TEMPORAL DYNAMICS FOR SPECTRAL SUB-BAND CENTROID AUDIO FINGERPRINTS

Minho Jin and Chang D. Yoo

Div. of EE, Dept. of EECS, Korea Advanced Institute of Science and Technology,
373-1 Guseong Dong, Yuseong Gu, Daejeon 305-701, Korea

ABSTRACT

Motivated by the effectual use of temporal information in speech recognition, we investigate the effectiveness of the temporal dynamics of the spectral sub-band centroid (SSC) fingerprints for audio fingerprinting. The SSC, which is known to be a robust audio fingerprint against various distortions, does not involve any temporal dynamics. Here, the temporal dynamics are defined as the difference between two neighboring SSCs. The robustness of the temporal dynamics against various distortions were compared to that of SSCs. The system using temporal dynamics showed similar performance to that using SSCs in various distortions except for time-scale modification and linear-speed change. This is to be expected since these distortions change the time correlation of an audio. In our experiment, the concatenation of SSCs and the temporal dynamics outperformed each of the individual fingerprints. This suggests that the SSCs and the temporal dynamics provide information which is mutually supplementary.

1. INTRODUCTION

The advancement of information technology has facilitated the development of software for copy and distribution of multimedia data, and for this reason the management and copyright protection of multimedia data are currently important issues. A fingerprinting system aims to identify the multimedia content using the fingerprint which is a summary of a multimedia signal. Some examples of fingerprinting systems that deal with aforementioned issue are automated indexing of multimedia database (DB), filtering for file-sharing service, broadcasting monitoring [1]. In this paper, we focus only on the audio data.

The performance of a fingerprinting system can be measured in terms of pairwise independence, robustness and DB search efficiency [1–3]: Perceptually different audio signals are expected to have different fingerprints while perceptually similar audio signals are expected to have the same fingerprints; The fingerprints must be distributed in such a way that is conducive for fast DB search.

This paper investigates the effectiveness of the temporal dynamics of spectral sub-band centroids (SSC) fingerprints for audio fingerprinting. The SSC is widely used in both the speech and the speaker recognition [4], and our previous work demonstrated that the system using SSC outperforms those systems using mel-frequency cepstral coefficients (MFCCs) and spectral flatness as reported in [2, 5]. The SSC of a frame is determined uniquely by the pertaining frame and not by the neighboring frames. Thus, it does not explicitly carry time correlation information. Motivated by the effective use of temporal information in speech recognition, we investigate the effectiveness of temporal dynamics. In this paper, the temporal dynamics are defined by the difference between two neighboring SSCs.

This paper is not the first work to investigate the effectiveness of the temporal dynamics for audio fingerprinting. Cano et al. [6] used a hidden Markov model to model the MFCCs, their delta and accelerated delta of the audio segments. Haitsma and Kalker [1] used an 1-bit quantizer to quantize the temporal difference of band-energy differences. We find that the temporal dynamics of SSCs are effective for audio fingerprinting and that using both the static audio fingerprints and the temporal dynamics can further improve the fingerprint-matching performance. The temporal dynamics can be sensitive to time-scale modification and linear-speed change since these distortions change the time correlation of an audio. However, it is shown experimentally that the concatenation of SSCs and temporal dynamics can improve the robustness against these distortions.

This paper is organized as follows: Section 2 describes how to incorporate temporal dynamics into audio fingerprinting system. Section 3 evaluates the proposed system. Finally, Section 4 concludes the paper.

This work was supported in part by MIC & IITA through IT Leading R&D Support Project in Korea.

