LETTER *Special Section on Enriched Multimedia —New Technology Trends in Creation, Utilization and Protection of Multimedia Information—*

A Resilience Mask for Robust Audio Hashing

Jin S. SEO[†a)], ***Member***

SUMMARY Audio hashing has been successfully employed for protection, management, and indexing of digital music archives. For a reliable audio hashing system, improving hash matching accuracy is crucial. In this paper, we try to improve a binary audio hash matching performance by utilizing auxiliary information, resilience mask, which is obtained while constructing hash DB. The resilience mask contains reliability information of each hash bit. We propose a new type of resilience mask by considering spectrum scaling and additive noise distortions. Experimental results show that the proposed resilience mask is effective in improving hash matching performance.

key words: *audio hashing, fingerprinting, resilience mask, power mask*

1. Introduction

With the huge volume of audio data available for protection, browsing, and indexing, there is a strong need to identify a given audio clip fast and reliably using its own features, which are called hash of the audio clip. Audio hashing can find the audio clips with the same origin but in a different shape or form by allowing for some modification of audio content to the amount that human auditory system cannot discern [1], [2]. Audio hashing is also known as audio identification or audio fingerprinting. The typical types of the perceptual-quality preserving distortions include filtering, compression, recording, resampling, time-scale modification, enhancement, and analog-to-digital conversion which are often encountered in the normal audio distribution chain.

Constructing a relevant hash function for audio signal is not a trivial task since the human auditory perception is intricate to model with a condensed representation. Including all the discriminant auditory information in a series of hash bits is almost impossible or at least not viable for now. As a way to supplement audio hash, the resilience-weighted hash matching method [3] has been proposed. The resilience weights are obtained from the remaining discriminant information which is not included in the hash bits and thus discarded. As shown in Fig. 1, the resilience mask, which encodes expected robustness of each hash bit, is obtained, while extracting hash from the original database (DB) songs, and stored along with the hash DB. Query hash, extracted from an unknown audio clip, is compared with the hash in the hash DB with the weights in the resilience mask. In [3], the resilience-weighted Hamming distance was proposed to improve robustness against additive noise over the conventional Hamming distance. This work extends the previous resilience-weighted hash matching method [3] by incorporating a novel resilience mask. To cope with a broader range of audio distortions, such as compression and filtering, this paper derives a mask which is resilient against spectrum scaling and additive noise. Since the performance of the weighted Hamming distance depends crucially on how much accurate the assessment of expected resilience of hash bits, study on deriving more accurate resilience weights is utmost important.

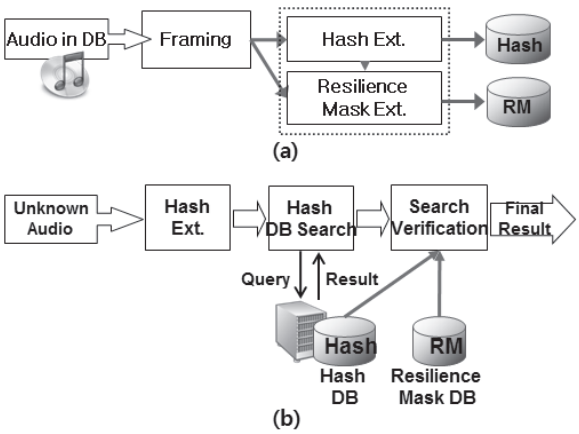

Fig. 1 The overview of (a) hash and resilience mask generation and (b) audio hashing system used for identification.

The baseline audio hashing method, we consider in this paper, is the Philips robust hash (PRH) [1] which has been widely accepted as one of the standard approaches in audio hashing [4]–[6]. The PRH takes the sign (phase) of the subband-energy difference as a fingerprint bit and discards the magnitude of the subband-energy difference. This paper incorporates the discarded information in deriving a resilience mask of each hash bit against the audio distortions with spectrum scaling and noise addition. By assuming that the hash bits with larger resilience mask is more likely to be robust, the weighted Hamming distance is used for hash matching. The proposed resilience mask showed better identification performance than the conventional power mask [3].

