Frequency-Temporal Filtering for a Robust Audio Fingerprinting Scheme in Real-Noise Environments

Mansoo Park, Hoi-Rin Kim, and Seung Hyun Yang

ABSTRACT—In a real environment, sound recordings are commonly distorted by channel and background noise, and the performance of audio identification is mainly degraded by them. Recently, Philips introduced a robust and efficient audio fingerprinting scheme applying a differential (high-pass filtering) to the frequency-time sequence of the perceptual filter-bank energies. In practice, however, the robustness of the audio fingerprinting scheme is still important in a real environment. In this letter, we introduce alternatives to the frequency-temporal filtering combination for an extension method of Philips' audio fingerprinting scheme to achieve robustness to channel and background noise under the conditions of a real situation. Our experimental results show that the proposed filtering combination improves noise robustness in audio identification.

Keywords— Music information retrieval, audio fingerprint, frequency filtering, temporal filtering.

I. Introduction

Recently content-based music information retrieval (MIR) has been noted as an attractive state-of-the-art application service in wire/wireless communication. For example, many companies nowadays offer application services that can not only provide information on songs being played over public loudspeakers, but can also be used to monitor broadcast music automatically or prevent the unauthorized sharing of music files over peer-to-peer networks [1], [2]. To apply commercial service, recent reports focus on the concurrent guarantee of both scalability and quality for MIR by content-based audio identification in a large database [3]-[7].

Philips' audio fingerprinting scheme, one of the most recent content-based audio identification techniques, is definitely suited to the above purpose [3]. In a real environment, however, problems still remain such as channel and background noise, speed-changes, arrangements, and so on. In particular, noise robustness is a major challenge in a real environment, just as in the general topic of speech recognition technique. Generally, sound recordings are easily corrupted by linear or non-linear distortion caused by channel and background noise in a real environment. Hence, false audio identification is mainly caused by a mismatch between the original audio signal and the distorted one. This letter concerns an audio fingerprinting scheme for noise-robustness, one of the major issues in MIR.

II. Audio Fingerprinting Scheme

1. Philips' Audio Fingerprinting Scheme

An overview of Philips' scheme is depicted in Fig. 1. A Mel or Bark scale filter-bank is commonly used to reflect the perceptual characteristics of an audio signal in this work. A sub-fingerprint for every frame is based on a sign of the power spectrum, differentiated simultaneously along the time and frequency axes. The differentiation of spectral parameters along the frequency or time axes corresponds to high-pass filtering. It may be possible to remove slowly varying components as undesired perturbations. In addition, the differentiated power spectrum is uncorrelated with its temporal and frequency neighbors. Here, a sub-fingerprint is typically a 32-bit code from 33 perceptually divided frequency bands for every frame. The 32-bit code is usually referred to as a hash value which acts as a direct addressing point for database lookup. That is, a sub-fingerprint can be hashed by a 32-bit code. The bit is assigned as

Manuscript received Dec. 11, 2005; revised May 06, 2006.

Mansoo Park (phone: +82 11 9746 4173, email: mansoo@icu.ac.kr) and Hoi-Rin Kim (email: hrkim@icu.ac.kr) are with Multimedia Group, Information and Communications University, Daejeon, Korea.

Seung Hyun Yang (email: yan@konantech.com) is with Information & Mining Research Team, Konan Technology Inc., Seoul, Korea.

$$H(n,m)=\begin{cases}1 \text{ if } E(n,m)-E(n,m+1)-(E(n-1,m)-E(n-1,m+1))>0\\ 0 \text{ if } E(n,m)-E(n,m+1)-(E(n-1,m)-E(n-1,m+1))\leq 0,\end{cases} \tag{1}$$

where $E(n, m)$ is the energy of the n-th frame and the m-th band. This scheme can be very efficient for database lookup since the hash value is highly unique. For fast database lookup, the matching candidates can be selected by hash values with pre-determined Hamming distance. In this scheme, the similarity measure is based on the Hamming distance between hash values. The best-matched result is determined on the basis of the bit error rate (BER) per fingerprint block over an audio clip.

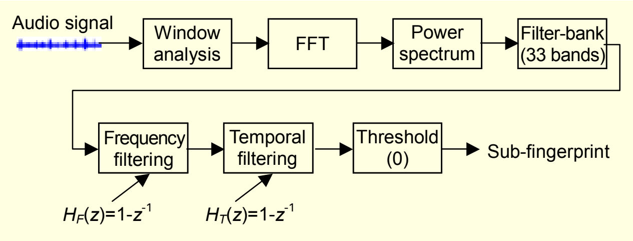

Fig. 1. Overview of Philips' audio fingerprinting scheme.

