Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 64960, 12 pages
doi:10.1155/2007/64960

Research Article

Automatic Genre Classification of Musical Signals

Jayme Garcia Arnal Barbedo and Amauri Lopes

Departamento de Comunicações, Faculdade de Engenharia Elétrica e de Computação (FEEC), Universidade Estadual de Campinas (UNICAMP), Caixa Postal 6101, Campinas 13083-852, Brazil

Received 28 November 2005; Revised 26 June 2006; Accepted 29 June 2006

Recommended by George Tzanetakis

We present a strategy to perform automatic genre classification of musical signals. The technique divides the signals into 21.3 milliseconds frames, from which 4 features are extracted. The values of each feature are treated over 1-second analysis segments. Some statistical results of the features along each analysis segment are used to determine a vector of summary features that characterizes the respective segment. Next, a classification procedure uses those vectors to differentiate between genres. The classification procedure has two main characteristics: (1) a very wide and deep taxonomy, which allows a very meticulous comparison between different genres, and (2) a wide pairwise comparison of genres, which allows emphasizing the differences between each pair of genres. The procedure points out the genre that best fits the characteristics of each segment. The final classification of the signal is given by the genre that appears more times along all signal segments. The approach has shown very good accuracy even for the lowest layers of the hierarchical structure.

1. INTRODUCTION

The advances experienced in the last decades in areas as information, communication, and media technologies have made available a large amount of all kinds of data. This is particularly true for music, whose databases have grown exponentially since the advent of the first perceptual coders early in the 90's. This situation demands for tools able to ease searching, retrieving, and handling such a huge amount of data. Among those tools, automatic musical genre classifiers (AMGC) can have a particularly important role, since they could be able to automatically index and retrieve audio data in a human-independent way. This is very useful because a large portion of the metadata used to describe music content is inconsistent or incomplete.

Music search and retrieval is the most important application of AGC, but it is not the only one. There are several other technologies that can benefit from AGC. For example, it would be possible to create an automatic equalizer able to choose which frequency bands should be attenuated or reinforced according to the label assigned to the signal being considered. AGC could also be used to automatically select radio stations playing a particular genre of music.

The research field of automatic music genre classification has got increasing importance in the last few years. The most significant proposal to specifically deal with this task was released in 2002 [1]. Several strategies dealing with related problems have been proposed in research areas such as speech/music discriminators and classification of a variety of sounds. More details about previous works are presented in Section 2.

The strategy presented here divides the audio signals into 21.3-milliseconds frames from which the following 4 features are extracted: bandwidth, spectral roll-off, spectral flux, and loudness. The frames are grouped into 1-second analysis segments, and the results of each feature along each analysis segment are used to calculate three summary features: mean, variance, and a third summary feature called "prevalence of the main peak," which is defined in Section 6. A pairwise comparison, using the Euclidean distance, for each combination of classes is made, using a set of reference vectors that specifically define the boundaries or differences between those classes. The procedure has some similarity with the "bag of frames" procedure used in [2].

The taxonomy adopted in this paper has a four-layer hierarchical structure, and the classification is firstly performed considering the 29 genres of the lowest layer. After that, the classes of the higher layers are determined accordingly. This bottom-up approach has resulted in very good

results, because the fine comparison carried out among the lower genres greatly improves the accuracy achieved for the upper levels. Moreover, it is important to note that the strategy works with only 12 summary features for each analysis segment. This relatively low-dimensional summary feature space makes the technique quite robust to unknown conditions.

Therefore, the strategy presents some interesting characteristics and novelties: a low-dimensional feature space, a wide and deep taxonomic structure, and a nonconventional pairwise classification procedure that compares all possible pairs of genres and explores such information to improve the discrimination. As a result, the technique allows the adoption of wider and deeper taxonomic structures without significantly harming the accuracy, since as more genres are considered, finer information can be gathered.

The paper is organized as follows. Section 2 presents a brief description of some of the most relevant previous related works. Section 3 discusses the problem of classifying musical signals and points out some of the main difficulties involved in such a task. Section 4 presents and describes the musical genre taxonomy adopted in this paper. Section 5 describes the extraction of the features from the signals. Section 6 describes the strategy used to classify the signals. Section 7 presents the design of the tests and the results achieved. Finally, Section 8 presents the conclusions and future work.

2. PREVIOUS WORK

Before 2002, there were few works dealing specifically with the problem of musical genre classification. Lambrou et al. [3] use a number of features extracted from both time and wavelet transform domains to differentiate between 3 musical genres. A graphical analysis of spectrograms is used by Deshpande et al. [4] to classify musical signals into 3 genres. Logan [5] studied the suitability of Mel frequency cepstral coefficients (MFCCs) for music classification. In 2002 Tzanetakis and Cook [1] released the most important work so far to specifically deal with the problem of musical genre classification. The authors used three sets of features representing timbral texture, rhythmic content, and pitch content, in order to characterize the signals. They used a number of statistical pattern recognition classifiers to classify the signals into 10 musical genres. The precision achieved was about 60%.

This last work has built most of the foundations used in subsequent researches. Some later proposals have succeeded in presenting new effective approaches to classify songs into genres. Agostini et al. [6] deal with the classification of musical signals according to the instruments that are being played. In the approach proposed by Pye [7], Mel frequency cepstral coefficients (MFCC) are extracted from music files in the MP3 format without performing a complete decoding. A Gaussian mixture model (GMM) is used to classify the files into seven musical genres. The database used in the experiments is relatively limited, and the procedure has achieved a precision of about 60%.

More recently, the number of publications in this area has grown fast, especially due to specialized events like the International Symposium on Music Information Retrieval (ISMIR), which occurs every year since 2000. In 2005, the first Music Information Retrieval Evaluation eXchange has taken place during the 6th ISMIR Conference, where the contestants should classify audio signals into one of 10 different genres; several algorithms were proposed, leading to accuracies between 60% and 82% [8].

Some recent works provide useful information. West and Cox [2] test several features and classification procedures, and a new classifier based on the unsupervised construction of decision trees is proposed. Gouyon et al. [9] carry out an evaluation of the effectiveness of rhythmic features. Hellmuth et al. [10] combine low-level audio features using a particular classification scheme, which is based on the similarity between the signal and some references. Pampalk [11] presents an extensive study about models for audio classification in his thesis. Dixon et al. [12] present a method to characterize music by rhythmic patterns. Berenzweig et al. [13] examine acoustic and subjective approaches to calculate the similarity between songs. McKay et al. [14] present a framework to optimize the music genre classification. Lippens et al. [15] made a comparison between the performance of humans and automatic techniques in classifying musical signals. Xu et al. [16] apply support vector machines to perform a hierarchical classification, and a clustering algorithm based on several features is used to structure the music content. Finally, a comparative study on the performance of several features commonly used in audio signal classification is presented by Pohle et al. [17].