This paper is organized as follows. Section 2 describes the proposed resilience-mask generation method for audio

Manuscript received April 2, 2016.
Manuscript revised July 19, 2016.
Manuscript publicized October 7, 2016.

[†]The author is with the Department of Electrical Engineering, Gangneung-Wonju National University, Gangneung, Rep. of Korea.
a) E-mail: jsseo@gwnu.ac.kr

DOI: 10.1587/transinf.2016MUL0003

hash. Section 3 compares the identification performance of the proposed resilience mask with that of the previous one in [3]. Finally, Sect. 4 concludes the paper.

2. Proposed Resilience Mask for Audio Hashing

This paper studies a way to improve audio hashing performance by coding resilience information of each hash bit in the form of binary mask.

2.1 Power Mask for Philips Robust Hash

The baseline audio hashing method, considered in this paper is PRH [1] which has been studied in depth and regarded as one of the standard approaches in audio hashing. Let $E_{n,m}$ be the subband energy of an audio signal at the band m of the n-th frame. The subbands lie in the range of 300Hz to 2000Hz and consists of 33 logarithmically spaced bands. Then the $E_{n,m}$ is filtered by a simple 2D difference filter (along both the frequency and the temporal axis) as follows:

$$F_{n,m} = E_{n,m} - E_{n,m+1} - E_{n-1,m} + E_{n-1,m+1} \quad . \tag{1}$$

Then the hash bit $H[n,m]$ of PRH is obtained by taking the sign of $F_{n,m}$ given by

$$H[n,m] = \begin{cases} 1 & F_{n,m} > 0 \\ 0 & F_{n,m} \le 0 \end{cases} \quad . \tag{2}$$

From the 33 subbands of each frame, we obtain 32 bits ($M = 32$). As the 32-bit hash from a frame does not contain enough information to identify the whole audio, a hash block, which is composed of N consecutive frames (typically $N = 256$, consequently $NM = 8192$ fingerprint bits), is used for hash matching. In the PRH, the Hamming distance D_H between the hash blocks H_1 and H_2 from the two different audio clips is used for hash matching as follows:

$$D_H(H_1, H_2) = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M} d[n,m]}{NM} \tag{3}$$

where $d[n,m] = \mathrm{XOR}(H_1[n,m], H_2[n,m])$. Two hash blocks are declared similar if their Hamming distance D_H, expressed as Bit Error Rate (BER), is below a certain threshold.

The Hamming distance used in PRH treats all the hash bits equally in comparison. However; obviously we could expect that the hash bit $H[n,m]$ with larger energy-difference magnitude $|F_{n,m}|$ is more likely to be robust against additive-noise distortions. The authors of [3] utilize this intuition in deriving power mask for PRH and propose the resilience-weighted Hamming distance for PRH matching. As shown in Fig. 1, the resilience weights are stored in the hash DB. Since it is infeasible to store the resilience weights of the large number of music clips in the hash DB due to storage restrictions, the resilience weight of each hash bit is binarized: the resilience of strong bits is assigned to one, and that of the weak bits is assigned to zero [3]. The strong bits are the T bits with the largest T energy-difference magnitude $|F_{n,m}|$, for each frame, and the weak bits are the other $M - T$ bits (typically $T = 24$). We set the resilience weight $w[n,m] = 1$ for strong bits and $w[n,m] = 0$ for weak bits. Then the weighted Hamming distance D_W between two hash blocks H_1 and H_2 is given as follows:

$$\begin{aligned} D_W(H_1, H_2) = & \frac{\sum_{n=1}^{N}\sum_{m=1}^{M} \alpha(1 - w[n,m])d[n,m]}{N(\alpha(M-T) + \beta T)} \\ & + \frac{\sum_{n=1}^{N}\sum_{m=1}^{M} \beta w[n,m]d[n,m]}{N(\alpha(M-T) + \beta T)} \end{aligned} \tag{4}$$

where α and β are weight factors to weak and strong bits respectively (typically $\alpha = 0.5$ and $\beta = 1$).

