Panel: New directions in Music Information Retrieval

Roger Dannenberg, Jonathan Foote, George Tzanetakis*, Christopher Weare (panelists)

*Computer Science Department, Princeton University
email: gtzan@cs.princeton.edu

Abstract

This paper and panel discussion will cover the growing and exciting new area of Music Information Retrieval (MIR), as well as the more general topic of Audio Information Retrieval (AIR). The main topics, challenges and future directions of MIR research will be identified and four projects from industry and academia are described.

1 Introduction

The internet is destined to become the dominant medium for disseminating recorded multimedia content. Currently, music downloads, such as MP3 files, are a major source of internet traffic. As more music is made available via networks, the need for sophisticated methods to query and retrieve information from these musical databases increases. The projected growth in musical databases parallels that of publishing: text databases have grown in size and complexity at a rapid pace.

One of the major technological consquences of content on the web has been the accelerated development of sophisticated search engines. As more content becomes available, users demand more sophisticated search processes. As users employ more sophisticated search processes, they quickly demand more content. The same user-driven model for information and retrieval has already started to develop for multimedia search and retrieval.

In the past, the majority of AIR and MIR research was conducted using symbolic representations of music like MIDI, because they are easy to work with and require modest amounts of processing power. There has been a large history of work in this area, and there are many existing tools that analyze and parse these representations. In recent years the large amounts of music available as raw or compressed digital audio and the improvements in hardware performance, network bandwidth and storage capacity have made working directly with digital audio possible.

Music search engines require an entirely different methodology than text search. These search engines require primarily a sonic interface for query and retrieval. Such an interface allows the user to explore the rich perceptual cues that are inherent in music listening.

Music is a multifaceted, multi-dimensional medium that demands new representations and processing techniques for effective search. Furthermore, constructing a music search engine with the scale and efficiency needed for the large amount of music available today requires fundamental research.

Music Information Retrieval (MIR) is not just interesting because of the commercial consumer applications that it enables. There are important applications to musicology, music theory, and music scholarship in general. Searching for examples of music features or analyzing a corpus of music for compositional techniques are just two examples of how MIR can assist music research.

Of even greater importance to the computer music community are the close ties between music information retrieval and other computer music research. MIR implies the use of analysis procedures for music in a variety of representations. What are good computer representations for music? What characterizes a style of music? What distinguishes one composer from another? Can we synthesize examples of style, genre, compositional techniques, rhythmic patterns, instruments and orchestration to render queries into sounds and to better understand our representations? These are fundamental questions for computer music research in general, not only for music information retrieval.

In this paper and panel we will provide an overview of current AIR research and topics. Areas relevant to MIR are Information Retrieval, Signal Processing, Pattern Recognition, AI, Databases, Computer Music and Music Cognition. A list of general references is given at the end of the paper. Many references are academic papers, but many are company web sites, reflecting the commercial interest and potential of MIR technology and applications. The list of company web sites and academic papers is representative of the increasing activity in MIR but it is by no means complete and exhaustive.

The paper is structured as follows: Section 2 provides a short descriptions of the main topics of MIR research as have been identified by academic and commercial work in this area with representative citations. Section 3, 4, 5 and 6 describe specific MIR projects, that the panelists have been involved, both from academia and from the industry.

2 MIR topics

Although still in its infancy, several different topics have been identified by the published papers on MIR research. These topics are related and would all be integrated in a full MIR system. For example, genre classification can inform play list generation or segmentation can improve classification results. This close relation is reflected in the papers than many times span more than one topic.

The following list provides a short description and rerpesentative references for each of these topics:

- ***Content-based similarity retrieval*.**

Given an audio file as a query, the system returns a list of similar files ranked by their similarity. The similarity measure is based on the actual audio content of the file. (Wold et al, 1996,1999, Foote 1997,1999)

- ***Play list generation*.**

Closely related to similarity retrieval. The input is a set of metadata constraints like genre, mood or beat. The result is a list of audio files that fulfil these constraints. Another play list generation method is to morph between audio file queries. In both cases smooth transitions between successive play list files are desired. (Algoniemy and Tewfik, 2000)

- ***Thumbnailing***

Given an audio file create a new file of smaller duration that captures the essential characteristics of the original file. Thumbnailing is important for presentation of multiple files, for example in similarity lists. (Logan, 2000, Tzanetakis and Cook, 2000)

- ***Fingerprinting***

The goal of this technique is to calculate a content-based compact signature that can be used to match an audio file in a large database of audio file signatures. No metadata information like the filename is used for the calculation of the signature. The calculated signatures must be compact, robust to different audio transformation like compression and must allow fast matching in a large database.

