STATISTICAL ANALYSIS OF FEATURES USED IN AUTOMATIC AUDIO GENRE CLASSIFICATION

J. G. A. Barbedo, A. Lopes

Abstract - This paper presents statistical models for some of the most important features used to classify audio signals into musical genres. The genres used here are selected according to a given taxonomy. The features are computed for each genre using the signals from a dataset and the results are grouped into histograms. Each proposed statistical model consists of an estimated Probability Density Function (PDF), optimized to best fit a determined histogram and the optimization criterion is the minimization of the Mean Square Error (MSE). Finally, the paper discusses how these models can be applied to classify audio signals into genres.

Keywords: audio signal classification, probability density function, feature extraction.

1. INTRODUCTION

In the last fifteen years, the human-music relationship has changed dramatically. The development of the first perceptual audio coders made possible to store large sets of music with relatively low memory requirements. At the same time, Internet has become available worldwide, making it possible to quickly exchange data and information. As such technologies evolve, the digitally available audio sets have developed into huge continuously growing databases.

In this context, the development of tools able to manipulate large audio datasets in a simple and fast way becomes essential. One of the most promising of such tools is the Automatic Audio Genre Classification (AAGC). The AAGC consists in the extraction of features capable to provide as much relevant information about the audio signals as possible, and in using such features to classify the signals according to a given taxonomy.

The first relevant works in audio classification were published in 1996 [1, 2]. Since then, several advances were achieved, but suitable results have only been reached for very restrict and specific applications. Researches aiming more general and robust classifiers have faced many difficulties. In such cases, the ratio of correct classifications rarely goes beyond 60% [3]. The difficulty to determine well-defined boundaries for the audio genres seems to be among the most important reasons for those relatively poor results [4].

This work presents a statistical study of some commonly used audio features, aiming to support the future development of a new AAGC system. The statistical analysis is divided into 5 main steps:

-Database construction: an audio database was implemented with 16 audio genres without any hierarchical relations. The signals of each genre were carefully chosen to be the as most representative as possible. A detailed description of the database is presented in Section 3.

- Features: eight features were considered – spectral centroid, zero-crossing rate, spectral flux, bandwidth, loudness, spectral roll-off, low energy ratio and high zero-crossing rate ratio. More details about the features and respective extraction are given in Section 4.

-Histogram construction: the values of each feature are grouped into histograms. This step is described in Section 5.

- PDF estimation: selection of the PDFs that best fit the histograms determined previously, according to a minimum mean square error criterion. The PDF estimation is addressed in Section 6.

- Performance measure: assessment of how precisely each PDF approximates the respective histogram. It has been observed that the value of the minimum mean square error alone is not enough to provide a conclusive assessment. Because of that, the MSE results are always analyzed together with the conclusions of a visual inspection of the curves. The performance analysis is detailed in Section 7.

Those five steps aim to outline a behavioral profile of each feature in order to guide their use in future researches. The conclusions about the tests are presented in Section 8.

2. MOTIVATIONS FOR THIS WORK

Using statistical models in AAGC is not a new idea. Actually, it is the main idea underlying several proposals [3, 5-13]. Among such statistical tools, the Gaussian Mixture Model (GMM) classifiers are the most largely used. These classifiers consist of a mixture of several multidimensional Gaussian distributions that tries to model the statistical behavior of an entire set of features at once. Other statistical classifiers follow the same basic ideas.

The approach presented in this work is somehow different from the principles used in previous works. Here, the statistical analysis is performed considering each feature individually and the result is an estimated PDF for each feature and genre. As a consequence, when a set of features is extracted from a signal, it will be possible to calculate automatically, for each feature, the probability that the signal belongs to a given genre. There are several possible ways to combine such probabilities. This fact assures a high degree of flexibility and the possibility of a finer manipulation of the data, increasing the chance of obtaining a more precise classifier. Preliminary studies have revealed that some strategies of combination lead to high degrees of correct classification. An example of classifier, with respective results, is briefly described in Section 7.1.

Another aspect to be considered is the fuzzy nature of

A. Lopes and J. G. A. Barbedo are with the Deptartment of Communications of the School of Electrical and Computer Engineer of the State University of Campinas. E-mails: {jgab,amauri}@decom.fee.unicamp.br. Phone: +55 19 3521 3813.

modern music, that is, much of the music produced nowadays has characteristics that derive from more than one music genre. The approach here proposed enables the development of a method capable to provide fuzzy classifications, of the type "70% genre 1 and 30% genre 2". This kind of approach provides a relative classification that, alone or combined to another information, can benefit several applications. For example, automatic equalizers could select the best equalization according to the relative composition of the signals. This line of research is also currently being carried out and a new classifier based on such principles is already being implemented. This approach and some results are presented in Section 7.2.

