“Content-Based Classification and Retrieval of Music: Overview and Research Trends”

Rui Pedro Paiva

1. Introduction

Music has always been present in the lives of human beings, both individually and socially, through the cultural, professional, leisure or religious aspects of life. Music is a way by which composers express their innermost feelings. As a means of celebration, music has always accompanied man’s festive moments. In the expression of human religiosity, music has always been regarded as a form of prayer. In sport activities, music can be used to keep an athlete motivated. Recently, music is being approached as an aid for the therapy of nervous disturbances or even for the improvement of student performance. Music is associated to the most marking moments of our life, brings us memories, arouses emotions, it is part of our individual and social imaginary.

For the reasons appointed, music plays an important role in the balance and development of world economy. In fact, the music industry runs, only in USA, an amount of money in the order of several billion dollars per year [Pampalk, 2001].

As a result of recent technological innovations, there has been a tremendous growth in the Electronic Music Distribution (EMD) industry. Factors like the widespread access to the Internet, bandwidth increasing in domestic accesses or the generalized use of compact audio formats with CD or near CD quality, such as mp3, have given a great contribution to that boom. Presently, it is expected that the number of digital music archives, as well as their dimension, grow significantly in the near future, both in terms of music database size and in number of genres covered. This situation poses new and exciting challenges.

1.1. Motivation

In spite of the growth of digital music libraries, any large music database, or, generically speaking, any multimedia database, is only really useful if users can find what they are seeking in an efficient manner. Furthermore, it is also important that the organization of such a database can be performed as objectively and efficiently as possible.

Presently, whether it is the case of a digital music library, the Internet or any music database, search and retrieval is carried out mostly in a textual manner, based on categories such as author, title or genre. This approach leads to a certain number of difficulties, both for service providers and general users, namely in what concerns music labeling or database search in a transparent and intuitive way, respectively.

Therefore, some efforts are now being conducted in order to make it possible to search music databases by content similarity [Logan and Salomon, 2001; Yang, 2001; Welsh *et al*, 1999]. In those systems, the goal is to allow the creation of musical queries through examples given by the user, e.g., by humming the melody to search for, or by specifying a song in some way similar to what is being looked for, in terms of certain searching criteria (theme, rhythm, genre, instrumentation, etc.). This approach can be very useful when users do not know or are not especially interested in the melody. Namely, an aerobic instructor can look for songs with a certain tempo or a truck driver can look for a song that keeps him alert [Huron, 2000], regardless of the melody or genre. This can be a daunting task if we think of the thousands or even millions of songs, organized sometimes in tens or hundreds of different and often non-uniform genres that many music libraries contain.

Another objective is to simplify the task of organizing musical databases via automatic classification systems, where similar songs may be found close to each other [Tzanetakis *et al*, 2001; Pampalk, 2001; Golub, 2000]. Such systems should overcome the limitations resulting from manual song labeling, which may be a highly time-consuming and subjective task.

1.2. Application Areas

Digital musical content analysis has a broad range of applicability, in spite of addressing several difficult and still open problems.

Regarding EMD, music web crawlers, which “traverse the web and index music-related files” [Huron, 2000], are applications with an enormous potential. Also, automatic classification systems should be extremely useful for the labeling and updating of huge music databases, as well as tool for content-based retrieval, as stated before. This also applies for multimedia databases and operating systems.

Besides the possibilities for EMD, systems for education and training can also gain from the results attained. For example, systems for automatic music transcription [Bello *et al*, 2000; Martin, 1996; Ellis, 1996] may simplify tasks such as manual transcription, music composition, music analysis or evaluation of musical performance. Also, professional composers may find useful tools that help detecting plagiarism.

People at Indiana University have been working on a project for creating a digital music library [Dunn, 2000]. The referred project, VARIATIONS, addresses both technical issues, such as content-based information retrieval, and educational ones, such as learning activities for music instruction and evaluation of learning impact, supported by the library.

Additionally, audio editors or audio browsers could become more intelligent with tools for automatic indexing of music/audio files [Wold *et al*, 1996].

Also, in cinema or advertising industries, it is often necessary to search for songs that induce a certain mood to the intended audience [Huron, 2000].

Video indexing and searching can gain a lot from music content analysis and, more generally, from audio content analysis. Instead of looking at image frames, audio frames can be analyzed. This is a much more efficient way to detect scene transitions, fundamental to video indexing. Furthermore, it can be useful to perform video segmentation. For example, romantic scenes (love inspiring song) or violence (shots, screams) can be detected by looking only at audio information [Pfeiffer *et al*, 1996].

1.3. Research & Development Projects

The field of music information retrieval has received recently a great amount of interest both from academia and industry. In fact, many research laboratories all over the world have set up research agendas in this area, e.g., MIT Media Lab, “Institut the

Recherche et Coordination Acoustique/Music” (IRCAM) in Paris, Center for Intelligent Information Retrieval (CIIR) at University of Massachusetts, Center for Computer Research in Music and Acoustics (CCRMA) at Stanford University, King’s College London (KCL) and many others. There are also some large-scale joint projects going on, like for example, OMRAS (Online Music Recognition and Searching), started in 1999, between CIIR and KCL. This project’s main goal “is to build a working prototype of a system, OMRAS, for content-based search of online music databases via an intuitive interface that uses music in a visual or aural form familiar to the user […] for both search-query construction and to display results” [OMRAS, 1999]. It encompasses many of the greater tasks necessary in any music retrieval system, such as usability, music representation, searching and search-surface reduction, type conversion, query construction and audio recognition. Furthermore, this project was the main catalyst for the ISMIR (International Symposium on Music Information Retrieval) conference, organized yearly since 2000, which was the first conference to congregate many of the most active researchers in the field.

In industry, there is an increasing number of commercial products like, for instance, Intelliscore from Innovative Music Systems, which aims at converting polyphonic music recordings to midi format [PRWeb, 2000]. Despite its usefulness for simple polyphonic music, results are still limited for “real-world” music. Also, Philips has developed a music recognition technology where users can receive information about songs they hear [Afterdawn, 2002]. Basically, users send their queries from a cell phone, putting it in front of the speaker for three seconds. Then, they receive an SMS with the song’s name, artist, album, etc. Philips plans to start licensing this technology by the end of the year.

