BOOSTED BINARY AUDIO FINGERPRINT BASED ON SPECTRAL SUBBAND MOMENTS

Sungwoong Kim and Chang D. Yoo

Div. of EE, Dept. of EECS, KAIST,
373-1 Guseong-dong, Yuseong-gu, Daejeon, 305-701, Korea
leehwiso@kaist.ac.kr, cdyoo@ee.kaist.ac.kr

ABSTRACT

An audio fingerprinting system identifies an audio based on a unique feature vector called the audio fingerprint. The performance of an audio fingerprinting system is directly related to the fingerprint that the system uses. To reduce both the DB size and the DB search time, binary fingerprints are often used. However converting a real-valued fingerprint into a binary fingerprint results in loss of information and leads to severe degradation in performance. In this paper, an algorithm known as boosting is used as a binary conversion method which minimizes the degradation. The experimental results showed that the proposed binary audio fingerprint obtained by boosting the spectral subband moments outperformed some of the state-of-the-art binary audio fingerprints in the context of both robustness and pair-wise independence (reliability).

Index Terms— Audio fingerprint, Spectral subband moment, Boosting

1. INTRODUCTION

An audio fingerprinting system identifies an audio by first extracting short feature vectors called audio fingerprints from the query audio clip and then identifying the audio whose fingerprints are closest to the query fingerprint. The audio fingerprints of all audio are initially stored in a database (DB). The performance of an audio fingerprinting system, which is often measured in terms of pair-wise independence (reliability) and robustness [1], is directly related to the fingerprint that the system uses. Since fast large-scale search is also essential in an audio fingerprinting system, the audio fingerprint also needs to be as compact as possible.

Recently, many audio fingerprinting systems based on a binary fingerprint have been proposed [2][3]. Haitsma and Kalker [2] calculate the subband energy differences between adjacent frames and then generate a binary fingerprint by quantizing the difference with a single bit. Ke *et al.* [3] generalize the Haitsma and Kalker's method by using the *pair-wise boosting* which is a variant of a well-known boosting technique called the AdaBoost [4]. The binary fingerprint is desirable for fast DB search since it enables direct indexing instead of a range search which is inevitable in most multimedia fingerprinting systems. Direct indexing uses a look-up table or a hash table to search the fingerprint in the DB and does not suffer from the curse of dimensionality. However the conversion of a real-valued fingerprint to a binary fingerprint results in loss of information and can lead to severe degradation in performance, which comes about due to the loss in pair-wise independence property.

In this paper, a binary audio fingerprint that is compact for a fast DB search while minimizing the performance degradation is proposed. The proposed audio fingerprint is based on the first-order normalized subband spectral moment [1]. The first-order normalized moment is known to be not only reliable but also robust against common audio processing steps including lossy compression, random start, equalization, etc. A well-known boosting technique known as the AdaBoost [4] is sometimes used in the audio classification problem to improve the performance of the classifier through learning [6][7]. The modified AdaBoost algorithm [3] is used in this paper to obtain a binary fingerprint from the first-order normalized subband spectral moment. The experimental results show that the proposed audio fingerprint outperforms other state-of-the-art binary fingerprint in the context of audio identification.

The rest of the paper is organized as follows. Section 2 describes the proposed audio fingerprint and boosting algorithm. Section 3 evaluates the performance of the proposed audio fingerprint for various distortions and compares the performance with that of other state-of-the-art audio fingerprints. Finally, Section 4 concludes this paper.

2. PROPOSED AUDIO FINGERPRINT

Fig. 1 shows the overall procedure to extract binary audio fingerprints from an audio clip. First, base audio features are extracted from the audio signal. As explained above, the first-order normalized spectral subband moment (η_1) is used as the base audio feature in this paper. Then, the base audio feature which consists of real-valued elements is converted to a binary fingerprint. The details of the proposed audio fingerprint are explained in the following subsections.

2.1. Normalized spectral subband moments

The νth-order moment of the mth subband in the nth frame is defined as

$$\zeta_\nu[n,m] = \sum_{k=C[m]+1}^{C[m+1]} k^\nu P[n,k] \tag{1}$$

where $C[m]$ and $P[n,k]$ denote the frequency boundary of the mth critical band and the short-time power spectrum of audio signal at frequency bin k of the nth frame, respectively. Then the first-order normalized spectral subband moment $\eta_1[n,m]$ of the mth subband in the nth frame is defined as

$$\eta_1[n,m] = \frac{\zeta_1[n,m]}{\zeta_0[n,m]}. \tag{2}$$

The η_1 has a range between -0.5 and 0.5 in all critical bands.

