GAUSSIAN MIXTURE MODELING USING SHORT TIME FOURIER TRANSFORM FEATURES FOR AUDIO FINGERPRINTING

Arunan Ramalingam and Sridhar Krishnan

Department of Electrical and Computer Engineering
Ryerson University, Toronto, ON, Canada, M5B 2K3
E-mail: (aramalin)(krishnan)@ee.ryerson.ca

ABSTRACT

In audio fingerprinting, an audio clip must be recognized by matching an extracted fingerprint to a database of previously computed fingerprints. The fingerprints should reduce the dimensionality of the input significantly, provide discrimination among different audio clips, and at the same time, invariant to the distorted versions of the same audio clip. In this paper, we design fingerprints addressing the above issues by modeling an audio clip by Gaussian mixture models (GMM) using a wide range of easy-to-compute short time Fourier transform features such as Shannon entropy, Renyi entropy, spectral centroid, spectral bandwidth, spectral flatness measure, spectral crest factor, and Mel-frequency cepstral coefficients. We test the robustness of the fingerprints under a large number of distortions. To make the system robust, we use some of the distorted versions of the audio for training. However, we show that the audio fingerprints modeled using GMM are not only robust to the distortions used in training but also to distortions not used in training. Using spectral centroid as feature, we obtain the highest identification rate of 99.1 % with a false positive rate of 10^{-4}.

1. INTRODUCTION

An audio fingerprint is a compact representation of perceptually relevant portion of the audio content. An audio fingerprint should be able to identify audio files even if they are severely distorted by perceptual coding or common signal processing operations. The type of distortions a fingerprint should withstand depends on the application. For example, audio fingerprints designed for broadcast monitoring should withstand distortions such as time compression, dynamic range compression, and equalization. An Audio fingerprinting system has two principle components: fingerprint extraction and matching algorithm. The fingerprint requirements include computational simplicity, robustness to distortions, smaller size, and discrimination power over a large number of other fingerprints [1]. The matching algorithms should be efficient to able to identify an audio item from a database of hundreds of thousands of audio songs in a few seconds. A large number of fingerprinting schemes have been proposed. For some recent work, please see [2] – [5].

The overview of the proposed fingerprinting scheme is shown in Fig. 1. First the incoming audio clip is preprocessed and features are extracted from them. Then using these features, the audio clip is modeled using Gaussian mixtures. In the training phase, the mixture models of all the audio clips are stored in the database along with the metadata information. In the identification phase, the features from an unknown audio clip are used to evaluate the likelihood of all the models in the database. Then the model that is most likely to generate the features is identified as the correct audio clip.

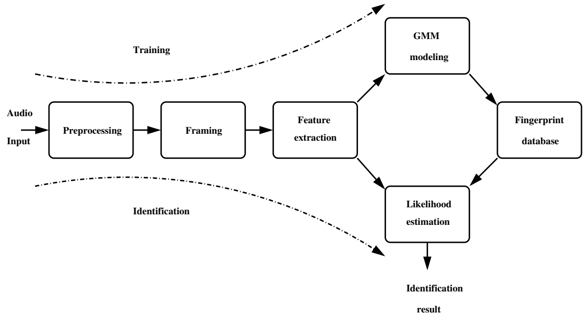

Fig. 1. Proposed Fingerprinting System

2. FEATURE EXTRACTION

In this work, we use the following features extracted from the short time Fourier transform (STFT) of the signal for fingerprint extraction. Let $F_i = f_i(u), u \in (0, M)$ be the Fourier transform of the i^{th} frame, where M is the index of the highest frequency band. To increase the robustness of the fingerprint, the features are not extracted on

We would like to acknowledge Micronet for their financial support.

0-7803-9332-5/05/$20.00 ©2005 IEEE

the whole spectrum but on non-overlapping logarithmically spaced bands. Let $F_{i,b} = f_i(u_b), u_b \in (l_b, u_b)$ where l_b and u_b are the lower and upper edges of the band b. In each of the frame, the following features are extracted. These features have been used successfully in audio fingerprinting [6] and music classification [7].