2.2 Proposed Resilience Mask for Philips Robust Hash

This paper proposes a resilience mask of PRH against the audio distortions with spectrum scaling and noise addition, which are commonly occurred during filtering and compression. The subband energy $E'_{n,m}$ subjected to the considered scaling distortion and an additive noise is given by

$$E'_{n,m} = \gamma_{n,m} E_{n,m} + \eta_{n,m} \tag{5}$$

where $\gamma_{n,m}$ is a scaling factor, and $\eta_{n,m}$ is an additive noise. Then the 2D difference-filter output $F'_{n,m}$ can be represented by the sum of $F_{n,m}$, the scaling distortion $K_{n,m}$, and the additive noise $\psi_{n,m}$ as follows:

$$F'_{n,m} = F_{n,m} + K_{n,m} + \psi_{n,m} \tag{6}$$

where

$$\begin{aligned} K_{n,m} = & (\gamma_{n,m} - 1)E_{n,m} + (1 - \gamma_{n,m+1})E_{n,m+1} \\ & + (1 - \gamma_{n-1,m})E_{n-1,m} + (\gamma_{n-1,m+1} - 1)E_{n-1,m+1} \end{aligned}$$

and $\psi_{n,m} = \eta_{n,m} - \eta_{n,m+1} - \eta_{n-1,m} + \eta_{n-1,m+1}$. For the sign of $F_{n,m}$ to be the same as that of $F'_{n,m}$ regardless of the value of scaling factors and the amount of noise, the following inequality should be satisfied:

$$|F_{n,m}| > |K^{\max}_{n,m}| + |\psi^{\max}_{n,m}| \tag{7}$$

where $|K^{\max}_{n,m}|$ and $|\psi^{\max}_{n,m}|$ are the maximum amount of expected scaling distortion and additive noise respectively which preserves audio perceptual quality. By assuming that the scaling factors change slowly along the time axis due to the large amount of overlap between adjacent frames of PRH [1], which leads to $\gamma_{n,m} = \gamma_{n-1,m}$ and $\gamma_{n,m+1} = \gamma_{n-1,m+1}$, the scaling distortion can be approximated by the sum of the temporal energy differences given by

$$K_{n,m} = (\gamma_{n,m} - 1)\Delta_{n,m} + (1 - \gamma_{n,m+1})\Delta_{n,m+1} \tag{8}$$

where $\Delta_{n,m} = E_{n,m} - E_{n-1,m}$. For a constant $\gamma^{\max}$, which is perceptually-tolerable maximum scaling factor, we set the $|K^{\max}_{n,m}|$ as

$$|K^{\max}_{n,m}| = (\gamma^{\max} - 1)(|\Delta_{n,m}| + |\Delta_{n,m+1}|) \quad . \tag{9}$$

By assuming that the magnitude of the additive noise might

not be greater than the subband energy itself, we make two different assumptions for $|\psi_{n,m}^{\max}|$; the first assumption is based on the local energy sum given by

$$|\psi_{n,m}^{\max 1}| = E_{n,m} + E_{n,m+1} + E_{n-1,m} + E_{n-1,m+1} \ , \qquad (10)$$

and the other one is based on the frame-level average energy given by

$$|\psi_{n,m}^{\max 2}| = \frac{1}{2(M+1)} \sum_{i=n-1}^{n} \sum_{j=1}^{M+1} E_{i,j} \ . \qquad (11)$$

