# CONTENT BASED AUDIO CLASSIFICATION AND RETRIEVAL USING JOINT TIME-FREQUENCY ANALYSIS

*S. Esmaili, S. Krishnan and K. Raahemifar*

Multimedia Information and Signal Analysis Research (MISAR) Laboratories
Department of Electrical and Computer Engineering
Ryerson University, Toronto, Ontario, Canada
e-mail: (sesmaili)(krishnan)(kraahemi)@ee.ryerson.ca

**ABSTRACT**

In this paper, we present an audio classification and retrieval technique that exploits the non-stationary behavior of music signals and extracts features that characterize their spectral change over time. Audio classification provides a solution to incorrect and inefficient manual labelling of audio files on computers by allowing users to extract music files based on content similarity rather than labels. In our technique, classification is performed using time-frequency analysis and sounds are classified into 6 music groups consisting of rock, classical, folk, jazz and pop. For each 5-second music segment, the features that are extracted include entropy, centroid, centroid ratio, bandwidth, silence ratio, energy ratio, and location of minimum and maximum energy. Using a database of 143 signals, a set of 10 time-frequency features are extracted and an accuracy of classification of around 93% using regular linear discriminant analysis or 92.3% using leave one out method is achieved.

## 1. INTRODUCTION

With the abundance of personal computers, advances in high speed modems operating at 100 Mbps and GUI based peer-to-peer (P2P) file-sharing systems that make it simple for individuals without much computer knowledge to download their favorite music, there has been an increase of digitized music available on the Internet and on personal computers. As such, there is also a rising need to manage and efficiently search the large number of multimedia databases available online which is difficult using text searches alone. Current multimedia databases are indexed based on song title or artist name which requires manual entry and improper indexing could result in incorrect searches. A more effective content based retrieval system, analyzes audio signals, selects and extracts dominant perceptual features and classifies the music based on these features. Stronger features provide a higher degree of separation between classes and thereby a higher classification accuracy. The aim is to make music search engines as effective as text-based ones and this is examined further in this paper.

In recent years, there has been many works on audio classification with various perceptual features and several classification algorithms. In one of the pioneer works done on audio classification and later commercialized into the "Muscle Fish" project, Wold et al [1] extracted an N dimensional vector consisting of several acoustical features such as loudness, pitch, brightness, bandwidth and harmonicity from each sound. A Euclidean (Mahalanobis) distance is then calculated between the input sound feature vector and the existing models in the database. Using the nearest neighbor (NN) rule, the signal is grouped into the class with the minimal Euclidean distance.

In a similar work to that of [1], Liu et al [2] extract 13 different audio features to separate audio clips into different scene classes such as advertisement, basketball, football, news and weather. Features consist of volume distribution, pitch contour, bandwidth, frequency centroid and energy. A neural network classifier with a one-class-in-one network (OCON) structure is used and an overall classification rate of 88% is achieved. Artificial neural networks (ANN) are effective in detecting complex nonlinear relationships while requiring little formal training. However, their process is computationally expensive and more importantly, the relation between the input and output variables is defined in a black box model that has no analytical basis. In terms of audio classification this means that it is difficult to deduce which acoustical features are significant in classifying each type of sound [1].

In a different technique, Lu and Hankinson [3] used a rule-based heuristic classification method to classify an audio signal into speech, music and noise. For each feature, a threshold is set to determine the segment type and the feature set includes silence ratio, centroid, harmonicity and pitch. Since the feature threshold must change for different audio inputs, this type of classifier is tedious and not ideal. A classification rate of 75% for speech, and 89% for music is reported.

Lu et al [4] proposed support vector machines (SVMs) as an alternative to current classification methods. Using a kernel-based SVM increases the classification rate by separating nonlinear cases. Here, a nonlinear kernel function maps the data to a high dimensional feature space where the data is linearly separable. The authors use a combination of a rule-based classifier and a kernel based SVM to distinguish between 5 different audio classes including silence, music, background sound, pure speech and non-pure speech. Their feature set include similar features to those reported in [1] and [5], such as MFCCs, zero-crossing rate (ZCR), short time energy (STE), sub-band powers, brightness, and bandwidth with some new features such as spectral flux (SF), band periodicity (BP), and noise-frame-ratio (NFR). An average classification accuracy of around 90% is achieved.