1. Spectral Centroid (SC): The spectral centroid is the center of gravity of the magnitude spectrum of the STFT and is a measure of spectral shape and "brightness" of the spectrum. Spectral centroid is defined as

$$SC_{i,b} = \frac{\sum_{u=l_b}^{u_b} u.\left|f_i(u)\right|^2}{\sum_{u=l_b}^{u_b} \left|f_i(u)\right|^2}. \tag{1}$$

2. Spectral Bandwidth (SB): The spectral bandwidth is measured as the weighted average of the distances between the spectral components and the spectral centroid. Spectral bandwidth is defined as

$$SB_{i,b} = \frac{\sum_{u=l_b}^{u_b} \left(u - SC_{i,b}\right)^2.\left|f_i(u)\right|^2}{\sum_{u=l_b}^{u_b} \left|f_i(u)\right|^2}. \tag{2}$$

3. Spectral Band Energy (SBE): The spectral band energy is the energy in the frequency bands normalized by the energy in the whole spectrum. Spectral band energy is calculated as

$$SBE_{i,b} = \frac{\sum_{u=l_b}^{u_b} \left|f_i(u)\right|^2}{\sum_{u=0}^{M} \left|f_i(u)\right|^2}. \tag{3}$$

4. Spectral Flatness Measure (SFM): The spectral flatness measure quantifies the flatness of the spectrum and distinguishes between noise-like and tone-like signal. Spectral flatness measure is defined as

$$SFM_{i,b} = \frac{\left[\prod_{u=l_b}^{u_b} \left|f_i(u)\right|^2\right]^{\frac{1}{u_b - l_b + 1}}}{\frac{1}{u_b - l_b + 1}\sum_{u=l_b}^{u_b} \left|f_i(u)\right|^2}. \tag{4}$$

5. Spectral Crest Factor (SCF): The spectral crest factor is also a measure of the tonality of the signal. Spectral crest factor is defined as

$$SCF_{i,b} = \frac{\max\left(\left|f_i(u)\right|^2\right)}{\frac{1}{u_b - l_b + 1}\sum_{u=l_b}^{u_b} \left|f_i(u)\right|^2}. \tag{5}$$

6. Shannon Entropy (SE): The Shannon entropy of a signal is a measure of its spectral distribution of the signal. Shannon entropy is defined as

$$SE_{i,b} = \sum_{u=l_b}^{u_b} \left|f_i(u)\right| \log_2 \left|f_i(u)\right|. \tag{6}$$

7. Renyi Entropy (RE): The Renyi entropy of a signal is also a measure of its spectral distribution. Renyi entropy is defined as

$$RE_{i,b} = \frac{1}{1-r} \log\left(\sum_{u=l_b}^{u_b} \left|f_i(u)\right|^r\right). \tag{7}$$

   We used Renyi Entropy of order $r = 2$.

8. Mel-frequency Cepstral Coefficients (MFCC): MFCC are perceptually motivated features based on the STFT. After taking the log-amplitude of the magnitude spectrum, the FFT bins are grouped and smoothed according to the perceptually motivated Mel-frequency scaling. Finally, in order to decorrelate the resulting feature vectors a discrete cosine transform is performed. In this work, we used 13 coefficients since this parameterization has been shown to be quite effective for speech recognition and speaker identification [8].

Let $\mathbf{X_i}$ be the set of features extracted for the frame i. $\mathbf{X_i}$ can be any one of the features described above. In order to better characterize the temporal variations of the signal, the first derivatives of the above features

$$\delta_i = \delta_i - \delta_{i-1} \tag{8}$$

are also used included in the feature matrix. In an audio clip, successive frames are related in time. To include this time dependency, a time vector is added to the feature matrix. This time vector is taken as an incremental counter from 0 to 1. Thus the feature matrix of the entire audio clip can be described as

$$\mathcal{F}'_{\mathcal{M}} = \left[\begin{array}{l} \mathbf{X_1}, \delta_1, t_1 \\ \mathbf{X_2}, \delta_2, t_2 \\ \vdots \\ \mathbf{X_N}, \delta_N, t_N \end{array}\right] \tag{9}$$

where N is the number of frames in the audio clip. Finally the feature matrix is mean subtracted and component wise variance normalized to get a normalized feature matrix $\mathcal{F}_{\mathcal{M}}$.