- ***Classification***

In classification an audio file is assigned to a class/category from a predefined set. Examples of possible classifications are: Genre, Male/Female voice, Singing vs. Instrumental etc. To express more complex relations, hierarchical classification schemes can be used (Wold et al., 1996, 1999, Tzanetakis and Cook, 2000, Scheirer and Slaney, 1997, Soltau et al. 1998).

- ***Segmentation***

Segmentation refers to the process of detecting segments when there is a change of “texture” in a sound stream. The chorus of a song, the entrance of a guitar solo, and a change of speaker are examples of segmentation boundaries. (Foote 2000a, Tzanetakis and Cook, 2000, Sundaram and Chang, 2000)

- ***Browsing***

In many cases the user does not have a specific search goal in mind. In those cases, browsing is used to explore the space of audio files in a structured and intuitive way.

- ***Beat detection***

Beat detection algorithms typically automatically detect the primary beat of a song and extract a measure of how strong the beat is. (Scheirer and Slaney, 1998, Guyon et al., 2000, Foote and Uchihashi 2001)

- ***Polyphonic transcription***

Polyphonic transcription systems are one of the bridges that can connect the world of symbolic analysis to real world audio. Unfortunately, despite various efforts at automatic transcription in restricted domains, a robust system that can work with real world audio signals has not yet been developed.

- ***Visualization***

Visualization techniques have been used in many scientific domains. They take advantage of the strong pattern recognition abilities of the human visual system in order to reveal similarities, patterns and correlation both in time and space. Visualization is more suited for areas that are exploratory in nature and where there are large amounts of data to be analyzed like MIR.

- ***User interfaces***

In addition to the standard design constraints, user interfaces for MIR must be able to work with sound, be informed by automatic analysis techniques and in many cases updated in real-time.

- ***Query synthesis***

An interesting direction of research is the automatic synthesis of queries. The query rather than being a sound file is directly synthesized by the user by manipulating various parameters related to musical style and texture. This research direction has close ties with automatic music style generation.

- ***Music Metadata***

In addition to content-based information other types of information like artist name, record label, etc. need to be supported in MIR. Standards like MPEG 7 are designed to provide researchers and industry with suggested attributes and tools for working with them. When sing musical metadata traditional text information retrieval techniques and databases can be used.

- ***Multimodal analysis tools***

An interesting direction of research is combining analysis information from multiple streams. Although speech analysis has been used with video analysis (Hauptmann and Witbrook, 1997) very little work has been done with music analysis.

3 The automation of the MSN Search Engine: a commercial perspective

(Christopher Weare, Microsoft)

Recent advances in the field of machine listening have opened up the possibility of using computer to create automated musical search engines. While several automated systems now exist for searching limited sets of recordings, much work remains to be done before a completely automated system that is suitable for searching the universe of recorded music is available.

The research at MSN music is focused on the development of a commercially-viable music search engine that is suitable for non-experts (MSN 2001). To be effective, the search engine must present a simple and intuitive interface, the database must contain a database of millions of songs, and searches should complete within a few seconds at most. Of course, the results need to be meaningful to the user.

3.1 Background

The MSN Music Search Engine (MMSE) interface is centered on the idea of musical similarity. The imagined user scenario is illustrated in the following: "I like this known piece of music, please give me other musical recordings that "Sound Like" this recording." The "Sound Like" metaphor implies some measure of similarity or metric.

The first challenge is to determine what actually constitutes distance between songs. At first glance this might not seem a difficult task. After all, most individuals can readily discern music of differing genres. They can even go on to describe various aspects of the music that they feel distinguishes songs of differing style. However, identifying the salient perceptual attributes that are useful in distinguishing a wide catalog of music is a non-trivial undertaking; just ask your local musicologist. Add to this task the constraint that there must be some hope of extracting said parameters from the musical recordings without human intervention and the task becomes even more difficult.

3.2 Perceptual Space

The perceptual attributes used by the MMSE were identified by musicologists at MongoMusic (acquired by Microsoft in the fall of 2000) and have been refined over time as user feedback comes in. The set of perceptual attributes form the perceptual space. Each musical recording is assigned a position in this perceptual space. The distance function that determines the distance between songs along with the perceptual space form a metric space (Mendelson, 1975).

