Privacy-enhanced Perceptual Hashing of Audio Data

Heiko Knospe
Institute of Communications Engineering, Cologne University of Applied Sciences, 50679 Cologne, Germany

Keywords: Perceptual Hashing, Audio Hashing, Audio Fingerprinting, Acoustic Fingerprint, Privacy, Security.

Abstract: Audio hashes are compact and robust representations of audio data and allow the efficient identification of specific recordings and their transformations. Audio hashing for music identification is well established and similar algorithms can also be used for speech data. A possible application is the identification of replayed telephone spam. This contribution investigates the security and privacy issues of perceptual hashes and follows an information-theoretic approach. The entropy of the hash should be large enough to prevent the exposure of audio content. We propose a privacy-enhanced randomized audio hash and analyze its entropy as well as its robustness and discrimination power over a large number of hashes.

1 INTRODUCTION

The increasing amount of multimedia data has led to a growing interest in fast and reliable identification techniques. Multimedia content can have various representations and is subject to transformations which preserve the perceptual content, but significantly alter the underlying data. It is obvious that cryptographic hash functions can not preserve similar content because of the *avalanche effect* of these functions. They are hence of limited use for the identification of multimedia data. Instead, robust *perceptual* hashes are required which are locality-sensitive (Slaney and Casey, 2008) and map similar input data to similar hashes. The hashes are usually represented by a sequence of binary vectors. The size of the original data is substantially reduced and similarity can be measured in the hash domain. Different copies (including their lossy representations) of the same multimedia document can then be identified by comparing their hashes. We note that content recognition (for example speech recognition and semantical correspondence) is not intended here and different recordings with identical or similar content should give different perceptual hashes.

The problem of audio identification can be considered as largely solved (Kurth and Müller, 2008) with commercial solutions available for large music collections (Wang and Smith III, 2008). But optimizations of the fingerprint are still sensible (Grutzek et al., 2012), e.g. for speech recordings, for very large repositories, fast searching, good robustness and a very low rate of false identifications.

Further aspects concern the security and privacy of the perceptual hash. Here, *security* refers in particular to content integrity and multimedia authentication. A key-dependent perceptual hash can authenticate the multimedia data: an adversary should not be able to produce perceptually different data with the same hash value. Different proposals for secure perceptual hashes exist and we refer to Section 2.2 for more details.

Privacy requirements for multimedia hashes have been examined less so far. Privacy is relevant for personal multimedia data, which is processed by distributed systems, for example telephone calls. Perceptual hashing can be used to identify similar copies, e.g. replayed spam calls. The main privacy concern thereby is that the hash may reveal information about the original content. Since the hash computation involves several reduction steps and the hash size is usually very small compared to the original data, it is generally impossible to reconstruct the complete multimedia content. But even a restricted information leakage, e.g. single words or characteristic properties of a speaker, would be critical. Ideally, an adversary should not be able to distinguish the hash from a random sequence.

In this paper, we present a privacy-enhanced perceptual hash for audio data. We are particularly interested in speech data where privacy is much more important than for music. The construction of the hash is based on the well-known work of (Haitsma and Kalker, 2002) and our contribution (Grutzek et al.,

Knospe H..
Privacy-enhanced Perceptual Hashing of Audio Data.
DOI: 10.5220/0004532605490554
In *Proceedings of the 10th International Conference on Security and Cryptography* (SECRYPT-2013), pages 549-554
ISBN: 978-989-8565-73-0

2012). The hash consists of a set of subhashes which are derived from spectral audio features and subsequently randomized by a cryptographic hash-based message authentication code (HMAC). We examine the capabilities of the hash with respect to different requirements including their robustness, discrimination performance and privacy properties.

This work is organized as follows: we review perceptual hashes and in particular the existing work on secure audio hashes in Section 2. The following Section 3 contains the privacy requirements for multimedia identification applications. Then we introduce a privacy-enhanced perceptual audio hash. Section 4 shows the performance of this hash and the conclusion is provided in Section 5.

2 RELATED WORK

2.1 Audio Fingerprinting Frameworks

Acoustic fingerprints, which are also called *audio fingerprints* or *audio hashes*, have been studied for some time (Cremer et al., 2001), (Clausen and Kurth, 2004), (Haitsma and Kalker, 2002), (Wang, 2003). There exists a number of different algorithms but usually the fingerprint is based on time-frequency features of the waveform. In a general framework, the fingerprint is computed in a number of steps (Cano et al., 2002): audio preprocessing, normalization, framing with overlap, spectral transformation and feature extraction, quantization and fingerprint modeling.