From the inequality (7), we propose a resilience mask $RM_{n,m}$ for the hash bit $H[n,m]$ as follows:

$$RM_{n,m} = \frac{|F_{n,m}|}{|K_{n,m}^{\max}| + |\psi_{n,m}^{\max}|} \qquad (12)$$

where $\gamma^{\max}$ in (9) (typically, $\gamma^{\max} = 15$) which controls the relative importance of scaling distortion and noise addition. The resilience masks corresponding to the two different assumptions on the additive noise, $|\psi_{n,m}^{\max 1}|$ and $|\psi_{n,m}^{\max 2}|$, are denoted by $RM_{n,m}^1$ and $RM_{n,m}^2$ respectively. As the value of $RM_{n,m}$ is greater, the corresponding hash bit $H[n,m]$ is expected to be more robust against considered distortions. As in [3], the T bits with the largest T resilience mask, $RM_{n,m}$, are the strong bits with resilience weight $w[n,m] = 1$, and the other $M - T$ bits are weak bits with $w[n,m] = 0$. The same weighted Hamming distance in (4) is used in hash comparison.

3. Experimental Results

The performance of the proposed resilience-mask based hash matching method was evaluated using the hash and the mask DB generated from thousand songs belonging to various genres, such as classic, jazz, pop, rock, and hiphop. We extract 32-bit hash and resilience mask for every frame of the songs from the energy differences of 33 bands which lie in the range from 300Hz to 2000Hz as in [1]. To test the effectiveness of the proposed resilience mask, the songs in the DB were subjected to various kinds of audio distortions (using Cool Edit Pro 2.1 software and Matlab) including telephone bandpass filter (BF), 3:1 expander below 10dB (EX), notch filter (NF), a filter emulating auditorium (AU), a filter emulating small room (SR), 10-dB white noise (WN), 48-kbps mp3 compression (MP). The hash matching was performed by the weighted Hamming distance in (4) which compares the hash block composed of 256 subsequent 32-bit hash vector (in total, 8192 bits) of the distorted songs with that of the corresponding original songs in the DB.

We consider both single and composite distortions in tests and compare the matching performance of the proposed resilience mask with that of the previous one [3]. To compare the matching performance of the proposed resilience mask with that of the previous one, we use the relative performance gain (RPG) defined by

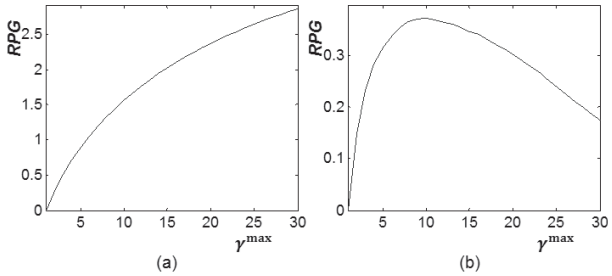

Fig. 2 RPG versus $\gamma^{\max}$ against the audio distortions; (a) filter emulating small room and (b) 20-dB white noise.

Table 1 Mean of the measured BER under various types of audio distortions. The $D_{W,PR}$ and $D_{W,RM}$ are the weighted Hamming distance with the weights from the power mask [3] and the proposed resilience mask $RM_{n,m}^1$ with $\gamma^{\max} = 15$ respectively.

Distortions	D_H	$D_{W,PR}$	$D_{W,RM}$	RPG (%)
AU	0.1066	0.0889	0.0806	9.10
EX	0.1400	0.1215	0.1152	5.19
NF	0.1114	0.0942	0.0899	4.87
SR	0.1254	0.1048	0.0980	6.96
MP	0.0741	0.0531	0.0519	3.15
EX+BF	0.1384	0.1181	0.1124	4.89
WN+MP	0.1330	0.1054	0.1077	-1.78

Table 2 Mean of the measured BER under various types of audio distortions. The $D_{W,PR}$ and $D_{W,RM}$ are the weighted Hamming distance with the weights from the power mask [3] and the proposed resilience mask $RM_{n,m}^2$ with $\gamma^{\max} = 15$ respectively.

