EXTRACTING QUALITY PARAMETERS FOR COMPRESSED AUDIO FROM FINGERPRINTS

P.J.O. Doets and R.L. Lagendijk
{p.j.doets, r.l.lagendijk}@ewi.tudelft.nl
Dept. of Mediamatics, Information and Communication Theory Group,
Faculty of Electrical Engineering, Mathematics and Computer Science,
Delft University of Technology, P.O. Box 5031, 2600 GA Delft

ABSTRACT

An audio fingerprint is a compact yet very robust representation of the perceptually relevant parts of audio content. It can be used to identify audio, even when of severely distorted. Audio compression causes small changes in the fingerprint. We aim to exploit these small fingerprint differences due to compression to assess the perceptual quality of the compressed audio file. Analysis shows that for uncorrelated signals the Bit Error Rate (BER) is approximately inversely proportional to the square root of the Signal-to-Noise Ratio (SNR) of the signal. Experiments using real music confirm this relation. Further experiments show how the various local spectral characteristics cause a large variation in the behavior of the fingerprint difference as a function of SNR or the bitrate set for compression.

1 INTRODUCTION

Identification of music on the Internet is usually done by searching in the metadata describing the music content. Metadata like song title, artist, etc., however, is often incoherent or misleading [1], especially on popular unmoderated Peer-to-Peer (P2P) file-sharing networks like KaZaA (www.kazaa.com) and eDonkey (www.edonkey.com). A solution is to identify the music based on the content.

Identification, however, is often not enough. The perceptual quality of a song compressed using MP3 at a bitrate of 32 kbps is totally different from the perceptual quality of the CD-recording of the same song. Therefore, a content-based indication for the perceptual quality is needed. The Music2Share project proposes to use audio fingerprints for both identification and quality assessment of unknown content on a P2P network [2].

Audio fingerprints are compact representations of the perceptually relevant parts of audio content that can be used to identify music based on the content. A fingerprinting system consists of two parts: fingerprint extraction and a matching algorithm. The fingerprints of a large number of songs are usually stored in a database. A song is identified by comparing its fingerprint with the fingerprints in the database. The procedure for music identification using fingerprints is schematically shown in Figure 1.

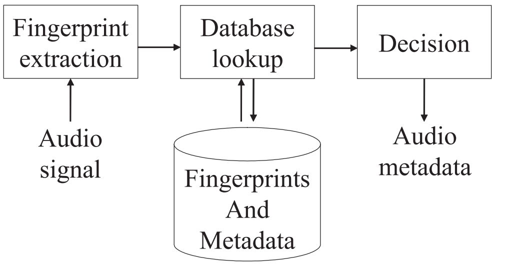


Figure 1: Music identification using an audio fingerprint. The extracted audio fingerprint is matched against a database with pre-computed fingerprints and metadata.

Fingerprinting applications are e.g. identification of songs or commercials being played on radio or television, music identification using a cell phone (e.g. Shazam [6]), and filtering for file sharing applications [3, 4].

Most audio fingerprinting systems derive their fingerprint from a time-frequency representation, e.g. using short-term Fourier transforms. They mainly differ in their choice of features to construct the fingerprint, e.g. spectral flatness features [5], spectral peaks [6], Fourier coefficients [7], Mel-Frequency Cepstrum Coefficients (MFCC) [8], and energy differences between frequency bands [3].

Fingerprints are robust to many kinds of processing: encoding using different coding schemes or bit rates, subsequent Digital-to-Analog (D/A) and Analog-to-Digital (A/D) conversions, small changes in play-out speed, etc. The fingerprints of two arbitrary pieces of music are very different, while fingerprints originating from the same music recording, but which differ due to a limited amount of processing or distortion, are only slightly different.

We aim to exploit the small fingerprint differences due to compression to assess the perceptual quality of the compressed audio file. For the time being we limit ourselves to compression using the popular MP3 format. This setup is shown schematically in Figure 2. $F_X(n,m)$ denotes the fingerprint bits of the original, undistorted recording, X, and $F_Y(n,m)$ denotes the fingerprint bits of the com-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

©2005 Queen Mary, University of London

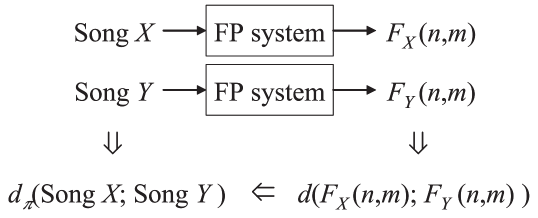


Figure 2: Relating differences in audio fingerprints of two versions of the same recording, X and Y, to differences in perceptual quality of these recordings.

pressed recording, Y. The difference between the fingerprints $d(F_X(n,m); F_Y(n,m))$ is related to the (perceptual) difference between the songs $d_\pi(\text{Song } X; \text{Song } Y)$.