Additionally, there are some cooperation projects between academia and industry like, for instance, CUIDADO, which is a European project started in 2000 that “tackles the problems of information overload and the inability to quick browse audio or search for similarities among songs” [CUIDADO, 2000]. The main goal of that project is to develop technologies for content-based music analysis using MPEG-7. Partners are IRCAM (Paris, France), Ben-Gurion University (Israel), Oracle (Spain), Cream@ware (Germany), Sony Computer Science Laboratory (Paris, France), Pompeu Fabra University (Barcelona, Spain) and ArtsPages (Norway).

2. General System Architecture

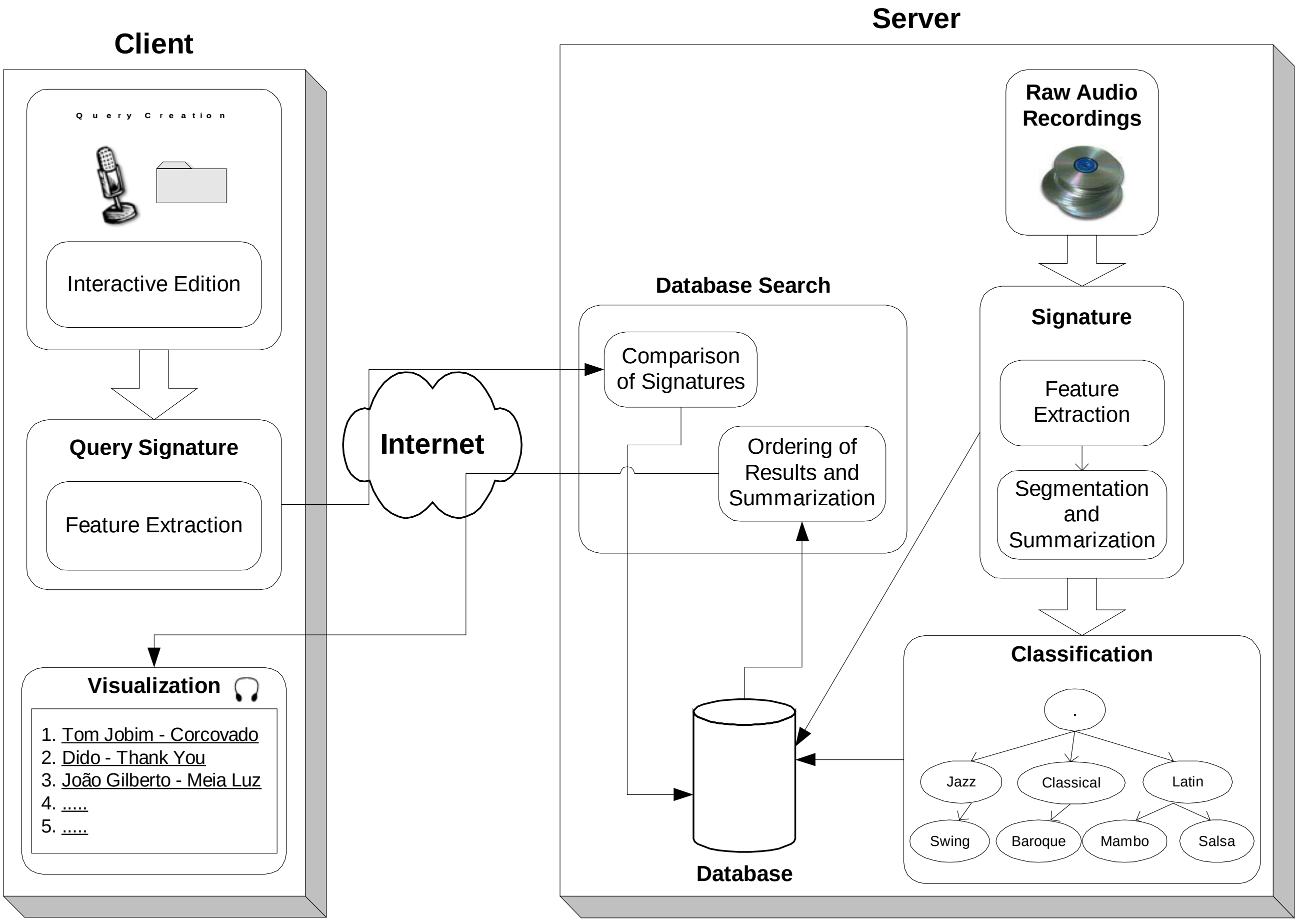

Figure 1. General System Architecture.

A general system for music retrieval is depicted in Figure 1, adapted from [Chai, 2001]. The architecture presented is a typical client-server one, suited for web-based applications, as it is the usual situation in music retrieval. However, such a general architecture can be easily adapted, for example, for standalone applications or music retrieval in local file systems or distributed databases. Below, an overview of the main tasks performed by both client and server will be described.

2.1. The Client-Side

The client-side of a music retrieval system is responsible for supporting the creation of musical queries to be sent to the server, as well as for the presentation of the results obtained.

When users are searching for songs, either they know what they want, e.g., song title, artist or genre, where a traditional text-based search is enough, or they want to find songs based on content similarity. They could also search for a song based on lyrics or other criteria.

In the case of searching by similarity, it must be possible for users to build their queries in an intuitive and interactive way [OMRAS, 1999]. One of the most intuitive ways to search music databases is by humming or singing some melody at the microphone, i.e., query-by-singing (QBS) or query-by-humming (QBH) [Chai, 2001; Bainbridge *et al*, 1999]. Another way of specifying the query is by selecting a song file, typically an mp3 file, similar to the desired song or songs in some way, e.g., rhythm, melody or genre, i.e., query-by-example (QBE) [Logan and Salomon, 2001; Yang, 2001; Welsh *et al*, 1999]. It is easy to notice that QBH and QBS can be regarded as particular cases of QBE. Either way, the client is responsible for building a query signature, which is sent to the server and then compared to the signatures present in the database, in the server-side.

