Distortion Estimation in Compressed Music Using Only Audio Fingerprints

Peter Jan O. Doets, *Student Member, IEEE*, and Reginald L. Lagendijk, *Fellow, IEEE*

***Abstract*—An audio fingerprint is a compact yet very robust representation of the perceptually relevant parts of an audio signal. It can be used for content-based audio identification, even when the audio is severely distorted. Audio compression changes the fingerprint slightly. We show that these small fingerprint differences due to compression can be used to estimate the signal-to-noise ratio (SNR) of the compressed audio file compared to the original. This is a useful content-based distortion estimate, when the original, uncompressed audio file is unavailable. The method uses the audio fingerprints only. For stochastic signals distorted by additive noise, an analytical expression is obtained for the average fingerprint difference as function of the SNR level. This model is based on an analysis of the Philips robust hash (PRH) algorithm. We show that for uncorrelated signals, the bit error rate (BER) is approximately inversely proportional to the square root of the SNR of the signal. This model is extended to correlated signals and music. For an experimental verification of our proposed model, we divide the field of audio fingerprinting algorithms into three categories. From each category, we select an algorithm that is representative for that category. Experiments show that the behavior predicted by the stochastic model for the PRH also holds for the two other algorithms.**

***Index Terms*—Audio fingerprinting, content-based identification, quality estimation, reduced-reference quality estimation, signal-to-noise ratio (SNR) estimation, stochastic model.**

I. Introduction

An AUDIO fingerprint is a compact low-level representation of an audio signal [1]. It has been used extensively for content-based identification of unlabeled audio [2]–[11]. Applications of audio fingerprinting include music identification using cell phones, identification of songs/commercials on the radio, television, and the Internet, and digital music library organization [1]. Fingerprints can be used in a watermarking context to obtain content-dependent (water)marks, to solve synchronization problems, and to use watermarks to check whether audio content has been altered [1], [12]. Snocap uses fingerprints for filtering in file sharing applications [13]. Its goal is to act as a middleman for music rights owners and legal online music distributors like iTunes [14] and specific peer-to-peer (P2P) networks. Peer Impact is a P2P network for legitimate multimedia distribution using different digital rights management (DRM) techniques [15]. Also centralized content-exchange platforms like Guba [16] and Soapbox [17] employ (audio) fingerprinting techniques to prevent copyrighted (video)material from being uploaded to their platforms. The use of fingerprinting in P2P for legal music distribution was presented by Kalker *et al.* in their Music2Share paper [18].

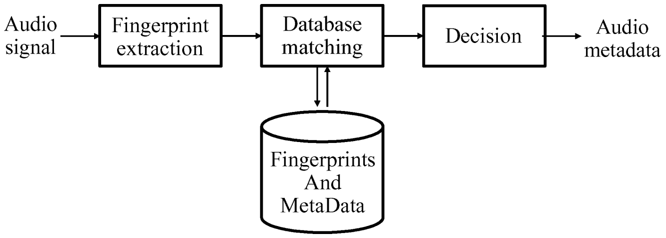


Fig. 1. Using fingerprints for music identification: the extracted audio fingerprint is matched against a database with precomputed fingerprints and metadata.

A fingerprinting system for identification consists of two phases: the enrollment phase and the identification phase. 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 Fig. 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, calling for approximate database matching procedures.

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 extent. 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 paper, 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.

Fig. 2(a) schematically shows the procedure proposed in this paper for estimating the SNR of compressed content using fingerprinting technology. After the song on the Internet has been

Manuscript received November 17, 2006; revised September 18, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Malcolm Slaney.

The authors are with the Information and Communication Theory Group, Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, 2600 GA Delft, The Netherlands (e-mail: p.j.o.doets@tudelft.nl; r.l.lagendijk@tudelft.nl).

Digital Object Identifier 10.1109/TASL.2007.911716

1558-7916/$25.00 © 2008 IEEE

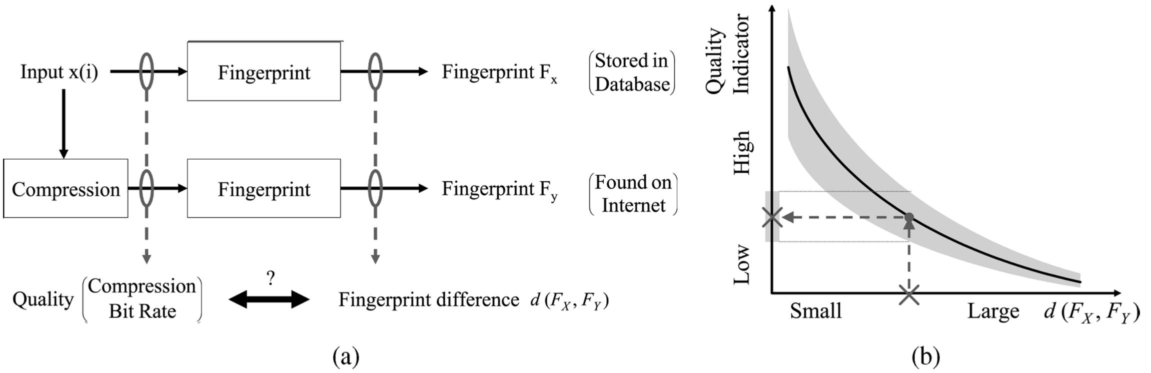


Fig. 2. Using fingerprints for music quality assessment. (a) Relating differences in audio fingerprints of two versions of the same recording, X and Y, to differences in perceptual quality of these recordings. (b) Example relationship between fingerprint differences and a quality indicator (compression bit rate).

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, $d(F_X, F_Y)$. Fig. 2(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. In this way, the fingerprints are used in a reduced-reference quality estimation; the fingerprint is the reduced reference. This in contrast to a full-reference quality estimation, where original audio file is fully available

At first sight, there are several alternatives to obtain the quality of compressed audio, e.g., the bit rate from the compressed audio file header and perceptual quality assessment algorithms [19]. The bit rate, however, like other metadata is unreliable. The bit rate is not a required parameter for decoding in every audio compression format (e.g., Ogg Vorbis [20]) and therefore not always present. Furthermore, the quality of the compressed audio content is a result of the selected compression bit rate, within the limits and settings of the specific implementation. Even compressing the same song with the same algorithm at the same bit rate 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 bit rate.

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 [21]–[24], some of which are used in the perceptual audio quality (PEAQ) measure adopted by the ITU [19]. These algorithms use elaborate psychoacoustic 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 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 paper is organized therefore as follows. Section II provides an overview of fingerprinting algorithms described in literature. Three algorithms which are considered representative for the field are reviewed. In Section III, 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 IV that relate the bit rate used for compression, and the resulting SNR, to the distance between the fingerprints. Section V draws conclusions and outlines directions for future research.

II. Audio Fingerprinting Algorithms

In the last decade, several fingerprinting systems have been developed. Cano *et al.* present a good survey of fingerprinting algorithms [1]. 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.

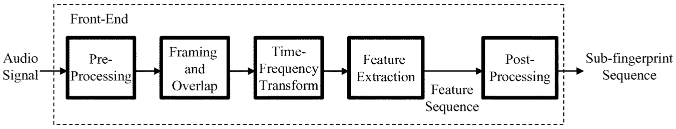


Fig. 3. Fingerprint extraction procedure.

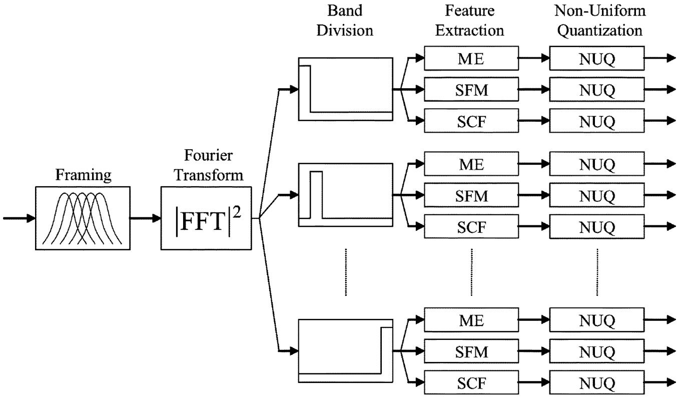


Fig. 4. Fingerprint extraction stage of Cefriel SSD [7].

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. Fig. 3 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 low-pass 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 postprocessing step. The compact representation of the time–frequency features of a single frame is called a subfingerprint. Due to the large overlap, subsequent subfingerprints are (strongly) correlated and vary slowly in time. The fingerprint of a song consists of a sequence of subfingerprints, which are stored in a database. A song-fragment is identified by matching a sequence of subfingerprints, 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 [1]. Based on the information used for extracting the feature sequence, we have divided fingerprinting algorithms into three categories [25]. 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 IV.

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 offline training.