The main differences of the algorithms are due to the combination of spectral information (Doets and Lagendijk, 2008). The resulting fingerprint is usually a sequence of vectors (subhashes) with one vector for each time frame. Adjacent frames often have identical or similar subhashes and redundancies can be reduced by fingerprint modeling. For an efficient search of a given fingerprint against a large repository of hashes, the comparison of individual fingerprints and the computation of their distances have to be avoided. *Index-based* search algorithms (Kurth and Müller, 2008) are computationally less expensive.

In the following, we consider audio fingerprints which preserve the time information, e.g. (Haitsma and Kalker, 2002) or (Wang, 2003). For each audio sample A and time window (frame) t, a subhash $h(A,t) \in V$ is computed, where V is a vector space (e.g. $V = \{0,1\}^{32}$) equipped with a distance function (metric) $d : V \times V \to \mathbb{R}_{\geq 0}$, e.g. the Hamming distance. The complete audio fingerprint is a collection of temporal positions and their associated subhashes: $h(A) = \{(t_1,v_1),(t_2,v_2),\ldots,(t_n,v_n)\}$. Two fingerprints are equivalent if they differ only by a global time shift. There are different possibilities to extend the distance d from the vector space of subhashes to equivalence classes of fingerprints. For example, d can be defined as reciprocal to the *maximum number of matches*. Two subsets $\{(t_1,v_1),(t_2,v_2),\ldots,(t_k,v_k)\}$ and $\{(t'_1,v'_1),(t'_2,v'_2),\ldots,(t'_k,v'_k)\}$ with k elements *match* if the temporal positions $t_1,\ldots,t_k$ coincide (after a possible global time shift) and $d(v_j,v'_j) \leq \delta$ (for example $\delta = 0$) for all $j = 1,\ldots,k$.

Standard requirements are given in (Wang, 2003), (Cano et al., 2002), (Doets and Lagendijk, 2008):

1. Robustness: perceptually similar audio samples $A \sim A'$ have hash vectors with a small distance $d(h(A),h(A')) \leq \varepsilon$, where $\varepsilon \geq 0$ is a threshold which controls the robustness of the algorithm.
2. Discrimination: perceptually different audio samples A and A' yield a large distance $d(h(A),h(A')) > \varepsilon$. The fingerprint must be sufficiently entropic to allow sufficient distinction and to prevent spurious matches.
3. Localization property and translation invariance: similar audio excerpts (e.g. only a few seconds long) can be identified independent of their absolute temporal position.

There exist various fingerprinting systems with the desired properties; robustness and discrimination are satisfied statistically (with sufficiently low error rate) for randomly chosen audio data. Important examples are the fingerprints defined in (Haitsma and Kalker, 2002) and (Wang and Smith III, 2008).

2.2 Secure Audio Fingerprinting

The presence of adversaries who deliberately manipulate the audio data or the hash gives rise to further requirements, compare (Thiemert et al., 2009):

1. Secure Robustness: it is hard to create perceptually similar audio data A and A' with $d(h(A),h(A')) > \varepsilon$.
2. Second Preimage Resistance: for a given audio sample A and hash value $h(A)$, it is hard to find perceptually different audio data A' with $d(h(A),h(A')) \leq \varepsilon$.
3. Collision Resistance: it is hard to create any perceptually different audio documents A and A' with $d(h(A),h(A')) \leq \varepsilon$.

The first requirement prevents adversaries from generating specifically manipulated versions of the audio content which can not be identified, e.g. for copyright protected music. An example of an attack against the secure robustness of the (Haitsma and Kalker, 2002) fingerprint is given in (Thiemert et al., 2009). The quantization properties of the algorithm are used to flip a number of *weak hash bits* without perceptually changing the audio data.

The second and the third requirement prevent that forged audio content is accepted as authentic. This is also relevant in connection with *watermarking* of audio files where a robust hash is used to protect the content integrity.