Regarding QBH / QBS, the main task of the client application is to extract the sequence of notes hummed or its melodic contour, i.e., the sequence of note transitions (up, down, equal) [Chai, 2001], especially suited for transposition-invariant searches [Lemström and Perttu, 2000]. In fact, many people can sing the same melody in different keys, e.g., "Happy Birthday", and the search engine should respond adequately (this point will be addressed later on, when the server-side is described). Also, rhythmic information should be captured [Chai, 2001], since two songs can be equal in terms of note sequence but differ in beat and tempo. In order to make the query as precise and robust as possible, its construction should also be interactive [OMRAS, 1999]. Namely, the notes extracted should be shown in a score, so that the user could evaluate and correct the sequence extracted by the application. The same applies if a melodic contour is extracted. Additionally, it should be possible to listen both to the songs hummed and to the notes extracted. Needless to say that any query could be reformulated if users are not satisfied with their own performance or the results obtained.

As for QBE, similarity criteria should defined by individual users. This can be accomplished by creating an interface where users are able to select relevant criteria for their searches, as well as weights given to those criteria [Wold *et al*, 1996]. Queries could encompass higher-level criteria such as melody [Goto, 2001; Klapuri *et al*, 2000], rhythm [Tzanetakis *et al*, 2001; Scheirer, 1998], timbre [Tzanetakis *et al*, 2001], loudness [Pampalk, 2001; Golub, 2000] or mood [Huron, 2000], as well as lower-level physical criteria such as energy, harmonicity, or bandwidth [Wold *et al*, 1996]. The signature of the query can also be made up of statistical data such as mean, variance, maximum or minimum, as well as histograms of volume, frequency and so forth.

Still regarding QBE, the system should also be extensible with user-defined concepts [Wold *et al*, 1996]. Methodologies for empirical learning, e.g., neural networks, could be applied to induce those concepts by means of training examples.

Finally, the signature of the query is sent to the server, which returns the search results ranked by similarity. For example, if a user had sent a query regarding some Bossa-Nova song, he/she could have obtained a list like the one in Figure 1 (visualization box). The server should also return a short summary for every song, which is useful for song recognition and validation of results [Huron, 2000].

2.2. The Server-Side

As for the server-side, the process starts with a set of raw music recordings, typically in mp3 format. Then, the server's main task is to obtain a signature for each of the songs stored, as a basis for comparison with the queries sent by the client.

The creation of signatures results from the necessity to obtain a meaningful representation from raw audio data. In fact, raw audio cannot be used directly for content analysis. For instance, it is not possible to look at a sequence of sample values and say, directly, what notes are present. Thus, it is essential to transform the crude information present in raw audio data into a meaningful representation suited for music content analysis.

The referred representation must also allow for data reduction by eliminating any redundant information present in the song [Huron, 2000]. Namely, most songs have a chorus, which is repeated several times throughout the song. Thus, it is important to segment the song in its most relevant components, e.g., introduction or chorus, and to eliminate all repeated segments. Then, a signature is obtained for each of the final non-redundant segments. However, in most of the cases found in literature, the authors simply extract a signature for the whole song or segments at regular intervals, without taking in consideration the aspects referred above. For instance, in [Pampalk, 2001], Pampalk extracts segments by dividing every piece of music into six-second's sequences and analyzing only every third segment. This is very useful for capturing different styles in music classification problems but does not work well for music summarization.

After segmentation, the signatures are then obtained based on features extracted from each of the segments. Features can be temporal sequences of fundamental frequency, energy, zero-crossing rate and so on, as stated before. Additionally, statistical information can be obtained from those sequences. The process of feature extraction will be described later on.

Based on the extracted features, each segment is then classified, typically in a genre hierarchy. In this situation, tools like neural networks, k-nearest neighbors, clustering techniques or support vector machines can be used, where classes such as jazz, baroque or pop are learned via training examples made up of such features. It is important to notice that segment classification rather than song classification can overcome some classification ambiguities in songs that encompass different styles. In this way, one song can be classified into several genres. This procedure can avoid many ambiguities in song labeling.

The methodology for signature creation and song classification is applied to all songs in the server, and all signatures are stored in the database. This task is carried out offline while setting up a new system and when new songs are added to the database. Therefore, no real-time requirements are imposed, though temporal efficiency is desired.

The situation is different when the server receives a query from a client. In fact, the signature of the query must then be compared with all signatures stored in the database. Clearly, a "world-wide-wait" is to be avoided and the system must give quick responses to the client. In huge databases, this can be a difficult problem. That is one of the reasons why it is important that the signatures stored are as short as possible, while keeping their relevance. When melody is not a main issue, comparison can be simply performed by using a metric distance. In fact, in such queries, signatures consist typically of feature-vectors with information regarding statistical data extracted from temporal sequences [Wold *et al*, 1996]. However, in query-by-humming and query-by-singing, algorithms for string matching must be used in order to compare sequences [Lemström and Perttu, 2000], e.g., melodic contour or sequence of notes, possibly in different keys. Therefore, such algorithms must be highly efficient and optimized, so as to minimize server response time in transposition-invariant searches [Lemström and Perttu, 2000]. Additionally, the system must be robust and flexible regarding query imprecision, e.g., a few notes out of tune, non-uniform tempo and beat. Furthermore, other procedures for search optimization are very important, e.g., database indexing.

Finally, the results are sent to the client, ranked by similarity, where the most similar song is the one containing the most similar segment. Additionally, song summaries are also sent, which are constructed using the segments obtained previously. As stated before, summaries are important for users, so that they can easily hear what the server returned and validate those results. One common and easy way of delivering summaries consists of getting only the incipit, i.e., “the initial few seconds” of the song [Huron, 2000]. This procedure is the same as in common web search engines, where the initial words from every page are returned. However, it is a rather limited approach since many songs have introductions that are very different from the chorus [Huron, 2000]. Furthermore, many songs have very distinct passages, all of them relevant, which should be present in the summary. Thus, summaries based on song segmentation and redundancy elimination are more effective. In conclusion, song summaries are relevant both for server and server sides: as a way to optimize searches, in the server, and as a way to evaluate results, in the client.