A. Systems That Use Features Based on a Single Band

Shazam uses the locations of peaks in the spectrogram to represent the fingerprint [2]. This algorithm does not reflect the distortions related to compression, especially at medium and high bit rates. Özer *et al.* use periodicity estimators and a singular value decomposition of the Mel frequency cepstrum coefficient (MFCC) matrix [3]. Sukkittanon and Atlas propose frequency modulation features [4]. These papers do not address the response to compression. MusicDNA uses global mean and standard deviation of the energies within 15 subbands of 15 s of music, thus creating a 30-dimensional vector [5]. 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) [6], [7]. 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 paper, we refer to this algorithm by the abbreviation SSD (spectral shape descriptors).

Fig. 4 shows the SSD fingerprinting algorithm proposed by Mapelli *et al.* [7]. 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 $\hat{X}(n,k)$

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

The extracted features are the mean energy (ME), the SFM, and the SCF. We follow the approach in [6] 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 $x(i)$, $i = 1,\ldots,N$ as

$$M_a\left(x(i)\right) = \frac{1}{N}\sum_{i=1}^{N} x(i) \tag{2}$$

and the geometric mean as

$$M_g\left(x(i)\right) = \sqrt[N]{\prod_{i=1}^{N} x(i)}. \tag{3}$$

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

$$\begin{aligned}\text{Feat}(m,n,1) &= ME(n,m)\\ &= M_a\left(S(n,k)\right),\ k \in \mathcal{K}_m \qquad (4)\\ \text{Feat}(m,n,2) &= SFM(n,m)\\ &= 10\log_{10}\left(\frac{M_g\left(S(n,k)\right)}{M_a\left(S(n,k)\right)}\right),\ k \in \mathcal{K}_m \qquad (5)\\ \text{Feat}(m,n,3) &= SCF(n,m)\\ &= 10\log_{10}\left(\frac{\max\left(S(n,k)\right)}{M_a\left(S(n,k)\right)}\right),\ k \in \mathcal{K}_m \qquad (6)\end{aligned}$$

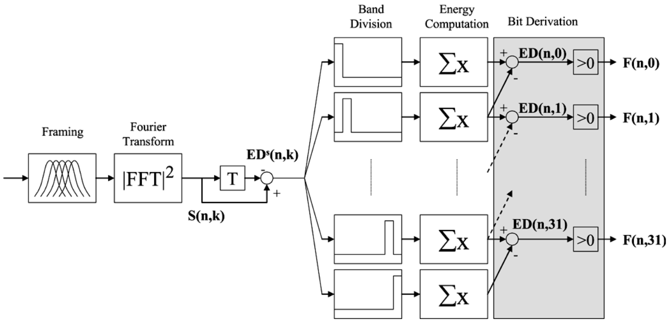


Fig. 5. Fingerprint extraction stage of Philips' PRH [8].

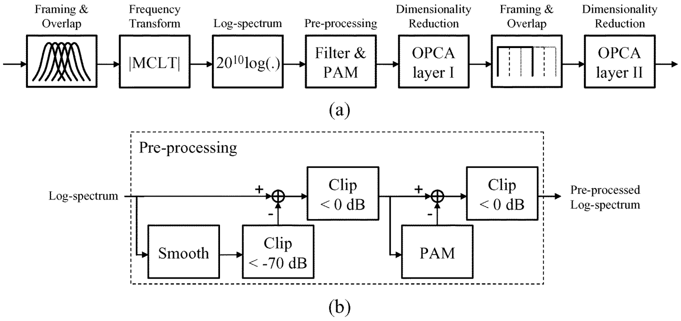


Fig. 6. Microsoft's Robust Audio Recognition Engine (RARE) [10]. (a) Fingerprint extraction. (b) Preprocessing.

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 nonuniform quantizer (NUQ). The fingerprint is thus defined as the quantization level index of each feature of the three features

$$F(n,m,p) = \text{NUQ}_p\left(\text{Feat}(m,n,p)\right), \quad p = 1,2,3. \tag{7}$$

The distance between two fingerprint blocks is computed using the mean square error (MSE)

$$\text{MSE} = \frac{1}{3MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\sum_{p=0}^{2}\left(F_X(m,n,p) - F_Y(m,n,p)\right)^2. \tag{8}$$

B. Systems That Use Features Based on Multiple Subbands

Philips' robust hash (PRH) uses the sign of the difference between energies in Bark-scaled frequency bands [8]. While it is reported to be highly robust against distortions [8], the difference between fingerprints of original and compressed content also reflects compression artifacts [26].

Fig. 5 shows an overview of the fingerprint extraction stage of the Philips system [8].[1] As in the SSD algorithm, features are extracted from strongly overlapping periodograms. To extract an M-bit subfingerprint for every frame, $M+1$ nonoverlapping frequency bands are selected from the periodogram. The difference between spectral values in the periodogram estimates [cf. (1)] for frame n and $n-1$, respectively, is computed as

$$ED^s(n,k) = S(n,k) - S(n-1,k). \tag{9}$$

Then, the energy difference between two neighboring subbands $ED(n,m)$ is computed as

$$ED(n,m) = \sum_{k\in\mathcal{K}_m} ED^s(n,k) - \sum_{k\in\mathcal{K}_{m+1}} ED^s(n,k). \tag{10}$$

Denoting the energy of frequency band m of frame n by

$$E^b(n,m) = \sum_{k\in\mathcal{K}_m} S(n,k) \tag{11}$$

it is easy to see that $ED(n,m)$ is equal to the difference between energies between successive frames and neighboring frequency bands

$$\begin{aligned}ED(n,m) = {} & E^b(n,m) - E^b(n,m+1)\\ & - \left(E^b(n-1,m) - E^b(n-1,m+1)\right).\end{aligned} \tag{12}$$

The bits of the subfingerprint are then derived from $ED(n,m)$ as follows:

$$F(n,m) = \begin{cases}1, & ED(n,m) > 0\\ 0, & ED(n,m) \le 0\end{cases} \tag{13}$$

where $F(n,m)$ denotes the mth bit of subfingerprint n (i.e., the fingerprint of frame n).

The distance between two realizations $FP_X(n,m)$ and $FP_Y(n,m)$ is computed as the bit-error rate (BER)

$$\text{BER} = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} F_{\text{diff}}(n,m) \tag{14}$$

where

$$F_{\text{diff}}(n,m) = \text{XOR}\left(F_X(n,m), F_Y(n,m)\right). \tag{15}$$

C. Systems Using Optimized Subband- or Frame-Combinations

Batlle *et al.* use hidden Markov models (HMMs) to describe their fingerprint [27]. 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" [9]. 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) [10]. 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.

Fig. 6(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 preprocessed to remove the

[1] In order to create a stochastic model in Section III, the time delay operation T is shifted forward yielding the equivalent arrangement (compare to [8, Fig. 1]).

effects of equalization and volume adjustment. A second preprocessing step removes the nonaudible frequency components from the spectrum based on a simple psycho-acoustic model (PAM) [28]. The entire preprocessing procedure is shown in Fig. 6(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, offline 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 preprocessed log-power spectra of the training data. The first OPCA projection is based on the preprocessed 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.

III. 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 [11], [13]. Second, it is well documented [8], 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 IV, we sketch a relation between the MSE and SNR for the log-magnitude spectrum for uncorrelated signals. This relation is easily extendible to the root mean square (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

$$y(i) = x(i) + w(i). \tag{16}$$

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 (13)]

$$\begin{aligned} P_e &= Pr\left[F_X(n,m) \neq F_Y(n,m)\right] \\ &= Pr\left[ED_X(n,m) \leq 0, ED_Y(n,m) > 0 \right. \\ &\qquad \left. \vee \quad ED_X(n,m) > 0, ED_Y(n,m) \leq 0\right]. \end{aligned} \tag{17}$$

Section III-A derives an expression between the SNR and P_e for the case that $x(i)$ is an uncorrelated signal. Section III-B extends this model to correlated signals $x(i)$. Section III-C uses the model from Section III-B to predict the behavior for music. Finally, Section III-D addresses the problem of the large variance in $d(F_X, F_Y)$ for a given bit rate or SNR level and proposes a modified distance measure to reduce this variance.

A. Uncorrelated Signals

We split the calculation of P_e into two parts. First, using (17), the following equation expresses P_e in terms of variances of $ED_X(n,m)$ and $ED_Y(n,m) - ED_X(n,m)$:

$$P_e = \frac{1}{\pi} \arctan\left(\sqrt{\frac{\mathrm{VAR}\left[ED_Y(n,m) - ED_X(n,m)\right]}{\mathrm{VAR}\left[ED_X(n,m)\right]}}\right). \tag{18}$$

This relation is based on Theorem 1 in Appendix I. Here, we assume that $ED_X(n,m)$ and $ED_Y(n,m)$ are drawn from normal distributions and have mean value zero. After the derivations, we motivate this assumption. Furthermore, Theorem 1 is based on the assumption that the signal and noise contributions $ED_X(n,m)$ and $ED_Y(n,m) - ED_X(n,m)$ are mutually uncorrelated.