We use the Philips audio fingerprinting system [3] because it is well documented, highly robust against compression and differences in fingerprint can be related to parameters used in MP3 compression. Recently, we have worked on a model of the fingerprint generation of an uncorrelated signal using the Philips system [9]. This paper focuses on the difference between fingerprints of a song and a distorted version of that same song.

Section 2 of the paper presents details of the audio fingerprinting system used for identifying songs, Section 3 analyzes the robustness of a fingerprint to MP3 compression of a song and presents experimental results, Section 4 draws conclusions and outlines future work.

2 PHILIPS AUDIO HASH

Figure 3 shows an overview of the fingerprint extraction stage of the Philips system [3]. The audio signal is first segmented into frames of 0.37 seconds with an overlap factor of 31/32, weighted by a Hanning window. The compact representation of a single frame is called a sub-fingerprint. In this way, it extracts 32-bit sub-fingerprints for every interval of 11.6 ms (370/32 ms). Due to the large overlap, subsequent sub-fingerprints have a large similarity and slowly vary in time. The fingerprint of a song consists of a sequence of sub-fingerprints, which are stored in a database.

To extract a 32-bit sub-fingerprint for every frame, 33 non-overlapping frequency bands are selected from the estimated Power Spectral Density (PSD). These bands range from 300 Hz to 2000 Hz and are logarithmically spaced.

Haitsma and Kalker report that experiments have shown that the sign of energy differences is a property that is very robust to many kinds of processing [3]. We denote the energy of frequency band m of frame n by $E(n,m)$. Energy differences are computed in time and frequency:

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

The bits of the sub-fingerprint are derived by

$$F(n,m) = \begin{cases} 1 & ED(n,m) > 0 \\ 0 & ED(n,m) \leq 0 \end{cases}, \quad (2)$$

where $F(n,m)$ denotes the m^{th} bit of sub-fingerprint n.

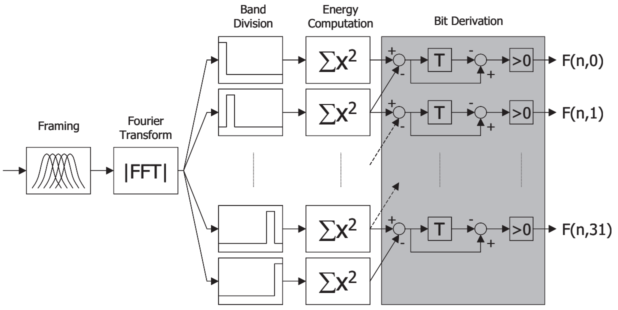


Figure 3: Philips audio fingerprinting extraction [3]. T indicates a unit-time delay.

Figure 4(a) shows an example of a fingerprint. White parts indicate positive energy differences (i.e. $F(n,m) = 1$). The small side of the fingerprint block is the frequency direction, consisting of the 32 bits corresponding to the differences between the 33 frequency bands. The long side of the block corresponds to the temporal dimension.

The system is capable of identifying a segment of about 3.3 seconds of music - generating 256 sub-fingerprints - in a large database, even if the segment is degraded due to a variety of signal processing operations. A match is found if the Bit Error Rate (BER) between the extracted fingerprint and the fingerprint in the database falls below a threshold of 0.35.

3 FINGERPRINT ROBUSTNESS ANALYSIS TO MP3 COMPRESSION

When the song is subject to compression, the fingerprint changes slightly. To indicate the effect of MP3 compression on the fingerprint extraction, Figures 4(c)-4(e) show the difference patterns of the fingerprint of a recording at different bit-rates relative to the fingerprint of the CD-quality recording of the same song. The difference between fingerprints can be defined as:

$$F_{diff}(n,m) = \text{XOR}\left(F_X(n,m), F_Y(n,m)\right) \quad (3)$$

The black sections mark the fingerprint differences, white sections indicate similarity between the fingerprints.