As mentioned previously, extracting adequate query signatures is a key issue for music classification and retrieval systems. In order to accomplish that goal, good features must be selected and extracted from audio data. This crucial aspect will now be addressed.

3. Feature Extraction

Content-based music analysis, or more generally audio analysis, relies heavily on feature extraction. In fact, raw audio gives no intelligible information, and so it is necessary to extract relevant information from it. Many features have been suggested recently and many other have been inherited from voice recognition research.

Before describing those features, it is important to know what characterizes a good feature. Intuitively, a good feature should carry meaningful information [Pampalk, 2001], such as melody, rhythm or timbre. Ideally the process of feature extraction should mimic the human brain. This has led to an attempt to apply psychoacoustics findings to the problem of music content analysis [Bregman, 1990; Slaney and Lyon, 1990]. However, many aspects of the behavior of the brain are still poorly understood. That is why some physical features, i.e., features extracted directly from data without perceptual concerns, can be useful, despite being less intuitive. They also have the advantage of being less costly in terms of computing time.

Below, some of the most relevant features for music content-analysis are described.

3.1. Melody-Related Features

When we think of music, melody is probably the first thing that comes to our heads. Therefore, it is intuitive to try to derive features that can somehow capture melody content from music signals.

Fundamental frequency

When we hear a song, it is amazing how our brains can distinguish instruments present and the melodic line followed by each of them (or, at least, part of it). In fact, in an orchestra we can try to follow the violins, flutes, trumpets and so on. Also, we can easily follow the global melodic line, instead of paying attention to particular instruments. However, no computer system can do anything similar yet.

The main issues here are, therefore, i) to build computer systems that can extract the dominant frequency or fundamental frequency at any instant of time and ii) to isolate each audio stream present in a music signal, e.g., vocals, guitar, violin, a problem known as source separation. More generally, the goal is to detect pitch at any instance and for every stream.

Often, pitch and fundamental frequency are used interchangeably, however, they are different concepts. Fundamental frequency is a physical variable that represent the base frequency present in harmonic sounds, i.e., sounds where all main frequencies present are multiple of a base frequency. Examples of harmonic sounds are the ones produced by most musical instruments, e.g., string or wind musical instruments. Inharmonic sounds often come from metallic objects, where the frequencies present are not multiple of a base frequency. Therefore, fundamental frequency is a physical variable. On the other hand, pitch is a perceptual variable that determines our individual perception of frequency. Pitch entails aspects like our brains' response to frequency and intensity. For now, only fundamental frequency will be discussed. Pitch detection is discussed, e.g., in [Slaney and Lyon, 1990].

There are presently hundreds of algorithms for fundamental frequency detection or estimation, many of them developed in the context of voice recognition research. However, most of them are only suited for monophonic music analysis, i.e., music with only one audio stream, such as folk music, "shower singing", humming, etc. As for polyphonic music, i.e., music with several audio streams, such as pop music or orchestras, algorithms for dominant frequency detection or estimation are still in their infancy.

In a monophonic audio signal, the fundamental frequency is the smallest harmonic frequency present. To illustrate this concept, imagine a signal described by Eq. (1). That signal has a fundamental frequency at 200 Hz and third and fifth harmonics at 600 and 800 Hz, respectively.

$$x(t) = A\sin(\omega_0 t) + A\sin(3\omega_0 t) + A\sin(5\omega_0 t), \quad A = 1;\ f_0 = 200Hz;\ \omega_0 = 2\pi f_0 \qquad (1)$$

One of the most used tools for frequency analysis of discrete signals is the Discrete Fourier Transform (DFT) [Rabiner and Schafer, 1978]. Through the DFT, signals are transformed from the time-domain to the frequency-domain, where their frequency spectrum, i.e., the range of frequencies covered, is showed up. The DFT for the signal in Eq. (1) is depicted in Figure 2.

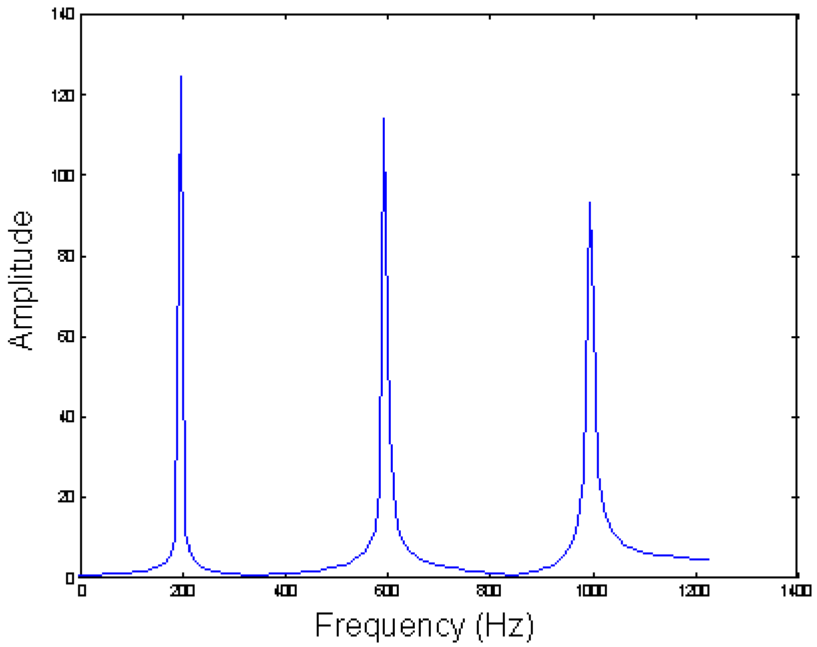

Figure 2. DFT of a simple signal.

In the previous figure, the three frequencies present in the signal are clearly shown. However, what happens when we have a signal whose frequency content varies through time? The DFT is only suited for stationary signals, i.e., signals that always have the same frequency content. This calls for a windowed version of the DFT: the Short-Time Frequency Transform (STFT) [Rabiner and Schafer, 1978]. Here, the main idea is to divide the signal into a set of time frames and calculate the DFT for each of those frames. Typically, frame length is around 20 ms, so that stationarity can be assumed.