In the next step, we have to relate $\mathrm{VAR}[ED_Y(n,m) - ED_X(n,m)]$ and $\mathrm{VAR}[ED_X(n,m)]$ 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 (1), (9), and (10), but now for the model in (16). First, the short-time Fourier transform $\hat{Y}(n,k)$ is computed for each frame n

$$\hat{Y}(n,k) = \hat{X}(n,k) + \hat{W}(n,k). \tag{19}$$

Second, the PSD is estimated using the periodogram

$$\begin{aligned} S_Y(n,k) &\triangleq \frac{1}{L}\left|\hat{Y}(n,k)\right|^2 \\ &= \frac{1}{L}\left(\left|\hat{X}(n,k)\right|^2 + \left|\hat{W}(n,k)\right|^2 \right. \\ &\qquad \left. + 2\mathrm{Re}\left(\hat{X}(n,k)\overline{\hat{W}(n,k)}\right)\right) \\ &= S_X(n,k) + S_W(n,k) + 2\mathrm{Re}\left(S_{XW}(n,k)\right) \end{aligned} \tag{20}$$

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

$$\begin{aligned} ED^s_Y(n,k) &\triangleq S_Y(n,k)-S_Y(n-1,k) \\ &= S_X(n,k)+S_W(n,k)+2\mathrm{Re}\left(S_{XW}(n,k)\right) \\ &\quad -\left(S_X(n-1,k)+S_W(n-1,k)\right. \\ &\qquad \left.+2\mathrm{Re}\left(S_{XW}(n-1,k)\right)\right) \\ &= ED^s_X(n,k)+ED^s_W(n,k)+2Q^s(n,k) \qquad (21) \end{aligned}$$

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

$$\begin{aligned} ED_Y(n,m) &\triangleq \sum_{k\in\mathcal{K}_m} ED^s_Y(n,k) - \sum_{k\in\mathcal{K}_{m+1}} ED^s_Y(n,k) \\ &= ED_X(n,m)+ED_W(n,m)+2Q(n,m) \qquad (22) \end{aligned}$$

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 (22), we obtain the following expression for the numerator under the square root in (18):

$$ED_Y(n,m)-ED_X(n,m) = ED_W(n,m)+2Q(n,m). \qquad (23)$$

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

$$\begin{aligned} &\mathrm{VAR}\left[ED_Y(n,m) - ED_X(n,m)\right] \\ &\quad = \mathrm{VAR}\left[ED_W(n,m)\right] + 4\mathrm{VAR}\left[Q(n,m)\right] \qquad (24) \end{aligned}$$

In Appendix III, we show that, if we assume $x(i)$ and $w(i)$ to be normally distributed, the variances in (24) are proportional to VAR$[ED_X(n,m)]$

$$\begin{aligned} &\mathrm{VAR}\left[ED_Y(n,m) - ED_X(n,m)\right] \\ &\quad = \left(\frac{\sigma_W^4}{\sigma_X^4} + 2\frac{\sigma_W^2}{\sigma_X^2}\right)\mathrm{VAR}\left[ED_X(n,m)\right]. \qquad (25) \end{aligned}$$

Finally, the combination of (18) and (25) results in

$$P_e = \frac{1}{\pi}\arctan\left(\sqrt{\frac{\sigma_W^4}{\sigma_X^4} + 2\frac{\sigma_W^2}{\sigma_X^2}}\right). \qquad (26)$$

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. Equation (26) was derived for Gaussian independent and identically distributed (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 (52).

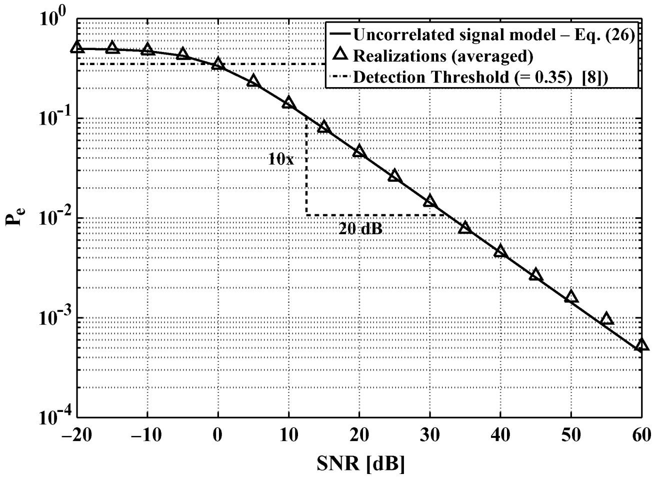


Fig. 7. Analytical relation between SNR and P_e for the PRH.

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.

Fig. 7 shows the SNR $-$ P_e relationship for model of (26) 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 versus P_e relation is a straight line. For these small distortions, the P_e as formulated in (26) is approximately inversely proportional to σ_X/σ_W

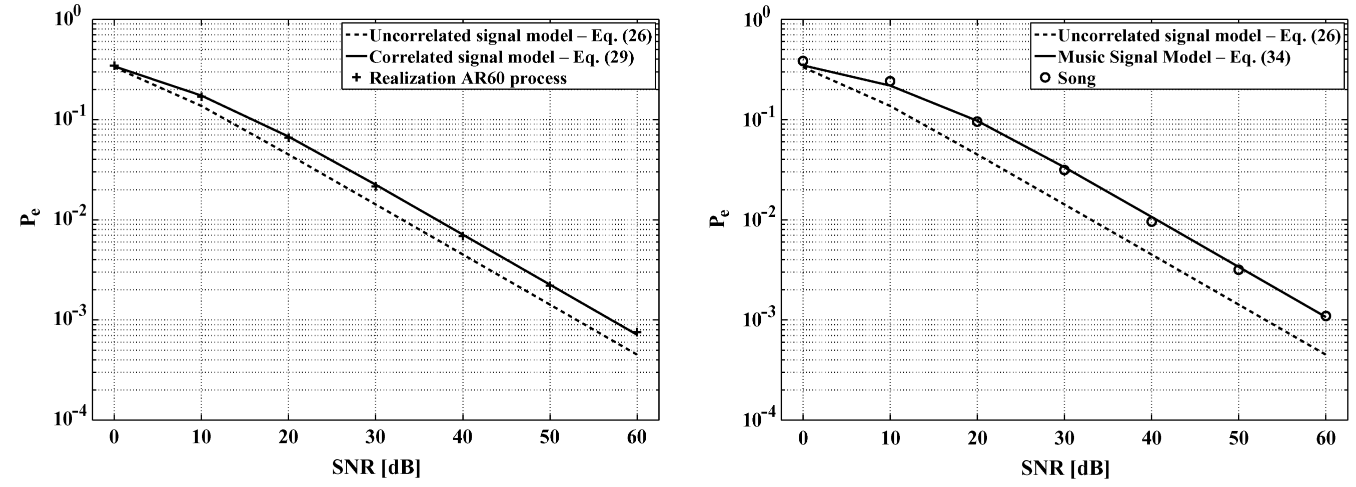
$$P_e \approx \frac{1}{\pi}\arctan\left(\sqrt{2}\frac{\sigma_W}{\sigma_X}\right) \approx \frac{\sqrt{2}}{\pi}\frac{\sigma_W}{\sigma_X}. \qquad (27)$$

In practice, this means that for a 20-dB increase of SNR, the P_e is expected to drop by a factor of 10. The region in the curve showing the "linear" SNR-P_e relation is of particular interest, since most audio compression algorithms 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.

B. Correlated Signals

The model outlined in Section III-A 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.



(a) (b)

Fig. 8. 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) and (b).

An extension to the model of (27) 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 corresponding to the signal and noise in band m and $m+1$ can be approximated by

$$P_e(m) \approx \frac{1}{\pi} \arctan\left(\sqrt{\frac{\sigma^4_{W_{m:m+1}}}{\sigma^4_{X_{m:m+1}}} + 2\frac{\sigma^2_{W_{m:m+1}}}{\sigma^2_{X_{m:m+1}}}}\right). \quad (28)$$

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

$$\begin{aligned} P_e(m) &\approx \frac{1}{\pi} \arctan\left(\sqrt{\frac{\sigma^4_W}{\sigma^4_{X_{m:m+1}}} + 2\frac{\sigma^2_W}{\sigma^2_{X_{m:m+1}}}}\right) \\ &= \frac{1}{\pi} \arctan\left(\sqrt{\frac{\sigma^4_W}{\sigma^4_X}\frac{\sigma^4_X}{\sigma^4_{X_{m:m+1}}} + 2\frac{\sigma^2_W}{\sigma^2_X}\frac{\sigma^2_X}{\sigma^2_{X_{m:m+1}}}}\right). \quad (29) \end{aligned}$$

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

$$\sigma^2_X = \sum_{m=0}^{M} N_m \sigma^2_{X_m} / \sum_{m=0}^{M} N_m. \quad (30)$$

In practical systems like the PRH, the subbands do not cover the entire spectral range; (30) 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

$$P_e = \frac{1}{M} \sum_{m=0}^{M-1} P_e(m). \quad (31)$$

The model in (29) assumes that the PSD of the signal is flat within two subsequent bands and the model in (31) that the probabilities are independent over m. Equation (31) again results in a more complicated $\arctan(\cdot)$ relation, since

$$\arctan(a) + \arctan(b) = \arctan\left(\frac{a+b}{1-ab}\right) + \begin{cases} 0 & ab < 1 \\ \pi & ab > 1, a > 0 \\ -\pi & ab > 1, a < 0. \end{cases} \quad (32)$$

As an illustration, Fig. 8(b) shows the modeled and experimental SNR-BER curves for a 60th order autoregressive (AR) process. The coefficients were obtained by fitting the AR model onto a frame of real music. This example shows a perfect fit.