Since the relation between time-frequency amplitudes and output hash bits is well localized and permits the computation of preimages, the desired properties can hardly be achieved without additional randomization. Diffusion operations similar to cryptographic hashes would destroy the required robustness. It is well known that already the feature extraction algorithm should be key-dependent (Fridrich and Goljan, 2000), (Swaminathan et al., 2006). Indeed, collisions and forged hashes would persist if the randomization would be applied *after* the feature extraction.

It was observed by (Swaminathan et al., 2006) that there is a trade-off between security and robustness. They analyzed several image hash functions and used the *conditional entropy* of the hash values for a given image and an unknown key. The entropy was surprisingly low with values between 6 and 16 bits.

Furthermore, an adversary could try to reveal the key from the given hashes. (Koval et al., 2008) and (Koval et al., 2009) analyzed the security of algorithms based on block random projections (Fridrich and Goljan, 2000) and used the conditional entropy of the key for a given media file and hash value. They discovered that information on the key is leaked, but the amount of information decreases with the input block size for the subhash computation.

(Weng and Preneel, 2011) proposed a secure image hash which provides block level protection and avoids collisions for malicious minor modifications. Their hash shows good robustness and discrimination properties but they did not analyze the security of the key.

(Zmudzinski and Steinebach, 2009) defined a so-called *rMAC* for audio data based on the (Haitsma and Kalker, 2002) fingerprint. The rMAC can be embedded as a watermark in the audio data. In their experiments, a 128-bit rMAC for audio samples of 7*s* length showed sufficient robustness and discrimination power. Possible open issues are the shift invariance, the entropy of the fingerprint and information leakage on the key.

In summary, there has been some work on secure robust hashing, but there exist relevant open issues on the security of various proposed algorithms. There are also indications that the required robustness impedes a high level of resistance against attacks.

3 PRIVACY-ENHANCED HASHES

3.1 Privacy Requirements

The use of fingerprinting techniques for multimedia identification can raise *privacy* concerns if personal information is processed. One of the main questions is whether the fingerprint leaks information on the original content. It is well known that the properties of *cryptographic hash functions* prevent any information gain other than the identification of exact copies. But *robust hashes* may leak partial information about the original data. For example, audio hashes usually contain quantized time-frequency features of the waveform. The compactness of most fingerprints prevents a complete reconstruction but it seems feasible to refine the probability distribution of the possible content and therefore gain partial information. A systematic analysis of the equivocation of fingerprints with respect to the multimedia data is still owing. In this situation, telecommunication privacy laws in many countries would not permit the use of fingerprints for telephone data.

We have the following information-theoretic requirements:

1. The entropy $H(h(A))$ of all hashes should be large enough to protect against frequency analysis and dictionary attacks. More specifically, the entropy of the subhashes shall be high enough to prevent the exposure of local audio content.
2. The conditional entropy $H(\mathcal{A}|h(A))$ of audio data for a given audio hash $h(A)$ shall be large enough to protect against information leakage. Furthermore, the conditional entropy of *local* audio data for a given subhash shall be high enough to prevent the exposure of local audio information.

Ideally, it should not be possible to distinguish the hash from random data but this can hardly be achieved with the current algorithms. In particular, the robustness and the shift invariance requires a large overlap of the audio frames. Hence the subhashes change only slowly with time and adjacent subhashes are clearly correlated.

Even a high entropy of the hash would not prevent a partial exposure of audio content if the relation between the input audio data and the output hash is easily traceable. It is well known that robust hashes require a secret key which obscures this relation. We remarked above that the current methods, which are based on a randomization of the feature extraction process, may leak information on the key. We therefore propose to randomize the subhashes by applying a *hash-based message authentication code* (HMAC) which can be used as a *pseudorandom function* prf_K, see (Bellare et al., 1996), but also (Bellare, 2006). This has several advantages compared to the randomization of features as discussed in Section 2.2 above: the values of prf_K do not leak information on the key and the original subhashes can not be reconstructed, even when the key is disclosed (only dictionary attacks). Furthermore, the entropy of the subhashes is preserved by prf_K and the application of prf_K does not generate new collisions.

Since collisions of the original subhashes are preserved by the prf_K-function, this construction is *not suitable for audio authentication*. In particular, an adversary may produce perceptually different multimedia data with the same hash value. But the *privacy is preserved* since the prf_K function is one-way. The overall protection depends on the distribution of subhashes and the number of known subhashes, i.e. information can only be gained if the entropy is low and the adversary has access to a large number of subhashes or is capable to generate a large number of them for given audio data.