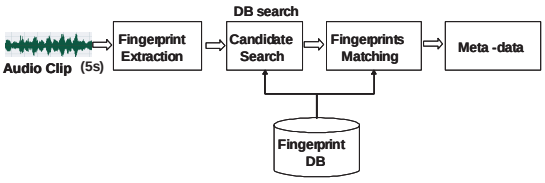
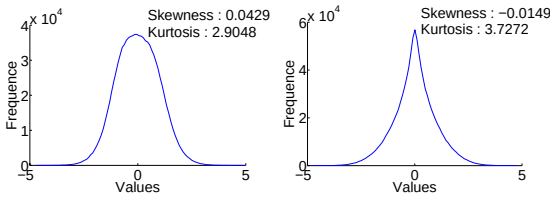


Fig. 1. Audio Fingerprinting System



(a) Histogram of the SSC (1^{st} dim.) (b) Histogram of the dynamic SSC (1^{st} dim.)

Fig. 2. Histograms of static and dynamic SSC fingerprints

2. TEMPORAL DYNAMICS FOR SPECTRAL SUB-BAND CENTROID AUDIO FINGERPRINTS

2.1. Audio Fingerprinting System

Fig. 1 illustrates our audio fingerprinting system. An audio fingerprint is first extracted from a query of a short audio clip. In our system, a query must be at least 5-seconds long, and every 185.8ms a 16-dimensional fingerprint is extracted. Thus, from a query, a 432-dimensional feature, which consists of 27 fingerprints, is obtained. To retrieve the meta-data associated with this feature, each fingerprint is searched in the DB to find the candidates within a certain Euclidean radius. In the verification, the candidate with the minimum distance is selected: The distance is compared to a threshold, and if the distance is below the threshold, the meta-data of the minimum-distance candidate are outputted; otherwise, the query is declared not to exist in the DB. In this paper, we will focus on the fingerprint extraction and matching.

2.2. Temporal Dynamics for Normalized Spectral Sub-band Centroids

Let $P_q[n,k]$ be the short-time power spectrum of an audio signal q at frequency bin k of the n^{th} frame, and let $C[m](1 \le m \le M)$ be the m^{th} critical band of Bark scale which is relevant to human perception [7]. Then, the first-order normalized moment $\eta_q[n,m](1 \le m \le M)$ of n^{th} frame of query q is given by

$$\eta_q[n,m] = \frac{\sum_{k=C[m]+1}^{C[m+1]} kP_q[n,k]}{\sum_{k=C[m]+1}^{C[m+1]} P_q[n,k]}. \quad (1)$$

Let the number of sub-frames in a query be N. We normalized (1) using empirical mean and standard deviation as follows:

$$\hat{\mathbf{q}} = [\hat{\eta}_q[1],\ \hat{\eta}_q[2],\ \cdots,\ \hat{\eta}_q[N]], \quad (2)$$

where $\hat{\eta}_q[i]$ is given by

$$\hat{\eta}_q[i] = [\frac{\eta_q[i,1]-\hat{\mu}_q[1]}{\hat{\sigma}_q[1]},\ \frac{\eta_q[i,2]-\hat{\mu}_q[2]}{\hat{\sigma}_q[2]},\ \cdots,\ \frac{\eta_q[i,M]-\hat{\mu}_q[M]}{\hat{\sigma}_q[M]}]^T, \quad (3)$$

and where

$$\hat{\mu}_q[m] = \frac{1}{N}\sum_{i=1}^{N}\eta_q[i,m], \quad (4)$$

$$\hat{\sigma}_q[m] = \sqrt{\Big(\frac{1}{N}\sum_{i=1}^{N}\eta_q^2[i,m]\Big) - \hat{\mu}_q^2[m]}. \quad (5)$$

We call (2) as the static audio fingerprint or SSC.

