World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:4, 2015

A Non-Parametric Based Mapping Algorithm for Use in Audio Fingerprinting

Analise Borg, Paul Micallef

***Abstract*—**Over the past few years, the online multimedia collection has grown at a fast pace. Several companies showed interest to study the different ways to organise the amount of audio information without the need of human intervention to generate metadata. In the past few years, many applications have emerged on the market which are capable of identifying a piece of music in a short time. Different audio effects and degradation make it much harder to identify the unknown piece. In this paper, an audio fingerprinting system which makes use of a non-parametric based algorithm is presented. Parametric analysis is also performed using Gaussian Mixture Models (GMMs). The feature extraction methods employed are the Mel Spectrum Coefficients and the MPEG-7 basic descriptors. Bin numbers replaced the extracted feature coefficients during the non-parametric modelling. The results show that non-parametric analysis offer potential results as the ones mentioned in the literature.

***Keywords*—**Audio fingerprinting, mapping algorithm, Gaussian Mixture Models, MFCC, MPEG-7.

I. Introduction

THE need to identify unknown songs emerged after people started to listen to music everywhere, i.e. on the radio, television, cinema, discotheques and even in the street. Moreover the amount of audio files available is increasing exponentially and therefore a solution to automatically identify audio is required. The current technologies help to identify the name of the artist, album, song and other facts by just recording a few seconds.

The idea behind searching of the unknown piece works similar to that of text-based search but a unique signature has to be extracted from the multimedia files. This signature is usually called fingerprint. A known set of audio pieces is processed and a fingerprint is extracted from each. Then, during classification, an unknown piece of audio, is first processed to extract its own fingerprint and then it is compared to the ones stored in a database.

During the fingerprint extraction process, a set of significant characteristics pertaining to a recording are extracted in an abridged and robust form. Each fingerprint extracted must be distinguishable from an amount of fingerprints, invariant to distortions, small in size and simple to compute. This means that there must be a balance between the dimensionality reduction and the amount of information lost. Some well-known algorithms used are MFCC, Linear Predictive Coding (LPC), spectral flatness, spectral sharpness, spectral crest factor, zero crossing, band energy, Karhunen-Loeve transform and Oriented Principal Component Analysis (OPCA) [1]. Some current systems make use of Chroma features and Pitch Class Profiles (PCP).

The set of features calculated during the feature extraction stage are passed to the fingerprint modelling block. The features are calculated on a frame-by-frame basis. The fingerprint modelling stage observes the near frames, the whole audio file and considers also the entire database to decrease the fingerprint size. Some of the popular fingerprint models include Linear Vector Quantisation (LVQ), Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs).

Section II gives a general background of audio fingerprinting systems. The feature extraction algorithms and the feature modeling algorithms are then described in Section III and a detailed explanation of the results obtained is given in Section IV. Finally, Section V concludes the paper and mentions some potential improvements in future work.

II. Background

Baluja and Covell [2] present their novel system for audio identification called Waveprint. It makes use of computer-vision techniques combined with large-scale-data-stream processing algorithms to achieve near-duplicate images from a large corpus of image data to the task of audio retrieval. The system achieves excellent results for small snippets of audio that have been degraded by competing noise, poor recording quality or mobile phone playback. In the Waveprint system, the audio input is converted into a spectrogram. It is then divided into frames with an overlap to create spectral images. The top Haar-wavelets are extracted according to their magnitude. A wavelet signature is then computed for each spectrogram image. The wavelet-image on its own is not resistant to noise or audio degradations. The top wavelets were then chosen to reduce the effects of noise while maintaining the major characteristics of the image. The Min Hash algorithm is then used to obtain a 100 byte fingerprint which can be compared directly with byte-wise Hamming distances.

Haitsma et al. [3] proposed audio fingerprinting based on the Bark Frequency Cepstrum Coefficients (BFCC). In this implementation, the frames are highly overlapped in order to ensure that the query probe can be detected at arbitrary time-alignment. Each fingerprint is 32 bits and it can be compared efficiently using the Hamming distances. Ke et al. [3] used as a base for their work. They made use of the AdaBoost technique from computer vision to improve the performance

Analise Borg was with the Department of Communications and Computer Engineering, University of Malta (e-mail: borg.analise@gmail.com).

Paul Micallef is with the Department of Communications and Computer Engineering, University of Malta (e-mail: paul.micallef@um.edu.mt).

of the fingerprinting scheme [4].

Ramalingam and Krishnan [5] make use of GMMs in their system. The audio clips are first pre-processed and features are extracted from each one, in this case using short-time Fourier transform (STFT). During the training process, the mixture models for all the audio clips together with the corresponding metadata information are stored in the database.

Battle et al. [6] make use of the double stochastic property of HMMs. Each descriptor in the HMM characterises a part of the whole song and the global HMM will then represent the whole song.