There are several works that have investigated other correlated problems, like speech/music discrimination (e.g., [18]) and classification of sounds (e.g., [19]), providing some ideas that can be extended to the subject treated here.

Next section presents a discussion on the difficulties and inconsistencies in classifying musical signals into genres.

3. DISCUSSION ON GENRE LABELING

Beside the inherent complexity involved in differentiating and classifying musical signals, the AGC have to face other difficulties that make this a very tricky area of research. In order to work properly, an AGC technique must be trained to classify the signals according to a predefined set of genres. However, there are two major problems involved in such a predefinition, which will be discussed next.

Firstly, the definition of most musical genres is very subjective, meaning that the boundaries of each genre are mostly based on individual points of view. As a result, each musical genre can have its boundaries shifted from person to person. The degree of arbitrariness and inconsistency of music classification into genres was discussed by Pachet and Casaly [20], where the authors compared three different Internet genre taxonomies: http://www.allmusic.com (531 genres), http://www.amazon.com (719 genres), and http://www.mp3.com (430 genres). The authors have drawn three major conclusions:

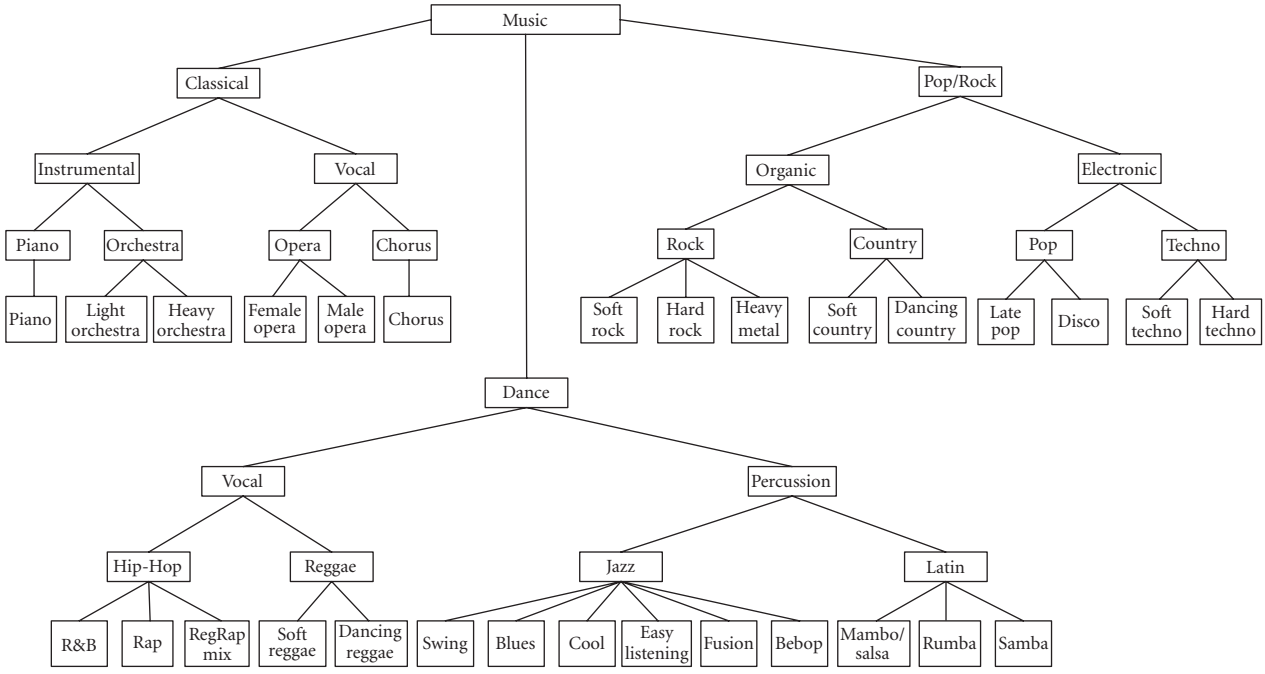

Figure 1: Musical genre taxonomy.

(i) there is no agreement concerning the name of the genres: only 70 words are common to all three taxonomies;
(ii) among the common words, not even largely used names, as "Rock" and "Pop," denote the same set of songs;
(iii) the three taxonomies have quite different hierarchical structures.

As pointed out by Aucouturier and Pachet [21], if even major taxonomic structures present so many inconsistencies among them, it is not possible to expect any degree of semantic interoperability among different genre taxonomies. Despite those difficulties, there have been efforts to develop carefully designed taxonomies [20, 21]. However, no unified framework has been adopted yet.

To deal with such a difficulty, the taxonomy adopted in this paper was carefully designed to use, as much as possible, genres and nomenclatures that are used by most reference taxonomies, and hence that are most likely to be readily identified by most users. This procedure reduces the inconsistencies and tends to improve the precision of the method, as will be seen in Section 6. However, it is important to emphasize that some degree of inconsistency will always exist due to the subjectiveness involved in classifying music, limiting the reachable accuracy.

The second major problem is the fact that a large part of modern songs have elements from more than one musical genre. For example, there are some jazz styles that incorporate elements of other genres, as fusion (jazz+rock); there are also recent reggae songs that have strong elements of rap; as a last example, there are several rock songs that incorporate electronic elements generated by synthesizers. To deal with this problem, the strategy used in this paper is to divide basic genres into a number of subgenres able to embrace those intermediate classes, as will be described in the next section.

4. TAXONOMY

Figure 1 shows the structure of the taxonomy adopted in the present work, which has 4 hierarchical layers and a total of 29 musical genres in the lowest layers. The description of each box is presented next. It is important to emphasize that the taxonomy and exemplar songs were crafted by hand, which is a major difference between this work and existing works.

The taxonomy shown in Figure 1 was created aiming to include as many genres as possible, improving the generality of the method, but keeping at the same time the consistency of the taxonomy, as commented in Section 3. With such a great number of musical genres in the lowest levels of the hierarchical structure, the classification becomes a very challenging issue. However, as will be seen later, the strategy proposed here is very successful in overcoming the difficulties. Moreover, the accuracy achieved for the higher layers increases greatly with such a complex structure, as described in Section 7.

From this point to the end of the paper, all musical classes of the lowest hierarchical level in Figure 1 are called "genres," while the divisions of higher levels are called "upper classes" or simply "classes."

4.1. *Classical*

The songs of this class have the predominance of classical instruments like violin, cello, piano, flute, and so forth. This

class has the following divisions and subdivisions (with examples for the lowest-level genres).

(i) *Instrumental*: the songs of this genre do not have any vocal elements.

(1) *Piano*: songs dominated by or composed exclusively for piano (e.g., Piano Sonata, Chopin).
(2) *Orchestra*: songs played by orchestras.

(a) *Light orchestra*: light and slow songs (e.g., Air, Bach).
(b) *Heavy orchestra*: fast and intense songs (e.g., Ride of Valkyries, Wagner).