2. Alternatives to Frequency Filtering

As a matter of fact, the above scheme is insufficient for highly robust audio hashing in a real-noise condition since the filter-bank energies (FBEs) are still correlated. When some bands are corrupted by noise, the Hamming distance could be greater between an original and a distorted sub-fingerprint owing to the correlation of the FBEs. Thus, the audio fingerprint would be much more robust to noise if the FBEs were decorrelated. Frequency filtering is generally used to decorrelate the FBEs, which is somewhat verified in the speech recognition system [8]-[10]. The typical frequency filtering techniques are defined as

$$H_{F1}(z)=1-z^{-1}, \tag{2}$$

$$H_{F2}(z)=\frac{\eta\cdot(1-z^{-1})}{(\eta+1)\cdot\left(1+\left(\frac{\eta-1}{\eta+1}\right)\cdot z^{-1}\right)} \quad where \quad \eta=0.5, \tag{3}$$

$$H_{F3}(z)=z-z^{-1}, \tag{4}$$

where H_{F1} is a high-pass first-order FIR filter, H_{F2} is a high-pass IIR filter, and H_{F3} is a band-pass second-order FIR filter. The H_{F1} used in the Philips' scheme could lose a significant amount of spectral information because it has a very steep slope, as shown in Fig. 2. Even if H_{F2} has a proper slope, it still weights high-quefrency components. The frequency filter should reduce noise characteristics that are slowly varying. As well, high-quefrency components should be cut off because, theoretically, they contain less information on audio characteristics. Thus, the band-pass filter may be more relevant to frequency filtering in this work. Our main goal in this letter is to find the most relevant frequency filter to extract a noise-robust audio fingerprint in a real environment. For that reason, H_{F3} would be more effective under real-noise conditions since it has the most suitable filter shape as shown in Fig. 2.

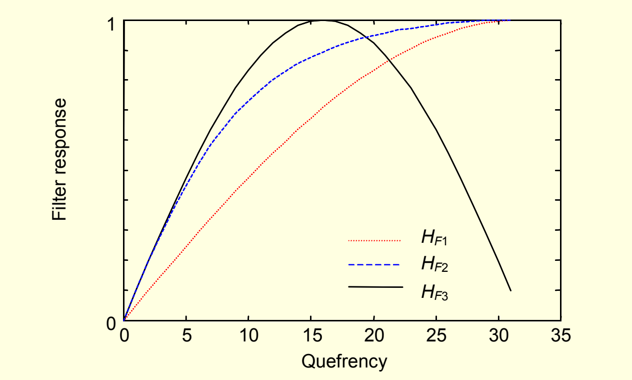

Fig. 2. Filter response of various frequency filters.

3. Alternatives to Temporal Filtering

Temporal filtering of the FBEs is for the purpose of removing D.C. and slowly varying components caused by the undesired perturbations of linear distortion such as channel noise [10], [11]. The audio fingerprints would be much more immune to channel distortion if the FBEs were uncorrelated with their temporal neighbors. The typical temporal filtering techniques are defined as

$$H_{T1}(z)=1-z^{-1}, \tag{5}$$

$$H_{T2}(z)=\sum_{k=1}^{K}k(z^{k}-z^{-k}), \tag{6}$$

$$H_{T3}(z)=0.1\times\frac{2+z^{-1}-z^{-3}-2z^{-4}}{z^{-4}\times(1-\alpha z^{-1})}, \tag{7}$$

where α = 0.94, H_{T1} is a first-order FIR filter, H_{T2} is a typical regression formula, and H_{T3} is the RASTA filter. H_{T2} when K=2 is used in this work. The H_{T1} used in the Philips' scheme may be expected to suppress the effects of convolutional noise by the channel. However, low-pass filtering is necessary for smoothing out the fast spectral change. For that reason, H_{T2} or H_{T3} would be more relevant to temporal filtering for removing channel-distortion.

III. Experiments

1. Audio Data

For experiments in a real environment, an audio query clip was captured using inexpensive microphones which were placed 10 to 20 cm from a 2.1-channel loudspeaker connected to an mp3 player. The audio query is converted to standard PCM format which is sampled at 11.025 kHz and quantized with 16 bits in a mono-channel. Music items for references consisted of 5,000 popular songs in mp3 format (192 kbps, 44.1 kHz, stereo) converted from audio CDs. They were down-sampled to 11.025 kHz in consideration of portable devices such as PDAs or mobile phones. They include various genres such as rock/ballad, pop/dance, hip-hop, and classical.

To evaluate the proposed techniques, the query set consists of five types of audio query data according to the device and the recording environment. The noise data consists of real-noise signals recorded in a real environment. Audio query data are captured from 50 randomly selected songs per set. Each song is played at randomly set offsets 30 times. Only set IV has a 7 second duration; the others have an 8 second duration.

