanden uit een collectie te halen die conceptueel op het voorbeeld fragment lijken, bijvoorbeeld omdat ze tot hetzelfde genre behoren, van dezelfde componist zijn, of door dezelfde artiest worden uitgevoerd. Audio fingerprinting kan uitsluitend gebruikt worden om dezelfde opname, mogelijk verstoord of in een andere digitale representatie, te herken-

nen.

Uiteraard hangen de gewenste eigenschappen van een fingerprinting systeem sterk af van de toepassing. De eigenschappen die in dit proefschrift centraal staan zijn de robuustheid van de fingerprint tegen verstoringen van het audio signaal, het onderscheidend vermogen van de fingerprint, de nauwkeurigheid waarmee identificatie plaatsvindt, en de grootte van de fingerprint.

Drie bijdragen staan centraal in dit proefschrift. Ten eerste modelleren we de statistische structuur van een specifiek audio fingerprinting algoritme, de Philips Robust Hash (PRH) [44]. De PRH fingerprint is gebaseerd op energie kenmerken van het onderliggende audiosignaal, en wordt compact in binaire vorm gerepresenteerd. Deze representatie vormt een 'samenvatting' van de temporele en spectrale karakteristieken van het onderliggende audio signaal, heeft een eigen, karakteristieke structuur. Deze karakteristieke structuur wordt vrijwel geheel bepaald door enkele parameters in het fingerprint algoritme.

Het model dat we ontwikkeld hebben beschrijft de structuur van de PRH fingerprint [35] als functie van een aantal parameters. Het model kan worden gebruikt om de structuur en eigenschappen van de fingerprint beter te begrijpen, en potentieel om deze te optimaliseeren. We valideren het ontwikkelde model aan de hand van kunstmatige ingangssignalen, waarvan de samples onderling onafhankelijk zijn en alle gegenereerd zijn op basis van dezelfde Gaussische kansverdeling. Deze analyse is eveneens uitgevoerd, herformuleerd en uitgebreid door Balado, Hurley, McCarthy en Silvestre [15, 52].

Ten tweede observeren we dat verstoringen in het audio signaal resulteren in verstoringen in de audio fingerprint. De verstoringen in het audio signaal tasten doorgaans de kwaliteit van het audio signaal aan. Het idee is nu om de verstoring in het audio signaal te schatten door de fingerprint van het verstoorde audio signaal te vergelijken met de fingerprint van een kopie van hetzelfde audio signaal, maar dan op hoge kwaliteit [38]. Op deze wijze zou de functionaliteit van een audio fingerprinting systeem kunnen worden uitgebreid. We hebben een aantal uit de literatuur bekende audio fingerprinting algoritmes geimplementeerd en vergeleken. De vergelijking laat zien dat de verschillen tussen de fingerprints die ontstaan ten gevolge van compressie vergelijkbare karakteristieken hebben.

We modelleren de effecten van verstoringen in de PRH fingerprints ten gevolge van compressie van, of het toevoegen van witte ruis aan, het onderliggende audio signaal. Het voornaamste resultaat is een wiskundige formule die voor PRH fingerprints het gemiddelde verschil tussen de fingerprint van het origineel en van de verstoorde versie relateert aan signaal-ruis verhouding van het ingangssigaal dat eveneens uit witte ruis bestaat [36, 38]. We valideren de gevonden relatie door middel van simulaties. Het model past perfect op de experimentele data voor het type ingangssignalen waarvoor het is ontwikkeld, en het model is een goede indicatie voor het gedrag dat we hebben waargenomen voor een groter aantal algoritmes op daadwerkelijke muziek bestanden.