(ii) *Vocal*: classical music with presence of vocals.

(1) *Opera*: songs with strong presence of vocals, normally with few people singing at a time.

(a) *Female opera*: opera songs with female vocals (e.g., Barcarolle, Offenbach).
(b) *Male opera*: opera songs with male vocals (e.g., O Sole Mio, Capurro and Capua).

(2) *Chorus*: classical songs with chorus.

4.2. Pop/Rock

This is the largest class, including a wide variety of songs. It is divided according to the following criteria.

(i) *Organic*: this class has the prevalence of electric guitars and drums; electronic elements are mild or not present.

(1) *Rock*: songs with strong predominance of electric guitars and drums.

(a) *Soft rock*: slow and soft songs (e.g., Stairway to Heaven, Led Zeppelin).
(b) *Hard rock*: songs of this genre have marked beating, strong presence of drums and a faster rhythm than soft rock (e.g., Livin' on a Prayer, Bon Jovi).
(c) *Heavy metal*: this genre is noisy, fast, and often has very aggressive vocals (e.g., Frantic, Metallica).

(2) *Country*: songs typical of southern United States; have elements both from rock and blues. Electric guitars and vocals are predominant.

(a) *Soft country*: slow and soft songs (e.g., Your Cheating Heart, Hank Williams Sr.).
(b) *Dancing country*: the songs of this genre have fast and dancing rhythm (e.g., I'm Gonna Get Ya Good, Shania Twain).

(ii) *Electronic*: most of the songs of this class have the predominance of electronic elements, usually generated by synthesizers.

(1) *Pop*: the songs of this class are characterized by the presence of electronic elements. Vocals are usually present. The beating is slower and less repetitive than techno songs, and vocals often play an important role.

(a) *Late pop*: pop songs released after 1980, with strong presence of synthesizers (e.g., So Hard, Pet Shop Boys).
(b) *Disco*: songs typical of late 70's with a very particular beating. Electronic elements are also present here, but they are less marked than in pop songs (e.g., Dancing Queen, Abba).

(2) *Techno*: this class has fast and repetitive electronic beating. Songs without vocals are common.

(a) *Soft techno*: lighter techno songs, like trance style (e.g., Deeply Disturbed, Infected Mushroom).
(b) *Hard techno*: extremely fast songs (up to 240 beats per minute) compose this genre (e.g., After All, Delirium).

4.3. Dance

The songs that compose this third and last general musical class have strong percussive elements and very marked beating. This class is divided according to the following rules.

(i) *Vocal*: vocals play the central role in this class of songs. Percussive elements also have strong presence, but not as significant as vocals.

(1) *Hip-Hop*: these songs have strong predominance of vocals and a very marked rhythm.

(a) *R & B*: the songs of this genre are soft and slow (e.g., Full Moon, Brandy).
(b) *Rap*: this genre presents really marked vocals, sometimes looking like pure speech (e.g., Lets Get Blown, Snoop Dogg).
(c) *RegRap mix:* these songs are actually reggae, but their characteristics are so close to rap that they fit best as a subgroup of Hip-Hop class. This genre is sometimes called "reggaeton" (e.g., I Love Dancehall, Capleton).

(2) *Reggae:* typical music of Jamaica that has a very particular beating and rhythm.

(a) *Soft reggae*: slow reggae songs (e.g., Is This Love, Bob Marley).
(b) *Dancing reggae*: songs with faster and more dancing rhythm (e.g., Games People Play, Inner Circle).

(ii) *Percussion*: the percussive elements are very marked and strong. Vocals may or may not be present. In some cases, the vocals are as important as the instrumental part of the song.

(1) *Jazz*: this class is characterized by the predominance of instruments like piano and saxophone.

Electric guitars and drums can also be present; vocals, when present, are very characteristic.

(a) *Swing:* the songs of this genre are vibrant and often have dancing rhythm. This style has popularized the big bands in the decades of 1930 and 1940. Several instruments are present, as drums, bass, piano, guitars, trumpets, trombones, and saxophones (e.g., Tuxedo Junction, Glenn Miller Orchestra).

(b) *Blues*: vocal and instrumental genre, it has strong presence of guitars, piano and harmonica. This style is the main predecessor of a large part of the music produced in the 20th century, including rock, which has several of its characteristics (e.g., Sweet Little Angel, B.B. King).

(c) *Cool*: this jazz style is light and introspective, with a very slow rhythm (e.g., Boplicity, Miles Davis).

(d) *Easy listening*: the songs of this class are soft and often orchestrated. The songs sometimes have dancing rhythm (e.g., Moon River, Andy Williams).

(e) *Fusion*: it is a mix of jazz and rock elements (e.g., Birdland, Weather Report).

(f) *Bebop*: it is a form of jazz characterized by fast tempos and improvisation based on harmonic structure rather than melody. Vocals are very marked (Ruby my Dear, Thelonius Monk).

(2) *Latin*: this class is composed of Latin rhythms like salsa, mambo, samba, and rumba; the songs of this genre have strong presence of instruments of percussion and, sometimes, guitars.

(a) *Mambo/Salsa*: dancing Caribbean rhythms with strong presence of percussive drums and tambours (e.g., Mambo Gozon, Tito Puente).

(b) *Rumba*: Spanish rhythm with strong predominance of guitars (e.g., Bamboleo, Gipsy Kings).

(c) *Samba*: strongly percussive Brazilian genre with predominance of instruments like tambourines, small guitars, and so forth (e.g., Faixa Amarela, Zeca Pagodinho).

5. FEATURE EXTRACTION

Before the feature extraction itself, the signal is divided into frames using a hamming window of 21.3 milliseconds, with superposition of 50% between consecutive frames. The signals used in this paper are sampled at 48 kHz, resulting in frames of 1024 samples. The extraction of the features is performed individually for each frame. The description of each feature is presented in the following.

5.1. Spectral roll-off

This feature determines the frequency R_i for which the sum of the spectral line magnitudes is equal to 95% of the total sum of magnitudes, as expressed by [22]:

$$\sum_{k=1}^{R_i} |X_i(k)| = 0.95 \cdot \sum_{k=1}^{K} |X_i(k)|, \tag{1}$$

where $|X_i(k)|$ is the magnitude of spectral line k resulting from a (discrete Fourier transform DFT) with 1024 samples applied to the frame i and K is half the total number of spectral lines (second half is redundant).

5.2. Loudness

The first step to calculate this feature is modeling the frequency response of human outer and middle ears. Such a response is given by [23]

$$W(k) = -0.6 \cdot 3.64 \cdot f(k)^{-0.8} - 6.5 \cdot e^{-0.6\cdot(f(k)-3.3)^2} + 10^{-3} \cdot f(k)^{3.6}, \tag{2}$$

where $f(k)$ is the frequency in kHz given by

$$f(k) = k \cdot d, \tag{3}$$

and d is the difference in kHz between two consecutive spectral lines (in the case of this work, 46.875).