C. Music

Previous sections considered synthetic signal models. Here, we will extend the analysis to real audio signals. Although the model in (29) and (31) assumes a stationary signal, it does reflect the influence of a nonflat 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 nonstationarity 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^2_X/σ^2_W by

$$P_e = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} P_e(n,m) \quad (33)$$

where

$$P_e(n,m) = \frac{1}{\pi} \arctan \left(\sqrt{2 \cdot \frac{\sigma_W^2}{\sigma_X^2} \cdot \frac{\sigma_X^2}{\sigma_{X_{n,m}}^2}} \right). \tag{34}$$

Here, $\sigma_{X_{n,m}}^2 / \sigma_W^2$ represents the SNR level corresponding to fingerprint bit $F(n,m)$. Equations (10) and (13) 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 (10). 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 (34) 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

$$\sigma_{X_{n,m}}^2 = \max_{i,j} E^b(i,j) \quad \begin{array}{l} i = n-r, \ldots, n, \ldots, n+r \\ j = m, m+1 \end{array}. \tag{35}$$

In our experiments, we used $r = 13$. The predicted and experimental curves for a 3-s music segment is shown in Fig. 8(b).

D. 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 to a semitone. If 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 therefore generate fingerprint bits which are unreliable. This is easily illustrated by setting $\sigma_{X_{n,m}}^2 \ll \sigma_X^2$ in the model in (34), 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 nonstationarity 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 nonstationary 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 Section III-C 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 (29). This estimation can be 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 only use those fingerprint bits $F(n,m), \{n,m\} \in \mathcal{L}$ to compute the distance between the fingerprints, such that the $(\text{SNR}, P_{e,est})$ behaves approximately the same as the theoretical (SNR, P_e)-curve for white noise, i.e.,

$$\mathcal{L} \quad \text{s.t.} \quad P_{e,est}(\text{SNR}|\mathcal{L}) \approx P_e(\text{SNR}) \tag{36}$$

where $P_{e,est}$ denotes the average probability of bit error estimated for a specific song, obtained using the model in (34) and (35). 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 $\mathcal{L}$

$$\text{BER}_W = \frac{1}{|\mathcal{L}|} \sum_{\{n,m\} \in \mathcal{L}} F_{\text{diff}}(n,m) \tag{37}$$

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

We now focus on how to obtain the set of usable fingerprint bits $\mathcal{L}$. Using (29), the behavior of a small fragment of $2r+1$ frames can be predicted from the spectrum. Let us denote the averaged behavior within a number of frames explicitly by the function

$$P_{e,est}(\text{SNR}, \mathcal{L}) = \frac{1}{\mathcal{L}} \sum_{\{n,m\} \in \mathcal{L}} P_e(n,m,\text{SNR}). \tag{38}$$

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

$$\mathcal{L} \quad \text{s.t.} \quad P_{e,est}(\text{SNR}|\mathcal{L}) \approx P_e(\text{SNR}). \tag{39}$$

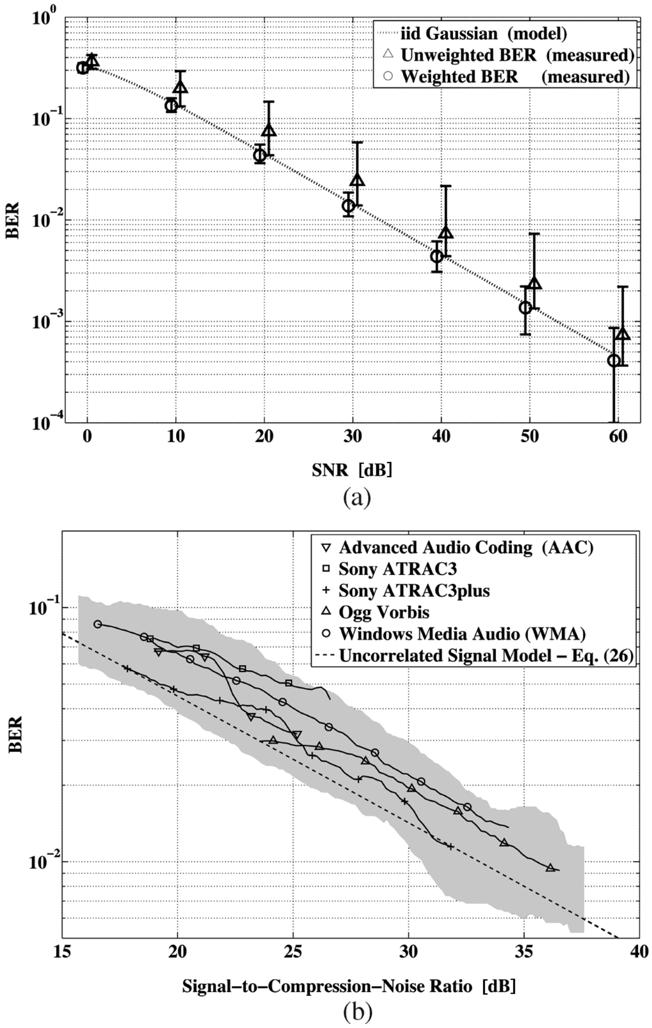
The set $\mathcal{L}$ is obtained in the following iterative way. Since the strongest spectral peaks generate the most reliable bits, in iteration i we select the bits corresponding to the $|\mathcal{L}_i|$ strongest spectral components. One can see that for a given SNR level, adding a spectral component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

$$P_{e,est}(\text{SNR}|\mathcal{L}_i) < P_{e,est}(\text{SNR}|\mathcal{L}_{i+1}) \quad \mathcal{L}_i \subset \mathcal{L}_{i+1}. \tag{40}$$

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

$$L_i = \int_{-\infty}^{\text{SNR}_{\max}} \{\log(P_e(\text{SNR})) - \log(P_{e,est}(\text{SNR}, \mathcal{L}_i))\}^2 \, d\text{SNR}. \tag{41}$$

The cost function expressed the distance between the two curves $P_{e,est}(\text{SNR}|\mathcal{L})$ and $P_e(\text{SNR})$. Due to the increasing nature of (40), 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 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 $\lim_{\text{SNR} \to -\infty} P_{e,est} = 0.5$.



(a)

(b)

Fig. 9. (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. Error bars indicate lower and upper 10% BER values for a given SNR. The curves have been shifted slightly horizontally in order not to overlap. The i.i.d. model is shown as a reference ("$-\ -$"). (b) SNR-BER relation for nine songs, comparing the behavior of the PRH algorithm in its original form [8] for five different compression algorithms: AAC ("▽"), Sony ATRAC ("□"), Sony ATRAC3plus ("+"), Ogg Vorbis ("△") and WMA ("o"), and the curve for the uncorrelated signal model (26).

Fig. 9(a) shows the result of applying this strategy to music and additive noise. The variance in BER for a given SNR level is greatly reduced.

IV. Experiments Using Music

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

Section IV-A discusses the details of the comparison process. Sections IV-C and IV-B compare the algorithms in a compression bit rate-versus-$d(F_X, F_Y)$ and a signal-to-compression-noise (SNCR)-versus-$d(F_X, F_Y)$ setting.

A. Enabling Algorithmic Comparison

The fingerprinting systems described in Section II 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 s;
- 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, and 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 I compares the settings for the different systems.

We have used 275 song fragments of 40 s 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 [10]. For each of these 100 song fragments, we have generated nine distorted versions. These distortions are mainly nonlinear 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 [29]. The selected bit rates for MP3 compression range from 32–256 kbit-per-second (kb/s) using constant bit rate. To test the variability over different compression algorithms, we have conducted a small-scale experiment shown in Fig. 9(b) (for the PRH algorithm only) with a number of different, widely used audio codecs, including Advanced Audio Coding (AAC) [30], Sony ATRAC(plus) [31], [32], Ogg Vorbis [20], and Windows Media Audio (WMA) [33]. 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.[2] 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}.

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

TABLE I
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 per 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

B. Experimental Relation Between Bit rate and $d(F_X, F_Y)$

Fig. 10 compares the relation between compression bit rate 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 bit rate from the fingerprint differences, the spread in the curves for a given bit rate 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 bit rate estimation.