Ten derde beschouwen we een informatie-theoretisch raamwerk dat door Westover en O'Sullivan (WOS) is ontwikkeld [104]. De vraag die daarin centraal staat is 'hoeveel verschillende signalen kun je in een bepaalde setting met een fingerprinting

systeem van elkaar onderscheiden'. De setting slaat hier op de eigenschappen van de fingerprint (de grootte van de fingerprint, en de wijze waarop deze gerepresenteerd wordt), en de eigenschappen van de omgeving waarin het systeem functioneert (de eigenschappen van de ingangssignalen en met hoeveel verstoring van de signalen het systeem overweg moet kunnen). We gebruiken ons voor de PRH fingerprint ontwikkelde model om te schatten maximaal hoeveel verschillende signalen van elkaar kunnen worden onderscheiden met een binaire fingerprint zoals de PRH. Tenslotte bekijken we of de waargenomen verstoringen in de fingerprint ten gevolge van verstoringen in het ingangssignaal in het informatie-theoretische raamwerk van Westover en O'Sullivan passen. We benoemen de verschillen tussen het WOS-model en de waarnemingen uit experimenten met praktische fingerprint algoritmes.

We sluiten af met aanbevelingen om de ontwikkelde modellen uit te breiden met modellen die de robuustheid en het onderscheidend vermogen van de fingerprints in samenhang modelleren; om meerdere soorten verstoringen te beschouwen en de modellen uit te breiden naar beeld en video fingerprinting; om een evaluatie raamwerk op te zetten specifiek voor audio fingerprinting; om psycho-acoustische modellen mee te nemen in het berekenen van de audio fingerprints; en om een theoretisch framewerk te ontwikkelen waarin specifieke algoritmes afgezet kunnen worden tegen de capaciteits grenzen.

Bibliography

[1] Lame, December 2005. http://lame.sourceforge.net.

[2] Gracenote, August 2006. http://www.gracenote.com.

[3] Snocap, August 2006. http://www.snocap.com.

[4] Digimarc, November 2007. http://www.digimarc.com.

[5] Ogg Vorbis Specification, June 2007. http://www.xiph.org/vorbis/doc/Vorbis_I_spec.html.

[6] Thomsom content security, November 2007. http://contentsecurity.thomson.net.

[7] CIvolution, March 2010. http://www.civolution.com.

[8] I-dash: Investigator's dashboard, March 2010. http://www.i-dash.eu/.

[9] Musictrace, March 2010. http://www.musictrace.de/.

[10] Nielsen broadcast data systems, March 2010. http://www.nielsen.com/.

[11] Testing youtube's audio fingerprinting, March 2010. http://www.csh.rit.edu/ parallax/.

[12] Viacom press release: Viacom files federal copright infringement complaint against youtube and google, March 2010. http://www.viacom.com/news/Pages/newstext.aspx?RID=1009865.

[13] Youtube, March 2010. http://www.youtube.com/.

[14] E. Allamanche, J. Herre, O. Hellmuth, B. Fröba, T. Kastner, and M. Cremer. Content-based identification of audio material using mpeg-7 low level description. In *2nd International Symposium on Music Information Retrieval (ISMIR)*, October 2001.

[15] F. Balado, N. J. Hurley, E. P. McCarthy, and G. C. M. Silvestre. Performance analysis of robust audio hashing. *IEEE Transactions on Information Forensics and Security*, 2(2):254 – 266, June 2007.

[16] F. Balado, N. J. Hurley, E. P. McCarthy, and G. C. M. Silvestre. Performance of philips audio fingerprinting under additive noise. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 209 – 212, April 2007.

[17] S. Baluja and M. Covell. Content fingerprinting using wavelets. In *3rd European Conference on Visual Media Production (CVMP)*, pages 198 – 207, November 2006.

[18] S. Baluja and M. Covell. Audio fingerprinting: Combining computer vision & data stream processing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 213 – 216, April 2007.

[19] J. Barr, B. Bradley, and B. Hannigan. Using digital watermarks with image signatures to mitigate the threat of the copy attack. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pages 69–72, April 2003.

[20] E. Batlle, J. Masip, and E. Guaus. Automatic song identification in noisy broadcast audio. In *IASTED International Conference on Signal and Image Processing*, August 2002.

[21] R. J. Beaton, J. G. Beerends, M. Keyhl, and W. C. Treurniet. *Objective Perceptual Measurement of Audio Quality*. Audio Eng. Society, 1996.