The goal is to relate the perceptual quality of the compressed version of the song (relative to the original recording) to features of the observed difference in the corresponding fingerprints, $f\left(F_{diff}(n,m)\right)$. The intended use is illustrated in Figure 4(f). The central research question is how to define the quality distance measure and the function $f(\cdot)$ operating on $F_{diff}(n,m)$ (e.g. BER).

Section 3.1 presents a simple model to analyze the relation between Signal-to-Noise Ratio (SNR) and BER for uncorrelated signals, Section 3.2 discusses the relation between the spectral content of a song and the robustness of the fingerprint bits, Section 3.3 presents details about two fingerprint distance measures used in the experiments presented in Section 3.4.

3.1 Analysis using uncorrelated signals

In previous work we have modeled the Philips fingerprint extraction for uncorrelated, stationary data sources [9].

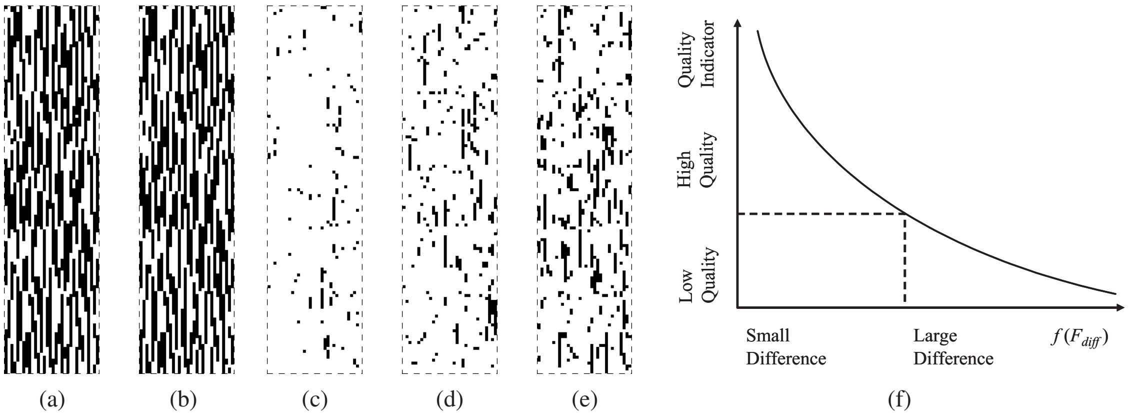


Figure 4: Fingerprints for an excerpt of ’Anarchy in the U.K.’ by the Sex Pistols. (a)-(b) Fingerprints of (a) the original and (b) of an MP3 compressed version encoded at 128 kbps; white indicates $F(n, m) = 1$ (c-e) Differences between the fingerprints of the original and an MP3 compressed version encoded at (c) 128 kpbs (d) 80 kpbs and (e) 32 kbps. The black positions mark the differences. (f) Relating differences between two fingerprints F_{diff} to a quality indication.

Of course, music is strongly correlated and highly non-stationary. These models, however, help to understand the effect of key signal and fingerprint parameters such as the frame length and the amount of frame overlap.

Here we extend the analysis to the situation where both the signal and the additive noise are assumed to be zero-mean Gaussian, independent identically distributed (iid) data sources. Such an analysis relates the BER to SNR. Although SNR is not a realistic real-life quality measure, it is a suitable distortion measure at the abstraction level of this analysis.

To simplify the analysis, the fingerprints are subjected to two constraints. First, for these data sources the BER is independent of the number of frequency bands. Therefore, without loss of generality, we limit the analysis to two frequency bands. Second, for these data sources, the amount of frame overlap has no influence on the BER when the fingerprints are perfectly aligned. Therefore, we assume non-overlapping windows.

Our first analysis starts with a simple model for the energy differences, $ED(n, m)$, that lead to the fingerprint bits using Eq. (2). Index m is omitted, since the analysis assumes two frequency bands, resulting in energy differences $ED(n, m)$ having just one frequency index. Let $ED_X(n)$ denote the energy differences of signal X, and $ED_W(n)$ denote the energy differences of the noise, W.