The frequency response is used as a weighting function that emphasizes or attenuates spectral components according to the hearing behavior. The loudness of a frame is calculated according to

$$\mathrm{ld}_i = \sum_{k=1}^{K} |X_i(k)|^2 \cdot 10^{W(k)/20}. \tag{4}$$

5.3. Bandwidth

This feature determines the frequency bandwidth of the signal, and is given by [19]

$$\mathrm{bw}_i = \sqrt{\frac{\sum_{k=1}^{K} \left[(\mathrm{ce}_i - k)^2 \cdot |X_i(k)|^2\right]}{\sum_{k=1}^{K} |X_i(k)|^2}}, \tag{5}$$

where ce_i is the spectral centroid of frame i, given by

$$\mathrm{ce}_i = \frac{\sum_{k=1}^{K} k \cdot |X_i(k)|^2}{\sum_{k=1}^{K} |X_i(k)|^2}. \tag{6}$$

Equation (5) gives the bandwidth in terms of spectral lines. To get the value in Hz, bw must be multiplied by d.

5.4. Spectral Flux

This feature is defined as the quadratic difference between the logarithms of the magnitude spectra of consecutive analysis frames and is given by [1]

$$fe_i = \sum_{k=1}^{K} \left\{\log_{10}\left[X_i(k)\right] - \log_{10}\left[X_{i-1}(k)\right]\right\}^2. \tag{7}$$

The purpose of this feature is to determine how fast the signal spectrum changes along the frames.

5.5. Considerations on feature selection

Although this work uses the four features just presented, early versions of the technique included nine other features: spectral centroid, zero-crossing rate, fundamental frequency, low energy frames ratio, and the first five cepstral coefficients. In a first step, those features were applied individually to differentiate between musical classes, and the results were used to generate a ranking from the worse to the better feature. Next, the performance of the strategy presented in this paper was determined using all 13 features. After that, the features were eliminated one by one according to the ranking, starting with the worse ones. Every time a feature was eliminated, the tests using the strategy were repeated and the results were compared to the previous ones. The procedure was carried out until only two features left. It was observed that, under the conditions described in this paper, if more than four features are used, the summary feature space dimension becomes too high and the accuracy starts to decrease. On the other hand, if less than four features are used, essential information is lost and the accuracy is also reduced. Taking all those factors into account, the adopted set of four features has shown the best overall results. However, it is important to note that the number of four features is not necessarily optimal for other strategies.

6. CLASSIFICATION STRATEGY

The features extracted for each frame are grouped according to 1-second analysis segments. Therefore, each group will have 92 elements, from which three summary features are extracted: mean, variance, and main peak prevalence which is calculated according to

$$p_{ft}(j) = \frac{\max\left[ft(i,j)\right]}{(1/I)\cdot\sum_{i=1}^{I} ft(i,j)}, \tag{8}$$

where $ft(i,j)$ corresponds to the value of feature ft in the frame i of segment j, and I is the number of frames into a segment. This summary feature aims to infer the behavior of extreme peaks with relation to the mean values of the feature. High p_{ft} indicate the presence of sharp and dominant peaks, while small p_{ft} often means a smooth behavior of the feature and no presence of high peaks.

As a result of this procedure, each segment will lead to 12 summary features, which are arranged into a test vector to be compared to a set of reference vectors. The determination of the reference vectors is described next.

6.1. Training phase: determination of the reference vectors

The reference vectors were determined according to the following steps.

(a) Firstly, 20 signals with length of 32 seconds were carefully selected to represent each of the 29 genres adopted in this paper, resulting in a training set with 580 signals. The signals were selected according to the subjective attributes expected for each genre, and were taken from the database described in Section 7. It is important to highlight that tests were also carried out picking the reference signals at random; in this case, 10% (best case analyzed) to 60% (worst case) more signals were necessary to achieve the same performance. In average, it was observed that a random reference database requires about 30% more components to keep the performance.

(b) Next, the summary feature extraction procedure was applied to each one of the training signals. Since those signals have 32 seconds, 32 vectors with 12 summary features were generated for each signal, or 640 training vectors representing each genre.

(c) A comparison procedure was carried out taking two genres at a time. For example, the training vectors corresponding to the genres "piano" and "rap" were used to determine the 6 reference vectors (3 for each genre) that resulted in the best separation between those genres. Those reference vectors were chosen as follows. Firstly, a huge set of potential reference vectors was determined for each genre, considering factors as the mean of the training vectors and the range expected for the values of each summary feature, discarding vectors that were distant from the cluster. After that, for a given pair of genres, all possible six-vector combinations extracted from both sets of potential vectors were considered, taking into account that each set must contribute with three vectors. For each combination, a Euclidean distance was calculated between each potential vector and all training vectors from both genres. After that, each training vector was labeled with the genre corresponding to the closest potential vector. The combination of potential vectors that resulted in the highest classification accuracy was taken as the actual set of reference vectors for that pair of genres.

(d) The procedure described in item (c) was repeated for all possible pairs of genres (406 pairs for 29 genres). As a result, each genre has 28 sets of 3 reference vectors, resulting from the comparison with the other 28 genres.

The vector selection described in item (c) tends to select the vectors that are closer to the core of the respective class cluster. The number of reference vectors was limited to 3 in order to guarantee that a given class is represented by vectors close to its core. If more than 3 vectors are considered, the additional vectors tend to be closer to other clusters, increasing the probability of a misclassification. This is particularly true when the classes have a high degree of similarity. An alternative approach would be using an adaptive number of vectors—more vectors when considering very dissimilar classes, and fewer vectors for closely related classes. This option is to be tested in future research. Other well-known systems, like grouping vectors into clusters and comparing new vectors to the clusters were also tested, with poorer results.

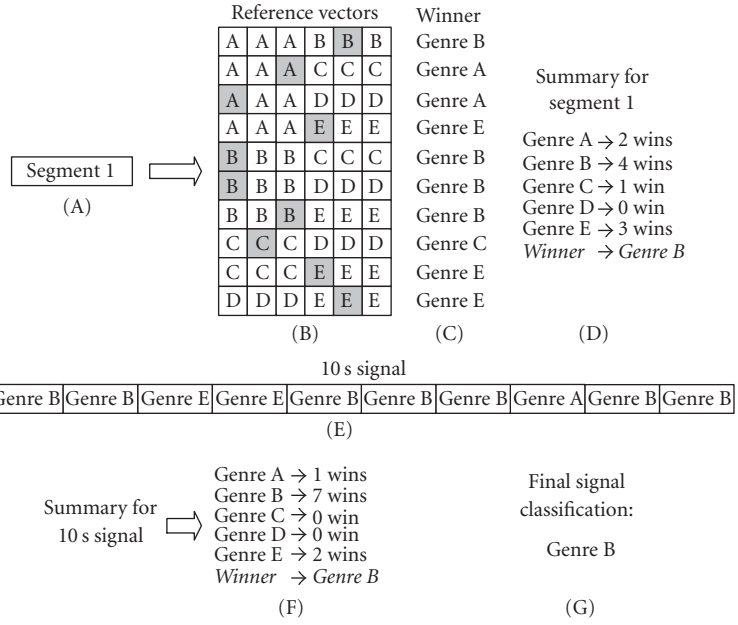

Figure 2: Classification procedure.