The base audio feature based on the η_1 is extracted as follows. First, an input audio is converted to mono and downsampled to 11025 Hz. Next, the downsampled signal is split into overlapping frames windowed by Hamming window. The window size is 4096 samples

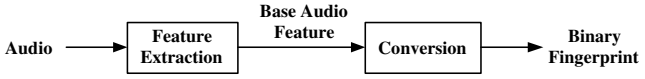


Fig. 1. Extraction binary fingerprint from audio

(0.372 s), and the adjacent frames are overlapped by 2048 samples (0.186 s). Then short-time Fourier transform (STFT) is applied to each frame to obtain the spectrum. The spectrum of each frame is divided into 16 critical bands from 300 to 5300 Hz, and finally the η_1s are computed at each critical band. As a result of the extraction, an 16 dimensional real-valued base audio feature is obtained every 2048 samples (0.186 s).

2.2. Candidate Set for The Binary Conversion Method

The binary conversion should preserve both the robustness and the pair-wise independence of the base audio feature. In [2], the binary fingerprint is obtained from the subband energies as follows:

$$F(n,m) = \begin{cases} 1 & \text{if } E(n,m) - E(n,m+1) \\ & -(E(n-1,m) - E(n-1,m+1)) > 0 \\ 0 & \text{if } E(n,m) - E(n,m+1) \\ & -(E(n-1,m) - E(n-1,m+1)) \leq 0 \end{cases} \quad (3)$$

where $F(n,m)$ and $E(n,m)$ denotes the m-th bit of fingerprint of the n-th frame and the energy of the m-th band of the n-th frame, respectively. The binary conversion by (3) decides the binary value as the sign of the difference between features at adjacent location in the direction of both frequency and time.

For a more resilient and discriminative binary fingerprint, it may be necessary to consider taking differences between adjacent features extracted from a wider time-frequency region: the binary conversion can be generalized to cover features from time-frequency region of different bandwidth and time-width. The difference of these features can be taken along the time, the frequency, and both time and frequency. These three types constitute a candidate set for finding the optimum binary fingerprint. Mathematically, the three types of binary-fingerprint candidates for the m-th bit of the n-th frame, $F_i(n,m)$ $(i = 1,2,3)$, can be defined as

$$F_i(n,m) = \begin{cases} 1 & \text{if } d_i > T_i \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where d_i and T_i denote the i-th difference formula and the i-th threshold. The three types of difference formulas are given by

$$d_1 = R(n,m_s;N,M) - R(n+N,m_s;N,M) \quad (5)$$

$$d_2 = R(n,m_s;N,M) - R(n,m_s+M;N,M) \quad (6)$$

$$d_3 = \big[R(n,m_s;N,M) + R(n+N,m_s+M;N,M)\big] - \big[R(n+N,m_s;N,M) + R(n,m_s+M;N,M)\big] \quad (7)$$

where

$$R(n_s,m_s;N,M) = \sum_{\Delta m=0}^{M} \sum_{\Delta n=0}^{N} \xi(n_s+\Delta n, m_s+\Delta m) \quad (8)$$

and $\xi(n,m)$ denotes the subband feature of the m-th band of the n-th frame. The subscript i denotes the direction of difference ($i = 1$: time(frame), $i = 2$: frequency(band), $i = 3$: both). The proposed binary fingerprint uses the first-order normalized spectral subband moment as the subband feature ($\xi = \eta_1$).

Since the η_1s are obtained from 16 critical bands, both the bandwidth, M, and the start band, m_s, vary from 1 to 16. The time-width, N, is set to vary from 1 to 13 frames (2.4 s). By considering the three directions ($i = 1, 2,$ or 3) and different values taken by M, N, and m_s, there are about 4,264 possible difference formulas. All possible formulas and thresholds form a candidate set for binary conversions. As the start frame, n_s, is increased every 2048 samples, an 16-dimensional binary audio fingerprint is obtained using 16 conversion methods obtained from the candidate set.

The objective is to find the 16 most appropriate M, N, m_s, the direction of difference ($i = 1, 2,$ or 3), and T that give the most discriminative and robust 16-dimensional binary fingerprint, $F(n,m)$. A learning algorithm called Boosting is used to determine the parameters. The details of the learning algorithm are explained in the next subsection.