In case of additive noise, the signal $ED_Y(n)$ becomes:

$$ED_Y(n) = ED_X(n) + ED_W(n). \tag{4}$$

$F_X(n)$ is generated by taking the sign of $ED_X(n)$:

$$F_X(n) = \begin{cases} 1 & ED_X(n) > 0 \\ 0 & ED_X(n) \leq 0 \end{cases}. \tag{5}$$

The BER can now be expressed in terms of probabilities:

$$\begin{aligned} \text{BER}_{temp} &= P[F_X(n) \neq F_Y(n)] \\ &= 2P[ED_X(n) > 0, ED_Y(n) \leq 0] \\ &= 2P[ED_X(n) > 0, \\ &\qquad ED_W(n) \leq -ED_X(n)]. \end{aligned} \tag{6}$$

Since both $ED_X(n)$ and $ED_W(n)$ are mutually independent, zero-mean Gaussian iid data sources, all signals are fully characterized by their variance:

$$\begin{aligned} \text{VAR}[ED_X(n)] &= \sigma^2_{ED_X} && \propto \sigma^2_X \\ \text{VAR}[ED_W(n)] &= \sigma^2_{ED_W} && \propto \sigma^2_W \\ \text{VAR}[ED_Y(n)] &= \sigma^2_{ED_X} + \sigma^2_{ED_W} && \end{aligned} \tag{7}$$

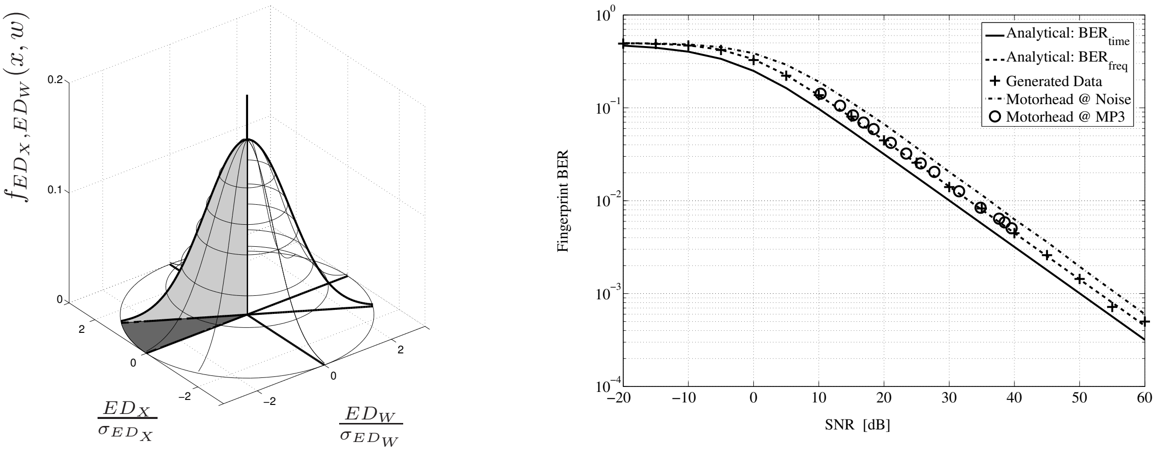
We will now express both SNR and BER in terms of the variances σ^2_X and σ^2_W. The SNR is defined as:

$$\text{SNR} = \frac{\text{VAR}[X]}{\text{VAR}[Y - X]} = \frac{\sigma^2_X}{\sigma^2_W} \tag{8}$$

By simple geometrical arguments using the joint PDF of $ED_X(n)$ and $ED_W(n)$, $f_{ED_X, ED_W}(x, w)$, it can be shown that the BER is equal to:

$$\begin{aligned} \text{BER}_{temp} &= 2P[ED_X(n) > 0, \\ &\qquad ED_W(n) \leq -ED_X(n)] \\ &= 2\int_{-\infty}^{0} f_{ED_W}(w) \left\{ \int_{0}^{-w} f_{ED_X}(x)dx \right\} dw \\ &= \frac{1}{\pi} \arctan\left(\frac{\sigma_{ED_W}}{\sigma_{ED_X}}\right) \\ &= \frac{1}{\pi} \arctan\left(\sqrt{\frac{\text{VAR}[ED_X]}{\text{VAR}[ED_Y - ED_X]}}\right) \qquad (9) \\ &= \frac{1}{\pi} \arctan\left(\frac{1}{\sqrt{\text{SNR}}}\right) \qquad (10) \end{aligned}$$

To illustrate the geometrical argument, Figure 5(a) shows the joint PDF $f_{ED_X, ED_W}(x, w)$. The axes of the ground plane represent the unit-variance variables $\frac{ED_X}{\sigma_{ED_X}}$ and $\frac{ED_W}{\sigma_{ED_W}}$. The PDF is now rotation-symmetric. The volume $Vol{=}P[ED_X{>}0, ED_W{\leq}{-}ED_X]$ can be computed by rotating the light shaded area around the f_{ED_X, ED_W}-axis over an angle ϕ. Since the total volume of the PDF is equal to 1, the relation between ϕ and the shaded volume is given by $Vol = \frac{\phi}{2\pi}$. The line $ED_X = -ED_W$ has an angle $\phi = \arctan\left(\frac{\sigma_{ED_W}}{\sigma_{ED_X}}\right)$ with the ED_W-axis.