This wide pairwise comparison of genres provides much better differentiation between the genres than using a single comparison considering all genres at a time. This is probably due to the ability of the approach to gather relevant information from the differences between each pair of genres.

6.2. Test procedure

Figure 2 illustrates the final classification procedure of a signal. The figure was constructed considering a hypothetical division into 5 genres (A), (B), (C), (D), and (E) and a signal of 10 seconds in order to simplify the illustrations. Nevertheless, all observations and conclusions drawn from Figure 2 are valid for the 29 genres and 32-second signals actually considered in this paper.

As can be seen in Figure 2, the procedure begins with the extraction of the summary feature vector from the first segment of the signal (Figure 2(A)). Such a vector is compared with the reference vectors corresponding to each pair of genres, and the smallest Euclidean distance indicates the closest reference vector in each case (gray squares in Figure 2(B)). The labels of those vectors are taken as the winner genres for each pair of genres (C). In the following, the number of wins of each genre is summarized, and the genre with most victories is taken as the winner genre for that segment (D); if there is a draw, the segment is labeled as "inconclusive." The procedure is repeated for all segments of the signal (E). The genre with more wins along all segments of the signal is taken as the winner (F); if there is a draw, the summaries of all segments are summed and the genre with more wins is taken as winner. If a new draw occurs, all procedures illustrated in Figure 2 are repeated considering only the reference vectors of the drawn genres; all other genres are temporarily ignored. The probability of a new draw is very close to zero, but if it occurs, one of the drawn genres is taken at random as winner. Finally, the winner genre is adopted as the definitive classification of the signal (G).

Normally, the last segment of a signal will have less than one second. In those cases, if the segment has more than 0.5 second, it is considered and the summary features are calculated using the number of frames available, which will be between 46 and 92. If such a segment has less than 0.5 second, its frames are incorporated to the previous segment, which will then have between 92 and 138 frames.

The classification procedure is carried out directly in the lowest level of the hierarchical structure shown in Figure 1. This means that a signal is firstly classified according to the basic genres, and the corresponding upper classes are consequences of this first classification (bottom-up approach). For example, if a given signal is classified as "Swing," its classification for the third, second, and first layers will be, respectively, "Jazz," "Percussion," and "Dance." This strategy was adopted because it was observed that the lower is the hierarchical layer in which the signal is directly classified the more precise is the classification of the signal into upper classes. In tests with a top-down approach, where the signals were classified layer by layer, starting with the topmost, the accuracy achieved was between 3% and 5% lower than that achieved using the bottom-up approach.

A strategy using linear discriminant analysis (LDA) to perform the pairwise classification was also tested. The performance was very close to that one achieved by the strategy presented in this section, but the greater computational effort has prevented its adoption.

Next section presents the results achieved by the proposal.

7. RESULTS

The results to be presented in this section derive from two different tests. Section 7.1 describes the results achieved using the test database especially developed for this work (complete test database). Such a database contains all 29 genres present in the lower hierarchical level of the taxonomy presented in Figure 1, and was especially designed to test the strategy face to difficult conditions. On the other hand, since this database has not been used by any previous approach, a direct comparison between the technique presented here and its predecessors is not possible. To allow a comparison, Section 7.2 shows the results obtained when the proposed approach was applied to the Magnatune dataset [24], which has been used to assess some previous approaches. Magnatune is a Creative Commons record label, which kindly allows the use of its material for academic research.

7.1. Complete test database

The complete test database is composed of 2266 music excerpts, which represent more than 20 hours of audio data (13.9 GB). Each genre is represented by at least 40 signals. The signals were sampled at 48 kHz and quantized with 16 bits. The audio material was extracted from compact discs, from Internet radio streaming and also from coded files (MP3, WMA, OGG, AAC). The music database was divided into a training set of 580 files, which was used to determine the reference vectors described in Section 6.1, and into a test set, which was used to validate the technique and is composed of the remaining 1686 files. All results reported in this section were obtained using the second set.

It is important to emphasize that some precautions were taken in order to avoid biased results. Firstly, the database was mostly built taking only one song by each artist (in less than 1% of the cases two songs were allowed). This avoids well-known over-fitting problems that arise when songs of a particular artist are split across the training and test sets. Moreover, each of the 2266 excerpts derives from a different song, which assures that the training and test sets are completely diverse, avoiding in this way biased results. Finally, it is worth noting that the training set does not include any file submitted to a perceptual codec. This procedure is important because perceptual coders can introduce characteristics that can be over-fitted by the model, which could result in low accuracies face to signals coming from sources not found in the test set.

Figure 3 shows the confusion matrix associated to the tests, in terms of relative values. First column shows the target genres, and first row shows the genres actually estimated by the technique. Taking the first line (light orchestra) as example, it can be seen that 82% of light orchestra songs were correctly classified, 8% were classified as heavy orchestra, 5% as piano, 3% as female opera, and 2% as swing.

The main diagonal in Figure 3 shows the correct estimates, and all values outside the main diagonal are errors. Also, as darker is the shading of an area, the lower is the hierarchical layer. As can be seen, most errors are concentrated inside a same class. Figure 4 shows the accuracy achieved for each genre and class in all layers of the hierarchical structure, and Table 1 shows the overall accuracy for each layer. It is important to observe that several alternative definitions of the training set were tested, with mild impact in the overall accuracy (in the worst case, the accuracy has dropped about 2%).

As expected, the accuracy is higher for upper classes. The accuracy achieved for the first layer is higher than 87%, which is a very good result. The accuracy of 61% for the basic genres is also good, especially considering that the signals were classified into 29 genres; a number that is greater than those in the previous works. It is also useful to observe that layer 3 has 12 classes; a number that is compatible with the best proposals presented so far. In this case, the technique has reached an accuracy of 72%. The good performance becomes even more evident when one compares the performance of the technique with the results achieved in subjective classifications. As discussed in Section 3, classifying musical signals into genres is a naturally fuzzy and tricky task, even when subjectively performed. The performance of humans in classifying musical signals into genres was investigated in [21], where it was asked from college students to classify musical signals into one of 10 different genres. The subjects were previously trained with representative samples of each genre. The students were able to correctly judge 70% of the signals. Although a direct comparison is not possible due to differences in the taxonomy and databases, it can be concluded that the technique proposed here has achieved a performance at least as good as that achieved in the subjective tests, even with 2 more genres in the third layer.