The set of perceptual attributes can be broken into two groups: objective and subjective. The objective attributes include elements such as tempo and rhythmic style, orchestration, and musical style. The subjective attributes focus on elements that are more descriptive in nature, such as the weight of the music, is it heavy or light, the mood of the music, etc. The subjective attributes can be described as terms that non-experts might use to describe music.

After identifying the salient perceptual attributes, their relative weights were determined. By far the most important attribute identified by the musicologists is the musical style. The weights of the remaining attributes were iteratively hand tuned over the period of several months as the database at MongoMusic grew in size.

Once the perceptual attributes were identified the process of manually classifying a catalog of music was begun. Additional musicologists were brought as full-time employees to classify a catalog of music that eventually contained a few hundred-thousand songs. This process took about 30 man-years. Special attention was paid to the training of the musicologists and a rigorous quality assurance procedure was put in place.

While the classification efforts of the musicologists yield excellent results, the process does not scale well. The goal of classifying several million records is simply not feasible using the above described process alone. In addition, the process is extremely fragile. In order to add a new parameter, one must re-analyze the entire catalog. Clearly, an automated approach is needed.

3.3 Parameter Space

The human-classified results form an excellent corpus of data with which to train an automated system. First, however, one must determine what parameters need to be extracted from sound files and fed into a mapping system so that the mapping system can enable estimations of perceptual distance. Once the parameters are identified one can attempt the construction of a suitable mapping system. In practice, the two steps are intertwined since one cannot know, in general if the proper parameters have been extracted from the sound file until the mapping system has some results.

The purpose of the parameterization phase is to remove as much information from the raw audio data as possible without removing the "important" pieces of data, i.e., the data that allows a mapping from parameters to perceptual distance. This is necessary because current machine learning algorithms would be swamped by the sheer amount of data represented by the raw PCM data of audio files. The prospects of training a system under such a torrential downpour of data are not bright. The approach of parameterization also takes place, in an admittedly more sophisticated fashion, in the human hearing system, so the approach has some precedence.

The mapping of the parameter space to the perceptual space is carried out by the mapping system using traditional machine learning techniques. It is important to note that systems which map similarity based on parameterization alone do not perform well across a wide range of music. What these systems are not able to capture is the subtle interdependence between the parameters that the human hearing system uses to determine perceptual similarity. Because of this, it is the opinion of this researcher that a successful MIR system must include a model of perceptual similarity.

3.4 Results

The performance of the automated system is comparable to that of the human musical experts over most of the perceptual parameters. The human musical experts still have a slight edge but that gap is closing. Accuracy in this context refers to the percentage of classifications made by either a musical expert or the automated classification system that agree with a second musical expert. The human experts typically have an accuracy rating of about 92 to 95% depending on the parameter in question. The automated system has an accuracy range of about 88 to 94%.

Musical style, however, is not even addressed by the system. At this point humans must still be used to assign musical style. Early attempts at classifying musical style showed little promise.

3.5 Future directions

Automating the assignment of musical style is a major goal of future research at MSN Music. The task, however, is daunting. Recent results in musical classification using a small catalog of music (Soltau, 1998), while important contributions, illustrate how much more work needs to be done. Currently, there exist over one thousand musical style categories in the MMSE. In order for an automated system to replace humans, it would have to accurately recognize a significant subset of these styles. Accurate here mean close to 100% accuracy with graceful errors. In other words, if the style is wrong, it is not so bad if an “East Coast Rap” song is classified as “Southern Rap” song but if that song is mistakenly classified as “Baroque” than the error is quite painful.

4 The Musart Project

(William P. Birmingham, Roger B. Dannenberg, Ning Hu, Dominic Mazzonni, Colin Meek, William Rand and Gregory Wakefield, University of Michigan and Carnegie Mellon)

The University of Michigan and Carnegie Mellon are collaborating on Music Information Retrieval research. The work draws upon efforts from both universities to deal with music representation, analysis, and classification, with support from the National Science Foundation (Award #0085945). There are around a dozen faculty and students working together on a number of projects. Musart is an acronym for MUSic Analysis and Retrieval Technology.