Then, a simple algorithm for fundamental frequency detection would consist of performing STFT analysis for the signal and determining the fundamental frequency for each frame. Filter banks can be used to extend this approach by measuring signal energy in each frequency band. This technique tries to mimic the behavior of the inner ear, which acts also as a bank of filters. Therefore, this strategy is the most valuable one in terms of human audio perception.

Another possibility is to perform cepstral analysis [Rabiner and Schafer, 1978; Logan and Salomon, 2001]. In this technique, Fourier coefficients are first determined, then their logarithms are calculated and finally the inverse Fourier transform is applied to them. The result is a large peak at the frequency of the original signal.

One other strategy consists of computing the autocorrelation function [Rabiner and Schafer, 1978]. There, if a signal is periodic (or pseudo-periodic, i.e., “almost” periodic), its autocorrelation function will also be a periodic signal with the same period as the original one. Then, the period found will be equal to the distance between peaks. Alternatively, the average-magnitude difference function can be used, where the distance between valleys must be determined [Rabiner and Schafer, 1978].

As for polyphonic music, many more difficulties are present. To illustrate such difficulties, let’s look at the spectra resulting from playing middle C on a flute, piano and trumpet, in Figure 3 (adapted from [Davis, 2002]).

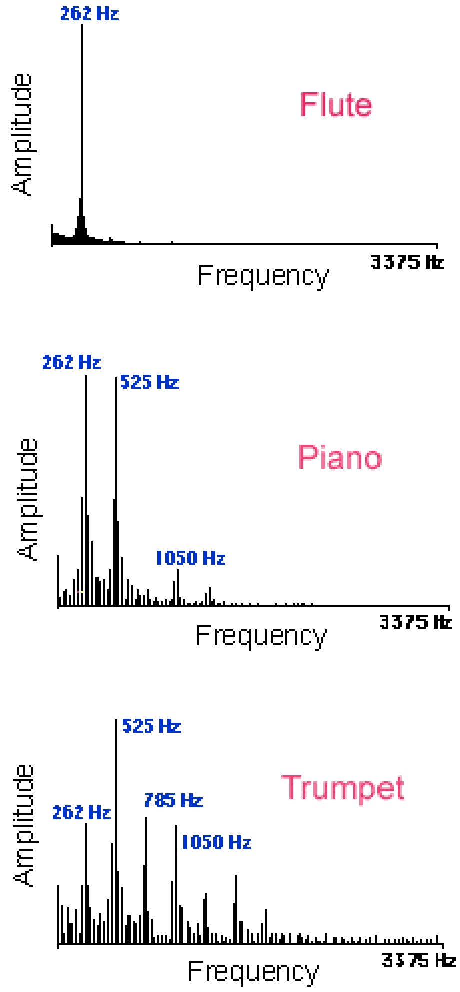

Figure 3. Spectra of middle C played on a flute, piano and trumpet [Davis, 2002].

In the previous figure, we can see that the same note played in different instruments originates very different spectra. In fact, flute is almost a pure tone (only a peak at the fundamental frequency), piano has clear peaks at the second and fourth harmonics and trumpet is the richer of the three instruments in terms of harmonic content. Also, there are overtones present, i.e., spectral components which are not multiple of the fundamental frequency (these overtones contribute to the particular timbre of each instrument). Now, if we imagine a song where those instruments are present, it is easy to conclude that the signal spectrum will be very complex, because of the interaction between fundamental frequencies, harmonics and overtones from each of them. So, extracting each audio streams or the dominant frequency is not an easy task any longer.

Some efforts are being conducted in order to attack the problems of source separation and dominant frequency detection/estimation.

Source separation is a major concern for polyphonic music analysis and automatic music transcription systems, and has no general solution yet. The human brain processes auditory information in a process called "auditory scene analysis" [Bregman, 1990]. As an attempt to replicate human behavior, some work has been carried out so as to develop computational auditory scene analysis systems. The results obtained are not yet very accurate and are only acceptable for simpler or well-constrained problems. Namely, Ellis [Ellis, 1996] tries to analyze a sound signal by means of competitive theories, where each of them proposes a combination of sounds that might produce the sound obtained. Sound source models are used as a basis for the proposed method. Bello *et al* [Bello *et al*, 2000] and Martin [Martin, 1996] have used computational blackboard systems for simple automatic music transcription. The blackboard system is composed of a global database, where hypotheses are proposed and developed, a scheduler, which determines how hypotheses are developed, and knowledge sources, corresponding to experts. Scheirer [Scheirer, 2000] proposes a model based on perceptual issues, using dynamic clustering of comodulation data. In contrast to the other systems referred, this model is designed for analysis of complex music. Other models impose constraints in the number of instruments present or the harmonic interaction between them, as it is referred in [Goto, 2001].

Dominant frequency detection/estimation, possibly a less difficult problem, consists of detecting only the main melodic line in a song. For instance, when we hear a pop song, we have vocals, guitar, bass, percussion and so forth. Yet, in spite of all that information, our brains still can retain the main melodic line. Klapuri *et al* [Klapuri *et al*, 2000] proposed a method for predominant pitch estimation where the musical signal is analyzed at separate frequency bands. Namely, 18 logarithmic distributed bands from 50 Hz to 6 kHz are used. Then at each band, a fundamental frequency likelihood vector is calculated. Finally, the results from each band are combined to yield global pitch likelihoods. They report results that outperform the average of ten trained musicians. Goto [Goto, 2001] uses a probabilistic model for the detection of melody and bass lines. The signal is first band-pass filtered and then a probability density function (PDF) is computed for each signal component. The PDFs are generated from a weighted-mixture model of tone models of all possible fundamental frequencies. The more dominant a model is in the PDF, the more likely the fundamental frequency belongs to that model. The author compared the dominant frequencies extracted with hand-labeled marked notes and reports an average rate of 88.4% for the melodic line.