Set I: Directly crop mp3 files.

Set II: Use a stand microphone and 2.1-channel loudspeakers in a very quiet environment.

Set III: Use a pair of stand/pin microphones and 2.1-channel loudspeakers in a noisy environment with TV sounds and human voices.

Set IV: Use a pair of stand/pin microphones and 2-channel laptop PC-speakers, which poorly reproduce sound, in a noisy environment with TV sounds, human voices, and other sporadic noises. In addition, some cases have overflown into the amplitude range of 16-bit PCM due to very loud music sounds.

Set V: Directly extract audio clips from a video file recorded from a TV music show. It is mixed with background noises such as the applauding or cheering sounds of the audience. Here, the vocals are live but the music is pre-recorded.

Noise Data: Record real-noise signals by MD (Sharp: IM-DR 580H) in a real-space such as a department store, restaurant, street, underground shopping center, or home.

In the signal processing step, the audio frame is parameterized into a 0.37 s rate and shifted at an 11.6 ms rate. Considering the human auditory system, the selected frequency bands lie in a range from 300 Hz to 3,000 Hz.

2. Performance Evaluation

In an ideal case, the sub-fingerprint is reliable and there is no bit error. However, it is not perfect when an audio signal is corrupted by a linear or non-linear distortion. To improve this defect, the candidate positions for the database lookup are expanded into hash values with a Hamming distance of a one-bit error [3]. Thus, the system needs 33 times more lookup for audio identification. However, it does not check all hash candidates since it sets the threshold for breaking the database lookup. Empirically, it took only 3 to 4 times more lookup when we set the threshold to 0.33 over fingerprint block.

As shown in Tables 1 and 2, in the cases where the lookup candidates are expanded or not, the alternatives to frequency or temporal filtering generally outperform and are useful for real-application when the query set is corrupted by noise and channel distortion. As expected, H_{F2} is a little better than H_{F1} which is used in the Philips' scheme. On the other hand, H_{F3} is superior to other frequency filtering methods in the noisiest conditions. As expected, the RASTA filter, H_{T3}, is more effective with regard to channel-distortion such as set IV.

Table 1. Recognition performance evaluation of alternatives to frequency filtering when H_{T1} is used as a temporal filter.

Frequency filter / Query	Database lookup candidates					
	Hamming distance = 0			Hamming distance ≤ 1		
	H_{F1}	H_{F2}	H_{F3}	H_{F1}	H_{F2}	H_{F3}
Set I	100%	100%	100%	100%	100%	100%
Set II	98.2%	98.6%	99.6%	100%	100%	100%
Set III	96.2%	97.7%	97%	100%	100%	100%
Set IV	78.7%	77.4%	82.5%	97.2%	98.3%	98.5%
Set V	56.3%	58.2%	68.8%	92.8%	94.7%	98.1%

Table 2. Recognition performance evaluation of alternatives to temporal filtering when H_{F1} is used as a frequency filter.

Temporal filter / Query	Database lookup candidates					
	Hamming distance = 0			Hamming distance ≤ 1		
	H_{T1}	H_{T2}	H_{T3}	H_{T1}	H_{T2}	H_{T3}
Set I	100%	100%	100%	100%	100%	100%
Set II	98.2%	98.2%	99.2%	100%	100%	100%
Set III	96.2%	97.5%	97.7%	100%	100%	100%
Set IV	78.7%	79.1%	86.3%	97.2%	97.2%	99%
Set V	56.3%	58%	63.1%	92.8%	94.5%	95.8%

Table 3. Recognition performance comparison by the frequency-temporal filtering combinations when the lookup candidates are expanded.

Query set	Hamming distance ≤ 1			
	$H_{F1}+H_{T1}$	$H_{F1}+H_{T3}$	$H_{F3}+H_{T1}$	$H_{F3}+H_{T3}$
Set I	100%	100%	100%	100%
Set II	100%	100%	100%	100%
Set III	100%	100%	100%	100%
Set IV	97.2%	99%	98.5%	99.2%
Set V	92.8%	95.8%	98.1%	95.4%

Table 3 shows the performance comparison according to the frequency-temporal filtering combinations. As shown in Table 3, the combination of H_{F3} and H_{T1} is more effective for real-noise such as set V. On the other hand, in the case of the channel-noise such as set IV, the combination of H_{F3} and H_{T3} has the best quality. That is, the RASTA filter is effective in normalizing the channel-effects, and H_{F3} is effective in smoothing out real-noise.

