Audio Information Retrieval (AIR) Tools

George Tzanetakis
Computer Science Department
Princeton University
gtzan@cs.princeton.edu

Perry Cook
Computer Science and Music Department
Princeton University
prc@cs.princeton.edu

Abstract

Most of the work in music Information Retrieval (MIR) and analysis has been performed using symbolic representation like MIDI. The recent advances in computing power and network connectivity have made large amounts of raw digital audio data available in the form of unstructured monolithic sound files. In this work the focus is on tools that work directly on real world audio data without attempting to transcribe the music. To distinguish from symbolic−based music IR for the remainder of the paper we use the term audio IR (AIR) to refer to techniques that work directly on raw audio signals. Obviously these signals can contain music as well as other types of audio like speech. We describe a series of tools based on current and newly developed techniques for AIR integrated under MARSYAS, our framework for audio analysis. For related work refer to (Foote, 1999).

The tools developed are based on Signal Processing, Pattern Recognition and Visualization techniques. Finally, due to the immature state of the available techniques and to the inherent complexity of the task it is important to take advantage of the human user in the system. We have developed two user interfaces to integrate and improve our techniques: an augmented sound editor and *TimbreGrams* a novel graphical representation for soundfiles. The previously unpublished contributions of this paper are the genre classification method, the segmentation−based retrieval and summarization, and the definition of the *TimbreGram*.

Feature−based audio analysis

The developed analysis tools are based on the calculation of short−time feature vectors. The signal is processed in small chunks so that its statistical characteristics are relativily stable. For each chunk some form of spectral analysis is performed and based on that analysis a vector of feature values is calculated. In our system features based on FFT (Fast Fourier Transform) analysis, MPEG filterbank analysis, LPC (Linear Predictive Coding) and MFCC (Mel−Frequency cepstrum coefficients) are supported. In addition derivatives and running statistics are used to express temporal changes. The flexible architecture of MARSYAS allows the easy integration and experimentation of new features. Based on the calculated features different types of audio analysis processes can be performed.

Classification

In Classification the extracted feature vectors are assigned to one of **c** classes. In MARSYAS, the Gaussian (MAP), Gaussian Mixture Model (GMM) and K−NN families of statistical Pattern Recognition classifiers are supported. Two case studies of classification have been implemented and evaluated in our system: a music/speech classifier and a genre classifier. The music/speech classifier achieves a 90.1% classification accuracy. The genre classifier uses three classes/genres to describe the data: classical, modern (rock, pop) and jazz. It achieves classification of 75%. These results are calcualted using a robust frame based evaluation to ensure that the performance is indicative of real world data. The dataset for the development and evaluation of these classifiers consists of two hours of representative audio broken into 30 sec sound files.

Segmentation

Segmentation refers to the process of detecting when there is a change of "texture" in a sound stream. For example the chorus of a song, the entrance of a guitar solo, or the change from music to speech are all examples of segmentation boundaries. MARSYAS supports the general segmentation methodology described in (Tzanetakis & Cook, 99) that uses tracking of multiple features in time. Intuitively the signal is viewed as a trajectory of points (feature vectors) in a high−dimensional space. Abrupt changes in this trajectory indicate segmenation boundaries.

Retrieval

In content−based AIR the query is a sound file and the result is a list of sound files ranked by their similarity. Three approaches are used in MARSYAS to represent a sound file for retrieval. In the first approach the sound file is represented by a single feature vector. In the second approach the sound file is initially segmented resulting in a variable length list of feature vectors, and in the final approach the whole trajectory is used. Content−based retrieval is subjective and the only way to properly evaluate a system is through user studies. We developed an infastructure for conducting AIR user evaluation studies over the Web and conducted a small−scale evaluation.

Summarizaion

Summarization refers to the process of creating a short summary sound file from a large sound file in such a way that the summary best captures the essential elements of the original sound file. Summarization is important for AIR especially for the presentation of the returned ranked list. A clustering−based and a segmenation−based summarization algorithms are supported in MARSYAS. User experiments comparing the two methods are planned for the future.

Augmented SoundEditor

The MARSYAS sound editor offers the same functionality as a traditional sound editor like Waveform and Spectogram displays, zooming etc. In addition to these typical features, a sound file can be segmented with each region displayed with a different color. For quick browsing the user can move by regions and each region can be annotated with text. In addition, regions can easily be added or deleted. Different classification and summarization schemes can be applied to each segmented region or to arbitrary selections. Integrating all the developed techniques under a common interface and letting the user affect the results makes an effective robust system that combines the strengths of different analysis tools and manual annotation.

TimbreGrams

TimbreGrams are a new graphical representation of sound. The main idea is to use the color perception and pattern recognition of the human visual system to depict timbral and temporal information. A *TimbreGram* is a series of vertical color stripes where each stripe corresponds to a feature vector. Time is mapped from left to right. The mapping of the vectors to color is performed using Principal Component Analysis (PCA) described in (Jollife, 86). Sound textures that are similar have similar colors. In addition, time periodicity (like ABA form) is shown in color. Without any explicit class model Speech, Rock, and Classical separate easily visually.

Suggested Reading

Duda, R., Hart, P. 1973. *Pattern Classification and Scene Analysis*. John Willy & Sons.

Foote, J. 1999 An oveview of audio information retrieval. *ACM Multimedia Systems*, 7:2−10.

Jollife, L.T. 1986. *Principal Component Analysis*. Springer Verlag.

Scheirer, E. 1998. Tempo and beat analysis of acoustic musical signals. *Journal .Acoustical Society of America (JASA)*, 103(1):588,601.

Tzanetakis, G.,Cook, P. 1999 Multifeature audio segmentation for browsing and annotation.*Proc.IEEE Workshop on Apps of Signal Proc to Audio and Acoustics (WASPAA'99).*