Several factors can explain those good results. Firstly, the division into 29 genres has allowed that several different nuances of a given musical genre could be properly considered without harming the performance of the strategy. This is particularly true for classes like jazz, which encompass quite diverse styles that, if considered together in a single basic genre, could cause problems to the classifier. On the other hand, such a wide taxonomy causes some genres to have very similar characteristics. To face that situation, the pairwise comparison approach was adopted, emphasizing the differences between genres and improving the classification process.

Another important reason for the good results was the bottom-up approach. In the lowest level, the differences between the genres are explored in a very efficient way. Hence, the different nuances of more general classes can be correctly identified. As a result, the classification accuracy achieved for the higher levels of the taxonomic structure is greatly improved. Moreover, even when a signal is misclassified in the lowest level, it is likely that the actual and estimated genres pertain to the same upper class. Therefore, an incorrect classification in the lowest level can become a correct classification when a higher level is considered. This explains the excellent results obtained for the higher levels of the taxonomic structure.

It is also important to emphasize that the summary feature space used in this paper has a relatively low dimension.

	LO	HO	PI	FO	MO	CH	SR	RO	HM	SC	DC	PO	DI	ST	HT	RB	RA	RR	RE	DR	SW	BL	CO	EL	FU	BE	MA	RU	SA
LO	.82	.08	.05	.03	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.02	0	0	0	0	0	0	0	0
HO	.07	.76	0	0	.02	.04	0	0	0	.02	0	0	0	0	0	0	0	0	0	0	.02	0	.03	0	.01	.03	0	0	0
PI	0	0	.97	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.03	0	0	0
FO	0	.06	0	.89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.03	0	.02	0	0	0	0	0	0
MO	0	.07	0	0	.85	.02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.02	.02	.02	0	0	0	0	0	0
CH	.06	.19	0	0	.03	.69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.03	0	0	0	0	0	0	0	0
SR	0	.01	0	0	0	0	.46	.08	0	.12	.04	.06	0	0	0	.03	.01	0	.01	.01	0	.01	.01	.06	0	.04	.01	.01	.03
RO	0	.01	0	0	0	0	.05	.57	.13	.01	.04	.05	.01	.01	.02	0	0	0	0	.02	0	.04	0	0	.01	.01	.01	0	.01
HM	0	0	0	0	0	0	.01	.25	.65	0	.02	.01	0	0	.02	0	0	0	0	0	0	.02	0	0	0	0	.01	0	.01
SC	0	.01	0	0	0	0	.07	.01	0	.59	.11	0	0	0	0	0	0	0	0	0	0	0	.01	.05	.03	.06	.03	.03	0
DC	0	.01	0	0	.02	0	.09	.15	0	.04	.34	.05	0	0	0	.02	0	0	0	.02	0	.01	0	.01	.04	.07	.07	.02	.04
PO	0	0	0	0	0	0	.04	.08	0	0	.03	.70	.07	.03	0	0	.01	0	0	.01	0	0	0	0	0	0	0	0	.03
DI	0	0	0	0	0	0	0	.11	0	0	.06	.06	.57	0	0	0	.03	0	0	.02	0	.03	0	0	0	.06	.03	0	.03
ST	0	0	0	0	0	0	0	.04	0	0	0	.06	.02	.72	.10	0	0	0	.02	.04	0	0	0	0	0	0	0	0	0
HT	0	0	0	0	0	0	0	.01	.05	.02	0	.03	0	.13	.62	.01	.05	.01	0	.07	0	0	0	0	0	0	0	0	0
RB	0	0	0	0	0	0	.03	0	0	0	0	.03	0	0	0	.72	.03	.04	.03	.09	0	0	0	0	0	0	0	0	.03
RA	0	0	0	0	0	0	0	0	0	0	0	.01	0	.02	.02	.09	.53	.19	0	.12	0	0	0	0	0	0	.02	0	0
RR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.06	.10	.68	.03	0	0	0	0	0	0	0	.10	0	.03
RE	0	0	0	0	0	0	0	0	0	0	0	.03	0	0	0	0	0	.06	.55	.27	0	0	0	0	0	.09	0	0	0
DR	0	0	0	0	0	0	0	0	0	0	0	0	.04	.02	.02	.04	0	.09	.08	.63	0	0	0	0	0	.06	.02	0	0
SW	0	.03	0	0	.04	0	.03	0	0	0	.03	.03	0	0	0	0	0	0	0	0	.64	0	.03	0	.07	.03	0	0	.07
BL	0	0	0	0	0	0	.05	.13	0	.03	0	.03	0	0	0	0	0	0	0	.05	0	.57	0	.08	.03	0	0	0	.03
CO	0	.05	.03	0	0	0	.05	0	0	.08	0	0	.05	0	0	0	0	0	.03	.03	0	0	.50	.10	0	.08	0	0	0
EL	0	.03	0	0	0	0	.08	.03	0	.10	0	0	.03	0	0	0	0	0	0	0	.02	0	.11	.50	0	.05	0	.02	.03
FU	0	0	0	0	0	0	.03	.05	0	0	.02	.05	.03	0	0	0	0	0	0	0	0	0	.03	0	.74	.05	0	0	0
BE	0	.06	0	0	0	0	.03	0	0	.03	.03	0	0	0	0	0	.03	0	.03	.03	0	.03	0	.03	0	.58	.09	.03	0
MA	0	0	0	0	0	0	0	0	0	0	.07	0	.01	.01	0	.03	.03	.07	0	.04	0	.01	0	.01	.01	.12	.50	0	.09
RU	0	.02	0	0	0	0	.10	.02	0	.03	.02	.02	0	0	0	.02	0	0	0	.05	0	0	0	.05	.03	.12	0	.45	.07
SA	0	0	0	0	.02	0	.03	.04	0	0	.04	.08	.04	0	0	0	.02	.04	0	.04	0	0	0	0	0	.04	.04	0	.57

BE: Bebop
BL: Blues
CH: Chorus
CO: Cool
DC: Dancing country
DI: Disco
DR: Dancing reggae
EL: Easy listening
FO: Female opera
HM: Heavy metal
FU: Fusion
HO: Heavy orchestra
HT: Hard techno
LO: Light orchestra
MA: Mambo
MO: Male opera
PI: Piano
PO: Late pop
RA: Rap
RB: R&B
RE: Soft reggae
RO: Hard rock
RR: RegRap mix
RU: Rumba
SA: Samba
SC: Soft country
SR: Soft rock
ST: Soft techno
SW: Swing

Figure 3: Confusion matrix.