It is important to emphasize that the main goals of the present work is to present the guidelines to properly determine Probability Density Functions for individual audio genres and features, and to show the advantages and flexibility of using such individual PDFs in the classification of audio signals. Because of that, no optimized classifier is presented. The two classifiers described in Section 7 are relatively crude, and were included to inspire possible future strategies using the PDFs, and also to provide some cues about the performance of the PDFs when used in actual audio classifiers. Better overall results are expected for carefully designed classifiers.

3. AUDIO DATABASE CONSTRUCTION

The 2,587 database audio files used in this work are in the *wav* format, with a 16-bit quantization and sampled at 48 kHz. The whole database has 13.5 GB, corresponding to more than 20 hours of audio. The signals were extracted from Compact Discs, from Internet radio streaming and also from coded files.

The database can be divided into two main groups: speech files and music files. The only difference between these two groups, besides their content, is the duration of the signals. Speech signals have lengths between 9 s and 21 s, while all music signals have duration of 32 s.

As said before, the signals of each genre of the database were carefully chosen according to the specific characteristics of the genre. After a first selection, all signals have been submitted to a second pruning process, in order to guarantee maximum uniformity inside each genre dataset.

It is important to highlight that such fine selection excludes from the tests most of the audio signals usually present in audio datasets. This is so because the objective of this work is to determine the behavior of the features face to typical elements of each genre. The results obtained for those "pure" elements can therefore be used to classify genre signals with unclear definitions, according to a given criterion. For instance, a given signal could receive a fuzzy classification of the type "70% genre 1 and 30% genre 2", as commented in Section 2. The strategy presented in Section 7.2 shows a possible way to implement such fuzzy approach.

Table 1 shows the composition of the database. The second column shows the number of files after the first selection, while third column shows this number after the second pruning. A brief description of each genre is presented in the following.

Genre	Files after 1st trial	Files after 2nd trial
Classical	128	98
Light Country	68	61
Danc. Country	69	62
Heavy Metal	114	77
Jazz	148	88
Latin	137	101
New Age	127	96
Opera	120	66
Pop	128	99
Rap	114	82
Reggae	126	82
Rock	142	91
Soft Rock	117	78
Soft	125	67
Techno	140	89
Female Speech	180	61
Male Speech	304	119

Table 1. Composition of the audio database.

Classical: songs that are generally played by symphonic and philharmonic orchestras, with predominance of classical instruments as violins, violoncellos, piano, flutes, etc. Solo works are also included. There are no vocals in this genre.

Country: songs typical from the south of USA. This group has some degree of similarity with rock, with strong presence of electric guitar, producing a particular kind of timbre and rhythm that characterizes this kind of music. This genre can be subdivided into two smaller groups, one composed by soft and slow songs, and other composed by songs with dancing characteristics.

Heavy Metal: this group is a ramification of rock, where electric guitars, drums and vocals generate accelerated, intense and aggressive music, with predominance of bass tones (low frequencies). The vocals are often expressed in the form of screams and grumbles.

Jazz: includes jazz and blues songs. This genre is quite difficult to be characterized, because different jazz styles can have very diverse characteristics. According to the definition found in the wikipedia website [14], Jazz is characterized by blue notes (notes sung or played at a lower pitch than those of the major scale for expressive purposes), syncopation (stressing of a normally unstressed beat in a bar or the failure to sound a tone on an accented beat), swing (rhythmic device in which the duration of the initial note in a pair is augmented and that of the second is diminished), call and response (succession of two distinct phrases usually played by different musicians, where the second phrase is heard as a direct commentary on or response to the first), polyrhythms (simultaneous sounding of two or more independent rhythms), and improvisation (act of making something up as it is performed). There is a predominance of instruments like piano and saxophone. Solo works are included.

Latin: group composed by Latin rhythms like salsa, mambo, samba and rumba. The songs of this genre have strongly percussive and dancing characteristics, with

intensive use of drums and percussive instruments. Guitars are also often used.

New Age: this group includes several songs considered "non-conventional". In general, most songs of this genre are soft and very related to classical music, but using elements of electronic music and several instruments considered exotic. Some songs of this group include very soft vocals.

Opera: this group is similar to the classical genre, but with a strong presence of vocals.

Pop: generally, the denomination "pop" is normally adopted for a wide range of songs, not always having a clearly bounded definition. In the context of this work, a song is classified as pop when there is a strong presence of electronic elements (as synthesizers), with a markedly dancing beat, but with an intensity and a number of beats by time unit smaller than those ones found in techno songs, as will be seen later. The presence of vocals is common, but not mandatory.

Rap: this group includes rap, hip-hop and funk; it is characterized by a regular and marked percussive beat, which is normally generated by electronic instruments. There are always vocals, which sometimes look like speech.