2.3. Pairwise Boosting

The AdaBoost [4] is a learning algorithm to produce a strong classifier from a number of weak classifiers. We modified the AdaBoost to select 16 binary conversion methods from the candidate set described above. A weak classifier is given by

$$h(x_1,x_2) = \text{sgn}\,[(d(x_1)-T)(d(x_2)-T)] \quad (9)$$

where $d(x)$ and $\{x_i\}_{i=1}^2$ denote respectively the output of the difference formula on x and base features extracted from an audio. Start frame of x is the same with n_s in the difference formula. When $d(x_1)$ and $d(x_2)$ are on the same side of threshold T, the weak classifier outputs '+1', predicting two audios as the same audios. When these are on the different side of threshold, the weak classifier outputs '-1', predicting two audios as different audios. This variant of the AdaBoost is called *Pairwise Boosting* in [3].

Selecting a weak classifier is equivalent to selecting a binary conversion method. In other words, selecting a weak classifier entails choosing one difference formula out of 4,264 possible candidates and selecting an appropriate threshold T. The *Pairwise Boosting* is used to select both robust and discriminative 16 difference formulas and those thresholds to generate an 16-dimensional binary audio fingerprint through 16 rounds. The *Pairwise Boosting* is described below:

Input

n training examples $(\mathbf{x}_{11},\mathbf{x}_{21}),...,(\mathbf{x}_{1n},\mathbf{x}_{2n})$ with $y_i = +1$ or -1

Initialize

weighted densities $\omega_{0,i} = \frac{1}{n}, i = 1...n$

Do for $t = 1,...,16$:

1. Choose M, N, m_s, the direction of difference ($i = 1, 2,$ or 3), and T that minimizes the weighted error :
   $\epsilon_t = \sum_{i=1}^{n} \omega_i \mathbf{I}[h_t(\mathbf{x}_{1i},\mathbf{x}_{2i}) \neq y_i]$
   where $\mathbf{I}[\varepsilon] = 1$, if the event ε occurs; $\mathbf{I}[\varepsilon] = 0$, otherwise.
2. Calculate weak hypothesis weight : c_t = log $(\frac{1-\epsilon_t}{\epsilon_t})$
3. Update weighted densities for matching pairs: if $y_i = 1$
   $\omega_{t+1,i} = \omega_{t,i} \cdot \exp[-c_t h_t(\mathbf{x}_{1i},\mathbf{x}_{2i}) y_i/2]$
4. Normalize weighted densities
   $\sum_{i:y_i=-1}^{n} \omega_{t+1,i} = \sum_{i:y_i=1}^{n} \omega_{t+1,i} = \frac{1}{2}$

All 16 selected weak classifiers are treated with same weights as each weak classifier determines each binary component of an 16-dimensional binary fingerprint. An experimental result shows that

the weights of all 16 weak classifiers selected by the *Pairwise Boosting* is almost same as shown in Fig. 2(a). The *Pairwise Boosting* is asymmetric in that only the matching audio pairs are boosted [3]. So, only the weighted densities of matching pairs are updated, in which the weighted density is increased if the prediction by a selected weak classifier is incorrect and the weighted density is decreased otherwise.

3. COMPARATIVE TEST

The performance of the proposed binary audio fingerprint are evaluated by comparative test. The following five binary audio fingerprints are used in the comparative test.

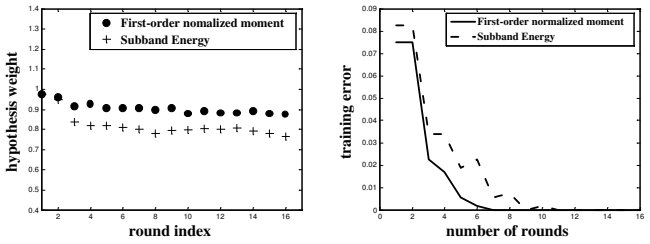
1. The first-order normalized moment with Boosting (η_1+Boosting) : A base audio feature is the first-order normalized spectral subband moment and the binary conversion is learned by the *Pairwise boosting*.
2. The subband energy with Boosting (E+Boosting) : A base audio feature is the subband energy and the binary conversion is learned by the *Pairwise boosting*.
3. The subband energy with (3) (E+3) : A base audio feature is the subband energy and the binary conversion is defined by (3).
4. The first-order normalized moment with (3) (η_1+(3)) : A base audio feature is the first-order normalized spectral subband moment and the binary conversion is defined by (3).
5. The first-order normalized moment with the sign (η_1+sign) : A base audio feature is the first-order normalized spectral subband moment and the binary value is determined by the sign of each element of the base feature vector.