We believe that Music Information Retrieval is interesting because it cuts across many music problems. One of our guiding principles is that music abstraction is necessary for effective and useful music search. Abstraction refers to qualities of music that reside beneath the “surface” level of melody and other directly accessible properties. We believe that search systems must understand and deal with deeper musical structure including style, genre, and themes. Searching based on abstract musical properties requires sophisticated techniques for analysis and representation.

These problems are not unique to music search. Composition systems, interactive music systems, and music understanding systems all deal with problems of music representation, analysis, and abstraction. Thus, some of the most fundamental problems in music search are shared by many other areas of computer music research.

4.1 Theme abstraction

A good example of abstraction is the theme extraction program, which is based on the observation that composers tend to repeat important musical themes. The program extracts all sub-sequences of notes up to a given length from a MIDI representation of a composition. The program then searches for common sub-sequences. Although the approach is simple in musical terms, the performance is quite good. To evaluate the system, output was compared to themes from Barlow's *A Dictionary of Musical Themes* (1983). Barlow and the program agree in 95.6% of test pieces.

4.2 Markov models and style

We are currently studying the use of Markov models to capture compositional style. As these models seem to be useful for melodic representation, we are also applying them to problems of melodic search.

States, called *concurrencies*, are defined as a collection of pitch classes and a duration. Scanning a score from beginning to end, each point in time corresponding to a note beginning or ending defines the start of a new concurrency. Zero-order and first-order Markov models are constructed from concurrencies and transitions from one concurrency to another.

Markov models are compared by computing their correlation. One can also compute the probability that a query is generated by the Markov model for a particular piece. It turns out that models constructed from large works such as piano concertos do impressively well at characterizing the “style” of different composers. Smaller works such as a simple melody have much less precise information, but these are still useful for music search.

4.3 Query synthesis

One way to assist in the formation of music queries is to synthesize music that is representative of the query. We have constructed a demonstration in which users can dial in various parameters to generate a variety of popular music rhythmic styles. The resulting set of dimensions along which we placed musical styles is interesting and indicates some of the features we might want to identify from recorded audio in a database. We also want to synthesize sung queries. For example, it might help to apply a female pop-singer’s voice to a sung or hummed query, and we are working on using research on voice analysis for this task. Speech analysis for searching lyrics and for time-aligning lyrics with audio is another task where we have made some progress.

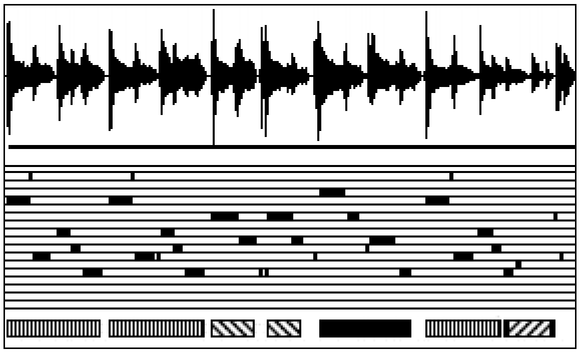

Fig. 1: Audio Analysis example

4.4 Audio Analysis

While much of our work has taken place in the symbolic domain of MIDI and music notation, we are also very interested in audio data. We have applied various machine learning techniques to classify audio and MIDI data according to style and genre. This work has produced good classifiers for small numbers of genres, but it is clear that we need more sophisticated features, especially for audio data.

Toward this goal, we have looked at the problem of machine listening to real examples. Figure 1 illustrates a jazz ballad ("Naima," composed and performed by by John Coltrane) in audio form at the top. In the middle of the figure, pitch analysis has extracted most of the notes. At the bottom of the figure, notes are grouped into recurring patterns. Thus, we not only have a rough transcription of the piece, but we have an analysis that shows the structure, e.g. AABA.

Another audio analysis effort has been to detect the chorus of a pop song by looking for repeating patterns of *chroma*. Chroma is essentially an amplitude spectrum folded into a pitch-class histogram. (Wakefield, 1999) This approach has worked well for finding choruses. A practical application is "audio thumbnailing," or choosing salient and memorable sections of music for use in browsing music search results.

4.5 Frame-based contour searching

One of the difficulties of dealing with audio is that music is difficult to segment into notes, so even a simple hummed query can be difficult to transcribe. We have developed a new technique for melodic comparison in which the melodic contour is compared rather than individual notes. The advantage of this method is that audio is not segmented. This means that there are no segmentation errors that could lead to an indication of "wrong notes." Unfortunately, contour comparison proceeds frame-by-frame using small time steps, which is more expensive even than note-by-note matching. Future work may look at more efficient implementations. Preliminary results indicate that this form of search is better than string-matching methods.