[22] J. G. Beerends. *Applications of Digital Signal Processing to Audio and Acoustics*, volume 437 of *The Int. Series in Engineering and Computer Science*, chapter Audio Quality Determination Based on Perceptual Measurement Techniques, pages 1 – 38. Kluwer Academic Publishers, 2002.

[23] J. G. Beerends and J. A. Stemerdink. A perceptual audio quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, 40(12):963 – 978, December 1992.

[24] W. Birmingham, R. Dannenberg, and B. Pardo. An introduction to query by humming with the vocalsearch system. *Communications of the ACM*, 49(8):49–52, August 2006.

[25] M. Bosi. *Multimedia Security Technologies for Digital Rights Management*, chapter Digital Rights Management Systems, pages 23 – 50. Internet and Communications. Academic Press, 2006.

[26] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, and M. Dietz. Iso/iec mpeg-2 advanced audio coding. *Journal of the Audio Engineering Society*, 45(10):789 – 813, October 1997.

[27] C. J. C. Burges, J. C. Platt, and S. Jana. Distortion discriminant analysis for audio fingerprinting. *IEEE Transactions on Speech and Audio Processing*, 11(3):165 – 174, May 2003.

[28] P. Cano, E. Batlle, T. Kalker, and J. Haitsma. A review of audio fingerprinting. *Journal of VLSI Signal Processing*, 41(3):271–284, November 2005.

[29] P. Cano, E. Batlle, H. Mayer, and H. Neuschmied. Robust sound modeling for song detection in broadcast audio. In *112th AES Convention*, 2002.

[30] I. J. Cox, M. L. Miller, and J. A. Bloom. *Digital Watermarking*. Morgan Kaufmann, 2002.

[31] D. Delannay and B. Macq. Watermarking relying on cover signal content to hide synchronization marks. *IEEE Transactions on Information Forensics and Security*, 1(1):87 – 101, March 2006.

[32] J. Dittmann. Content-fragile watermarking for image authentication. In *Security, steganography, and watermarking of multimedia contents III*, volume 4314 of *Proceedings of the SPIE*, pages 175 – 184, January 2001.

[33] J. Dittmann, A. Steinmetz, and R. Steinmetz. Content-based digital signature for motion pictures authentication and content-fragile watermarking. In *International Conference on Multimedia Computing and Systems (ICMCS)*, volume 2, pages 209 – 213, 1999.

[34] P. J. O. Doets, M. Menor Gisbert, and R. L. Lagendijk. On the comparison of audio fingerprints for extracting quality parameters of compressed audio. In *Security, steganography, and watermarking of multimedia contents VII*, Proceedings of the SPIE, January 2006.

[35] P. J. O. Doets and R. L. Lagendijk. Stochastic model of a robust audio fingerprinting system. In *5th International Conference on Music Information Retrieval (ISMIR)*, pages 349 – 352, October 2004.

[36] P. J. O. Doets and R. L. Lagendijk. Theoretical modeling of a robust audio fingerprinting system. In *4th IEEE Benelux Signal Processing Symposium*, pages 101 – 104, April 2004.

[37] P. J. O. Doets and R. L. Lagendijk. Extracting quality parameters for compressed audio from fingerprints. In *6th International Conference on Music Information Retrieval (ISMIR)*, pages 498 – 503, September 2005.

[38] P. J. O. Doets and R. L. Lagendijk. Distortion estimation in compressed music using only audio fingerprints. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):302 – 317, February 2008.

[39] R. P. W. Duin. Compactness and complexity of pattern recognition problems. In *International Symposium on Pattern Recognition "In Memoriam Pierre Devijver"*, pages 124 – 128, February 1999.

[40] T. Ericson. *Topics in Coding Theory*, volume 128 of *Lecture Notes in Control and Information Sciences (LNCIS)*, chapter Bounds on the Size of a Code, pages 45–72. 1989.

[41] A. M. Eskicioglu and E. J. Delp. *Multimedia Security Handbook*, chapter Protection of Multimedia Content in Distribution Networks, pages 3 – 62. Internet and Communications. CRC Press, 2005.