(a)

(b)

Figure 5: Robustness to additive noise (analytical and experimental) (a) Illustrating the geometrical argument used to compute $P[F_X(n) = 1, F_Y(n) = 0]$, which is a volume in the joint-PDF $f_{X,W}(x,w)$ (b) SNR vs. BER for a simple models using iid Gaussian random variables.

Figure 5(b) shows the SNR vs. BER$_{temp}$ plot for both experimental fingerprint of Gaussian zero-mean iid data and the analytical results of Eq. (10). The analytical curve is shifted with respect to the experimental curve. This deviation is caused by the computation of the fingerprints in the frequency domain instead of the time domain.

In the frequency domain, the spectrum of $Y(n)$ is related to the spectra of $X(n)$ and $W(n)$:

$$\begin{aligned} |\hat{Y}(k)|^2 &= |\hat{X}(k) + \hat{W}(k)|^2 \\ &= |\hat{X}(k)|^2 + |\hat{W}(k)|^2 + 2\mathrm{Re}\left(\hat{X}(k)\overline{\hat{W}(k)}\right) \end{aligned} \quad (11)$$

where $\hat{Y}(k)$ denotes the Fourier transforms of $Y(n)$ and $\overline{\hat{W}(k)}$ denotes the complex conjugate of $\hat{W}(k)$.

In order to compute the BER using Eq. (9), the variances of ED_X and ED_Y are expressed in terms of frequency and time domain signal variances [10]:

$$\begin{aligned} \mathrm{VAR}\left[ED_X\right] &\propto \mathrm{VAR}\left[|\hat{X}(k)|^2\right] \\ &\propto \sigma_X^4 \\ \mathrm{VAR}\left[ED_Y - ED_X\right] &\propto \mathrm{VAR}\left[|\hat{Y}(k)|^2 - |\hat{X}(k)|^2\right] \\ &\propto \sigma_W^4 + 2\sigma_X^2\sigma_W^2 \end{aligned}$$

Now the BER can be expressed as:

$$\mathrm{BER}_{freq} = \frac{1}{\pi}\arctan\left(\sqrt{\frac{\sigma_W^4}{\sigma_X^4} + 2\frac{\sigma_W^2}{\sigma_X^2}}\right) \quad (12)$$

Figure 5(b) also shows the curve for the BER$_{freq}$ expression, which fits perfectly to the experimental data. For SNR $\gg 1$ the expression can be further simplified to:

$$\mathrm{BER}_{freq} \approx \frac{1}{\pi}\arctan\left(\sqrt{2}\frac{\sigma_W}{\sigma_X}\right) \approx \frac{\sqrt{2}}{\pi}\frac{\sigma_W}{\sigma_X} \quad (13)$$

This implies that for sufficiently large SNR, the fingerprint BER is reduced by a factor 10 when the SNR is increased by 20 dB.