4.6 Scaling Issues

We are also concerned with the problems of scaling up to larger databases. This concern includes the problems of melodic search: simple abstract queries of relatively few notes will tend to match many database entries. Identifying themes and more robust melodic similarity measures will help, but ultimately, we need to search more dimensions, so style will become very important.

A second issue is efficiency in a large database. We clearly need sub-linear algorithms, that is, algorithms whose runtimes do not increase linearly with the size of the database. Some sort of indexing scheme may be possible, but we think that good search will require multiple levels of refinement, with fast but imprecise search used to narrow the search, combined with increasingly sophisticated (but increasingly expensive) search techniques to narrow the search results further. Searchable abstractions are a key to progress in this area.

Third, we hope to evaluate our results and techniques in terms of precision and recall. Toward this goal, we have assembled a test database of music, and we are implementing a modular search system architecture to facilitate experimentation.

5 Just what problem are we solving?

(Jonathan Foote, FX Pal Alto, Fuji Xerox)

In the "Cranfield" model of information retrieval, users approach a corpus of "documents" with an "information need", which is expressed in a "query" typically composed of "keywords". This is appropriate and can work surprisingly well for text as shown in web search engines. It is not often obvious what these terms mean when considering music. Several music IR (MIR) systems take the approach of using humming or musical input as a query (Ghias et al. 1995, Bainbridge 1999). This is completely appropriate for many kinds of music, but not as useful for some other genres (rap and electronic dance music spring to mind). Even if a "relevant" document is found, there is no guarantee that it satisfies the "information need". As an anecdotal example, there is a recording of "New York, New York" that was played on a collection of automobile horns (Chambers, 2001). Though the notes are correct, it can be imagined that this "document" would not be satisfactory as a search result for a user seeking a Sinatra performance. Undoubtedly the reader knows of similar examples that have the correct note sequence but the wrong "feel." Thus there is room for many other types of "queries" and other definitions of "relevance".

An alternative approach attempts to capture the “feel” of a musical recording with data-driven signal processing and machine learning techniques. One of the first music retrieval-by-similarity systems was developed by one of the panelists (Foote, 1997) while at the Institute of Systems Science in Singapore. In this system, audio is first parameterized into a spectral representation (mel-frequency cepstral coefficients). A learning algorithm then constructs a quantization tree that attempts to put samples from different training classes into different bins. A histogram is made for each audio sample by looking at the relative frequencies of samples in each quantization bin. If histograms are considered vectors, then simple Euclidean or cosine measures can be used to rank the corresponding audio files by similarity (Foote, 1997). David Pye at ATT Research has compared this approach with Gaussian distance measures on the same corpus (Pye, 2000). Gaussian models improve retrieval performance slightly but at a higher computational cost. In these experiments, "relevance" was assumed to be by artist, in other words all music by the same artist was considered similar. Although this has obvious disadvantages, it simplifies experimentation, as relevance can be easily determined from metadata.

As above, retrieval strategies are often predicated on the relevance classes, which may be highly subjective. One experimental strategy is to choose relevance classes that are not subject to debate, such as different performances of the same orchestral work. This approach was used in another retrieval system, dubbed ARTHUR (after Arthur P. Lintgen, an audiophile who can determine the music on LP recordings by examining the grooves). ARTHUR retrieves orchestral music by characterizing the variation of soft and louder passages. The long-term structure is determined from envelope of audio energy versus time in one or more frequency bands. Similarity between energy profiles is calculated using dynamic programming. Given a query audio document, other documents in a collection are ranked by similarity of their energy profiles. Experiments were presented for a modest corpus that demonstrated excellent results in retrieving different performances of the same orchestral work, given an example performance or short excerpt as a query (Foote, 2000b). However it is not clear that this is solving a particularly pressing information need, or one that couldn't be satisfied by even the most rudimentary metadata, such as the name of the orchestral work.

Recent research at FX Palo Alto Laboratory is based on self-similarity analysis. This is a relatively novel approach that characterizes music and audio by a measure of its self-similarity over time. Rather than explicitly determining particular features such as pitch, timbre, or energy, the location and degree of repetition is analyzed. Because of its independence from particular acoustic attributes, this has proved to be robust across a wide range of genres: in essence, the audio is used to model itself. In addition, it provides some interesting visualizations of structure and rhythm (for examples, see (Foote, 2001b) in this volume).