[42] R. Ge, G. R. Arce, and G. DiCrescenzo. Approximate message authentication codes for n-ary alphabets. *IEEE Transactions on Information Forensics and Security*, 1(1):56–67, March 2006.

[43] E. Gómez, P. Cano, L. Gomes, E. Batlle, and M. Bonnet. Mixed watermarking-fingerprinting approach for integrity verification of audio recordings. In *IEEE International Telecommunications Symposium*, September 2002.

[44] J. Haitsma and T. Kalker. A highly robust audio fingerprinting system. In *3rd International Conference on Music Information Retrieval (ISMIR)*, October 2002.

[45] J. Haitsma and T. Kalker. Speed-change resistant audio fingerprinting using auto-correlation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 728 – 731, April 2003.

[46] J. Haitsma, T. Kalker, and J. Oostveen. Robust audio hashing for content identification. In *Content-Based Multimedia Indexing*, September 2001.

[47] O. Harmanci, M. Kucukgoz, and M. K. Mihçak. Temporal synchronization of watermarked video using image hashing. In *Security, steganography, and watermarking of multimedia contents VI*, volume 5681 of *Proceedings of the SPIE*, pages 370 – 380, January 2005.

[48] J. Herre, O. Hellmuth, and M. Cremer. Scalable robust audio fingerprinting using mpeg-7 content description. In *5th IEEE Workshop on Multimedia Signal Processing (MMSP)*, pages 165 – 168, October 2002.

[49] C. Herrero. Subjective and objective assessment of sound quality: solutions and applications. In *CIARM Conference*, pages 1 – 20, 2005.

[50] F. Holm and W. T. Hicken. Audio fingerprinting system and method, September 2003.

[51] N. J. Hurley, F. Balado, E. P. McCarthy, and G. C. M. Silvestre. Performance of philips audio fingerprinting under desynchronisation. In *8th International Conference on Music Information Retrieval (ISMIR)*, October 2007.

[52] N. J. Hurley, F. Balado, and G. C. M. Silvestre. Markov modelling of fingerprinting systems for collision analysis. *EURASIP Journal on Information Security*, 2008.

[53] A. K. Jain, R. Bolle, and S. Pankanti. *Biometrics: Personal identification in a Networked Society*, chapter Introduction to Biometrics, pages 1 – 41. Kluwer Academic Publishers, 2002.

[54] W. Jonker and J.-P. Linnartz. Digital rights management in consumer electronics products. *IEEE Signal Processing Magazine*, 21(2):82 – 91, March 2004.

[55] T. Kalker. Applications and challenges for audio fingerprinting. In *111th AES convention*, sheets, December 2001.

[56] T. Kalker, D. H. J. Epema, P. H. Hartel, R. L. Lagendijk, and M. van Steen. Music2share - copyright-compliant music sharing in p2p systems. *Proceedings of the IEEE*, 92(6):961 – 970, June 2004.

[57] T. Kastner, E. Allamanche, J. Herre, O. Hellmuth, M. Cremer, and H. Grossmann. Mpeg-7 scalable robust audio fingerprinting. In *112th AES Convention*, May 2002.

[58] Y. Ke, D. Hoiem, and R. Sukthankar. Computer vision for music identification. In *Computer Vision and Pattern Recognition (CVPR)*, pages 597 – 604, June 2005.

[59] S. Kim and C. D. Yoo. Boosted binary audio fingerprint based on spectral subband moments. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 241 – 244, April 2007.

[60] F. Kurth. A ranking technique for fast audio identification. In *5th IEEE Workshop on Multimedia Signal Processing (MMSP)*, pages 186 – 189, December 2002.

[61] M. Kutter, S. Voloshynovskiy, and A. Herrigel. The watermark copy attack. In *Security, steganography, and watermarking of multimedia contents II*, volume 3971 of *Proceedings of the SPIE*, pages 371 – 379, January 2000.

[62] R. Lancini, F. Mapelli, and R. Pezzano. Audio content identification by using perceptual hashing. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 739 – 742, June 2004.