III. Design And Implementation

Audio fingerprinting involves four main stages, the pre-processing of the audio signal, the feature extraction, the feature modeling and classification. Pre-processing and feature extraction form the front end processing of the audio fingerprinting model. This part of the model deals with signal processing. During pre-processing, raw audio is digitised, filtered and divided into frames. Afterwards during feature extraction, the frames are converted to features vectors. Mel Spectrum coefficients and MPEG-7 Basic Spectral Descriptors are the feature vectors computed. Then the feature coefficients were mapped using non-parametric modeling and parametric modeling. During non-parametric analysis, the data binning technique was used while during parametric analysis, GMMs were used. An overall representation can be seen in Fig. 1.

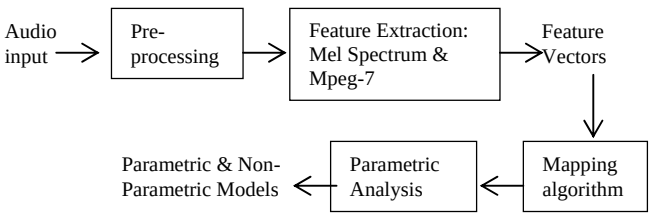

Fig. 1 Overall Process

A. Feature Extraction

The MFCC and the MPEG-7 basic descriptors were the chosen features to be extracted in the developed system. Mel-Frequency Cepstrum (MFC) is used to represent the short-term power spectrum of a sound. The MFCC are the coefficients that together set up an MFC. These are computed from a cepstral representation of the audio file which is a nonlinear mapping of the spectrum. The Mel scale is a perceptual scale based on human auditory perception. On the other hand, the MPEG-7 basic parameters contain elements that are based directly on the spectrum of the original audio signal.

For the MFC, each individual frame was windowed using a Hamming window [7]. The windowed frames were then converted into the power spectrum by employing a 256-point Fast Fourier Transform (FFT). The powers of the spectrum are mapped to the mel scale using triangular overlapping windows. The log of the powers at each of the mel frequencies was obtained.

The basis of the calculation of the MPEG-7 basic descriptors is a short-term power spectrum within overlapping time frames. The calculation is done in the frequency domain. The initial procedure used is similar to that used for MFCC parameters. The frames were windowed and the resultant output of the window was passed through the FFT algorithm to calculate the DFT. The MPEG-7 basic descriptors are composed of four parameters, the Audio Spectrum Envelope (ASE), Audio Spectrum Centroid (ASC), Audio Spectrum Spread (ASS) and Audio Spectrum Flatness (ASF).

The ASE is the audio spectrum defined in the logarithmic frequency scale. It is mainly used to get a concise spectrogram containing the most meaningful data of the original audio signal. The ASE calculation involves the summation of energies of the original power spectrum over a set of frequency bands [8].

The Audio Spectrum Centroid (ASC) gives a description of the centre of gravity of a log-frequency power spectrum. The spectrum centroid gives a brief depiction of the structure of the power spectrum. It also shows whether the power spectrum is controlled by low or high frequencies and whether a major perceptual dimension of the timbre has affected it.

The Audio Spectrum Spread (ASS) is another calculation that defines the spectral shape of the audio signal. It shows whether the power is concentrated around its centroid or if it is spread throughout the whole spectrum.

The Audio Spectrum Flatness (ASF) shows the flatness characteristics that exist in the power spectrum. A series of values are outputted for each frame showing the divergence of the signal's power spectrum from a flat shape inside a predefined frequency band. This will define whether an audio signal is similar or correlated to white noise.

B. Non-Parametric Modeling

During non-parametric analysis, no assumptions are made regarding the type of distribution the probability density function (PDF) belongs to or how the PDF is shaped by the data.

The mapping algorithm transfers the original feature vectors to the integer classification of the bins in an efficient way to get a compact fingerprint model. This was performed for the two feature extraction methods, Mel Spectrum Coefficients and MPEG-7 basic descriptors.

The mapping algorithm considered uses a data binning technique. It processes the data and replaces the actual values by a bin number. Thus, the detail of actual coefficients is mapped into a range represented by a bin number, which is an integer. The data bins are easier to work with, the system remains robust and the searching time is reduced. Initially, the set of feature coefficients extracted from each audio file result in frame vectors that are stored in a database. Each vector, C, consists of a number of elements c_i, where i ranges from 1 to k, where k is the number of elements in a vector. For the whole audio set there is a considerably long set of vectors that characterise every frame. The mapping of these vectors to a reduced integer set is done as follows.

1) Every element c_i, from each vector is reorganised in ascending order so that there is an ordered array c^1_i to c^N_i

where N is the total number of vectors in the audio set.