Tonal Histograms and Transitions

The frequency content of a music signal can also be analyzed by means of tonal histograms and transitions. In [Welsh *et al*, 1999], histograms of frequency amplitudes across the notes of the Western music scale are proposed. This information can be used to detect dominant chords, as well as the key where the song is played in.

The same authors suggest tonal transitions for QBE. Basically, music signals can be seen as sequences of frequency transitions over time. Therefore, they propose an extractor that measures the number of tonal transitions in a given frequency range for ten seconds’ samples. There, five feature-vectors are obtained, each of them containing 306 values.

Zero-Crossing Rate

Measuring the zero-crossing rate (ZCR) in an audio signal consists of counting the number of times the sound wave crosses the zero axis [Rabiner and Schafer, 1978]. Thus, this feature gives frequency-related information. Namely, a high ZCR indicates a signal with high-frequency content, whereas a low ZCR suggests the opposite.

ZCR is a feature imported from voice recognition systems. In fact, it has been used there as a robust measure to detect unvoiced speech. Also, in general audio classification, it has been used for music/speech discrimination [Wold *et al*, 1996].

Regarding music content analysis, ZCR-based features, namely statistical ones, are present in feature-vectors for music signal classification.

3.2. Rhythm-Related Features

Melody is normally regarded as the most important feature in music retrieval tasks. However, as was referred previously, rhythm is also important for query matching. Furthermore, rhythmic information is essential for music genre classification. In fact, the same melody can be performed according to many different styles, as it is often the case of song versions. On the other hand, rhythm is a very important attribute for music genre classification [Pampalk, 2001].

Rhythm analysis encompasses aspects such as beat and tempo analysis, which are described below.

Beat and Tempo

Regarding beat and tempo analysis, the main idea is to find periodicities in the signal amplitude envelope.

In [Tzanetakis *et al*, 2001], a bank of filters is used to divide the signal into a number of bands, each of them representing an octave. Then, the amplitude envelope of the signal at each band is extracted, by means of full wave rectification, low pass filtering and downsampling. Next, the envelopes at each band are summed up. Finally, periodicities are detected by finding peaks in the envelope autocorrelation function.

In [Scheirer, 1998], a filterbank is also used, which divides the signal into six bands. For each band, the amplitude envelope is calculated, as well as its derivative. Next, each of the derivatives is passed to a set of comb filter resonators, where only one of them will phase-lock. The output of those resonator filterbanks is then summed across the frequency bands. Then, the energy output from each resonator channel is examined. The tempo of the signal is selected as the frequency of the resonator with the maximum energy output. Finally, beats are detected by looking back to the peak phase points in the phase-locked resonators.

Energy

Signal energy, also called volume, is also useful for rhythmic analysis. In fact, signal energy can be used as a basis for amplitude envelope extraction.

Energy is obtained by computing the sum of squares of signal amplitude values for each frame [Rabiner and Schafer, 1978]. This feature has been largely used in voice recognition systems for silence detection and voiced/unvoiced speech discrimination.

Loudness is the perceptual correspondent to volume [Pampalk, 2001]. As it was the case with fundamental frequency and pitch, volume and loudness are sometimes used interchangeably. As before, loudness is a perceptual variable that determines our individual perception of volume. Loudness involves issues like our brains' response to frequency and intensity and will not be discussed now.

3.3. Timbre-Related Features

Besides rhythm, the instruments present in a song are also important for genre classification. For instance, rhythm & blues is a variation of blues that gets its identity from the massive presence of brass instruments [Pachet and Cazaly, 2000].

Instrument detection must be grounded in timbre analysis. Physically speaking, timbre is a feature related to the sound wave. This idea is illustrated in Figure 4 (extracted from [Davis, 2002]), where sounds waves for middle C played on a flute, piano and trumpet are depicted.

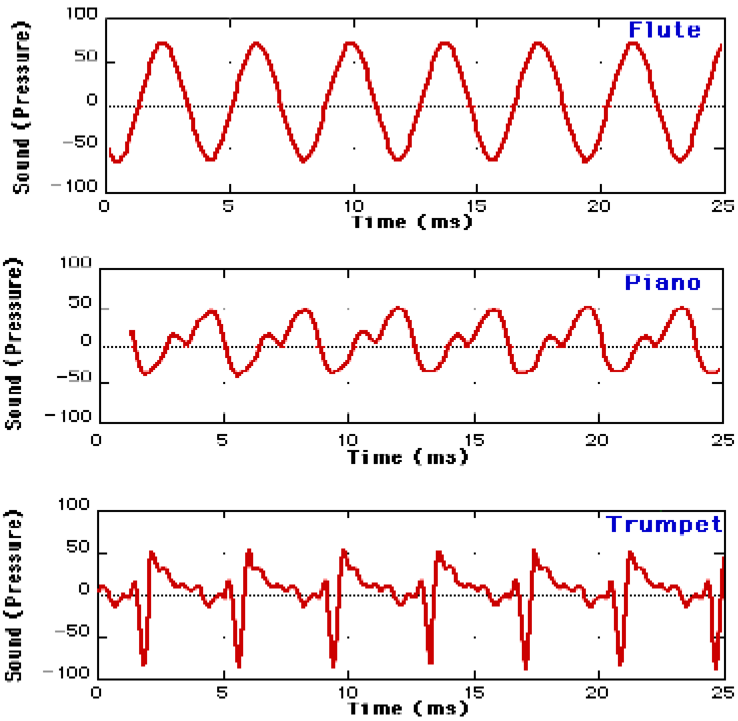

Figure 4. Sound waves of middle C played on a flute, piano and trumpet [Davis, 2002].

As can be seen, the same note has a different "color" or "texture" for each instrument. This is also reflected in the spectral content, shown previously in Figure 3. Therefore, timbre has much to do with a signal's harmonic content. The number of harmonics and overtones present, as well as their intensity, strongly influence the richness of the sound.

Deriving perceptually-inspired features for capturing timbre from an audio signal is not a trivial task. Some physical features such as harmonicity, uniformity and others were suggested and are described below.

Harmonicity

Harmonicity is a physical feature that tries to derive timbre information from the analysis of harmonics present in an audio signal [Welsh *et al*, 1999].