C. 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. Fig. 11 shows the signal-to-compression-noise for the three algorithms. Fig. 11(b) and (c) 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 Fig. 11. It is quite clear that all curves have approximately the same gradient in the SNR plots. Although the SNR $- P_e$ in (26) was derived for an uncorrelated signal in the presence of additive, uncorrelated noise, the experimental SNCR-$d(F_X, F_Y)$ for all three algorithms follow the $\arctan(\cdot)$-regime. RARE and SSD make use of the log-magnitude spectrum. In Appendix IV, we roughly outline the relation between MSE and the SNR for i.i.d. Gaussian data.

Due to the fact that in compression the bit rates are chosen, and the SNR levels are a result of the selected bit rate, 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.

V. CONCLUSION AND DISCUSSION

A. 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

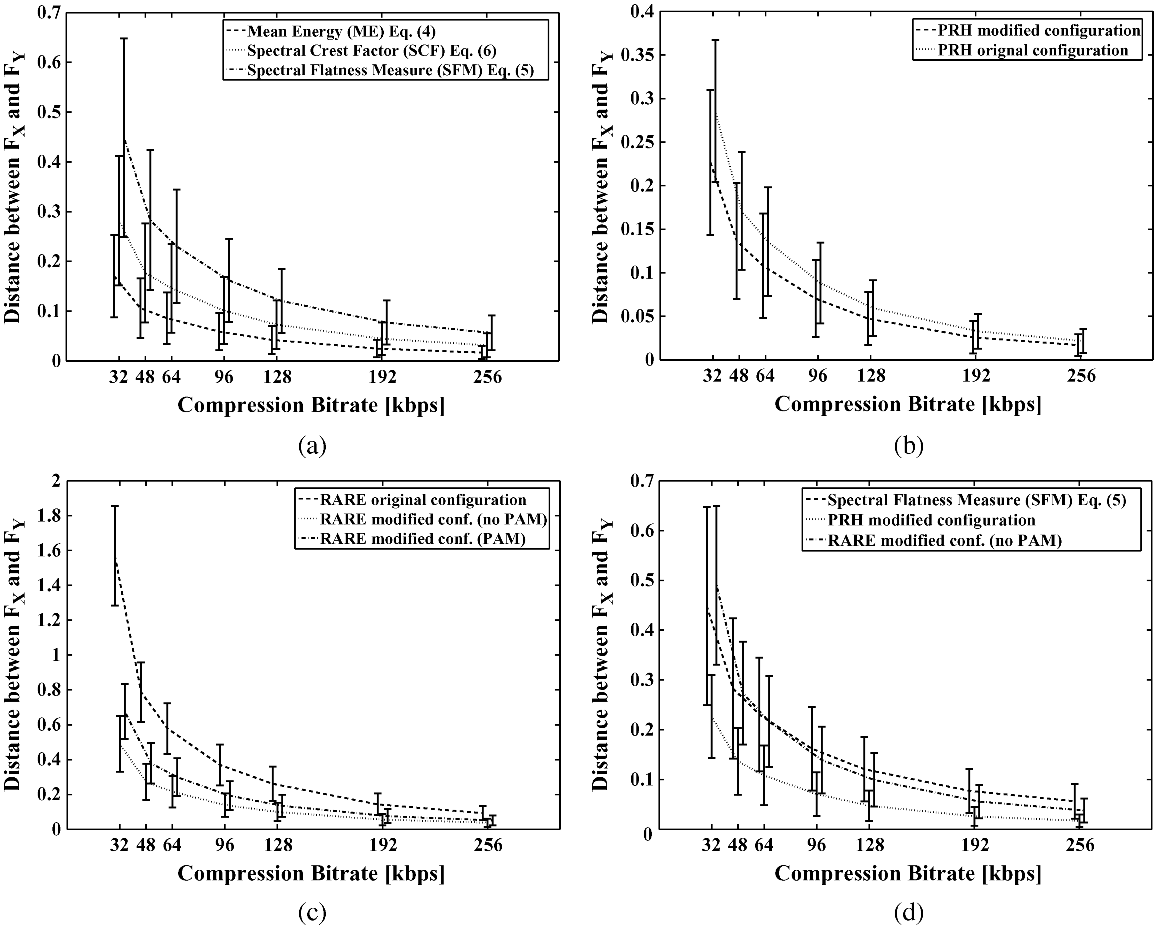


Fig. 10. Compression bit rate versus fingerprint differences. The curves have been shifted such that there is no overlap. (a) The features in the SSD algorithm: from top to bottom: energy ("— —"), SCF ("· · ·"), SFM ("—."). (b) PRH: modified ("— —"), original ("· · ·"). (c) RARE: original ("— —"), modified, no psycho-acoustic model ("· · ·"), modified, using a psycho-acoustic model ("—."). (d) comparison between the modified versions of SFM ("— —"), PRH ("· · ·"), RARE ("—.").

fingerprint differences behave in a comparable fashion as a function of SNR or compression bit rate. This behavior matches the behavior predicted by the models presented in Section III. 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 difference 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 relatively large. This limits the classification possibilities to three, maybe five, 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.

B. 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 HAS reacts to sounds. Spectral components that fall below this masking threshold are not audible and are therefore considered irrelevant.

The 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.

In the first approach, the fingerprint extraction procedure outline in Fig. 3 is changed to estimate the sound representation inside the human ear using the masking threshold, shown in Fig. 12(a). Spectral components that exceed the masking threshold are scaled by it; components that fall below the

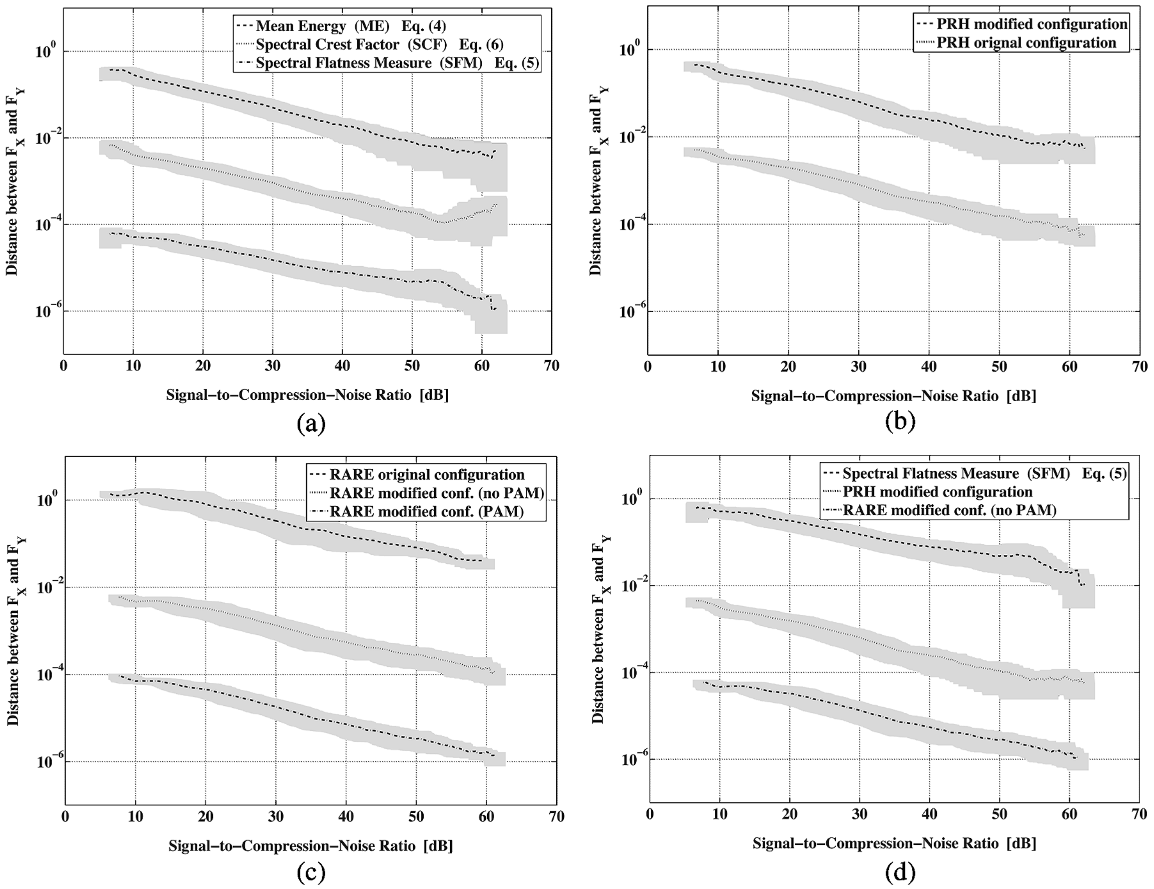


Fig. 11. Compression SNCR versus fingerprint distances. The lines mark the average behavior; the shaded areas indicate the spread. The curves have been scaled such that there is no overlap. (a) The features in the SSD algorithm: from top to bottom: energy ("— —", not scaled), SCF ("· · ·", scaled by factor 10^{-2}), SFM ("—.", scaled by factor 10^{-4}), (b) PRH: modified ("— —", not scaled), original ("· · ·", scaled by factor 10^{-2}). (c) RARE: original ("— —", not scaled), modified, no psycho-acoustic model ("· · ·", scaled by factor 10^{-2}), modified, using a psycho-acoustic model ("—.", scaled by factor 10^{-4}). (d) Comparison between the modified versions of SFM ("——", not scaled), PRH ("· · ·", scaled by factor 10^{-}), RARE ("—.", scaled by factor 10^{-4}).