Locating times where audio ceases to be highly self-similar has proved to be a good way of segmenting complex audio (Foote and Uchihashi, 2000). This approach is currently being used to automatically generate music videos by aligning video shots with musical events.

It is possible to generate a measure of self-similarity versus time lag that we call the "beat spectrum." Analyzing the beat spectrum gives an excellent way of measuring tempo (Foote and Cooper, 2001). Additionally, the beat spectrum can characterize different rhythms or time signatures even at the same tempo. For example, the following figure shows a "beat spectrogram" with time on the X axis and repetition lag on the Y axis. Bright horizontal bars show periodicities at those lag times. In the figure a transition from 4/4 to a 7/4 time signature is visible as an increase of repetition intensity at the lag time labeled "7".

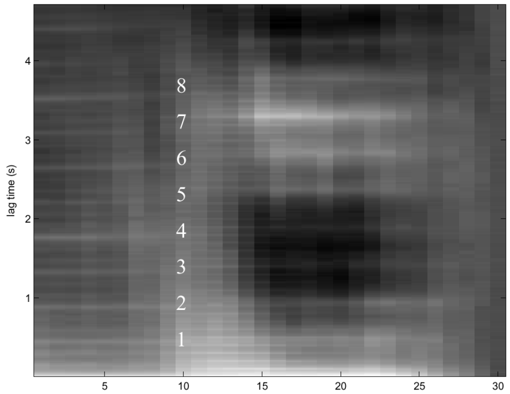

Fig. 2: Beat spectrogram showing transition from 4/4 to 7/4 time in an excerpt of Pink Floyd's *Money*

A retrieval system based on beat-spectral similarity is currently under development at FXPAL; early results indicate that the beat spectrum captures rhythmic "feel" much better than purely tempo-based approaches (Scheirer 1998, Cliff, 2000).

6 MARSYAS

(George Tzanetakis and Perry Cook, Princeton University)

MARSYAS (**M**usical **A**nalysis and **R**etrieval **SY**stems for **A**udio **S**ignals) is a software framework, written in C++, for rapid prototyping of computer audition research. In addition, a graphical user interface for browsing and editing large collections of audio files, written in JAVA, is provided. The primary motivation behind the development of MARSYAS has been research in content-based audio information retrieval. As a consequence, a significant number of AIR related tools have been implemented and integrated into this framework. A frequent problem with current MIR implementations is that typically a single analysis technique is developed and evaluated. Since the field of MIR is still in its infancy it is very important to use as much information as possible and allow the user to interact with the system at all stages of retrieval and browsing. This is achieved in MARSYAS by interacting with all the developed tools and algorithms under a common graphical user interface and allowing the exchange of information between different analysis techniques. The main design goal has been to implement a system for researching and developing new MIR algorithms and techniques, rather than focusing on a single approach.

6.1 Feature extraction and classification

The core of MARSYAS is short time audio feature extraction. The available features families are based on the following time-frequency analysis techniques: Short Time Fourier Transform (STFT), Linear Prediction Coefficients (LPC), Mel Frequency Cepstral Coefficients (MFCC), Discrete Wavelet Transform (DWT) and the MPEG analysis filterbank (used for compressing mp3 files). Complicated features can be constructed by creating arbitrary graphs of signal processing blocks. This flexible architecture facilitates the addition of new features and experimentation with the currently available features. The supported features represent timbral, rhythmic and harmonic aspects of the analyzed sounds without attempting to perform polyphonic transcription. Multiple feature automatic segmentation and classification and similarity retrieval of audio signals are supported.

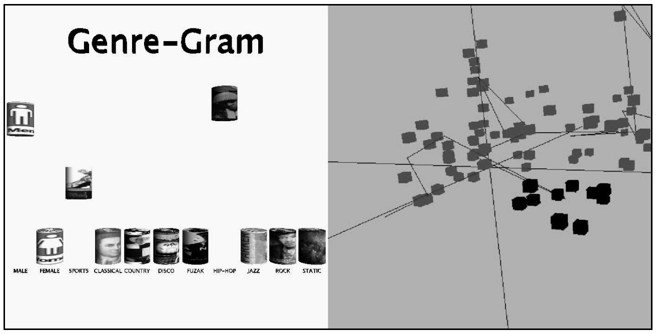

Fig. 3 Genregram and Timbrespace

The following classification schemes have been evaluated: Music/Speech, Male/Female/Sports announcing, 7 musical genres (Classical, Country, Disco, Easy Listening, Hip Hop, Jazz, Rock), Instruments and Sound Effects. In addition it is easy to create other classification schemes from new audio collections. The currently supported classifiers are Gaussian Mixture Model (GMM), Gaussian, K Nearest Neighbor (KNN) and K-Means clustering. Content-based similarity retrieval and segmentation-based thumbnailing are also supported.