To consider a more realistic situation, the noisy query set was made by adding set II to each noise data in accordance with the signal-to-noise ratio (SNR). Figure 3 shows the effectiveness of the frequency-temporal filtering combinations in real environments. Of course, the results in Fig. 3 were achieved when the lookup candidates were expanded. As expected, the filtering combination of H_{F3} and H_{T1} generally had a very high quality in real environments everywhere. In the cases of street and home noise, the combination of H_{F3} and H_{T3} had the best quality. In these cases it could achieve synergy effects because some of the noise data had time-stationary characteristics like channel-noise. That is, H_{T3} as a temporal filter is only effective with regard to channel-distortion. On the other hand, H_{F3} as a frequency filter is much more effective in extracting the audio fingerprints highly robust to real-noise everywhere.

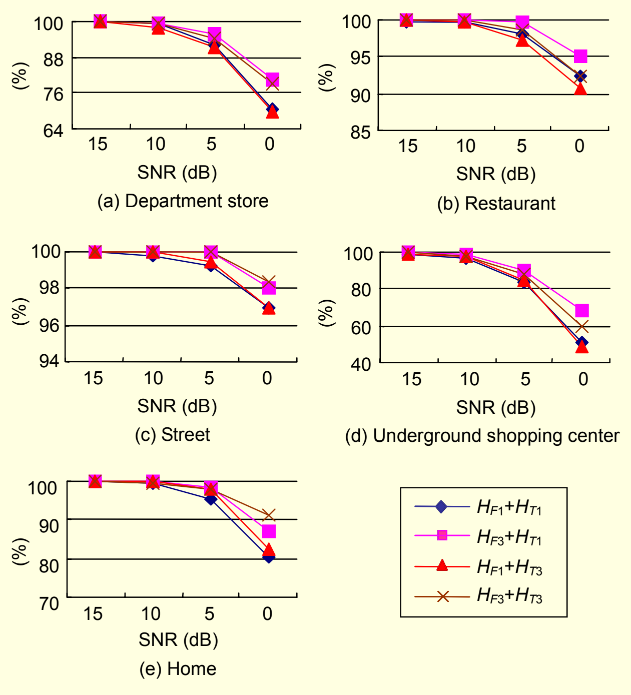

Fig. 3. Performances of the frequency-temporal filtering.

IV. Conclusions

In our experiments, we observed that the alternatives to frequency or temporal filtering, in terms of noise robustness, are generally effective in the case of recording a query signal in a real situation. In particular, the band-pass second-order FIR filter is superior to the other frequency filtering techniques under real-noise conditions. However, the RASTA filter as a temporal filter is robust to only channel-distortion. In this work, there was no synergy of the filtering combination of H_{F3} and the RASTA filter anywhere. For further study, we will seek other methods of frequency-temporal filtering in order to achieve better synergy effects in the above cases.

References

[1] Shazam Entertainment, http://www.shazam.com.

[2] Gracenote, http://www.gracenote.com.

[3] J. Haitsma and T. Kalker, "A Highly Robust Audio Fingerprinting System," *Proc. ISMIR 2002*, 2002, pp. 144-148.

[4] C. Burges, J. Platt, and S. Jana, "Distortion Discriminant Analysis for Audio Fingerprinting," *IEEE Trans. Speech and Audio Processing*, vol. 11, Mar. 2003, pp. 165-174.

[5] M. L. Miller, M. A. Rodriguez, and I. J. Cox, "Audio Fingerprinting: Nearest Neighbor Search in High-dimensional Binary Space," *IEEE Multimedia Signal Processing Workshop*, Dec. 2002, pp. 182-185.

[6] D. Kirovski and H. Attias, "Beat-ID: Identifying Music via Beat Analysis," *IEEE Multimedia Signal Processing Workshop*, Dec. 2002, pp. 190-193.

[7] M. K. Mihcak and R. Venkatesan, "A Perceptual Audio Hashing Algorithm: A Tool for Robust Audio Identification and Information Hiding," *LNCS*, vol. 2137, 2001, pp. 51-65.

[8] J. Chen, K. Paliwal, and S. Nakamura, "Cepstrum Derived from Differentiated Power Spectrum for Robust Speech Recognition," *Speech Communication*, vol. 41, Oct. 2003, pp. 469-484.

[9] H.-Y. Jung, "Filtering of Filter-Bank Energies for Robust Speech Recognition," *ETRI J.*, vol. 26, no. 3, June 2004, pp. 273-276.

[10] C. Nadeu, D. Macho, and J. Hernando, "Time and Frequency Filtering of Filter-Bank Energies for Robust HMM Speech Recognition," *Speech Communication*, vol. 34, Apr. 2001, pp. 93-114.

[11] H. Hermansky et al., "Compensation for the Effect of the Communication Channel in the Auditory-Like Analysis of Speech (RASTA-PLP)," *Proc. Eurospeech*, 1991, pp. 1367-1370.