2) Given a set of L integers, the ordered set is partitioned, such that thresholds, T^l are found. The number of elements between any two thresholds T^l and T^{l-1} is the same. This is similar in principle to non-uniform quantization thresholds. For any c_i such that

$$T^{l-1} < c_i < T^l$$

the element c_i is mapped to bin element v_i.

The process of ordering and setting appropriate thresholds is performed independently for each element i in all the vectors. Every vector C with elements c_1 to c_k maps into an integer vector V with elements v_1 to v_k, where the value of v_i is

$$1 < v_i < L$$

There are bins that have a big difference between the start value and the last value while there are bins with a small difference. This depends mainly on the mode since values that are more common are stored in a bin having a smaller range. Consequently the actual feature coefficient elements are preserved, but mapped.

3) A list storing the range for each bin is kept so that the feature coefficients of a testing file can be changed to bin numbers.
4) All the audio files in the database were passed through the process and the feature vector values were mapped to the number of the bin they fall into.

C. Parametric Modelling

In parametric analysis, a kind of distribution is first chosen and then its parameters and form are determined [9]. Its parameters include the mean and the standard deviation of the distribution. The GMMs were chosen as the parametric model. This approximates the feature vector elements into a sum of multiple Gaussians. A model is required for each of the five types of features. Four mixtures are used for each implementation. Then, using the built-in function in MATLAB, gmdistribution, a Gaussian mixture distribution is created based on the features and the number of mixtures. It creates a multivariate distribution which is made up of a mixture of four Gaussian distribution components. Since four mixtures are used, there are four distributions, each of which is described by a variance, mean and weight. The probability density for an element x, is denoted by the function:

$$f(x, \mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right] \quad (1)$$

Then each feature is mapped using the GMM by:

$$F(x, \mu, \sigma_i^2, \omega_i) = \sum_{i=1}^{N} \varpi_i f(x, \mu_i, \sigma_i^2) \quad (2)$$

The sum of the weights w_i is equal to one.

D. Arithmetic Measure

For the non parametric analysis, the Arithmetic Measure was the chosen classification method since it gives a quicker response. This is possible since the feature vectors within the audio database are mapped into integers. The Arithmetic distance is defined as:

$$d(test, train) = \sum_{i=1}^{size_{trdt}} \left|(train_i - test_i\right| \quad (3)$$

IV. Testing and Results

The experiments were carried out on 155 audio files. The audio files are classical music using different instruments like piano, violin, organ, harp, oboe, clarinet, flute and bassoon. All the feature coefficients were modelled using both parametric analysis and non-parametric analysis. Initially ten bins were used during mapping of non-parametric analysis, with the aim of reducing the search time during comparison of the test data. However this gave poor results especially for the MPEG-7 parameters.

The number of bins used in the mapping technique was incremented to fifteen. Table I shows the results for the non-parametric binning technique. The majority of the results for ASC and ASS still remained 0%. On the other hand, the accuracy results for ASE and ASF increased while that of the Mel Spectrum coefficients remained approximately the same. The results show that the audio parameter type is fundamental for proper classification. While the mapping technique used is quite simple, the results are satisfactory. Analysis of the whole set to obtain the bin model takes considerable time. But this is performed off-line. The classification, based on integer arithmetic, is however very fast.

TABLE I
Percent Age Accuracy Results Using 15 Bins

Testing File	ASE	ASC	ASS	ASF	Mel Spectrum
op10_01	100.00	10.00	0.00	85.00	100.00
op10_04_2	100.00	0.00	0.00	94.44	100.00
op10_08	88.46	0.00	0.00	100.00	100.00

Apart from doing non-parametric analysis, parametric analysis was also performed. GMMs were the chosen model. All of the audio files stored in the database were transformed into five Gaussian models, each representing one of the five features. GMMs perform quite similar to the binning technique for the ASE coefficients and Mel Spectrum Coefficients. On the other hand, the recognition accuracy for ASF coefficients was significantly reduced. This shows that the ASF resultant coefficients are much more sensitive during the modelling stage. The idea of PDFs in GMM affects the way the feature coefficients represent the flatness that exists in the audio pieces. On the other hand in the binning technique, the structure of the ASF feature vectors remains closer to the actual values since it is not being approximated by the parametric model. The results can be seen in Table II.