This is important because, although the training data have been carefully selected, there are variations from song to song, even if they belong to the same genre. As the dimension increases, the degrees of freedom of the reference vectors also increase and they become more adapted to the training data. As a consequence, it is not possible to foresee which will be the behavior of the strategy face to new signals that almost certainly will have characteristics that are at least a little diverse from the data used to train the technique. If the summary feature space dimension is kept adequately low, the vectors generated by each signal will not be as diverse or particular, and the strategy has its generality and robustness improved. Those observations have motivated a careful selection of the features, as described in Section 5.5.

7.2. Magnatune database

The Magnatune database consists of 729 pieces labeled according to 6 genres. In order to determine the accuracy of the strategy applied to such a database, a mapping between the lower-level genres of Figure 1 and the 6 genres of the Magnatune database must be performed. Table 2 shows how this mapping is performed.

The mapping does not provide a perfect match between the characteristics of the classes in the first and second columns of Table 2, because the taxonomic structure used here was not created having in mind the content of the Magnatune database. This can cause a little drop in the accuracy, but significant conclusions can still be inferred. Additionally, the Magnatune's "World Music" set contains a bunch of ethnical songs (Celtic, Indian, Ukrainian, African, etc.) that have no counterpart in the database used to train the strategy. Because of that, a new set containing world music segments was built and a new training was carried out in order to take into account those different genres. Table 3 shows the confusion matrix associated to the Magnatune database.

As can be seen, the accuracy for all classes but classical lies between 70% and 80%. The overall accuracy of the strategy was 82.8%, which is competitive with the best results achieved by the proposals of ISMIR [25]. This value is

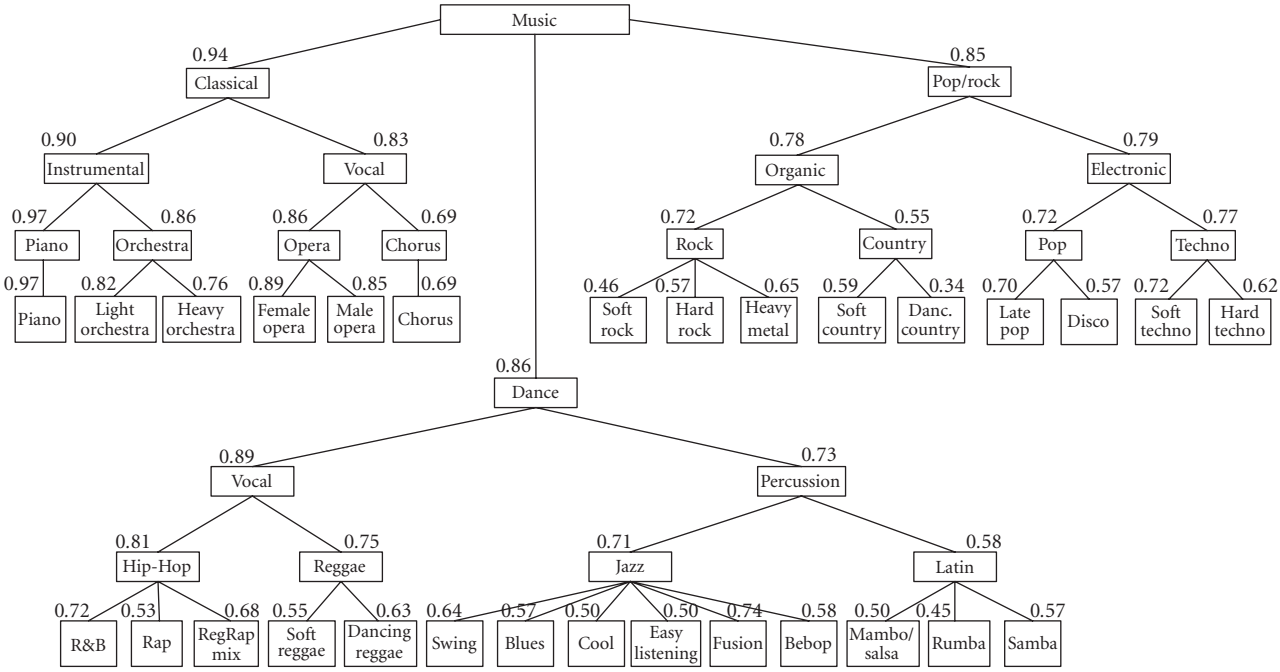

Figure 4: Accuracy for each genre and class in all layers.

Table 1: Overall accuracy.

Layer	Accuracy
1	87 %
2	80 %
3	72 %
4	61 %

Table 2: Mapping between class structures.

Complete database classes	Magnatune database classes
Piano, light orchestra, heavy orchestra, female opera, male opera, chorus	Classical
Soft techno, hard techno, rap, regRap mix	Electronic
Swing, blues, cool, easy listening, fusion, bebop	Jazz/blues
Hard rock, heavy metal	Metal/punk
Soft rock, soft country, dancing country, late pop, disco, R&B	Rock/pop
Mambo/salsa, rumba, samba, soft reggae, dancing reggae, world*	World

* This class was created to embrace the several ethnical genres present in the Magnatune's world music set.

higher than the simple average of individual class accuracies because almost half of the Magnatune database consists of classical songs, for which the accuracy is higher. The performance is also competitive with other recent works [26, 27]. Those results validate all observations and remarks stated in Section 7.1, making the strategy presented here a good option for music classification.

7.3. Computational effort

Finally, under the point of view of computational effort, the strategy has also achieved relatively good results. The program, running on a personal computer with an AMD Athlon 2000+ processor, 512 MB of RAM, and Windows XP OS, has taken 15.1 seconds to process an audio file of 32 seconds: 8.9 seconds for extracting the features and 6.2 seconds for the classification procedure. The time required varies practically linearly with the length of the signal.

The training cycle has taken about 2 hours to be completed. Such a value varies linearly with the number of training signals and exponentially with the number of genres. Since the training can be normally precomputed, and is in general carried out only once, the time required by this step is not as critical as the runtime of the final trained program.

Table 3: Results using the Magnatune database.

	Class.	Elect.	Jazz	Metal	Rock	World
Class.	**0.93**	0	0.01	0	0	0.06
Elect.	0	**0.73**	0	0.09	0.16	0.02
Jazz	0.04	0	**0.80**	0	0.08	0.08
Metal	0	0.06	0	**0.79**	0.15	0
Rock	0	0.02	0.02	0.13	**0.76**	0.07
World	0.02	0.06	0.08	0	0.11	**0.73**

8. CONCLUSIONS AND FUTURE WORK

This paper presented a new strategy to classify music signals into genres. The technique uses four features that are integrated over 1-second analysis segments, generating 12-summary feature vectors for each segment. Those summary feature vectors are used in a classification procedure that considers all possible pairs of genres separately, gathering information that is used to infer the correct genre of the signal. The taxonomy adopted has a four-layer hierarchical structure, with 29 genres in the lowest layer. The classification procedure follows a bottom-up approach, meaning that the signals are firstly classified according to the division of the lowest hierarchical level, and the upper classes are determined simply as a consequence of such first classification.