MARSYAS has been designed to be flexible and extensible. New features, classifiers, and analysis techniques can be added to the system with minimal effort. In addition, utilities for automatic and user evaluation are provided. User studies in segmentation, thumbnailing and similarity retrieval have been performed and more are planned for the future.

6.2 Graphical User Interfaces

Several different browsing and visualization 2D and 3D displays are supported. All these interfaces are informed by the results of the feature based analysis. Some examples of novel user interfaces developed using MARSYAS are:

1. An augmented waveform editor that in addition to the standard functionality (mouse selection, waveform and spectogram display, zooming) is enhanced with automatic segmentation and classification. The editor can be used for "intelligent" browsing and annotation. For example the user can jump to the first instance of a female voice in a file or can automatically segment a jazz piece and then locate the saxophone solo.
2. Timbregram : a static visualization of an audio file that reveals timbral similarity and periodicity using color. It consists of a series of vertical color stripes where each stripe corresponds to a feature vector. Time is mapped from left to right. Principal Component Analysis (PCA) is used to map the feature vectors to color.
3. Timbrespace (Figure 3): a 3D browsing space for working with large audio collections based on PCA of the feature space. Each file is represented as a single point in a 3D space. Zooming, rotating, scaling, clustering and classification can be used to interact with the data.
4. GenreGram: a dynamic real-time display of the results of automatic genre classification. Different classification decisions and their relative strengths are combined visually, revealing correlations and classification patterns. Since the boundaries between musical genres are fuzzy, a display like this is more informative than a single all or nothing classification decision. For example, most of the time a rap song will trigger Male Speech, Sports Announcing and HipHop.

6.3 Architecture – Implementation

The software follows a client server architecture. All the computation-intensive signal processing and statistical pattern recognition algorithms required for audio analysis are performed using a server written in C++. The code is optimized resulting in real time feature calculation, analysis and graphics updates. For further numerical processing utilities for interfacing MARSYAS with numerical packages like MATLAB or OCTAVE are provided. The use of standard C++ and JAVA makes the code easily portable to different operating systems. It is available as free software under the GNU public license. It can be obtained from:

http://www.cs.princeton.edu/~marsyas.html

This work was funded under NSF grant 9984087, by the state of New Jersey Commission on Science and Technology grant 01-2042-007-22 and from gifts from Intel and Arial Foundation.

7 Summary

Music information retrieval is becoming increasingly important as digital audio and music are becoming a major source of internet use. In this paper, the main topics and directions of current research in MIR were identified. Four specific projects from industry and academia were described. These projects show the increasing interest in the evolving field of MIR and the diversity of different approaches to the problem. A panel discussion about the current state, challenges and future directions of MIR by the authors of this paper will be conducted during the conference and we hope that this paper will serve as a foundation for discussion during this panel.

References

Algoniemy, M., and Tewfik, A. 2000. "Personalized Music Distribution", *Proceedings of the International Conference on Audio, Speech and Signal Processing ICASSP 2000.*

Barlow, H. 1983 "A dictionary of musical themes". Crown Publishers.

Bainbridge, D., Nveill-Manning, C., Witten, L, Smith, L., and McNab, R. (1999). "Towards a digital library of popular music", *Proceedings of ACM Digital Libraries (DL) Conference.* 161-169

Cantametrix. *http://www.cantametrix.com*

Chambers, Wendy Mae. 2001. "The Car Horn Organ sound samples" at http://www.wendymae.com/carhornorgan.html

Chen, A., et al. 2000. "Query by music segments: an efficient approach for song retrieval". *Proceedings International Conference on Multimedia and Expo*.

Cliff, David. 2000. "Hang the DJ: Automatic Sequencing and Seamless Mixing of Dance Music Tracks", *HP Technical Report HPL-2000-104*, Hewlett-Packard Laboratories. http://hpl.hp.com/techreports/2000/HPL-2000-104.html

Etanttrum. *http://www.etantrum.com*

Foote, J. 1997. "Content-based retrieval of music and audio." In *Multimedia Storage and Archiving Systems II*, Proc. SPIE, Vol 3229.