It would be desirable to extend the randomization operation beyond the subhashes, to add dependencies between the blocks or to use salt values, but the required shift-invariance and the localization properties impede this. But we obtain an additional randomization by dropping the time position of the subhashes, removing repeated entries and finally permuting them. We remark that this method could also be combined with a randomization during the feature extraction as described in Section 2.2.

3.2 Implementation

The proposed construction is based on (Haitsma and Kalker, 2002), our work (Grutzek et al., 2012) and several privacy-related enhancements.

For identification purposes, audio samples of several seconds suffice. The audio data is extracted every 11.8 ms with overlapping frames of length 370 ms. Silent sections are skipped and only the first 100 frames with sufficient energy are processed. A Fourier transform is applied to the frames and the spectral coefficients are filtered by a mel filter bank in order to determine the energy in each sub-band. The bands are equally distributed on a logarithmic frequency axis between 300 Hz and 1800 Hz. For the privacy-enhanced hash, we extract for each frame 41 spectral and 21 cepstral coefficients (so-called MFCCs). The spectral coefficients are differentiated in time and frequency direction, and the cepstral coefficients only in frequency direction. This information is quantized by only considering the sign while disregarding magnitudes (compare (Grutzek et al., 2012)). This yields a binary subhash vector of length $40+20=60$ bits for each frame. Other common algorithms use bit-lengths of approximately 32 bits, but we can show (Section 4) that 60 bits provide additional entropy while still ensuring sufficient robustness. The hash has the following structure:

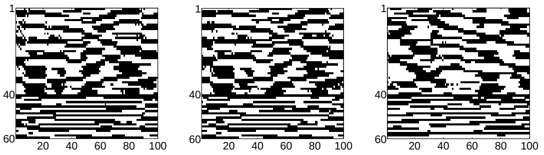

Figure 1: Binary hash matrix of three speech samples. The left and central sample are similar, while the right is dissimilar to both other samples. The upper 40 bit correspond to spectral and the lower 20 bits to cepstral coefficients.

$$h(A) = \{(t_1, v_{t_1}), (t_2, v_{t_2}), \ldots, (t_{100}, v_{t_{100}})\}$$

The subhashes are vectors $v_i \in \{0,1\}^{60}$ and the complete hash can be represented by a binary matrix of size 60×100 (see Figure 1).

Then a key-dependent pseudorandom function function prf_K is applied to the vectors v_i, the time positions are dropped and the resulting randomized hash $h_K(A)$ is a *set* of binary vectors:

$$h_K(A) = \{prf_K(v_{t_1}),\ prf_K(v_{t_2}), \ldots, prf_K(v_{t_{100}})\}$$

Hence duplicates are removed and the ordering is not relevant. The size of the hash $h_K(A)$ is only approximately 2 kBytes, depending on the keyed hash function used as pseudorandom function.

We assume that the randomized hashes $prf_K(v_i)$ are *computationally indistinguishable* from random output and do not leak information on the key. Then the security of our hash depends solely on the distribution of subhashes v_i. If they have sufficient entropy, then an adversary obtains few information from observing $prf_K(v_i)$. Ideally, the v_i's would be long enough (say more than 100 bits) and uniformly distributed. In practice, it is hard to construct *robust* audio hashes with such a large binary length and their distribution is biased.

4 ANALYSIS

4.1 Entropy

We analyzed our hash with 5,530 real audio samples

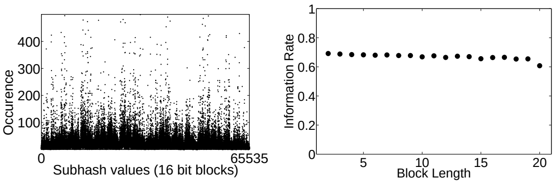

Figure 2: Frequency distribution of 450,000 randomized speech data subhashes (left) and estimated information rate for different block lengths (right).