Once again, this is a rather complex task for polyphonic music. In [Wold *et al*, 1996], harmonicity is measured as the deviation of the signal's spectrum from a perfectly harmonic spectrum. Tzanetakis *et al* [Tzanetakis *et al*, 2001] propose fundamental frequency histograms for the analysis of harmonic content. Another possible approach could be to measure the number and percentage of harmonic peaks in the spectrum.

In polyphonic signals, the harmonic content can be used as basis for measuring noise levels. The noise content of a music signal can be useful to discriminate between soft/aggressive songs, e.g., ambient songs or heavy metal [Welsh *et al*, 1999].

Uniformity

Uniformity is a simpler approach for harmonic analysis. Here, energy levels in different frequency bands are calculated and their similarity is compared. In this way, highly pitched sounds, where most of the energy is concentrated in few frequency bands, can be distinguished from unpitched sounds, where energy is spread across more frequency bands [Golub, 2000].

Centroid

Centroid is usually used as an indicator of signal brightness, i.e., “the higher frequency content of the signal” [Wold *et al*, 1996]. This feature is calculated as the energy-weighted mean of frequencies of the short-time Fourier magnitude spectra.

Bandwidth

This feature is used as a measure of the frequency range of the signal. It is computed as the “magnitude-weighted average of differences between the spectral components and the centroid” [Wold *et al*, 1996]. Bandwidth is equivalent to frequency standard deviation.

Rolloff and Flux

In [Tzanetakis *et al*, 2001], spectral rolloff and flux are suggested as a measure of spectral shape and change respectively, useful to capture features related to music texture and instrumentation.

Low Energy

In measuring the amount of bass in a song, it is useful to use a feature called low energy [Tzanetakis *et al*, 2001]. This feature consists of calculating the percentage of frames that have less energy than the average energy in all frames.

3.4. Other Features

For most of the features described, analysis can be conducted in three ways, as referred in [Golub, 2000].

The first method consists of deriving short-term features from raw audio data. This is the case of fundamental frequency estimation methods, for example. Analysis is performed in short-time windows, as it happens in STFT analysis. Therefore, such features consist of temporal trajectories of some meaningful variables. Trajectory variations are also obtained, e.g., by measuring signal derivates (first-differences) [Golub, 2000].

Based on such short-term features, medium-term or long-term features can be obtained. Here, statistical data such as means and standard deviations [Wold *et al*, 1996], as well as histograms [Wold *et al*, 1996] are derived. What distinguishes medium from long-term features is the window size where the statistical analysis is performed. Clearly, the time window is wider for long-term features.

4. Research Results and Open Problems

The discipline of music information retrieval is still in its infancy. Most of the research conducted in this area deals with searching databases of MIDI songs, e.g., [Chai, 2001; Lemström and Perttu, 2000; Bainbridge *et al*, 1999]. This is a direct consequence of the difficulties posed by “real-world” music data, e.g., mp3 files.

However, we are now assisting to a strong interest in the issues of searching and classifying audio music signals. Some systems for QBE, e.g., [Logan and Salomon, 2001; Yang, 2001; Welsh *et al*, 1999] and music classification and clustering, e.g., [Tzanetakis *et al*, 2001; Pampalk, 2001; Golub, 2000] have been developed. In those systems, different authors use different subsets of the features described above. Typically, feature-vectors are constructed with statistical data obtained from features such as cepstral coefficients, energy, ZCR, harmonicity, centroid, bandwidth, spectral rolloff or tonal transitions.

Regarding classification, most results reported indicate 60 to 80% accuracy in relatively simple problems, where two to seven classes are separated (jazz, pop and classical music are the most common). Namely, Golub [Golub, 2000] refers an average classification accuracy of 77% in a problem involving two highly similar genres, 82% for three highly dissimilar genres and 64% for seven genres with different similarity levels. There, a database of 1714 songs was used. Classifiers evaluated were the generalized linear model, the multilayer perceptron and the k-nearest neighbors algorithm.

As for QBE systems, evaluation is usually carried out by counting the average number of similar songs in the first 5, 10 or 20 in the list of results. Typically, a set of users is chosen to personally evaluate the results obtained. This is clearly a subjective metric, since similarity is a rather vague concept. In [Logan and Salomon, 2001], a database of over 8000 songs encompassing several genres is used. In order to evaluate the matches obtained, these were judged by two users. The results reported indicate that, in the first 5 matches returned by the system, on an average 2.5 are similar to the query. For the first 10 and 20, 4.7 and 8.2 are said to be similar, respectively.

As stated before, the analysis carried out above is very subjective. Furthermore, results are not comparable since many different databases of songs are used. Regarding classification, empirical supervised learning is normally used to train music classifiers (one exception is [Pampalk, 2001], where clustering is performed, so that similar songs are found close to each other in a self-organizing map). Therefore, training examples must have been labeled previously. However, if we look at the taxonomies used in some music libraries, it is difficult to find any uniformity there. In fact, there are many semantic discrepancies in the classes defined, both in horizontal and vertical terms, i.e., the number of classes used (horizontal dimension) and the number of subclasses derived from them (vertical dimension) is very different. Also, the same song appears with different labels in different libraries. These and other problems are analyzed in [Pachet and Cazaly, 2000]. As a consequence, it is urgent to define standard test collections and benchmark problems, as well as a uniform taxonomy. This problem is the subject of the Workshop on the "Creation of Standardized Test Collections, Tasks and Metrics for Music Information Retrieval (MIR) and Music Digital Library (MDL) Evaluation", to be held at the Second Joint Conference on Digital Libraries (JCDL' 2002).

As for QBH/QBS systems, the bulk of the research deals with databases of MIDI songs, as stated before. QBH/QBS systems for real-world music must be grounded on robust and accurate strategies for polyphonic music analysis, which is still an open problem. In the same way, systems for automatic wave-based music segmentation and summarization are still non-existent.

In conclusion, content-based classification and retrieval of music is a fascinating research problem, with a broad range of possible commercial applications. Nevertheless, we still have many years of hard research ahead, before robust and accurate products are available.