Foote, J. 1999. "An overview of audio information retrieval", *ACM Multimedia Systems 1999, vol.7*.

Foote, J. 2000a. "Automatic audio segmentation using a measure of audio novelty". *Proceedings of the International Conference on Multimedia and Expo*.

Foote, J. 2000b. "ARTHUR: retrieving orchestra music by long term structure". *Proceedings of International Symposium on Music Information Retrieval*. http://ciir.cs.umass.edu/music2000/papers/foote_paper.pdf

Foote, J. and Uchihashi, S. 2001. "The beat spectrum: a new approach to rhythm analysis". *Proceedings of International Conference in Multimedia and Expo*. (in press) http://www.fxpal.com/people/foote/papers/ICME2001.htm

Foote, J. and Cooper, M. 2001. "Visualizing musical structure and rhythm via self-similarity". *Proceedings of the International Computer Music Conference*. International Computer Music Association. (this volume)

Ghias, A., Logan, J., Chamberlin, D., and Smith, B.C. 1995. "Query by humming-musical information retrieval in an audio database." *Proceedings of the ACM Multimedia.*

Guuyon, F., Pachet, F., and Delerue, O. 2000. "On the use of zero-crossing rate for an application of classification of percussive sounds", *Proceedings of COST G6 Workshop on Digital Audio Effects, DAFX 2000.*

Hauptmann, A. and Witbrook, M. 1997. "Informedia: News on demand multimedia information acquisition and retrieval", *Intelligent Multimedia Information Retrieval*, MIT Press, 1997.

Hewlett, W., and Eleanor Selfridge-Field, E., eds 1999. *Melodic similarity: concepts, procedures and applications* (Computing in Musicology, 11) MIT Press.

Logan, B. 2000. "Music summarization using key phrases", *Proceedings of the International Conference on Audio, Speech and Signal Processing ICASSP 2000.*

Mendelson, B. 1975. "Introduction to topology", Dover Publications Inc. 1975.

Martin, K. 1999. "Sound-source recognition: A theory and computational model", *PhD thesis, MIT http://www.sound.media.mit.edu/~kdm*

MSN. http://music.msn.com

MongoMusic. *http://www.mongomusic.com*

Moodlogic. *http://www.moodlogic.com*

MPEG 7. *http://www.darmstadt.gmd.de/mobile/MPEG7*

Mubu. *http://mubu.com*

Pye,D. 2000. "Content-based methods for the management of digital music" *Proceedings of the International Conference on Audio, Speech and Signal Processing ICASSP 2000.*

Relatable. http://www.relatable.com

Scheirer, E., Slaney, M. 1997. "Construction and evaluation of a robust multifeature Speech/Music discriminator", *Proceedings of the International Conference on Audio, Speech and Signal Processing ICASSP 1997.*

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

Soltau, H., Schultz, T., Westphal, M., Waibel, A., "Recognition of Music Types", Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing.

Sundaram, H., Chang, F. 2000. "Audio scene segmentation using multiple features, models and time scales", *Proceedings of the International Conference on Audio, Speech and Signal Processing ICASSP 2000.*

Tzanetakis, G., and Cook, P. 2000 "Audio information retrieval (AIR) tools", *International Symposium on Music Information Retrieval*, 2000.

Tuneprint. http://www.tuneprint.com

Uitdenbogerd, A. and Zobel, J. 1999. "Manipulation of music for melody matching." *Proceedings ACM Multimedia*

Wakefield, G.H. 1999. "Mathematical representations of joint time-chroma distributions". In Intl.Symp. on Opt.Sci., Eng., and Instr., SPIE 1999. Denver.

Weare, C. and Tanner, T. 2001 "In search of a mapping from parameter space to perceptual space", Proceedings of the 2001 AES International Conference on Audio and Information Appliances, 2001.

Wold, E., Blum, T., Keislar, D. and Wheaton, J. 1996 "Content-based classification, search and retrieval of audio", *IEEE Multimedia*, vol.3(2), 1996.

Wold, E., Blum, T., Keislar, D., and Wheaton, J. 1999. "Classification, Search and Retrieval of Audio", *Handbook of Multimedia Computing*, ed. B. Furht, CRC Press.