The temporal dynamics can be captured using the difference between two neighboring SSCs as follows:

$$\mathbf{\Delta}\hat{\mathbf{q}} = [\mathbf{\Delta}\hat{\eta}_q[1],\ \mathbf{\Delta}\hat{\eta}_q[2],\ \cdots,\ \mathbf{\Delta}\hat{\eta}_q[N-W]], \quad (6)$$

where

$$\mathbf{\Delta}\hat{\eta}_q[i] = [\frac{\Delta\eta_q[i,1]-\Delta\hat{\mu}_q[1]}{\Delta\hat{\sigma}_q[1]}\ \frac{\Delta\eta_q[i,2]-\hat{\mu}_q[2]}{\Delta\hat{\sigma}_q[2]},\ \cdots,\ \frac{\Delta\eta_q[i,M]-\Delta\hat{\mu}_q[M]}{\Delta\hat{\sigma}_q[M]}]^T \quad (7)$$

and where

$$\Delta\eta_q[i,m] = \eta_q[i+W,m]-\eta_q[i,m], \quad (8)$$

$$\Delta\hat{\mu}_q[m] = \frac{1}{N}\sum_{i=1}^{N}\Delta\eta_q[i,m], \quad (9)$$

$$\Delta\hat{\sigma}_q[m] = \sqrt{\Big(\frac{1}{N}\sum_{i=1}^{N}\Delta\eta_q^2[i,m]\Big) - \Delta\hat{\mu}_q^2[m]}. \quad (10)$$

Here, W is the step size chosen beforehand.

Fig. 2(a) and Fig. 2(b) illustrate the histogram of the first dimension of the SSC and the dynamic SSC, respectively. The histograms of other dimensions of both the SSC and the dynamic SSC are omitted since these also have the same tendency as those of the first dimension. From the skewness and the kurtosis of Fig. 2(a), the SSC seems to follow a Gaussian distribution. If SSCs are independently identical distributed (i.i.d.), the dynamic SSC should follow a Gaussian distribution [8]. However, from Fig. 2(b), it is clearly seen that the dynamic SSC does

not follow a Gaussian distribution but a super-Gaussian distribution. This suggests that the neighboring SSCs are not independent but correlated. The purpose of this paper is to examine whether explicit use of the temporal information is useful for the audio fingerprint. In addition, this paper alos examines the acceleration, which was obtained by taking the normalized temporal difference of (8).

2.3. Distance Metric for Both SSC and Temporal Dynamics

Assume we have two fingerprints of length L, $\hat{\mathbf{x}} = [\mathbf{x}[1], \mathbf{x}[2], \cdots, \mathbf{x}[L]]$ and $\hat{\mathbf{y}} = [\mathbf{y}[1], \mathbf{y}[2], \cdots, \mathbf{y}[L]]$, where $\mathbf{x}[i]$ and $\mathbf{y}[i]$ are M by 1 vectors. Then, the fingerprint matching is a hypothesis testing below:

H0 : $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are from perceptually equivalent.

H1 : $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are from perceptually inequivalent.

In this paper, the hypothesis testing is performed using the sum of squares of Euclidean distances as follows:

$$D(\mathbf{x}, \mathbf{y}) = \frac{1}{LM} \sum_{i=1}^{L} (||\mathbf{x}[i] - \mathbf{y}[i]||_2^2) \underset{H_1}{\overset{H_0}{\lessgtr}} \tau \qquad (11)$$

where τ is a threshold chosen beforehand and $||\mathbf{x}[i] - \mathbf{y}[i]||_2$ denote the 2-norm of vector, $\mathbf{x}[i] - \mathbf{y}[i]$.

Let $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ be the static SSCs to be matched, and let $\mathbf{\Delta}\hat{\mathbf{p}}$ and $\mathbf{\Delta}\hat{\mathbf{q}}$ be corresponding dynamic SSCs. The performance of the SSCs is measured using $D(\hat{\mathbf{q}},\hat{\mathbf{p}})$ while that of the dynamic SSCs is measured using $D(\mathbf{\Delta}\hat{\mathbf{q}}, \mathbf{\Delta}\hat{\mathbf{p}})$. Their hybrid is obtained as the sum of two, $D(\hat{\mathbf{q}}, \hat{\mathbf{p}}) + D(\mathbf{\Delta}\hat{\mathbf{q}}, \mathbf{\Delta}\hat{\mathbf{p}})$.