3. GAUSSIAN MIXTURE MODELS

Gaussian Mixture Models (GMM) have been successfully used in audio classification [7] and content based retrieval [9]. In this work, the technique is used to model an audio fingerprint as a probability density function (PDF), using a weighted combination of Gaussian component PDFs (mixtures). During the training phase, the GMM parameters of an audio fingerprint are estimated to maximize the probability of the audio frames present in the audio fingerprint. We use the Baum-Welch (Expectation-Maximization) algorithm to estimate the GMM parameters with initialization by $k - means$ clustering. As the feature vectors in this work have reasonably uncorrelated components, computationally convenient diagonal covariance matrices can be used. We used GMM with 16 mixtures. Thus in the fingerprint extraction phase, each audio clip is modeled by GMM. During the matching phase the fingerprint from an unknown recording is compared with the database of pre-computed GMM and the GMM that gives the highest likelihood for the fingerprint is identified as correct match.

4. RESULTS

We used a database containing 250 five-second audio clips chosen from the categories of rock, pop, country, classical, and jazz. The audio clips are chosen from random portions of songs from Compact Discs.

4.1. Robustness to Distortions

We used several distorted versions of the audio clips to test the robustness of the proposed scheme. We used the following distorted versions in our tests.

I. Compression – 1) MP3 at 32 kbps, 2) AAC at 32 kbps, 3) WMA at 32 kbps, 4) Real encoding at 32 kbps.

II. Amplitude distortion – 1) 3 : 1 Compression above 30 dB, 2) 3 : 1 Expander below 10 dB, 3) 3 : 1 compression below 10 dB, 4) Limiter at 9 dB, 5) 'Superloud' amplitude distortion, 6) Noise gate at 20 dB, 7) De-esser, 8) Nonlinear amplitude distortion.

III. Frequency distortion – 1) Nonlinear bass distortion, 2) Midrange frequency boost, 3) Notch Filter, 750 - 1800 Hz, 4) Notch Filter 430 - 3400 Hz, 5) Telephone bandpass, 135 - 3700 Hz, 6) Bass cut, 7) Bass boost.

IV. Change in pitch – 1) Lower pitch 2 - 6 %, 2) Raise pitch 2 - 6 %.

V. Change in speed – 1) Linear speed increase 2 - 6%, 2) Linear speed decrease 2 - 6%.

VI. Resampling at 8 kHz

VII. Echo addition

To increase the robustness of the fingerprints, in addition to the original audio, some distorted versions of the audio are also used in training. We used the following distorted versions in our training: 1) Undistorted audio, 2) 3 : 1 Compression above 30 dB, 3) Nonlinear amplitude distortion, 4) Nonlinear bass distortion, 5) Midrange frequency boost, 6) Notch Filter, 750 - 1800 Hz, 7) Notch Filter 430 - 3400 Hz, 8) Raise Pitch 1%, 9) Lower Pitch 1%. The log-likelihood of the test clips are evaluated for all the models in the database. Then the model that gives the highest log-likelihood is taken as the correct match. Table 1 shows the percentage of clips that are correctly identified for different features for distortions used in training as well as for distortions not used in training. The results show that it is not necessary to train the model for all possible distortions. By training the model to some representative distortions, we can obtain robustness to a wide variety of distortions.

Table 1. Mean Recognition rate for distortions

	Train	**Test**	**Mean**
MFCC	99.0	98.5	98.7
Spectral centroid	99.4	99.1	99.2
Spectral bandwidth	99.4	98.9	99.1
Spectral band energy	98.8	98.8	98.8
Spectral flatness measure	99.4	98.6	98.9
Spectral crest factor	99.2	98.6	98.8
Shannon Entropy	99.4	98.8	99.0
Renyi Entropy	99.4	98.9	99.0

4.2. False Positive Analysis

In the previous section it was assumed that the test clip is present in the database. Hence the model that gives the highest log-likelihood value is identified as the correct match. However it is possible that the test clip may not be in the database. So there should be a criteria to reject the audio clips that are not in the database. A suitable threshold for log-likelihood can be used to vary the false positive and false negative rates. The false positive and the corresponding identification rate are shown in Figs. 2 and 3. The percentage of audio clips correctly identified at different false positive rates are shown in Table 2. Among the different features used, spectral centroid gives the highest identification rate of 99.1% with a false positive rate of 10^{-4}. MFCC performs poorly with an identification rate of 13 %. All the features except the spectral flatness measure give an identification rate of more than 90 % with a false positive rate of 10^{-3}.