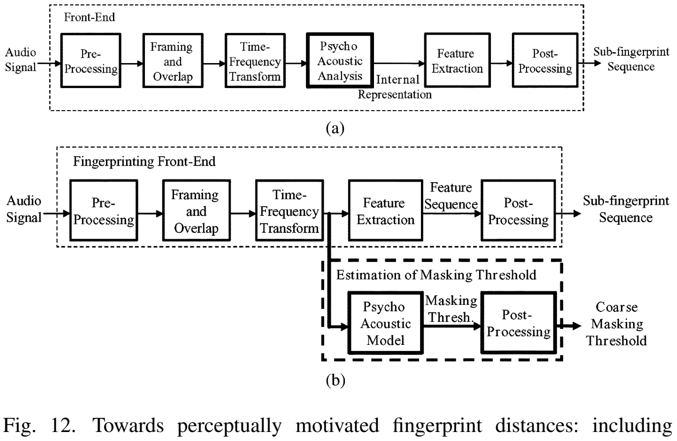


Fig. 12. Towards perceptually motivated fingerprint distances: including psycho-acoustical models (a) in the audio fingerprint extraction stage and (b) parallel to the fingerprint extraction stage.

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 Fig. 12(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 [34]. 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 paper. 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 be 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.

C. 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.* [36]. 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 bit rate 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.

Appendix I
Relation Between $ED_X(n,m)$, $ED_Y(n,m)$, and P_e

Equation (18) 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 (18).

Theorem 1: 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

$$\begin{aligned} P_e &= Pr[A \leq 0, C > 0 \quad \vee \quad A > 0, C \leq 0] \\ &= \frac{1}{\pi}\arctan\left(\frac{\sigma_B}{\sigma_A}\right) \\ &= \frac{1}{\pi}\arctan\left(\sqrt{\frac{\text{VAR}[C-A]}{\text{VAR}[A]}}\right). \end{aligned} \tag{42}$$

Proof: Due to symmetry, $Pr[C > 0|A \leq 0] = Pr[C \leq 0|A > 0]$ and $Pr[A \leq 0] = Pr[A > 0] = 1/2$. Therefore

$$\begin{aligned} P_e &= Pr[A \leq 0, C > 0 \quad \vee \quad A > 0, C \leq 0] \\ &= Pr[C > 0|A \leq 0]Pr[A \leq 0] \\ &\quad + Pr[C \leq 0|A > 0]Pr[A > 0] \\ &= Pr[C \leq 0|A > 0] \\ &= Pr[B \leq -A|A > 0]. \end{aligned} \tag{43}$$

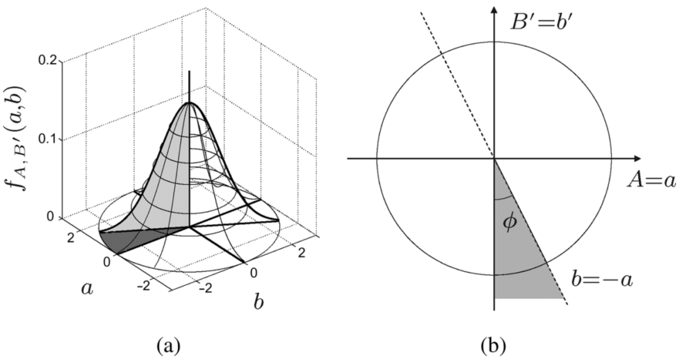


Fig. 13. Probability density function $f_{A,B'}(a,b)$. (a) 3-D visualization. (b) projection onto the ground plane (contour line).

Define $\alpha = \sigma_B/\sigma_A$ and introduce the normalized version of B, *viz.* $B' = (\sigma_A/\sigma_B) \cdot B = (1/\alpha) \cdot B$, $B' \in \mathcal{N}(0,\sigma_A^2)$. Due to the scaling factor α, the joint-probability density function (pdf) $f_{A,B'}(a,b)$ is rotation symmetric with respect to the origin, as illustrated in Fig. 13(a). P_e is related to $f_{A,B'}(a,b)$ by

$$\begin{aligned} P_e &= Pr[B \leq -A|A > 0] \\ &= \int_0^\infty \int_{-\infty}^{-a} f_{A,B}(a,b)dbda \\ &= \int_0^\infty \int_{-\infty}^{-\alpha a} f_{A,B'}(a,b)dbda. \end{aligned} \tag{44}$$

The angle between the vertical axis and the integration boundary is denoted by the angle ϕ, where $\phi = \arctan(\alpha)$, as illustrated in Fig. 13(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,[3] the probability P_e is proportional to the $0 \leq \phi < \pi$. We can now express P_e in terms of ϕ as follows:

$$P_e = \int_0^\infty \int_{-\infty}^{-\alpha a} f_{A,B'}(a,b)dbda = \frac{\phi}{\pi} = \frac{1}{\pi}\arctan\left(\frac{\sigma_B}{\sigma_A}\right).$$

■

Appendix II
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 III-A to derive (24).

Theorem 2: The variables $ED_W(n,m)$ and $Q(n,m)$ are mutually uncorrelated, and as a result

$$\begin{aligned} &\text{VAR}\left[ED_Y(n,m) - ED_X(n,m)\right] \\ &\quad = \text{VAR}\left[ED_W(n,m) + 2Q(n,m)\right] \\ &\quad = \text{VAR}\left[ED_W(n,m)\right] + 4\text{VAR}\left[Q(n,m)\right]. \end{aligned} \tag{45}$$

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 $\text{COV}[ED_W^s(n,k), Q^s(n,k+l)] = 0$.

[3] The result in (45) 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.

Using the short-hand notation $R_{\hat{X}}(n,k) = \mathrm{Re}(\hat{X}(n,k))$ and $I_{\hat{X}}(n,k) = \mathrm{Im}(\hat{X}(n,k))$, we can express $Q^s(n,k)$ in terms of the two input components $\hat{X}(n,k)$ and $\hat{W}(n,k)$

$$\begin{aligned} Q^s(n,k) &= \frac{1}{L}\mathrm{Re}\left(\hat{X}(n,k)\overline{\hat{W}(n,k)}\right. \\ &\qquad \left. -\hat{X}(n-1,k)\overline{\hat{W}(n-1,k)}\right) \\ &= \frac{1}{L}\left(R_{\hat{X}}(n,k)R_{\hat{W}}(n,k)\right. \\ &\qquad \left. -R_{\hat{X}}(n-1,k)R_{\hat{W}}(n-1,k)\right) \\ &\quad + \frac{1}{L}\left(I_{\hat{X}}(n,k)I_{\hat{W}}(n,k)\right. \\ &\qquad \left. -I_{\hat{X}}(n-1,k)I_{\hat{W}}(n-1,k)\right). \end{aligned} \tag{46}$$

The covariance can now be computed as

$$\begin{aligned} &\mathrm{COV}\left[ED^s_W(n,k), Q^s(n,k+l)\right] \\ &= \frac{1}{L}\left(\mathrm{COV}\left[ED^s_W(n,k), R_{\hat{X}}(n,k+l)R_{\hat{W}}(n,k+l)\right]\right. \\ &\qquad + \mathrm{COV}\left[ED^s_W(n,k), I_{\hat{X}}(n,k+l)I_{\hat{W}}(n,k+l)\right] \\ &\qquad - \mathrm{COV}\left[ED^s_W(n,k), R_{\hat{X}}(n-1,k+l)\right. \\ &\qquad\qquad \left.\times R_{\hat{W}}(n-1,k+l)\right] \\ &\qquad -\mathrm{COV}\left[ED^s_W(n,k), I_{\hat{X}}(n-1,k+l)\right. \\ &\qquad\qquad \left.\left.\times I_{\hat{W}}(n-1,k+l)\right]\right) \\ &= \frac{1}{L}\left(E\left[ED^s_W(n,k)R_{\hat{W}}(n,k+l)\right]E\left[R_{\hat{X}}(n,k+l)\right]\right. \\ &\qquad +E\left[ED^s_W(n,k)I_{\hat{W}}(n,k+l)\right]E\left[I_{\hat{X}}(n,k+l)\right] \\ &\qquad -E\left[ED^s_W(n,k)R_{\hat{W}}(n-1,k+l)\right] \\ &\qquad \times E\left[R_{\hat{X}}(n-1,k+l)\right] \\ &\qquad - E\left[ED^s_W(n,k)I_{\hat{W}}(n-1,k+l)\right] \\ &\qquad \left.\times E\left[I_{\hat{X}}(n-1,k+l)\right]\right) \\ &= 0. \end{aligned} \tag{47}$$

■

Appendix III
Relation Between $\mathrm{VAR}[ED_W(n,m)]$, $\mathrm{VAR}[Q(n,m)]$, and $\mathrm{VAR}[ED_X(n,m)]$

Equation (25) in Section III-A relates the variance $\mathrm{VAR}[ED_Y(n,m) - ED_X(n,m)]$ to the variance $\mathrm{VAR}[ED_X(n,m)]$.