3. EXPERIMENT

3.1. Experimental Setup

The evaluation of the fingerprint was performed using an audio fingerprint DB of 1,000 songs of various genres which amounts to 62-hours playing time. The audio data were extracted from compact discs and compressed into 128kpbs or 196kbms MP3 files. The audio signals were first re-sampled at 11.025kHz, and each SSC was extracted every 185.8 ms using a window length of 371.5ms, i.e., 50% overlap. And we used critical bands between 300Hz and 5300Hz, which gave SSCs of $M = 16$ dimension. Thus, each query of 5s-long audio signals amounts to $N = 27$ SSCs. Four sets of audio query were made by distorting the original query by four kinds of distortions (using Cool Edit Pro 2.1 software). Table. 3.2 describes these four kinds of distortions.

Set	Distortions
Test set 1	Filter-emulating old-time radio + pitch increment by 1% + 92.9ms delay.
Test set 2	Filter-emulating ambient metal room + pitch decrement by 1% + 92.9ms delay.
Test set 3	Filter-emulating super loud + linear speed change (LSC) by 1%.
Test set 4	Filter-emulating rich chamber + time-scale modification (TSM) by 4%.

Table 1. Test set

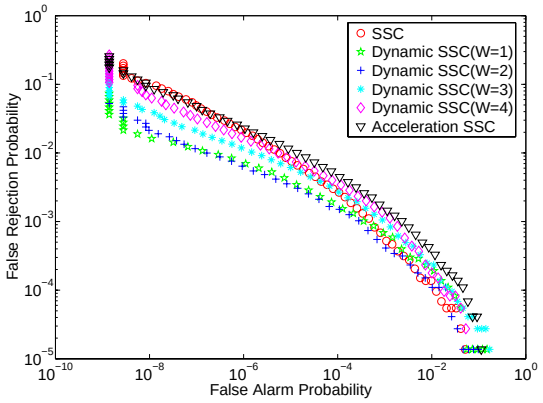
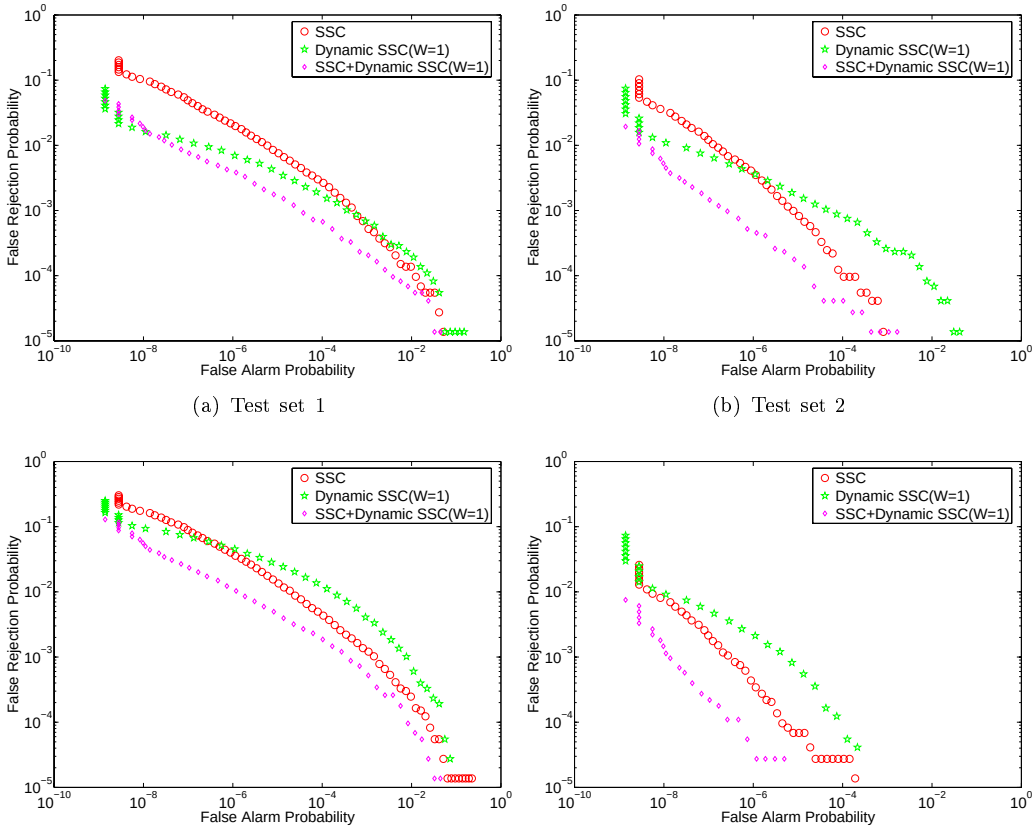