(see Section 4.2) and 450,000 randomized subhashes. Figure 2 shows the approximate distribution and the information rate (entropy per bit-symbol). The entropy is estimated by counting the number of occurrences for blocks of length between 2 and 20 bits, and the frequency distribution is computed for words of length 16 bits. There may be dependencies between the blocks, but for computational reasons it is not possible to estimate the entropy for the given block-length of 60, since this would require a multiple of 2^{60} subhashes. Our computations show an information rate of approximately 0.65 for the given audio data. Additionally, the concatenated file of binary subhashes was compressed with different algorithms and parameters; the file size could be reduced by at most 43%. We conclude that the subhashes provide at least 34 bits of entropy. We therefore expect that any information gain from the frequency of the randomized subhashes requires at least several million subhashes. Changing the key K prevents such an attack, but only fingerprints which were randomized with the same key can be identified.

4.2 Hypothesis Testing

The performance of the hash is analyzed with respect to its capability to identify resp. to discriminate audio samples. We assume a repository with a large number of hashes when a new hash arrives. There are two possible decisions:

- H_0: The audio sample is perceptually different from all the given ones.
- H_1: The audio sample is perceptually similar to one or more samples in the database.

This can be considered as an hypothesis testing problem where the decision depends on the distance $d(h_K(A), h_K(A'))$ of hashes. In our case, the distance is reciprocal to the number m of matching subhashes. If none of the subhashes match, i.e. $h_K(A) \cap h_K(A') = \varnothing$, then the decision is clearly H_0. Otherwise, the decision for either H_0 or H_1 depends on a threshold T. A low threshold provides good robustness but less discriminative power. Higher thresholds deteriorate the robustness but also decrease the number of false identifications.

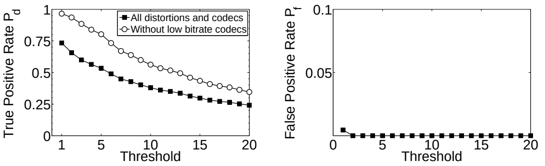

Figure 3: True positive rate P_d (left) and false positive rate P_f (right) for different thresholds.

We analyzed 5,330 different audio samples from Verbmobil II corpus of German telephone dialogs (Bavarian Archive for Speech Signals, 1998) and 200 additional telephone spam files with perceptual similar copies. These files are based on 20 real telephone spam recordings which were intentionally altered by noise, audio- and telephone codecs. The following types of alterations and distortions were considered (compare (Grutzek et al., 2012)): MP3-codec at 32 and 96 kbps, GSM fullrate, G.726 codec at 16 and 32 kbps, 5% and 10% packet loss, white and pink noise with 20dB SNR.

The performance can be characterized by the *true positive rate* $P_d = P(m \geq T \mid H_1)$ and the *false positive* rate $P_f = P(m \geq T \mid H_0)$ where m is the number of matching subhashes and T a threshold. For the true positive rate, the hashes of all telephone spam recordings and their distorted versions are compared. P_d is the quotient of the number of positive identifications and the number of expected identifications. With subhashes of length 60 bits, the recognition rate is relatively low ($\approx 73\%$ for $T = 1$) compared to the common 32-bit hashes. But the identification mainly fails for audio samples which are encoded with low bit rate codecs (G.726 at 16 kBit/s and GSM fullrate at 13 kBit/s). We observe that the hit rate is much higher ($\approx 97\%$ for $T = 1$) if these two codecs are not incorporated. The true positive rate for various thresholds is depicted in Figure 3.

The false positive rate P_f is computed relative to a given repository of audio hashes. Hence P_f depends on the number of hashes in the repository, but this reflects the *error of first kind* in an identification scenario. We used $N = 5{,}330$ perceptually different audio samples from the above corpus and performed $N(N-1)/2$ pairwise hash comparisons. A hash is considered a false positive if it has at least T common subhashes with any of the other $N-1$ hashes. We observed only 12 false positives for $T = 1$ and even not a single false positive for $T \geq 2$ (see Figure 3). This advantageous property is mainly due to the large bit-length and the entropy of our subhashes. For the usual 32-bit subhashes, random collisions occur much more often. On the other hand, shorter subhashes provide more robustness and better true positive rates.

For the identification of telephone spam, a significant rate of false negatives can be accepted since the audio data will be replayed a number of times. But false positive identifications of telephone spam should be avoided, even for large hash repositories.

5 CONCLUSIONS

We studied the security and privacy requirements of audio fingerprints and analyzed the existing approaches and algorithms. There exist various powerful fingerprinting frameworks which permit an efficient identification of audio samples. Some work has been done on the security of audio hashes, but open issues remain if the hash is used for multimedia authentication and watermarking. This contribution analyzes the privacy issues which are relevant for speech data, for example to identify replayed telephone data (spam calls). The fingerprint should not leak information on the original audio data.