TABLE II
PERCENT AGE ACCURACY RESULTS USING GMM

Testing File	ASE	ASC	ASS	ASF	Mel Spectrum
op10_01	100.00	2.50	5.00	47.50	100.00
op10_04_2	100.00	0.00	0.00	50.00	100.00
op10_08	100.00	0.00	0.00	64.71	92.31

In order to get more realistic scenarios and check the robustness, the non-parametric system was tested in four different scenarios as can be seen in Table III. The first scenario included the addition of random noise to the test files where the accuracy rate for ASF reduced. This occurs since the flatness captures the characteristics that exist in the audio spectrum. During the second scenario, the speed of the testing sample was altered by dropping a sample every fifty samples. The flatness accuracy rate dropped since the flatness works out an average of the samples. Another testing scenario was up sampling of the testing files. This means that every fifty samples, an additional sample was inserted. As a result, the overall envelope of each frame changes and consequently the ASE accuracy is approximately 0% for the majority of the files. One important scenario that was tested was that a similar copy of the audio files played by a different musician was downloaded from the Internet. The Mel Spectrum coefficients and the ASF coefficients identified the majority of the files but none of the files were identified with the ASE coefficients. The files used had different formats and these had to be converted to wav files for analysis. In general the downloaded files will have different characteristics resulting in some loss of data precision as well as a change of data. Another factor that affects the performance is the number and type of instruments being used in the testing files. This might differ from those files stored in the database.

TABLE III
PERCENTAGE OVERALL RESULTS FOR ONE TEST FILE

	10 bins	15 bins	Addition of noise	Slow-Speed	Fast-Speed	Files from Internet
Mel Spectrum	100.00	100.00	100.00	100.00	0.00	39.35
ASE	98.53	100.00	100.00	100.00	0.00	0.01
ASF	79.41	100.00	77.94	82.09	79.41	25.16

Published results include [10], that make use of four seconds test items and search in a 15,000 song database. Their recognition rate varies between 90% and 100%, including tests for robustness on known songs. Allamanche et al make use of the Spectral Flatness Measure (SFM) using a standard Vector Quantization approach. The ASF using the data binning technique achieves comparable result to those in [11]. For no distortion the results are identical. For robustness, the results show that there are problems, especially with unknown music taken down from the internet.

V. CONCLUSION

This paper has presented the implementation of an audio fingerprinting system using a non-parametric based modeling algorithm.

The experiments were carried out on 155 audio files. The audio files are classical music using different instruments like piano, violin, organ, harp, oboe, clarinet, flute and bassoon. The Mel Spectrum and the MPEG-7 basic descriptors were chosen as the two feature extraction methods. The features were modeled using a non-parametric based model algorithm which made use of fifteen bins and parametric modeling using GMMs.

The data-binning feature modelling method using fifteen bins performed better than the GMM method since it reflects better the actual values lying in the feature coefficients. This shows that non-parametric models offer potential results as the classic ones mentioned in the literature. Moreover, the number of calculations involved for non-parametric models is low.

Future work can include scaling the database using non-uniform ranges, as well as the use of parallel binning sets aiming to increase the robustness of the system.

REFERENCES

[1] P. Cano, E. Batlle, T. Kalker and J. Haitsma, “A Review of Audio Fingerprinting,” Journal of VLSI Signal Processing Systems, vol. 41, no. 3, pp. 271-284, Nov. 2005.

[2] S. Baluja and M. Covell, “Waveprint: Efficient Wavelet-Based Audio Fingerprinting,” in Pattern Recognition, pp. 3467-3480, Nov. 2008.

[3] J. Haitsma and T. Kalker, “A Highly Robust Audio Fingerprinting System,” in Proc. Of ISMIR, 2002.

[4] Y. Ke, D. Hoiem, and R. Sukthankar, “Computer vision for music identification,” in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June, 2005.

[5] A. Ramalingam and S. Krishnan, “Gaussian Mixture Modeling of Short time Fourier Transform Features for Audio Fingerprinting,” IEEE Trans. Inf. Forens. Security, vol. 1, no. 4, pp. 457-463, Dec. 2006.

[6] E. Battle, J. Masip, E. Guaus and P. Cano, “Scalability issues in an HMM-based audio fingerprinting,” in Multimedia and Expo 2004. ICME ’04. 2004 Int. Conf., vol. 1, 2004, pp. 735-738.

[7] J.W. Picone, “Signal modeling techniques in speech recognition,” in Proc. of IEEE, vol. 81, no. 9, 1993, pp. 1215-1247.

[8] M. Babtan, (2009, December 23). MPEG-7 (Online). Available: http://www.cs.bilkent.edu.tr/~bilmdg/bilaudio-7/MPEG7.html.

[9] J. Bercher and C. Vignat, “Estimating the entropy of a signal with applications,” IEEE Transactions on Signal Processing, vol. 48, no. 6, pp. 1687–1694, June 2000.

[10] J. Herre, O. Hellmuth and M. Cremer, “Scalable Robust Audio Fingerprinting Using MPEG-7 Content Description,” Multimedia Signal Processing, 2002 IEEE Workshop, pp. 165-168, Dec. 2002.

[11] E. Allamanche, J. Herre, O. Helmuth, B. Fröba, T. Kasten, and M. Cremer, “Content-Based Identification of Audio Material Using Mpeg-7 Low Level Description,” Proc. of the Int. Symp. Of Music Information Retrieval, pp. 197-204, Oct. 2001.