Fig. 3. Dynamics of various W and acceleration in test set 1

3.2. Detection Error Trade-off Curves

Fig. 3 illustrates the detection error trade-off (DET) curves various W values and acceleration using test set 1. The horizontal axis and the vertical axis denote the false alarm probability and the false rejection probability, respectively. When $W = 1$ and $W = 2$, the best performance was obtained. The performance degraded as W increases. It supports the result in [9] that the audio signal is strongly correlated only within a short interval. The acceleration performed the worst. It is natural since the one acceleration requires at least three concatenated SSCs.

Fig. 4 illustrates the DET curves for four kinds of test sets. In test set 1, the system using dynamic SSC outperformed that using SSC. However, the system using dynamic SSC performed worse than that using SSC in other sets, especially in test set 3 and 4. It was expected since TSM and LSC change the time correlation of SSCs. The hybrid of the SSCs and the dynamic SSCs consistently outperformed each of the individual in all test sets. This implies that the SSC and the dynamics SSC provide mutually supplementary information.



(a) Test set 1 (b) Test set 2

(c) Test set 3 (d) Test set 4

Fig. 4. DET curves drawn for four test sets

4. CONCLUSION

This paper investigates the effectiveness of the dynamic SSCs in the audio fingerprinting. The use of the dynamic SSC in addition to the SSCs improves the robustness of fingerprinting matching system against various distortions. Our future work will be focused on estimating robust temporal dynamics against the TSM and LSC distortions which degrade the performance of the system using the dynamic SSCs: the weighted sum of the scores from the SSC and the dynamic SSC can be a possible solution, where the weights are determined by the estimated time scale modification.

5. REFERENCES

[1] J.A. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *Proc. ICMIR*, 2002.

[2] 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, no. 4, 2006.

[3] J.S. Seo, J. Haitsma, T. Kalker, and C.D. Yoo, "A robust image fingerprinting system using the Radon transform," *Signal Processing: Image Communication*, vol. 19, pp. 325–339, 2004.

[4] J. Chen, Y. Huang, Q. Li, and K.K. Paliwal, "Recognition of noisy speech using dynamic spectral subband centroids," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 258–261, 2004.

[5] Jin S. Seo, Minho Jin, Sunil Lee, Dalwon Jang, Seungjae Lee, and Chang D. Yoo, "Audio fingerprinting based on normalized spectral subband centroids," in *Proc. ICASSP*, 2005.

[6] Pedro Cano, Eloi Batlle, Harald Mayer, and Helmut Neushmied, "Robust sound modeling for song detection in broadcast audio," in *Proc. the 112th AES Int. Conv.*, 2002.

[7] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer-Verlag, 1999.

[8] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, New York:McGrow-Hill, 2002.

[9] Matthew L. Miller, Manuel Acevedo Rodriguez, and Ingemar J. Cox, "Audio fingerprinting: Nearest neighbor search in high dimensional binary spaces," *Journal of VLSI Signal Processing*, vol. 41, no. 3, pp. 285–291, 2005.