By modifying well known audio fingerprinting algorithms and combining them with a cryptographic message authentication code, we defined a randomized audio hash which consists of a set of binary vectors. We estimated the entropy of the subhash values which is important for the security properties of the proposed method. Furthermore, we analyzed the performance in terms of robustness and discrimination power. We showed that the hash has adequate robustness, at least if the audio samples have sufficient audio quality, and excellent discrimination capabilities. The hash permits an efficient identification of speech signals in large databases and prevents the exposure of audio content.

Future work will incorporate additional audio material and extend the study of the security properties of robust keyed hash functions.

REFERENCES

Bavarian Archive for Speech Signals (1998). Verbmobil II.

Bellare, M. (2006). New proofs for NMAC and HMAC: Security without collision-resistance. *Advances in Cryptology-CRYPTO 2006*, pages 602–619.

Bellare, M., Canetti, R., and Krawczyk, H. (1996). Keying hash functions for message authentication. In *Advances in Cryptology—CRYPTO'96*, pages 1–15. Springer.

Cano, P., Batlle, E., Kalker, T., and Haitsma, J. (2002). A Review of Algorithms for Audio Fingerprinting. In *Multimedia Signal Processing, IEEE Workshop on*, pages 169–173.

Clausen, M. and Kurth, F. (2004). A unified approach to content-based and fault-tolerant music recognition. *IEEE Transactions on Multimedia*, 6(5):717–731.

Cremer, M., Froba, B., Hellmuth, O., Herre, J., and Allamanche, E. (2001). AudioID: Towards Content-Based Identification of Audio Material. In *Audio Engineering Society Convention 110*.

Doets, P. J. O. and Lagendijk, R. L. (2008). Distortion Estimation in Compressed Music Using Only Audio Fingerprints. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2).

Fridrich, J. and Goljan, M. (2000). Robust Hash Functions for Digital Watermarking. In *Information Technology: Coding and Computing, International Conference on*, pages 178–183.

Grutzek, G., Strobl, J., Mainka, B., Kurth, F., Poerschmann, C., and Knospe, H. (2012). Perceptual hashing for the identification of telephone speech. *Speech Communication; 10. ITG Symposium; Proceedings of*, pages 1–4.

Haitsma, J. and Kalker, T. (2002). A highly robust audio fingerprinting system. In *Proc. ISMIR*, volume 2, pages 13–17.

Koval, O., Voloshynovskiy, S., Bas, P., and Cayre, F. (2009). On security threats for robust perceptual hashing. In *IS&T/SPIE Electronic Imaging 2009*.

Koval, O., Voloshynovskiy, S., Beekhof, F., and Pun, T. (2008). Security analysis of robust perceptual hashing. In *IS&T/SPIE Electronic Imaging 2008*.

Kurth, F. and Müller, M. (2008). Efficient Index-Based Audio Matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):382–395.

Slaney, M. and Casey, M. (2008). Locality-sensitive hashing for finding nearest neighbors [lecture notes]. *Signal Processing Magazine, IEEE*, 25(2):128–131.

Swaminathan, A., Mao, Y., and Wu, M. (2006). Robust and Secure Image Hashing. *IEEE Transactions on Information Forensics and Security*, 1(2):215–230.

Thiemert, S., Nurnberger, S., Steinebach, M., and Zmudzinski, S. (2009). Security of robust audio hashes. In *Information Forensics and Security, 2009. First IEEE International Workshop on*, pages 126 –130.

Wang, A. L.-C. (2003). An Industrial-Strength Audio Search Algorithm. *ISMIR 2003, 4th Symposium Conference on Music Information Retrieval*, pages 7–13.

Wang, A. L.-C. and Smith III, J. O. (2008). Methods for recognizing unknown media samples using characteristics of known media samples.

Weng, L. and Preneel, B. (2011). A secure perceptual hash algorithm for image content authentication. In *Communications and Multimedia Security*, pages 108–121.

Zmudzinski, S. and Steinebach, M. (2009). Perception-based Authentication Watermarking for Digital Audio Data. In *IS&T/SPIE Electronic Imaging 2009*.