Theorem 3: The variance $\mathrm{VAR}[ED_Y(n,m) - ED_X(n,m)]$ is proportional to $\mathrm{VAR}[ED_X(n,m)]$ and is equal to

$$\begin{aligned} &\mathrm{VAR}\left[ED_Y(n,m) - ED_X(n,m)\right] \\ &\qquad = \left(\frac{\sigma^4_W}{\sigma^4_X} + 2\frac{\sigma^2_W}{\sigma^2_X}\right)\mathrm{VAR}\left[ED_X(n,m)\right]. \end{aligned} \tag{48}$$

Proof: Theorem 2 expressed the variance on the left-hand side of the equation as

$$\begin{aligned} &\mathrm{VAR}\left[ED_Y(n,m) - ED_X(n,m)\right] \\ &\qquad = \mathrm{VAR}\left[ED_W(n,m)\right] + 4\mathrm{VAR}\left[Q(n,m)\right]. \end{aligned} \tag{49}$$

Since $ED_X(n,m)$, $ED_W(n,m)$, and $Q(n,m)$ are based on summations of $ED^s_X(n,k)$, $ED^s_W(n,k)$, and $Q^s(n,k)$, respectively, over index k, it is sufficient to relate $\mathrm{COV}[ED^s_W(n,k), ED^s_W(n,k+l)]$ and $\mathrm{COV}[Q^s(n,k), Q^s(n,k+l)]$ to $\mathrm{COV}[ED^s_X(n,k), ED^s_X(n,k+l)]$.

In the following, we only consider these covariances. We first express the covariance $\mathrm{COV}[ED^s_X(n,k), ED^s_X(n,k+l)]$ in terms of $R_{\hat{X}}(n,k)$ and $I_{\hat{X}}(n,k)$

$$\begin{aligned} &\mathrm{COV}\left[ED^s_X(n,k), ED^s_X(n,k+l)\right] \\ &= \mathrm{COV}\left[S_X(n,k) - S_X(n-1,k),\right. \\ &\qquad \left. S_X(n,k+l) - S_X(n-1,k+l)\right] \\ &= 2\left(\mathrm{COV}\left[S_X(n,k), S_X(n,k+l)\right]\right. \\ &\qquad \left. -\mathrm{COV}\left[S_X(n,k), S_X(n+1,k+l)\right]\right) \\ &= \frac{2}{L^2}\mathrm{COV}\left[R^2_{\hat{X}}(n,k) + I^2_{\hat{X}}(n,k),\right. \\ &\qquad \left. R^2_{\hat{X}}(n,k+l) + I^2_{\hat{X}}(n,k+l)\right] \\ &\quad - \frac{2}{L^2}\mathrm{COV}\left[R^2_{\hat{X}}(n,k) + I^2_{\hat{X}}(n,k), R^2_{\hat{X}}(n+1,k+l)\right. \\ &\qquad \left. +I^2_{\hat{X}}(n+1,k+l)\right] \\ &= \frac{4}{L^2}\left(\mathrm{COV}\left[R^2_{\hat{X}}(n,k), R^2_{\hat{X}}(n,k+l)\right]\right. \\ &\qquad \left. -\mathrm{COV}\left[R^2_{\hat{X}}(n,k), R^2_{\hat{X}}(n+1,k+l)\right]\right) \\ &= \frac{8}{L^2}\left(\mathrm{COV}\left[R_{\hat{X}}(n,k), R_{\hat{X}}(n,k+l)\right]^2\right. \\ &\qquad \left. -\mathrm{COV}\left[R_{\hat{X}}(n,k), R_{\hat{X}}(n+1,k+l)\right]^2\right). \end{aligned} \tag{50}$$

Here we used two properties of the Fourier transform of an uncorrelated signal: first, the real part $R_{\hat{X}}(n,k)$ and imaginary part $I_{\hat{X}}(n,k)$ 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:

$$\mathrm{COV}\left[X^2_1, X^2_2\right] = 2\mathrm{COV}[X_1, X_2]^2. \tag{51}$$

Since the autocorrelation functions of $R_{\hat{X}}(n,k)$ and $R_{\hat{W}}(n,k)$ are proportional to the variances σ^2_X and σ^2_W, respectively, it is straightforward to relate these to each other

$$\begin{aligned} &\mathrm{COV}\left[R_{\hat{W}}(n,k), R_{\hat{W}}(n+p,k+l)\right] \\ &\qquad = \frac{\sigma^2_W}{\sigma^2_X}\cdot\mathrm{COV}\left[R_{\hat{X}}(n,k), R_{\hat{X}}(n+p,k+l)\right]. \end{aligned} \tag{52}$$

Hence, we can express $\mathrm{COV}[ED^s_W(n,k), ED^s_W(n,k+l)]$ as

$$\begin{aligned} &\mathrm{COV}\left[ED^s_W(n,k), ED^s_W(n,k+l)\right] \\ &\qquad = \frac{\sigma^4_W}{\sigma^4_X}\cdot\mathrm{COV}\left[ED^s_X(n,k), ED^s_X(n,k+l)\right]. \end{aligned} \tag{53}$$

We can now relate $\text{COV}[Q^s(n,k), Q^s(n,k+l)]$ to $\text{COV}[ED_X^s(n,k), ED_X^s(n,k+l)]$

$$\begin{aligned}
&\text{COV}\left[Q^s(n,k), Q^s(n,k+l)\right] \\
&= \frac{4}{L^2}\big(\text{COV}\left[R_{\hat{X}}(n,k), R_{\hat{X}}(n,k+l)\right] \\
&\qquad \times \text{COV}\left[R_{\hat{W}}(n,k), R_{\hat{W}}(n,k+l)\right] \\
&\qquad - \text{COV}\left[R_{\hat{X}}(n,k), R_{\hat{X}}(n+1,k+l)\right] \\
&\qquad \times \text{COV}\left[R_{\hat{W}}(n,k), R_{\hat{W}}(n+1,k+l)\right]\big) \\
&= \frac{4}{L^2}\frac{\sigma_W^2}{\sigma_X^2}\Big(\text{COV}\left[R_{\hat{X}}(n,k), R_{\hat{X}}(n,k+l)\right]^2 \\
&\qquad\quad -\text{COV}\left[R_{\hat{X}}(n,k), R_{\hat{X}}(n+1,k+l)\right]^2\Big) \\
&= \frac{1}{2}\frac{\sigma_W^2}{\sigma_X^2}\text{COV}\left[ED_X^s(n,k), ED_X^s(n,k+l)\right]. \qquad (54)
\end{aligned}$$

Combining (22), (53), and (54) results in

$$\begin{aligned}
&\text{VAR}\left[ED_Y(n,m) - ED_X(n,m)\right] \\
&\qquad = \left(\frac{\sigma_W^4}{\sigma_X^4} + 2\frac{\sigma_W^2}{\sigma_X^2}\right)\text{VAR}\left[ED_X(n,m)\right]. \qquad (55)
\end{aligned}$$

■

Appendix IV
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

$$\begin{aligned}
E[\text{MSE}] &\propto E\left[(FP_Y - FP_x)^2\right] \\
&\propto E\left[\left(\log\left(S_Y(n,k)\right) - \log\left(S_X(n,k)\right)\right)^2\right] \\
&= E[Z^2]
\end{aligned}$$

where $Z = \log(S_Y(n,k)/S_X(n,k))$.

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 $f_{U,V}(u,v)$

$$\begin{aligned}
f_{U,V}(u,v) &= \int_0^{2\pi}\int_0^{2\pi} f_{U,V,\Phi,\Theta}(u,v,\phi,\theta)d\phi d\theta \\
&= \frac{4uv}{\sigma_X^2\sigma_W^2}\exp\left(-\frac{u^2}{\sigma_X^2} - \frac{u^2+v^2}{\sigma_W^2}\right) \\
&\quad \times \sum_{l=0}^{\infty}\frac{1}{(l!)^2}\left(\frac{uv}{\sigma_W^2}\right)^{2l}.
\end{aligned}$$

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 for $f_Z(z)$ 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 pth 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 $n = 2$

$$Li_n(x) = \sum_{k=1}^{\infty}\frac{x^k}{k^n} \qquad |x| \leq 1.$$

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

$$E[Z^2] \approx \frac{1}{3}\pi^2 - 2Li_2(1 - 10^{-\text{SNR}/10}).$$

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

$$E[Z^2] \approx \frac{\ln(10)}{5\ln(2)^2} \cdot \mathrm{SNR} \cdot 10^{-\mathrm{SNR}/10}.$$

On a log-scale this works out into

$$\log_{10}\left(E[Z^2]\right) \approx \log_{10}(\mathrm{SNR}) - \frac{\mathrm{SNR}}{10} + \mathrm{const.}$$

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.