In the majority of the previous work in this area, audio is examined in either the time or frequency domain where it is assumed that the signals are wide sense stationary. In reality, sounds are non-stationary and multi-component signals consisting of series

Thanks to Micronet and NSERC for funding.

of sinusoids with harmonically related frequencies. Our algorithm considers the short-time Fourier transform (STFT) of an audio signal to extract parameters that will be used to classify signals. Our retrieval technique is less computationally intensive than those that use ANN, SVM, or Hidden Markov Models (HMM). Also, the efficiency of features can be examined which is not feasible in ANNs. Note that while HMM can be used to examine spectral change over time, past works have shown that HMM needs to be coupled with external features such as Cepstral or perceptual features to be efficient [6]. Finally, our method also offers the added improvement that it is not specific to certain audio files and can be applied without adjusting the algorithm or thresholds such as in rule-based models.

Our work on content-based audio classification is presented as follows. Section 2 presents the application of time-frequency analysis to feature selection and analysis for audio classification. In Section 3 we present our classification results for the system and our conclusions are provided in Section 4.

## 2. METHODOLOGY

### 2.1. Short-time Fourier transform (STFT) algorithm

Since speech and audio signals have spectral characteristics that vary over time, they require a non-stationary signal model such as the STFT to describe them. Ultimately, we would like to imitate the capability of the ear and provide simultaneous information about time and frequency of the music. STFT uses a sliding window to compute the Fourier transform thereby providing an estimate of the "local frequency" at a given time. The STFT of a signal $x[n]$ is given by,

$$STFT(n,f) = \sum_{m=-\infty}^{\infty} x[n+m]w[m]e^{-j(2\pi f)m} \qquad (1)$$

where $w[m]$ is the window function and the spectrogram of $x$ is defined as $SPEC(n,f) = |STFT(n,f)|^2$. For a given signal $x$, $SPEC(n,f)\Delta n\Delta f$ represents the energy in the time interval $[n, n+\Delta n]$ in the frequency band $[f, f+\Delta f]$. In STFT analysis, we can improve the frequency resolution by decreasing the spectral width $\Delta f$ at the expense of increasing the temporal width $\Delta n$ (poor time resolution). Also the shape of the window $w[n]$ is important as a window with a sharp cutoff will introduce artificial discontinuities. Hanning windows are mainly used in audio classification techniques as they reduce spectral leakage.

### 2.2. Audio feature extraction

The set of features extracted are critical as they need to be strong enough to clearly separate the classes of signals. This procedure requires perceptual features that model the human auditory system. Discriminating music from speech is less complex than between different classes of music. The latter may only require a small number of features such as zero crossing rate or energy envelope and since the spectral characteristics are not very similar, high accuracy rates are achieved.

In this paper, we examine the similarities of 143 audio signals and classify them under six different genres. Each audio signal is 5 seconds, mono-channel, 16 bits per sample and sampled at 44.1 kHz. The length of the audio samples was chosen to be 5 seconds in relevance with the human neurological behavior which was examined by Perrot et al in [7]. They found that human beings require at least 3 second excerpts to identify different musical genres with a 70% accuracy rate while the accuracy decreases to 53% for a 250 ms excerpt.

We start by transforming our audio signal into a spectrogram with a window size of 1024 samples which corresponds to about 23 ms at 44.1 kHz. This window size is similar to that used in [4] and [8]. A Hanning window with 50% overlap is used and the DFT is calculated in each window. The audio features extracted from the two-dimensional time-frequency distribution (TFD) are explained below.

#### *2.2.1.* Entropy

The entropy of a signal is a measure of its spectral distribution and portrays the noise-like or tone-like behavior of the signal. The entropy of a signal in time frame $n$ can be calculated as:

$$H(n) = \sum_{f=0}^{F_m} P_f(TFD(n,f)) \log_2 P_f(TFD(n,f)), \qquad (2)$$

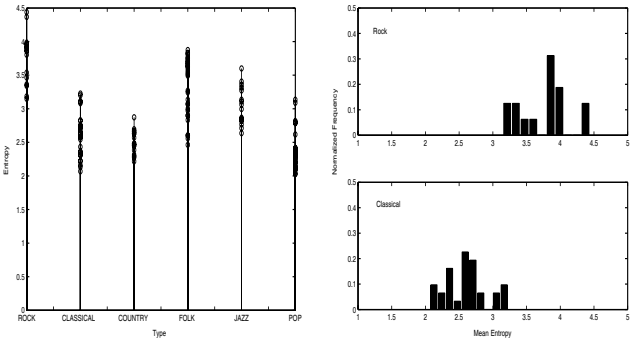
where

$$P_f(TFD(n,f)) = \frac{TFD(n,f)}{\sum_{f=0}^{f=F_m} TFD(n,f)}. \qquad (3)$$

Here, $TFD(n,f)$ represents the energy of the signal at time frame $n$ and frequency index $f$ (it is equivalent to $SPEC(n,f)$ defined in Section 2.1). Also, $F_m$ refers to the maximum frequency.