The technique has reached very good results, even for the lower hierarchical layers. Particularly, analyzing the results achieved for the third layer, which is composed of 12 classes, the strategy appears to perform at least as well as humans, as indicated in [21]. The good performance results from the combination of a very deep and wide taxonomy, a quite effective strategy to stress the differences between the genres, a simple and efficient hierarchical classification, and a relatively low-dimensional summary feature space.

Despite the good results achieved by the proposed technique, further improvement is still possible. The first and more obvious direction for new research is the development of new features able to extract more useful information from the signals. Those new features could be based, for example, on psychoacoustic properties of human hearing. Another direction for future research is expanding the number of genres and the number of hierarchical levels, taking into account musical genres that were not considered in the present work. Moreover, it is expected that a deeper hierarchical structure can improve even more the accuracy achieved for the upper hierarchical layers. Another interesting line of research is the extraction of features directly from the compressed domain of songs submitted to perceptual coders like MP3, AAC, WMA and Ogg-Vorbis.

ACKNOWLEDGMENT

Special thanks are extended to FAPESP for supporting this work under Grant 04/08281-0.

REFERENCES

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

[2] K. West and S. Cox, "Features and classifiers for the automatic classification of musical audio signals," in *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR '04)*, Barcelona, Spain, October 2004.

[3] T. Lambrou, P. Kudumakis, R. Speller, M. Sandler, and A. Linney, "Classification of audio signals using statistical features on time and wavelet transform domains," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '98)*, vol. 6, pp. 3621–3624, Seattle, Wash, USA, May 1998.

[4] H. Deshpande, R. Singh, and U. Nam, "Classification of musical signals in the visual domain," in *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX '01)*, Limerick, Ireland, December 2001.

[5] B. Logan, "Mel-frequency cepstral coefficients for music modeling," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR '00)*, Plymouth, Mass, USA, October 2000.

[6] G. Agostini, M. Longari, and E. Pollastri, "Musical instrument timbres classification with spectral features," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 1, pp. 5–14, 2003.

[7] D. Pye, "Content-based methods for the management of digital music," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '00)*, vol. 4, pp. 2437–2440, Istanbul, Turkey, June 2000.

[8] 2005 MIREX Contest Results - Audio Genre Classification, http://www.music-ir.org/evaluation/mirex-results/audio-genre/index.html.

[9] F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer, "Evaluating rhythmic descriptors for musical genre classification," in *Proceedings of the 25th International AES Conference*, London, UK, June 2004.

[10] O. Hellmuth, E. Allamanche, J. Herre, T. Kastner, N. Lefebvre, and R. Wistorf, "Music genre estimation from low level audio features," in *Proceedings of the 25th International AES Conference*, London, UK, June 2004.

[11] E. Pampalk, "Computational models of music similarity and their application to music information retrieval," Doctoral thesis, Vienna University of Technology, Vienna, Austria, 2006.

[12] S. Dixon, F. Gouyon, and G. Widmer, "Towards characterisation of music via rhythmic patterns," in *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR '04)*, Barcelona, Spain, October 2004.

[13] A. Berenzweig, D. Ellis, B. Logan, and B. Whitman, "A large scale evaluation of acoustic and subjective music similarity measures," in *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR '04)*, Barcelona, Spain, October 2004.

[14] C. McKay, R. Fiebrink, D. McEnnis, B. Li, and I. Fujinaga, "ACE: a framework for optimizing music classification," in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR '05)*, London, UK, September 2005.

[15] S. Lippens, J. P. Martens, T. De Mulder, and G. Tzanetakis, "A comparison of human and automatic musical genre classification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 4, pp. 233–236, Montereal, Quebec, Canada, May 2004.

[16] C. Xu, N. C. Maddage, and X. Shao, "Automatic music classification and summarization," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 441–450, 2005.

[17] T. Pohle, E. Pampalk, and G. Widmer, "Evaluation of frequently used audio features for classification of music into perceptual categories," in *Proceedings of the 4th International Workshop on Content-Based Multimedia Indexing (CBMI '05)*, Riga, Latvia, June 2005.

[18] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 504–516, 2002.

[19] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, vol. 3, no. 3, pp. 27–36, 1996.

[20] F. Pachet and D. Casaly, "A taxonomy of musical genres," in *Proceedings of the 6th Conference on Content-Based Multimedia Information Access (RIAO '00)*, Paris, France, April 2000.

[21] J.-J. Aucouturier and F. Pachet, "Representing musical genre: a state of the art," *Journal of New Music Research*, vol. 32, no. 1, pp. 83–93, 2003.

[22] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '97)*, vol. 2, pp. 1331–1334, Munich, Germany, April 1997.

[23] T. V. Thiede, *Perceptual audio quality assessment using a non-linear filter bank*, Ph.D. thesis, Technical University of Berlin, Berlin, Germany, 1999.

[24] ISMIR 2004 Magnatune Genre Classification Training Set, http://ismir2004.ismir.net/.

[25] ISMIR 2004 Contest Results - Audio Genre Classification, http://ismir2004.ismir.net/genre_contest/results.htm.

[26] T. Lidy and A. Rauber, "Evaluation of feature extractors and psycho-acoustic transformations for music genre classification," in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR '05)*, pp. 34–41, London, UK, September 2005.

[27] E. Pampalk, A. Flexer, and G. Widmer, "Improvements of audio-based music similarity and genre classification," in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR '05)*, pp. 628–633, London, UK, September 2005.

Jayme Garcia Arnal Barbedo received a B.S. degree in electrical engineering from the Federal University of Mato Grosso do Sul, Brazil, in 1998, and M.S. and Ph.D. degrees for research on the objective assessment of speech and audio quality from the State University of Campinas, Brazil, in 2001 and 2004, respectively. From 2004 to 2005 he worked with the Source Signals Encoding Group of the Digital Television Division at the CPqD Telecom & IT Solutions, Campinas, Brazil. Since 2005 he has been with the Department of Communications of the School of Electrical and Computer Engineering of the State University of Campinas as a Researcher, conducting postdoctoral studies in the areas of content-based audio signal classification, automatic music transcription, and audio source separation. His interests also include audio and video encoding applied to digital television broadcasting and other digital signal processing areas.

Amauri Lopes received B.S., M.S., and Ph.D. degrees in electrical engineering from the State University of Campinas, Brazil, in 1972, 1974, and 1982, respectively. He has been with the Electrical and Computer Engineering School (FEEC) at the State University of Campinas since 1973, where he has served as a Chairman in the Department of Communications, Vice-Dean of the Electrical and Computer Engineering School, and currently is a Professor. His teaching and research interests include analog and digital signal processing, circuit theory, digital communications, and stochastic processes. He has published over 100 refereed papers in some of these areas and over 30 technical reports about the development of telecommunications equipment.