Acknowledgment

The authors like to thank the anonymous reviewers for their helpful comments and suggestions.

References

[1] P. Cano *et al.*, “A review of audio fingerprinting,” *J. VLSI Signal Process.*, vol. 41, no. 3, pp. 271–284, Nov. 2005.

[2] A. Wang, “An industrial strength audio search algorithm,” in *Proc. 4th Int. Conf. Music Inf. Retrieval (ISMIR)*, Oct. 2003, pp. 7–13.

[3] H. Özer, B. Sankur, N. Memon, and E. Anar, “Perceptual audio hashing functions,” *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 12, pp. 1780–1793, Sep. 2005.

[4] S. Sukittanon, L. E. Atlas, and J. W. Pitton, “Modulation-scale analysis for content identification,” *IEEE Trans. Signal Process.*, vol. 52, no. 10, pp. 3023–3035, Oct. 2004.

[5] V. Venkatachalam *et al.*, “Automatic identification of sound recordings,” *IEEE Signal Process. Mag.*, vol. 21, no. 2, pp. 92–99, Mar. 2004.

[6] J. Herre, O. Hellmuth, and M. Cremer, “Scalable robust audio fingerprinting using mpeg-7 content,” in *Proc. 5th IEEE Workshop Multimedia Signal Process. (MMSP)*, Oct. 2002, pp. 165–168.

[7] F. Mapelli, R. Pezzano, and R. Lancini, “Robust audio fingerprinting for song identification,” in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2004, pp. 2095–2098.

[8] J. Haitsma and T. Kalker, “A highly robust audio fingerprinting system,” in *Proc. 3rd Int. Conf. Music Inf. Retrieval (ISMIR)*, Oct. 2002, pp. 107–115.

[9] H. Neuschmied, H. Mayer, and E. Batlle, “Content-based identification of audio titles on internet,” in *Proc. 1st IEEE Int. Conf. Web Delivering Music (WEDELMUSIC)*, Nov. 2001, pp. 96–100.

[10] C. J. C. Burges, J. C. Platt, and S. Jana, “Distortion discriminant analysis for audio fingerprinting,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 165–174, May 2003.

[11] “Gracenote,” 2006 [Online]. Available: http://www.gracenote.com.,

[12] S. Beauget, M. van der Veen, and A. Lemma, “Informed detection of audio watermark for resolving playback speed modifications,” in *Proc. Workshop Multimedia Security (MM&Sec)*, 2004, pp. 117–123.

[13] “Snocap.” [Online]. Available: http://www.snocap.com 2006

[14] “Apple iTunes.” [Online]. Available: http://www.apple.com/itunes

[15] “Peerimpact,” 2006 [Online]. Available: http://www.peerimpact.com

[16] “Guba,” 2007 [Online]. Available: http://www.guba.com

[17] “MSN Soapbox,” 2007 [Online]. Available: http://soapbox.msn.com

[18] T. Kalker *et al.*, “Music2share—Copyright-compliant music sharing in P2P systems,” *Proc. IEEE*, vol. 92, no. 6, pp. 961–970, Jun. 2004.

[19] T. Thiede *et al.*, “PEAQ—The ITU standard for objective measurement of perceived audio quality,” *J. Audio Eng. Soc. (JAES)*, vol. 29, no. 1/2, pp. 3–29, Jan./Feb. 2000.

[20] “Ogg Vorbis Specification,” 2007 [Online]. Available: http://www.xiph.org/vorbis/doc/Vorbis_I_spec.html

[21] J. G. Beerends and J. A. Stemerdink, “A perceptual audio quality measure based on a psychoacoustic sound representation,” *J. Audio Eng. Soc.*, vol. 40, no. 12, pp. 963–978, Dec. 1992.

[22] J. G. Beerends, “Audio quality determination based on perceptual measurement techniques,” in *Applications of Digital Signal Processing to Audio and Acoustics*. Norwell, MA: Kluwer, 2002, pp. 1–38.

[23] T. Thiede and E. Kabot, “A new perceptual quality measure for bit rate reduced audio,” in *Proc. 100th AES Convention*, May 1996, preprint 4280.

[24] C. Herrero, “Subjective and objective assessment of sound quality: Solutions and applications,” in *Proc. CIARM Conf.*, 2005, pp. 1–20.

[25] P. J. O. Doets, M. M. Gisbert, and R. L. Lagendijk, “On the comparison of audio fingerprints for extracting quality parameters of compressed audio,” in *Proc. Security, Steganography, Watermarking Multimedia Contents VII*, Jan. 2006, vol. 6072, pp. 228–239, ser. Proc. SPIE.

[26] P. J. O. Doets and R. L. Lagendijk, “Extracting quality parameters for compressed audio from fingerprints,” in *Proc. 6th Int. Conf. Music Inf. Retrieval (ISMIR)*, Sep. 2005, pp. 498–503.

[27] E. Batlle, J. Masip, and E. Guaus, “Automatic song identification in noisy broadcast audio,” in *Proc. Signal Image Process. (SIP)*, Aug. 2002.

[28] H. S. Malvar, “Auditory masking in audio compression,” in *Audio Anecdotes*. Wellesley , MA: A. K. Peters, 2001, pp. 217–236.

[29] 2005, Lame [Online]. Available: http://lame.sourceforge.net

[30] M. Bosi *et al.*, “ISO/IEC MPEG-2 advanced audio coding,” *J. Audio Eng. Soc.*, vol. 45, no. 10, pp. 789–813, Oct. 1997.

[31] K. Tsutsui *et al.*, “Atrac adaptive transform acoustic coding for minidisc,” in *Proc. 93rd AES Conv.*, Oct. 1992, preprint 3456.

[32] “Sony ATRAC,” 2007 [Online]. Available: http://www.sony.net/Products/ATRAC3/index.html

[33] “Microsoft WMA on Wikipedia,” 2007 [Online]. Available: http://en.wikipedia.org/wiki/Windows_Media_Audio

[34] R. J. Beaton *et al., Objective Perceptual Measurement of Audio Quality*. New York: Audio Eng. Soc., 1996.

[35] P. J. O. Doets and R. L. Lagendijk, “Stochastic model of a robust audio fingerprinting system,” in *Proc. 5th Int. Conf. Music Inf. Retrieval (ISMIR)*, Oct. 2004, pp. 349–352.

[36] F. Balado *et al.*, “Performance analysis of robust audio hashing,” *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 2, pp. 254–266, Jun. 2007.

Peter Jan O. Doets (S’02) received the M.Sc degree in electrical engineering from Delft University of Technology, Delft, The Netherlands, in 2003.

In the same year, he joined the Information and Communication Theory Group at Delft University of Technology, where he has been working towards the Ph.D. degree. His research interests include signal processing, pattern recognition, watermarking and fingerprinting.

Reginald L. Lagendijk (S’87–M’90–SM’97–F’07) received the M.Sc. and Ph.D. degrees in electrical engineering from the Delft University of Technology, Delft, The Netherlands, in 1985 and 1990, respectively.

He became an Assistant Professor at the Delft University of Technology in 1987. He was a Visiting Scientist in the Electronic Image Processing Laboratories, Eastman Kodak Research, Rochester, NY, in 1991 and Visiting Professor at Microsoft Research and Tsinghua University, Beijing, China, in 2000 and 2003, respectively. Since 1999, he has been a Full Professor in the Information and Communication Theory Group, Delft University of Technology. He is the author of *Iterative Identification and Restoration of Images* (Kluwer, 1991) and a coauthor of *Motion Analysis and Image Sequence Processing* (Kluwer, 1993) and *Image and Video Databases: Restoration, Watermarking, and Retrieval* (Elsevier, 2000). He has been involved in the conference organizing committees of ICIP2001, 2003, 2006, and 2011. Currently, his research interests include multimedia signal processing theory and algorithms, with emphasis on audiovisual communications, compression, analysis, searching, and security. He is currently leading and actively involved in a number of projects in the field of intelligent information processing for ad hoc and peer-to-peer multimedia communications.

Prof. Lagendijk was a member of the IEEE Signal Processing Society’s Technical Committee on Image and Multidimensional Signal Processing. He was an Associate Editor of the IEEE Transactions on Image Processing and IEEE Transactions on Signal Processing’s Supplement on Secure Digital Media. He is currently an Associate Editor for the IEEE Transactions on Information Forensics and Security.