Distortions	D_H	$D_{W,PR}$	$D_{W,RM}$	RPG (%)
AU	0.1066	0.0889	0.0879	1.17
EX	0.1400	0.1215	0.1207	0.74
NF	0.1114	0.0942	0.0918	2.41
SR	0.1254	0.1048	0.1026	2.02
MP	0.0741	0.0531	0.0523	1.65
EX+BF	0.1384	0.1181	0.1166	1.46
WN+MP	0.1330	0.1054	0.1047	0.80

$$\text{RPG} = 100 \times \frac{D_{W,PR} - D_{W,RM}}{D_{W,PR}} \qquad (13)$$

where $D_{W,PR}$ and $D_{W,RM}$ are the weighted Hamming distance (4) with the weights from the power mask [3] and the proposed resilience mask (12) respectively. Figure 2 shows the behavior of the RPG versus the $\gamma^{\max}$ against two different audio distortions for $RM_{n,m}^2$. The RPG is increasing for the distortions with filtering, while it is increasing at first and then decreasing for the noise addition. If we have a knowledge on the expected audio distortions, we may adjust the value of the $\gamma^{\max}$ corresponding to them. However; even though we do not know the expected audio distortions, the value of $\gamma^{\max}$ between 5 and 20 was effective for most of the common audio distortions. Thus we chose $\gamma^{\max} = 15$ in our experiment. Table 1 and 2 show the hash matching results of the considered distortions for the resilience mask $RM_{n,m}^1$ and $RM_{n,m}^2$ respectively. The $RM_{n,m}^1$ was more effective in boosting hash matching performance under various filtering attacks than the $RM_{n,m}^2$. Except the white noise addition, the resilience mask $RM_{n,m}^1$ outperformed the con-

ventional power mask in terms of RPG by more than 3% (up to 9%). The $RM^2_{n,m}$ showed stable performance improvement over the power mask against both filtering and noise addition. The $RM^1_{n,m}$ is based on the local-energy assumption $|\psi^{\max 1}_{n,m}|$, that is somehow mingled with the local scaling distortion $|K^{\max}_{n,m}|$, while the $RM^2_{n,m}$ is based on the frame-level energy $|\psi^{\max 2}_{n,m}|$, that is clearly separated with $|K^{\max}_{n,m}|$. The $RM^1_{n,m}$ was more effective for filtering-based distortions, while the $RM^2_{n,m}$ showed better performance for noise addition.

4. Conclusion

This paper proposes a novel resilience mask for the weighted Hamming distance in audio hash matching. The resilience mask encodes the expected robustness of each hash bit and is stored in hash DB. To derive a reliable resilience mask against various audio distortions, we model the audio distortions as spectrum scaling and noise addition. In this paper, we applied the proposed method to PRH and tested its performance over thousands of audio clips with various audio distortions. For most of considered distortions, the proposed resilience mask improved matching performance over the previous power mask. This work shows that refined resilience mask is promising in boosting hash matching performance. Further study includes an extension of the proposed resilience mask to other types of hashing methods.

Acknowledgments

This research project was supported by Ministry of Culture, Sports and Tourism (MCST) and from Korea Copyright Commission in 2016. [Development of predictive detection technology for the search for the related works and the prevention of copyright infringement]

References

[1] J. Haitsma and T. Kalker, “A highly robust audio fingerprinting system,” Proc. International Conf. on Music Information Retrieval, 2002.
[2] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, “A review of algorithms for audio fingerprinting,” Proc. IEEE Workshop on Multimedia Signal Processing, pp.169–173, 2002.
[3] B. Coover and J. Han, “A power mask based audio fingerprint,” Proc. IEEE ICASSP, pp.1394–1398, 2014.
[4] M. Park, H.R. Kim, Y.M. Ro, and M. Kim, “Frequency filtering for a highly robust audio fingerprinting scheme in a real-noise environment,” IEICE Trans. Inf. & Syst., vol.89, no.7, pp.2324–2327, 2006.
[5] P. Doets and R. Lagendijk, “Distortion estimation in compressed music using only audio fingerprints,” IEEE Transactions on Audio, Speech, and Language Processing, vol.16, no.2, pp.302–317, Feb. 2008.
[6] J. Seo, “An asymmetric matching method for a robust binary audio fingerprinting,” IEEE Signal Process. Lett., vol.21, no.7, pp.844–847, 2014.