Consider the case where there are $L$ number of frequency bins. Then the maximum entropy in time window $n$ is $\log_2 L$ which occurs if the frequency bins are equiprobable. First, we examined the entropies of 3 different types of signals. These signals were analyzed using 128 frequency bins, implying that the maximum entropy is 7 bits. The first signal consisted of a single sine wave, at a sampling frequency of 1 kHz. In this case, the mean entropy was 1.24 bits and the standard deviation at $5.636 \times 10^{-6}$. Next we considered the vowel "a" (a signal component with harmonic structure) and its entropy was calculated to be 2.84 bits with a standard deviation of 0.1. Finally, we considered white Gaussian noise and its mean entropy was 6.38 bits with a standard deviation of 0.06. As we expected, the sine wave had the lowest entropy and a standard deviation of almost zero while white noise had the largest entropy (approaching maximum) with a larger standard deviation.

From our database of music signals, we found that entropy was a dominant feature in classifying particularly rock or folk music. As shown in Figure 1a, rock signals possessed the highest entropy followed closely by folk music while classical, country, jazz and pop had low entropies. Figure 1b shows the distribution of entropy for rock music compared to classical. As can be seen, the entropy ranges for the two types of signals are quite different. In order to determine the strength of entropy from a different perspective, a receiver operating curve (ROC) was plotted. The ROC curve is a two dimensional measure of classification performance. The area under this curve measures discrimination, or the ability of a feature to correctly classify signals. An area of 1.0 represents a perfect test; where an area of 0.5 or less shows the feature is not useful in discrimination of that class. Rock, folk, jazz, classical, country and pop music had ROC areas of 0.933, 0.808, 0.644, 0.337, 0.294, and 0.145 respectively. These results show that although entropy is a strong feature, further features are required to improve classification.




(a) (b)

**Fig. 1**. Comparison of entropy values a) Results for different genres b) Distribution for classical and rock.

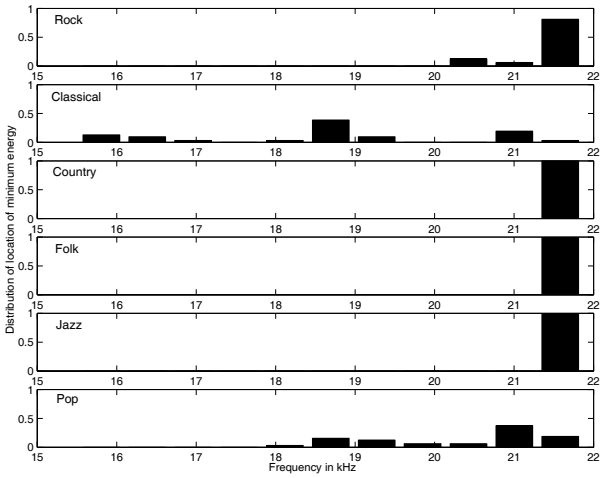



**Fig. 2**. Distribution of location of minimum energy

#### *2.2.2.* **Energy ratio**

The rate of change in the spectral energy over time was measured as the mean of the total energy in a frequency sub-band to the previous time window ($E[\frac{\sum_{f=f_{lower}}^{f_{upper}} TFD(n,f)}{\sum_{f=f_{lower}}^{f=f_{upper}} TFD(n-1,f)}]$). This was examined in three different sub-bands [0, 5 kHz], [5, 10 kHz], [10 kHz, $F_m$]. However, it was found empirically that the energy ratio in mid and high frequency bands did not improve the classification. This is probably because most energy activity in audio signals is in the low frequency band. Therefore, only the mean of energy in the low-band was used in our feature set.