3.2 Content dependence

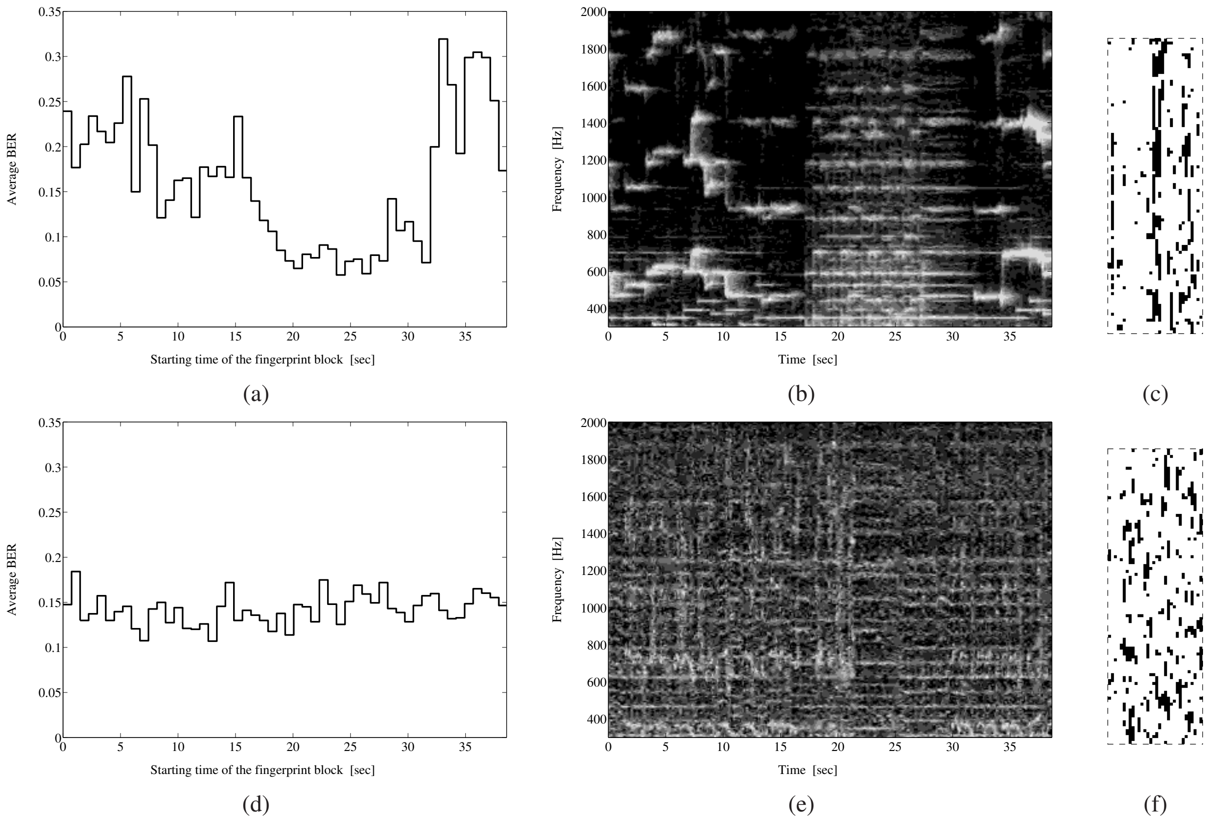
Figures 6(a) and 6(d) shows the BER between original and a compressed version for blocks of 64 sub-fingerprints of two songs: ‘Requiem - Pie Jesu’ composed by Fauré and ‘Motörhead’ by Motörhead, respectively. Two observations can be made from this graph: First, there is a large *inter*-song variance. Different songs compressed at the same bit-rate show different average behavior. Second, a song can have a large *intra*-song variance as well. Fauré shows a large spread in BER behavior, while the fingerprint blocks of Motörhead show a very small spread.

The spectrograms are shown in Figures 6(b) and 6(e). The horizontal axis shows the starting time of a frame, the vertical axis shows the frequency dimension and the gray-value indicates the magnitude in the energy spectrum of each frame in decibel (dB). Lighter values indicate larger magnitude. Comparing Figures 6(a) and 6(d) with Figures 6(b) and 6(e), respectively, clearly relates the BER of a block to spectral characteristics of that region in time. The spectrogram of Fauré shows distinct peaks and regions which have near-zero energy. These regions in the spectrogram containing near-zero energy generate near-zero $ED(n,m)$ signals which are rather sensitive to compression artifacts. Since these valleys in the spectrogram do not occur uniformly over time, they cause a large spread in time of the BER.

The difference in behavior of these regions is also reflected in the fingerprint blocks. Figures 6(c) and 6(f) show two differences between fingerprint blocks having the same number of bit errors. Figure 6(c) corresponds to a part of Fauré (having relatively little energy) while Figure 6(f) corresponds to a part of Motörhead.