Reggae: genre originated in Jamaica, it is the result of a rhythmic and strongly percussive combination of electric guitars, drums and electronic instruments. The peculiar rhythm of this kind of song is reinforced by their particular vocals.

Rock: this group is characterized by the dominance of electric guitars and drums. It has the same basic elements of heavy metal, but it is less aggressive and with a more regular beat. This genre is located in an intermediary point between heavy metal and soft rock. In the context of this work, songs with electronic elements are discarded from this group.

Soft Rock: in this case, electric guitar and drums are also dominant, but with a very soft and slow beat.

Soft: this group, like soft rock genre, is composed by very soft and slow songs. However, in this case electric guitars are not present, and the drums, if present, are used in a very discrete and soft way. Other instruments that may be present in this set are piano, soft percussive instruments and, in some cases, soft electronic elements.

Techno: in this kind of music only electronic instruments are used. This genre has a very accelerated, repetitive and clearly determined beat. The presence of vocals is rare. It includes sub-genres as trance, house and techno itself.

Female Speech: several languages are present in this group, as English, Portuguese, French, Spanish, German, Italian, Japanese, Chinese, Russian, and others. Files with environmental noise were eliminated after the second pruning.

Male Speech: this group has analogous characteristics of female speech genre.

4. FEATURE EXTRACTION

The literature presents several features that can be extracted from audio signals for classification purposes. Eight of the most used features are studied here. They can be divided into two main groups:

Short-term features: their extraction is performed after the division of signal into 21.3 ms analysis frames with a Hanning window. The frames are 50% superimposed. The features are individually extracted for each of those frames. Therefore, each signal will generate a set of parametric values for each feature to be tested.

Medium-term features: in this case, the signals are also divided into 21.3 ms frames. After this division, the resulting frames are grouped into one-second segments, named texture windows [3]. The features are calculated for each analysis frame, and then they are combined along each texture window according to a given criterion. Therefore, in this case the signals will generate a set of values that is smaller than that one obtained for the short-term features.

It is important to emphasize that it is a common practice to combine the values of the short and medium-term features into a single value for the whole signal, usually by means of mean and variance calculation. In the present work, the values obtained for the frames and texture windows are used to construct histograms, as will be seen in Section 4. It is also important to underline that some researchers propose a few long-term features, whose extraction is performed along the whole signal. Such features were not included in this phase of the work.

The extraction of each feature is described in the following.

4.1 SHORT-TERM FEATURES

1) *Spectral Centroid*

The spectral centroid has been used in several works [2, 3, 15, 16, 17]. It represents the "mass center" of the spectral energy distribution of the signals, and is given by

$$ce_i = \frac{\sum_{k=1}^{K} k \cdot |X_i(k)|^2}{\sum_{k=1}^{K} |X_i(k)|^2}, \qquad (1)$$

where $|X(k)|$ is the magnitude of the k^{th} spectral line resulting from an FFT (Fast Fourier Transform) applied to the frame i of the signal $x(n)$, and K is half the total number of spectral lines. High centroid values indicate more "bright" textures, with significant presence of high frequency components. The spectral centroid is given in terms of spectral lines. To obtain the value in Hz, *ce* must be multiplied by the difference in Hz between two consecutive spectral lines.

2) *Zero-Crossing Rate (ZCR)*

Examples of the use of this feature can be found in [3, 18, 19]. A zero crossing occurs whenever the amplitudes of two consecutive temporal samples have opposed signs, as indicated by the expression

$$zcr_i = 0.5 \cdot \sum_{n=1}^{N} \left| sgn\left[x_i(n)\right] - sgn\left[x_i(n-1)\right] \right|, \qquad (2)$$

where $x_i(n)$ represents the samples of frame i of signal $x(n)$ and $sgn(x)$ equals –1 or +1 as x is negative or positive, respectively. This feature is very effective to discriminate between music and speech.

3) *Spectral Roll-off*

This feature is defined as the frequency R_i below which 95% of the magnitude distribution is concentrated, as expressed by

$$\sum_{k=1}^{R_i}\left|X_i(k)\right| = 0.95\cdot\sum_{k=1}^{K}\left|X_i(k)\right|. \quad (3)$$

Spectral roll-off was adopted, with some variation, in [3, 20].

4) *Spectral Flux*

This feature quantifies the changes in the spectral shape between consecutive frames. It is defined as

$$fe_i = \sum_{k=1}^{K}\left\{\log_{10}\left|X_i(k)\right| - \log_{10}\left|X_{i-1}(k)\right|\right\}^2. \quad (4)$$

This feature was used, with some variation, in [3, 18, 20].