The frequency location with the lowest energy component was also computed. Although an estimate of the mean can be calculated from the frequency domain, it was included in our feature set as it improved the classification rate by 5%. In fact, using the mean and standard deviation of the location of minimum energy provided 100% classification rates for classifying country, folk and jazz music but low classification rates for the other three genres. When examining the histogram of the location of minimum energy for our database of signals (Figure 2), the frequency spread was smaller for country (21.4-21.5 kHz), folk (21.45-21.85 kHz), jazz (21.36-21.51 kHz) and a wider range for pop (18.1-21.5kHz), classical 15.5-21.5kHz) and rock (20-21.6 kHz).

#### *2.2.3.* **Brightness**

The brightness of a signal also referred to as its frequency centroid, shows the weighted midpoint of the energy distribution in a given frame. It is defined by:

$$f_i(n) = \frac{\sum_{f=0}^{F_m} f\, TFD(n,f)}{\sum_{f=0}^{F_m} TFD(n,f)}. \qquad (4)$$

The brightness feature could also be seen as the instantaneous mean frequency parameter, a typical non-stationary feature of a signal. The frequency centroid of the audio signal in the low frequency range (0-5KHz) is also examined as most of the frequency content of audio signals is concentrated in low frequency.

In addition, the mean of centroid ratio to previous window is a useful feature as it measures the spectral change over time. We found that rock, folk, pop and country music signals had the largest change in centroid frequency over time while classical and jazz signals had the lowest change. This is expected as classical and jazz music generally have less activity over time compared to the other 4 genres.

#### *2.2.4.* **Bandwidth**

Bandwidth is the magnitude-weighted average of the difference between the signal's spectral components and centroid. It can be defined as:

$$B(n) = \sqrt{\frac{\sum_{f=0}^{F_m} (f - f_i(n))\, TFD(n,f)}{\sum_{f=0}^{F_m} TFD(n,f)}}. \qquad (5)$$

Effectively, it shows the spectral shape and the spread of energy relative to the centroid, therefore it is also a non-stationary feature. For instance, a sine wave without noise has zero bandwidth.

#### *2.2.5.* **Silence ratio**

Silence ratio is the number of silent time window frames with total energy less than 0.01. This threshold is set empirically. Note that this feature could also be extracted from the time domain.

Bandwidth, brightness and silence ratio have been proven to be effective in previous audio classification papers including [1, 2] although an STFT approach showing the rate of change to previous windows has not been used.

## 3. AUDIO CLASSIFICATION

Using the above analysis, the 10 features extracted for each sample included mean and standard deviation of centroid frequency, mean centroid (low-frequency range), mean of centroid ratio to previous window, mean bandwidth, silence ratio, mean and standard deviation of the frequency location with the lowest energy, mean and standard deviation of entropy. Note that mean and variance of a feature are calculated over the entire time window. Once the features are extracted for the 143 audio signals, linear discriminant analysis (LDA) is then applied using SPSS software [9], to predict group classification of cases. This type of analysis tries to

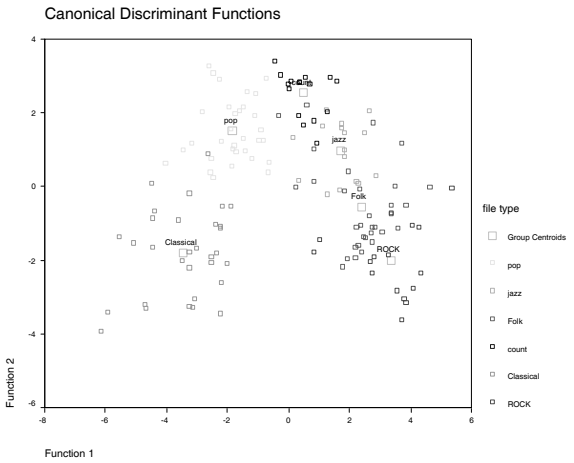



**Fig. 3**. All-groups scatter plot with the first two canonical discriminant functions

find a linear combination of those extracted features that best separate the group of cases. To represent this linear combination, a discrimination function is formed using the extracted features as discrimination variables and can be expressed as:

$$L = b_1 x_1 + b_2 x_2 + ..... + b_{10} x_{10} + c, \qquad (6)$$

where $b_1..b_{10}$ are the coefficients, $c$ is a constant and $x_1..x_{10}$ are the values of the extracted features. This technique finds the first function that separates the groups as much as possible and then finds further functions that improve the separation and are uncorrelated to previous ones. The number of functions is determined by the number of predictors or features and the number of groups available.