List of Abbreviations

DFT – Discrete Fourier Transform
EMD – Electronic Music Delivery
QBE – Query-By-Example
QBH – Query-By-Humming
QBS – Query-By-Singing
STFT – Short-Time Fourier Transform
ZCR – Zero-Crossing Rate

References

Afterdawn, 2002
Afterdawn (2002). “Philips has developed a music recognition technology”, URL: http://www.afterdawn.com/news/archive/2515.cfm, published on December 17, 2001, available by March 3, 2002.

Bainbridge *et al*, 1999
Bainbridge D., Nevill-Manning C., Witten I., Smith L. and McNab R. (1999). “Towards a Digital Library of Popular Music”, *4th ACM International Conference on Digital Libraries,* pp. 161-169.

Bello *et al*, 2000

Bello J. P., Monti G. and Sandler M. (2000). "Techniques for Automatic Music Transcription", *1st International Symposium on Music Information Retrieval - ISMIR'2000*.

Bregman, 1990

Bregman A. S. (1990). *Auditory Scene Analysis: the Perceptual Organization of Sound*. MIT Press.

Chai, 2001

Chai W. (2001). *Melody Retrieval on the Web*, MSc Thesis, Massachusetts Institute of Technology.

CUIDADO, 2000

CUIDADO (2000). "Interfaces de recherche par le contenu et descripteurs pour l'Audio et la Musique accessibles en ligne", URL: http://www.cuidado.mu/, available by February 2, 2002.

Davis, 2002

Davis D. (2002). "Characteristics of Sound", *Lecture notes on the Principles of Physics I course, Chapter 16*, Eastern Illinois University, Physics Department. URL: http://oldsci.eiu.edu/physics/DDavis/1150/16Waves/char.html.

Dunn, 2000

Dunn J. W. (2000). "Beyond VARIATIONS: Creating a Digital Music Library", *1st International Symposium on Music Information Retrieval - ISMIR'2000*, Invited speaker.

Ellis, 1996

Ellis D. (1996). *Prediction-Driven Computational Auditory Scene Analysis for Dense Sound Mixtures*, PhD Thesis, Massachusetts Institute of Technology.

Golub, 2000

Golub S. (2000). *Classifying Recorded Music*, MSc Thesis, University of Edinburgh, Division of Informatics.

Goto, 2001

Goto M. (2001). “A Predominant-F0 Estimation Method for CD Recordings: MAP Estimation Using EM Algorithm for Adaptive Tone Models”, *IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP’2001.*

Huron, 2000

Huron D. (2000). “Perceptual and Cognitive Applications in Music Information Retrieval”, *International Symposium on Music Information Retrieval ISMIR’2000,* Invited speaker.

Klapuri *et al*, 2000

Klapuri A., Virtanen T. and Holm J.-M. (2000). “Robust Multipitch Estimation for the Analysis and Manipulation of Polyphonic Music Signals”, *Conference on Digital Audio Effects* - *COST-G6.*

Lemström and Perttu, 2000

Lemström K. and Perttu S. (2000). “SEMEX – An Efficient Music Retrieval Prototype”, *1st International Symposium on Music Information Retrieval – ISMIR’2000.*

Logan and Salomon, 2001

Logan B. and Salomon A. (2001). “A Music Similarity Function Based on Signal Analysis”, *IEEE International Conference on Multimedia and EXPO – ICME’2001.*

Martin, 1996

Martin K. D. (1996). “Automatic Transcription of Simple Polyphonic Music: Robust Front End Processing”, *3rd Joint Meeting of the Acoustical Societies of America and Japan.*

OMRAS, 1999

OMRAS (1999). “Online Music Recognition and Searching”. URL: http://www.omras.org/.

Pachet and Cazaly, 2000

Pachet F. and Cazaly D. (2000). “A Taxonomy of Musical Genres”, *Computer-Assisted Information Retrieval International Conference – RIAO’2000*.

Pampalk, 2001

Pampalk E. (2001). *Islands of Music: Analysis, Organization and Visualization of Music Archives*, MSc Thesis, Vienna University of Technology.

Pfeiffer *et al*, 1996

Pfeiffer S., Fischer S. and Effelsberg W. (1996). “Automatic Audio Content Analysis”, *4th ACM International Conference on Multimedia*, pp. 21-30.

PRWeb, 2000

PRWeb (2000). “Music recognition software frees musicians to be more creative”, URL: http://www.prweb.com/releases/?14571, published on May 13, 2000, available by March 3, 2002.

Rabiner and Schafer, 1978

Rabiner L. R. and Schafer R. W. (1978). *Digital Processing of Speech Signals*, Englewood Cliffs, NJ: Prentice-Hall.

Scheirer, 1998

Scheirer E. D. (1998). “Tempo and beat analysis of acoustic musical signals”, *Journal of the Acoustic Society of America*, Vol. 103, No. 1, pp. 588-601.

Scheirer, 2000

Scheirer E. D. (2000). *Music-Listening Systems*, PhD Thesis, Massachusetts Institute of Technology.

Slaney and Lyon, 1990

Slaney M. and Lyon R. (1990). “A Perceptual Pitch Detector”, *IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP’90*, Vol. 1, pp. 357-360.

Tzanetakis *et al*, 2001

Tzanetakis G., Essl G. and Cook P. (2001). “Automatic Musical Genre Classification of Audio Signals”, *2nd International Conference on Music Information Retrieval – ISMIR’2001*.

Welsh *et al*, 1999

Welsh M., Borisov N., Hill J., von Behren R. and Woo A. (1999). *Querying Large Collections of Music for Similarity*, Technical Report, University of California, Berkeley, Computer Science Division.

Wold *et al*, 1996

Wold E., Blum T., Keislar D. and Wheaton J. (1996). “Content-Based Classification, Search and Retrieval of Audio”, *IEEE Multimedia*, Vol. 3, No. 3, Fall 1996, pp. 27-36.

Yang, 2001

Yang C. (2001). “Music Database Retrieval Based on Spectral Similarity”, *2nd International Conference on Music Information Retrieval – ISMIR’2001*, poster.