[63] G. C. Langelaar, I. Setyawan, and R. L. Lagendijk. Watermarking digital image and video data. a state-of-the-art overview. *IEEE Signal Processing Magazine*, 17(5):20 – 46, September 2000.

[64] J. Lebossé and L. Brun. Audio fingerprint identification by approximate string matching. In *8th International Conference on Music Information Retrieval (ISMIR)*, October 2007.

[65] J. Lebossé, L. Brun, and J. C. Pailles. A robust audio fingerprint extraction algorithm. In *IASTED Conference on Signal Processing, Pattern Recognition, and Applications (SPPRA)*, February 2007.

[66] J. Lebossé, L. Brun, and J. C. Pailles. A robust audio fingerprint's based identification method. In *3rd Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, volume 4477 of *Lecture Notes in Computer Science (LNCS)*, pages 185–192, June 2007.

[67] E. T. Lin, A. M. Eskicioglu, R. L. Lagendijk, and E. J. Delp. Advances in digital video content protection. *Proceedings of the IEEE*, 93(1):171 – 183, January 2005.

[68] J.-P. Linnartz, T. Kalker, and G.F.G. Depovere. Modelling the false alarm and missed detection rate for electronic watermarks. In *2nd International Information Hiding Workshop*, volume 1525 of *Lecture Notes in Computer Science*, pages 329 – 343, April 1998.

[69] H. S. Malvar. *Audio Anecdotes: Tools, Tips, and Techniques for Digital Audio*, chapter Auditory masking in audio compression, pages 217 – 236. A.K. Peters, Ltd., 2001.

[70] F. Mapelli, R. Pezzano, and R. Lancini. Robust audio fingerprinting for song identification. In *12th European Signal Processing Conference (EUSIPCO)*, pages 2095 – 2098, September 2004.

[71] E. P. McCarthy, F. Balado, G. C. M. Silvestre, and N. J. Hurley. A model for improving the performance of feature extraction based robust hashing. In *Security, steganography, and watermarking of multimedia contents VI*, volume 5681 of *Proceedings of the SPIE*, pages 59 – 67, January 2005.

[72] E. P. McCarthy, F. Balado, G. C. M. Silvestre, and N. J. Hurley. Statistical analysis of an audio fingerprinting scheme. In *Irish Signals and Systems Conference (ISSC)*, September 2005.

[73] E. P. McCarthy, F. Balado, G. C. M. Silvestre, and N. J. Hurley. Statistical model and error analysis of a proposed audio fingerprinting algorithm. In *Multimedia Content Analysis, Management, and Retrieval*, volume 6073 of *Proceedings of the SPIE*, January 2006.

[74] Microsoft WMA on Wikipedia, July 2007. http://en.wikipedia.org/wiki/Windows_Media_Audio.

[75] M. K. Mihçak and R. Venkatesan. A perceptual audio hashing algorithm: A tool for robust audio identification and information hiding. In *4th International Information Hiding Workshop*, volume 2137 of *Lecture Notes in Computer Science*, pages 51 – 65, April 2001.

[76] M. L. Miller, M. A. Rodriguez, and I. J. Cox. Audio fingerprinting: Nearest neighbor search in high dimensional binary spaces. *Journal of VLSI Signal Processing*, 41(3):285–291, November 2005.

[77] P. Moulin and R. Koetter. Data-hiding codes. *Proceedings of the IEEE*, 93(12):2083 – 2126, December 2005.

[78] P. Moulin and J. A. O'Sullivan. Information-theoretic analysis of information hiding. *IEEE Transactions on Information Theory*, 49(3):563 – 593, March 2003.

[79] H. Neuschmied, H. Mayer, and E. Batlle. Content-based identification of audio titles on internet. In *1st IEEE International Conference on Web Delivering of Music (WEDELMUSIC)*, pages 96 – 100, November 2001.

[80] J. Oostveen, T. Kalker, and J. Haitsma. Feature extraction and a database strategy for video fingerprinting. In *5th International Conference VISUAL*, volume 2314 of *Lecture Notes in Computer Science*, pages 117 – 128, October 2002.