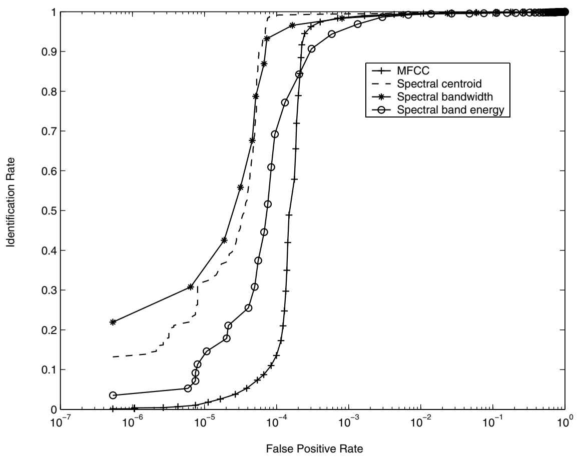

Fig. 2. Identification rates at different false positive rates for MFCC, Spectral centroid, Spectral bandwidth, and Spectral band energy

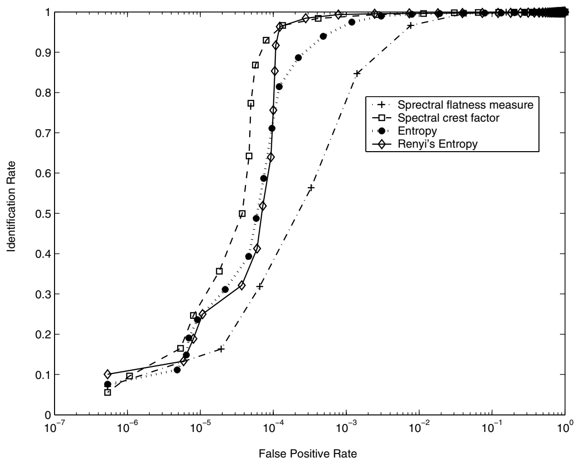

Fig. 3. Identification rates at different false positive rates for Spectral flatness measure, Spectral crest factor, Shannon Entropy and Renyi Entropy

Table 2. Identification Rate at different false positive rates

	10^{-4}	10^{-3}	10^{-2}
MFCC	13.5	98.4	99.3
Spectral centroid	99.1	99.5	99.8
Spectral bandwidth	93.2	98.4	99.3
Spectral band energy	69.2	94.3	99.2
Spectral flatness measure	31.8	56.4	96.6
Spectral crest factor	93.0	98.4	99.3
Shannon Entropy	71.1	93.9	99.4
Renyi Entropy	64.0	99.3	99.7

5. CONCLUSION

Gaussian Mixture Models have been successfully used in many classification and identification problems in audio. In this work, we modeled audio recordings for audio fingerprinting by Gaussian mixtures using features extracted from the STFT of the signal. Even though we use some distorted samples of the audio during training, the system is robust to distortions not used in training. Using spectral centroid as feature, we obtain the highest identification rate of 99.1 % with a false positive rate of 10^{-4}.

6. REFERENCES

[1] P. Cano, E. Batle, T. Kalker, and J. Haitsma, "A review of algorithms for audio fingerprinting," in *IEEE Workshop on Multimedia Signal Processing, 2002*, December 2002, pp. 169–173.

[2] J. Herre, O. Hellmuth, and M. Cremer, "Scalable robust audio fingerprinting using MPEG-7 content description," in *IEEE Workshop on Multimedia Signal Processing, 2002*, December 2002, pp. 165–168.

[3] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *Proc. of the 3rd Int. Symposium on Music Information Retrieval,*, October 2002, pp. 144–148.

[4] V. Venkatachalam, L. Cazzanti, N. Dhillon, and M. Wells, "Automatic identification of sound recordings," *IEEE Signal Processing Magazine*, vol. 21, no. 2, pp. 92 – 99, March 2004.

[5] C.J.C. Burges, J.C. Platt, and S. Jana, "Distortion discriminant analysis for audio fingerprinting," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 165–174, May 2003.

[6] E. Allamanche, B. Frba, J. Herre, T. Kastner, O.Hellmuth, and M. Cremer, "Cotent-based identification of audio material using MPEG-7 low level description," in *Proceeding of the International Symposium on Music Information Retrieval (ISMIR)*, Indiana, USA, October 2002.

[7] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Tran. on Speech and Audio Processing*, vol. 10, no. 5, pp. 293 – 302, July 2002.

[8] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ,, 1993.

[9] D. Pye, "Content-based methods for the management of digital music," in *Proceedings of ICASSP*, 2000, vol. 4, pp. 24–27.