Using Fisher's coefficients and prior probabilities of each group, a scatterplot (Figure 3) is created showing the discriminant scores of the cases on two discriminant functions. This plot shows the separation between different cases. Songs are categorized into six groups (rock, classical, country, folk, jazz and pop) and the confusion matrix depicted in Table 1 shows the classification performance. Using the original LDA, 93.0% of all original grouped cases are correctly classified with folk music having the lowest rate. A more accurate estimate is obtained through the cross-validated method where a portion of cases belong to the learning sample and the other cases belong to the test sample. In the leave-one-out method used, each case is classified by the functions derived by all cases except that one. This method yields a 92.3% classification rate revealing the discrimination strength of our feature set.

## 4. CONCLUSIONS

In this paper, we examined a technique where features used to classify music signals are derived directly from the time-frequency domain. Using six different genres for classification, we have shown that high accuracy rates can be obtained using features that reflect the non-stationarity properties of audio signals and are able to depict its spectral, energy and entropy change over time. In addition to the success rate, our algorithms have low computational complexity compared to other techniques and they offer versatility as they can be applied to any audio signal without alteration. Further work will include optimization of window size in the TF domain as well as examining other classification methods such as minimum classification error (MCE) to improve classification rate for a larger database of signals.

| Method | Type | RO | CL | CO | FO | JA | PO | CA% |
|---|---|---|---|---|---|---|---|---|
| **1. Original** | **RO** | **14** | 0 | 0 | 2 | 0 | 0 | **87.5** |
| | **CL** | 0 | **30** | 0 | 0 | 0 | 1 | **96.8** |
| | **CO** | 0 | 0 | **15** | 0 | 0 | 1 | **93.8** |
| | **FO** | 2 | 0 | 1 | **27** | 1 | 1 | **84.4** |
| | **JA** | 0 | 0 | 0 | 1 | **15** | 0 | **93.8** |
| | **PO** | 0 | **0** | 0 | 0 | 0 | **32** | **100** |
| | **Overall** | | | | | | | **93.0** |
| **2. Cross-** | **RO** | **14** | 0 | 0 | 2 | 0 | 0 | **87.5** |
| **Validated** | **CL** | 0 | **30** | 0 | 0 | 0 | **1** | **96.8** |
| | **CO** | 0 | 0 | **15** | 0 | 0 | 1 | **93.8** |
| | **FO** | 2 | 0 | **1** | **26** | 1 | 2 | **81.3** |
| | **JA** | 0 | 0 | 0 | 1 | **15** | 0 | **93.8** |
| | **PO** | 0 | **0** | 0 | 0 | 0 | **32** | **100** |
| | **Overall** | | | | | | | **92.3** |

**Table 1**. Classification results. Method: Original - Linear discriminant analysis, Cross - validated - Linear discriminant analysis with leave-one-out method (RO-Rock, CL-Classical, FO-Folk, Ja-Jazz, PO-Pop, CA% - Classification accuracy rate)

## 5. REFERENCES


[1] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, pp. 27–36, 1996.

[2] Z. Liu, J. Huang, Y. Wang, and T. Chuan, "Audio feature extraction and analysis for scene classification," in *IEEE Workshop on Multimedia Signal Processing*, June 1997, pp. 343–348.

[3] G. Lu and T. Hankinson, "A technique towards automatic audio classification and retrieval," in *Fourth International Conference on Signal Processing*, Beijing, China, October 1998, pp. 1142–1145.

[4] L. Lu, H. Zhang, and S. Li, "Content-based audio classification and segmentation by using support vector machines," *ACM Multimedia Systems Journal 8*, vol. 8, no. 6, pp. 482–492, March 2003.

[5] J. Foote, "Content-based retrieval of music and audio," in *Multimedia Storage and Archiving Systems II, Proc. of SPIE*, 1997, pp. 138–147.

[6] T. Zhang and C. Kuo, "Hierarchical classification of audio data for archiving and retrieving," in *Proc. ICASSP*, March 1999, pp. 3001–3004.

[7] D. Perrot and R.O. Gjedigen, "Scanning the dial: An exploration of factors in the identification of musical style," *Proceedings of the 1999 Society for Music Perception and Cognition*, p. 88, 1999.

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

[9] SPSS Inc., "SPSS advanced statistics user's guide," in *User manual, SPSS Inc., Chicago, IL*, 1990.