The first fingerprint is the proposed fingerprint. The second fingerprint is similar to the fingerprint proposed in [3] in which the subband energy is used as the base feature and the spatial domain filters selected by the Boosting is applied to the base feature to generate the binary fingerprint. The third fingerprint is same with that used in [2] except the extraction parameters of base features: In [2], the subband energies are extracted according to the following method: Down-sampling to 5512.5 Hz, 2048-sample window with succesive offset by 64 samples, and 33 critical bands from 300 to 2000 Hz. However, both the first-order normalized moments and the subband energies used in all above binary audio fingerprints are extracted from audio with down-sampling to 11025 Hz, 4096-sample window, 50% overlap, and 16 critical bands from 300 to 5300 Hz. Accordingly, for an audio snippet of 5 seconds, the dimension of base feature vector with the former extraction setting is 14,190 (430×33), whereas that with the latter is 416 (26×16). The fourth and fifth fingerprints are extracted from the same base feature as the proposed fingerprint uses but converted to the binary value by different methods.

3.1. Training with Boosting

To make the binary audio fingerprint robust against the audio processing steps common in the practical applications, the audio clips subjected to the following 'real world' distortions, referring to [8], are used in the training with boosting:

- Time shift : 92.9 ms shift.
- Volume change : Envelop tremors.



(a) (b)

Fig. 2. (a) Hypothesis weight determined in each round (b) Training error of positive pairs

- Octave band equalization : Adjacent band attenuations set to -6dB and +6dB in an alternating fashion.
- Noise addition : White noise (SNR : 25 dB).
- Echo : Filter-emulating old time radio.
- Perceptual audio coding : 96 kbps MP3 compression.

About 600 matching pairs and 600 non-matching pairs are used for the training. The ratios of respective distortions in total train pairs are even. Each pair consists of two music clips of 5 seconds (26 frames) considering a maximum N of 13.

Fig. 2(a) shows that all confidences of selected weak classifiers applied on first-order normalized spectral subband moments are higher than that of selected weak classifiers applied on subband energies. Fig. 2(b) shows that as the number of rounds increased, the training error of positive pairs with the first-order normalized spectral subband moment are reduced more rapidly compared with the subband energy. These two results indicate that the conversion from first-order normalized spectral subband moments to binary audio fingerprints is less degraded than the conversion from subband energies to binary audio fingerprints.

3.2. Performance Evaluation

Approximately 2,000 matching pairs and 7,000,000 non-matching pairs are tested in the performance evaluation. The test data is completely exclusive from the training data. Each pair consists of two music clips of 9.85 seconds (52 frames). The size of the proposed binary audio fingerprints from each clips of 9.85 seconds is just 54-bytes (27×16 bits). The audio fingerprint should satisfy both the robustness and the pair-wise independence. Therefore, for a impartial comparison among the five binary audio fingerprints, the receiver operating characteristic (ROC) curves are obtained for each binary audio fingerprint. The ROC curve plots the true positive (TP) rate versus the false positive (FP) rate according to the bit error rate. The TP rate is the rate at which matching pairs are correctly predicted as matching pairs and is related to the robustness. The FP rate is the rate at which non-matching pairs are incorrectly predicted as matching pairs and is related to the pair-wise independence.

Fig. 3 shows the ROC curves for the distortion set. The distortion set consists of all distortions that were used in the training. From these ROC curves, the proposed algorithm (η_1+Boosting) performed best : the TP rate is the highest for a given FP rate. Fig. 4 shows ROC curves for the respective distortions. Test data is commonly subjected to a white background noise and a 96 kbps MP3 with the respective distortions. All ROC curves for the proposed binary audio fingerprint (η_1+Boosting) and the comparable binary audio fin-

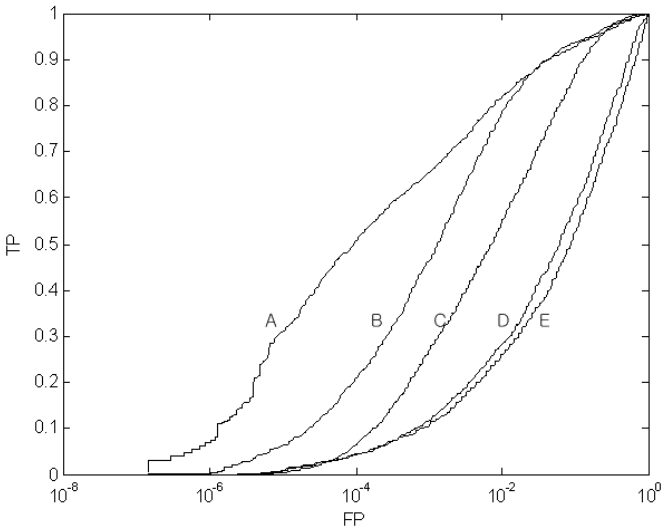


Fig. 3. ROC curves for the distortion set; A : η_1+Boosting, B : E+Boosting, C : η_1+sign, D : η_1+(3), E : E+(3)