3.3 Distance measures

The large intra-song variance might be reduced by using other distance measures than the BER, e.g. average length or average area of runs of fingerprint errors. In this paper we use two distance measures: BER based on the hamming distance and BER based on the weighted hamming



(a) (b) (c)

(d) (e) (f)

Figure 6: (a,d) Bit Error Rate as function of time, (b,e) Spectrograms and (c,f) fingerprint differences between original recording and MP3@32 kbps of (a-c) Fauré and (d-f) Motörhead.

distance. The former is defined as:

$$\text{BER} = \frac{1}{32N} \sum_{n=0}^{N-1} \sum_{m=0}^{31} F_{diff}(n,m), \tag{14}$$

the latter is defined as:

$$\text{BER}_W = \frac{\sum_{n=0}^{N-1} \sum_{m=0}^{31} H(n,m) F_{diff}(n,m)}{\sum_{n=0}^{N-1} \sum_{m=0}^{31} H(n,m)}. \tag{15}$$

Certain regions can be excluded by setting the weights $H(n,m)$ to 0. For the experiments, the weights are set based on the $ED(n,m)$ signal.

Let $ED_h^{max}(n,m)$ denote the maximum value of $|ED(n,m)|$ within a sliding window of size h. The binary weight $H(n,m)$ is zero if its corresponding $ED(n,m)$ value is smaller than a signal dependent threshold T:

$$H(n,m) = \begin{cases} 0 & ED_h^{max}(n,m) \leq T \\ 1 & ED_h^{max}(n,m) > T \end{cases} \tag{16}$$

3.4 Experiments

Experiments have been performed on fragments of 39 seconds for 11 songs. To keep the figures comprehensible the results presented here are limited to the 2 songs mentioned earlier, viz. Fauré and Motörhead. The fingerprints of these fragments were split into 13 non-overlapping fingerprint blocks consisting of 256 sub-fingerprints.

When compressing a song, the perceptual quality can be controlled by selecting the bitrate. This indirectly influences the difference between the fingerprints. Two quality measures have been used: the MP3 bitrate and the SNR.

The weights $H(n,m)$ were assigned using a threshold T equal to the global median value of $|ED(n,m)|$ and a window size h of 32 frames. Depending on (the part of) the song, the 8-35% of the bits was excluded.

Figure 7 shows the experimental results. Figures 7(a) and 7(c) use BER, while 7(b) and 7(d) use BER$_W$; Figure 7(a) and 7(b) use MP3 bitrate as quality indicator while 7(c) and 7(d) use SNR. The lines indicate the results averaged over the 13 fingerprint blocks, the errorbars and shaded regions indicate the corresponding standard deviations. Both SNR - in dB - and BER are displayed on a logarithmic scale. A straight line in a plot using logarithmical scales indicates a power law relation ship. From Figures 7(c) and 7(d), the relation between the expected value of the BER and the SNR confirms Eq. (13):

$$E[\text{BER}] \propto \frac{1}{\sqrt{\text{SNR}}} = \frac{\sigma_W}{\sigma_X}. \tag{17}$$

From the experiments it can be concluded that the BER between fingerprints originating from the same audio file is inversely proportional to the square root of the SNR of one song with respect to the other. Relating BER to bitrate is less straightforward, since compressing different songs at the same bitrate yield different SNR.