[81] H. Özer, B. Sankur, and N. Memon. Robust audio hashing for audio identification. In *12th European Signal Processing Conference (EUSIPCO)*, September 2004.

[82] H. Özer, B. Sankur, N. Memon, and E. Anar. Perceptual audio hashing functions. *EURASIP Journal on Applied Signal Processing*, 12:1780–1793, 2005.

[83] L. Pérez-Freire, F. Pérez-González, and S. Voloshynovskiy. An accurate analysis of scalar quantization-based data hiding. *IEEE Transactions on Information Forensics and Security*, 1(1):80–86, March 2006.

[84] Philips Content Identification. Philips cinefence - forensic watermarking solutions for digital cinema, 2007. http://www.contentidentification.philips.com.

[85] R. Radhakrishnan, C. Bauer, C. Cheng, and K. Terry. Audio signature extraction based on projections of spectrograms. In *IEEE International Conference on Multimedia and Expo (ICME)*, July 2007.

[86] A. Ramalingam and S. Krishnan. Gaussian mixture modeling of short-time fourier transform features for audio fingerprinting. *IEEE Transactions on Information Forensics and Security*, 1(4):457 – 463, December 2006.

[87] G. Richly, L. Varga, F. Kovàs, and G. Hosszú. Optimised soundprint selection for identification in audio streams. *IEE Proceedings - Communications*, 148(5):287 – 289, October 2001.

[88] G. R. Schmidt and M. K. Belmonte. Scalable, content-based audio identification by multiple independent psychoacoustic matching. *Journal of the Audio Engineering Society*, 52(3):366 – 377, March 2004.

[89] J. S. Seo, M. Jin, S. Lee, D. Jang, S. Lee, and C. D. Yoo. Audio fingerprinting based on normalized spectral subband moments. *IEEE Signal Processing Letters*, 13(4):209 – 212, April 2006.

[90] Shazam, March 2010. http://www.shazam.com.

[91] Shazam. Shazam targets brands and broadcasters with the launch of sara (shazam audio recognition advertising), March 2010. http://www.shazam.com/music/web/news.html.

[92] Sony ATRAC, July 2007. http://www.sony.net/Products/ATRAC3/index.html.

[93] S. R. Subramanya and B. K. Yi. Digital rights management. *IEEE Potentials*, 25(2):31 – 34, March / April 2006.

[94] S. Sukittanon, L. E. Atlas, and J. W. Pitton. Modulation-scale analysis for content identification. *IEEE Transactions on Signal Processing*, 52(10):3023 – 3035, October 2004.

[95] T. Thiede and E. Kabot. A new perceptual quality measure for bit rate reduced audio. In *100th AES Convention*, April 1996.

[96] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten. PEAQ - the ITU standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society (JAES)*, 29(1/2):3 – 29, January/Frebuary 2000.

[97] K. Tsutsui, H. Suzuki, O. Shimoyoshi, M. Sonohara, K. Akagiri, and R. M. Heddle. Atrac adaptive transform acoustic coding for minidisc. In *93rd AES Convention*, October 1992.

[98] E. Tuncel, P. Koulgi, and K. Rose. Ratedistortion approach to databases: Storage and content-based retrieval. *IEEE Transactions on Information Theory*, 50(6):953–967, June 2004.

[99] Tutanic, March 2010. http://www.wildbits.com/tunatic/.

[100] V. Venkatachalam, L. Cazzanti, N. Dhillon, and M. Wells. Automatic identification of sound recordings. *IEEE Signal Processing Magazine*, 21(2):92 – 99, March 2004.

[101] A. Wang. An industrial strength audio search algorithm. In *4th International Conference on Music Information Retrieval (ISMIR)*, October 2003.

[102] A. Wang and J. O. Smith III. System and methods for recognizing sound and music signals in high noise and distortion, February 2002.

[103] S. Ward and I. Richards. A system and method for acoustic fingerprinting, March 2001.

[104] M. B. Westover and J. A. O'Sullivan. Achievable rates for pattern recognition. *IEEE Transactions on Information Theory*, 54(1), January 2008.