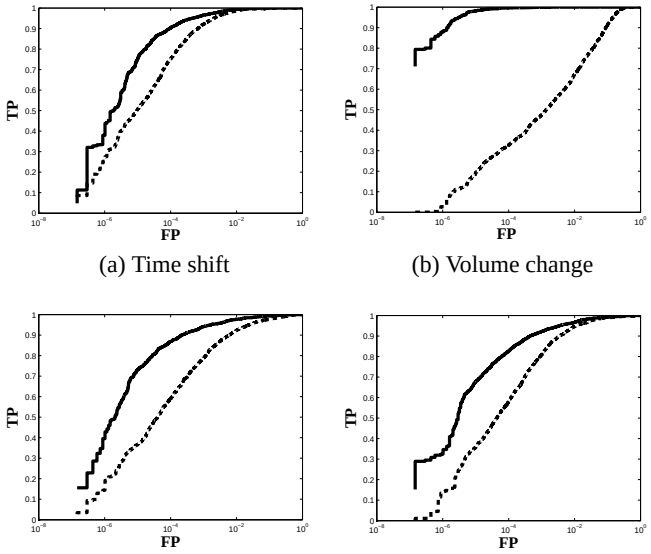
gerprint (E+Boosting) in Fig. 4 show that the proposed binary audio fingerprint (η_1+Boosting) outperforms the other.

4. CONCLUSION

Having fast search capability in a large-scale database is essential in most audio fingerprinting systems. A binary audio fingerprint is suitable for fast search owing to the compactness as well as an appropriateness for the direct indexing. In this paper, we presented a new binary audio fingerprint based on the spectral subband moment with boosted binary conversion. For minimizing the performance degradation during binary conversion, boosting algorithm learned the binary conversion from the candidate set of difference formulas. In the train procedure, typical 'real world' distortions were considered, and in the comparative test, the boosted binary audio fingerprint based on the first-order normalized spectral subband moment outperformed the other state-of-the-art binary audio fingerprints in terms of both robustness and pair-wise independence. Further works will focus on the fast search algorithm including a range search and the binary video fingerprint.

5. ACKNOWLEDGMENTS

This work was supported by grant No. R01-2003-000-10829-0 from the Basic Research Program of the Korea Science and Engineering Foundation, by University IT Research Center Project, and by grant IT-839 from the Institute for Information Technology Advancement.



(a) Time shift (b) Volume change

(c) Equalization (d) Old time radio

Fig. 4. ROC curves for the respective distortions; solid line : η_1+Boosting, dot line : E+Boosting

6. REFERENCES

[1] Jin S. Seo, Minho Jin, Sunil Lee, Dalwon Jang, Seungjae Lee, and Chang D. Yoo, "Audio fingerprinting based on normalized spectral subband moments," *IEEE Signal Processing Letters*, vol. 13, pp. 209–212, April 2006.

[2] J.A. Haitsma and T.Kalker, "A highly robust audio fingerprinting system," in *International Conf. on Music Information Retrieval*, 2002.

[3] Yan Ke, Derek Hoiem, and Rahul Sukthankar, "Computer vision for music identification," in *CVPR 2005. IEEE Computer Society Conference on*, 2005, vol. 1, pp. 597–604.

[4] Yoav Freund and Robert E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *European Conference on Computational Learning Theory*, 1995, pp. 23–37.

[5] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.

[6] G. Guo, H. Zhang, and S. Li, "Boosting for content-based audio classification and retrieval: an evaluation," in *G. Guo, H. Zhang, and S. Z. Li, Boosting for content-based audio classification and retrieval: an evaluation, Microsoft Research Tech. Rep. MSR-TR-2001-15.*, 2001.

[7] S. Ravindran and David V. Anderson, "Boosting as a dimensionality reduction tool for audio classification," in *ISCAS 2004, IEEE*, 2004, vol. 3, pp. 465–8.

[8] E. Allamanche, J. Herre, O. Helmuth, B. Frba, T. Kasten, and M. Cremer, "Content-based identification of audio material using mpeg-7 low level description," in *Int. Symposium on Music Information Retrieval*, 2005.