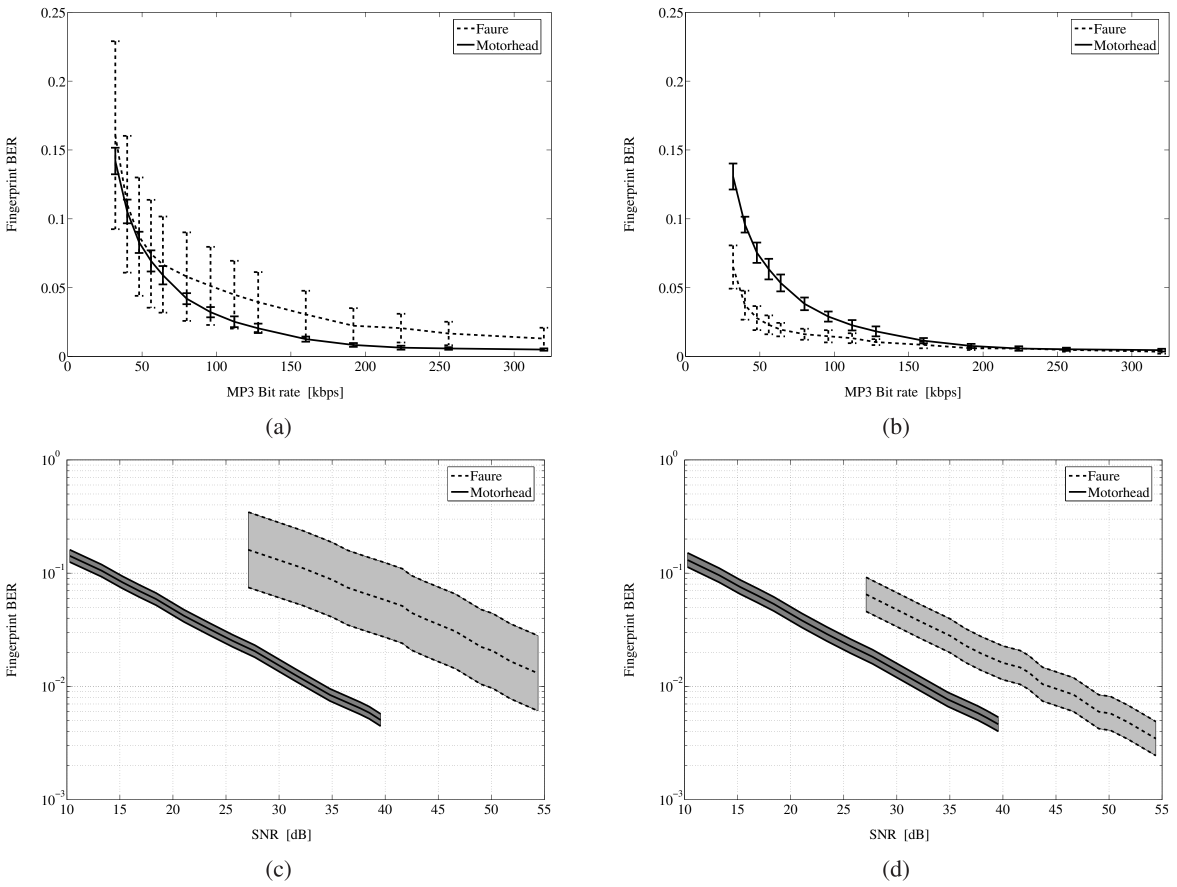


Figure 7: Average BER in 13 fingerprint blocks for two songs. The errorbars and shadings indicate the standard deviation of the BER at a specific bitrate or SNR level. (a-b) MP3 bitrate vs. (a) BER (b) BER_W; (c-d) SNR vs. (c) BER (d) BER_W

4 CONCLUSION AND FUTURE WORK

Experiments have indicated how differences in audio fingerprint due to compression are related to the spectral characteristics of the audio signal. Variations in these local spectral characteristics cause a large variation in the behavior of fingerprint differences for a given compression bitrate or SNR. We have shown that this variation can be reduced when the fingerprint bits related to spectral regions with near-zero energy are excluded. It was determined both theoretically and experimentally that the BER is approximately inversely proportional to the SNR of the signal.

Future work concerns the further exploration and theoretical foundation SNR-BER relationship, its expansion to bitrate-BER relations, and the definion of a similarity metric which is suitable for quality assessment using fingerprints.

REFERENCES

[1] J. Liang *et al.* Pollution in P2P file sharing systems. In *IEEE Infocom*, March 2005.

[2] T. Kalker *et al.* Music2Share copyright-compliant music sharing in P2P systems. *Proceedings of the IEEE*, 92(6):961 – 970, 2004.

[3] J. Haitsma and T. Kalker. A highly robust audio fingerprinting system. In *3rd Int. Symp. on Music Information Retrieval (ISMIR)*, October 2002.

[4] P. Cano, E. Batlle, T. Kalker, and J. Haitsma. A review of algorithms for audio fingerprinting. In *IEEE Int. Workshop on Multimedia Sig. Proc.*, Dec. 2002.

[5] E. Allamanche *et al.* Content-based identification of audio material using mpeg-7 low level description. In 2nd *ISMIR*, October 2001.

[6] A. Wang. An industrial strength audio search algorithm. In 4th *ISMIR*, October 2003.

[7] Y. Cheng. Music database retrieval based on spectral similarity. In 2nd *ISMIR*, October 2001.

[8] P. Cano, E. Batlle, H. Mayer, and H. Neuschmied. Robust sound modeling for song detection in broadcast audio. In *112th AES Convention*, 2002.

[9] P.J.O. Doets and R.L. Lagendijk. Stochastic model of a robust audio fingerprinting system. In 5th *ISMIR*, October 2004.

[10] A. Leon-Garcia. *Probability and Random Processes for Electrical Engineering*. ISBN 0-201-50037-X. Addison-Wesley, 2nd edition, 1994.