[105] F. Willems, T. Kalker, J. Goseling, and J.-P. Linnartz. On the capacity of a biometrical identification system. In *IEEE International Symposium on Information Theory (ISIT)*, page 82, July 2003.

[106] M. J. Wilson, L. Xie, G. R. Arce, and R. F. Graveman. Content-based searching of multimedia databases by use of approximate digital signatures. *SPIE Journal of Applied Optics*, 42(29):5855 – 5871, October 2003.

[107] E. H. Wold, T. L. Blum, D. F. Keislar, and J. A. Wheaton. Method and apparatus for creating a unique audio signature, November 2000.

[108] C.-P. Wu and C.-C. J. Kuo. Speech content integrity verification integrated with. itu g.723.1 speech coding. In *IEEE International Conference on Information Technology: Coding and Computing*, pages 680 – 684, April 2001.

[109] L. Xie, G. R. Arce, and R. F. Graveman. Approximate image messages authentication codes. *IEEE Transactions on Multimedia*, 3(2):242 – 252, June 2001.

Acknowledgements

I would like to take this opportunity to thank all those people who have supported and encouraged me throughout the years to start, continue and finish this work. I would like to thank the following people in particular.

Rosanne, thank you for your love, support and patience. Many evenings, weekends and holidays were spent on this thesis. Further, I would like to thank my parents for their continuous support and love. Finally, I would like to thank my friends and family. In particular, I would like to thank Jeroen, Rob, Matthé, Gert-Jan, Marc and Mark.

Inald, thank you for giving me the opportunity to start this work, the fruitful discussions, as well as for your patience. Loubna and Mario, it was a privilege to supervise your master's theses.

I would further like to thank my colleagues at the Information and Communication Group at TU Delft. I have had a great time with great people. In particular, I would like to thank my roommates Mark and Ioana for the good atmosphere; and Bartek, Ronald, Omar, Kathy, Eugene, Hasan, Jan, Ben, Robbert and Leo for the good times. Also a special word of thanks to Richard for starting the fine tradition of course evaluations in the pub.

Finally, a word of thanks to my colleagues at TNO. In particular, to Dick, Wessel and Stephan; to my roommates Frits, Omar, Menno, Erik who were always asking for the status of my 'proefwerk'; and to Harm for proofreading parts of this work.

Curriculum Vitae

Peter Jan Doets was born in Amsterdam, the Netherlands, on July 29, 1977. He obtained his Gymnasium diploma from the Murmellius Gymnasium in Alkmaar in 1995. He then went on to study Electrical Engineering at Delft University of Technology. In 2001, during a four month internship at Thales Navel Netherlands in Hengelo, the Netherlands, he worked on simulation models for Direct Digital Synthesis for radar waveform generation. In 2003, he graduated from the Information and Communication Theory Group, where he developed a new algorithm for correcting geometrical distortions in watermarked images.

During his period as a student at Delft University of Technology, he has been an active member of the student fraternity Delftsche Studenten Bond (DSB), the university political party AAG, and the student union VSSD. He served in several board functions in the before mentioned student bodies, including a full-time period in the board of the DSB, and part-time in the members' council of the VSSD.

In 2003, he started as a Ph.D. student at the Information and Communication Theory Group of Delft University of Technology. His research focusses on the development of stochastic models of audio fingerprints. During this period he served as a teaching assistant in several courses, including Multimedia Compression and Stochastic Processes. He contributed to the development of the lab work module for the Stochastic Processes course. During 2003-2006 he also assisted in the yearly course on Video Coding (VCODING) taught at Philips Center for Technical Training (Philips CTT).

Since 2008, he has been working as a consultant in the Multimedia Technology group of TNO Information and Communication Technology in Delft. Here, his work focusses on Digital Video Broadcasting (DVB) technology, IPTV, and the application of audio and video fingerprinting and media mining technology in the public safety domain. He currently chairs the Identity and Personalisation task force in the Commercial Module on IPTV of the DVB Forum.

He is a member of the IEEE Signal Processing and Circuits and Systems Societies. His personal interests include history, politics and traveling.

He is married to Rosanne de Geus and proud father of Douwe.