5) *Bandwidth*

This feature determines the bandwidth of the signal. There are several definitions for this feature, but all use spectral centroid (ce_i) as one of the variables [1, 16, 17, 21]. In the present work, the adopted expression is

$$lb_i = \sqrt{\frac{\sum_{k=1}^{K}\left[(ce_i - k)^2\cdot\left|X_i(k)\right|^2\right]}{\sum_{k=1}^{K}\left|X_i(k)\right|^2}}. \quad (5)$$

Equation 5 gives the bandwidth in terms of spectral lines. To get the value in Hz, *lb* must be multiplied by the difference in Hz between two consecutive spectral lines.

6) *Loudness*

This feature measures the signal intensity the way it is perceived by a human listener. The strategy here adopted is slightly different from those ones used in previous works [1, 22]. The first step to calculate this feature is modeling the frequency response of outer and middle ears of an average person. Such response is given by [23]

$$W(k) = -0.6\cdot 3.64\cdot f(k)^{-0.8} - 6.5\cdot e^{-0.6\cdot(f(k)-3.3)^2} + 10^{-3}\cdot f(k)^{3.6}, \quad (6)$$

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

$$f(k) = k\cdot d. \quad (7)$$

being d the difference between two consecutive spectral lines.

The frequency response is used in the calculation of the loudness as a weighting function to emphasize spectral components for which the ear is more sensible and to attenuate the less audible ones. The loudness of each frame is calculated according to

$$L_i = \sum_{k=1}^{K}\left|X_i(k)\right|^2\cdot 10^{\frac{W(k)}{20}}. \quad (8)$$

4.2 MEDIUM-TERM FEATURES

As commented before, the medium-term features are extracted from 1 s segments called texture windows. Since each frame has 21.3 ms and are 50 % superimposed, there will be 96 frames inside each segment. In the following equations, two variables are used to delimitate the texture window: n_1 and n_2 are the indices of the first and last frame of a given texture window. Therefore, for the first texture window, $n_1 = 1$ and $n_2 = 96$, for the second texture window, $n_1 = 97$ and $n_2 = 192$, and so on. Generalizing, $n_1 = 1 + 96\cdot(m-1)$ and $n_2 = 96\cdot m$, where m is the index of the texture window.

1) *High Zero-Crossing Rate Ratio (HZCRR)*

This feature derives directly from the zero-crossing rate feature and was proposed in [18]. It is defined as the number of frames whose number of zero crossings is greater than 1.5 times the average zero-crossing rate along a texture window, and is given by

$$tq_m = \frac{1}{2\cdot(n_2-n_1+1)}\cdot\sum_{i=n_1}^{n_2}\left\{\mathrm{sgn}\left[zcr_i - 1.5\cdot\left(\frac{1}{n_2-n_1+1}\cdot\sum_{i=n_1}^{n_2}zcr_i\right)\right]+1\right\}. \quad (9)$$

2) *Low Energy Ratio*

This feature measures the percentage of analysis frames whose energy E_i is less than half the average energy of the corresponding texture window. This approach was used in [18], and is given by

$$le_m = \frac{1}{2\cdot(n_2-n_1+1)}\cdot\sum_{i=n_1}^{n_2}\left\{\mathrm{sgn}\left[0.5\cdot\left(\frac{1}{n_2-n_1+1}\cdot\sum_{i=n_1}^{n_2}E_i\right) - E_i\right]+1\right\}\cdot(10)$$

5. HISTOGRAM CONSTRUCTION

As commented before, all features are individually extracted for each analysis frame or texture window of the signals. All the values of each feature of each audio genre are summarized into a histogram. The histograms have 2,000 bins, value adopted according to the number of data values available. An optimization procedure was carried out, having as main goal to find the best compromise between data quantization (bin widths) and meaningfulness of the histogram shape. If too large bin widths were adopted, too many data samples would be grouped in a same bin, and the shape of the histogram would not represent adequately the distribution of the data; on the other hand, if a too fine data quantization was used, the data would be excessively spread along the bins, resulting in a meaningless histogram shape. In this context, the value of 2,000 bins has shown the best characteristics. This results in a bin width that is 1/2000 of the expected range for the feature values. This is equivalent to quantize the feature values with steps of 1/2000 of its value range. The lower limit of a feature value range is always 0, and the upper limit is given by adding 10 % to the highest value obtained for that feature. A moving average of five data points is applied in order to smooth the shape of the histogram. The value of five points was adopted because it results in an effective smoothing of the histogram, keeping at the same time its basic shape, which is important to determine the most adequate PDF. Fig. 1 shows an example of a histogram, obtained for the feature centroid applied to the classical genre.

The histograms will be approximated by PDFs as described in the next section.

6. PDF ESTIMATION

The main goal of this work is to estimate Probability Density Functions (PDF) that best fit the histograms defined in last section. The process starts with a visual analysis of a given histogram and choosing a PDF that could fit that shape. For the example showed in Fig. 1 (spectral centroid), the best candidate appeared to be the lognormal PDF. After that, an algorithm that minimizes the MSE between the PDF and the histogram determines the best parameters for the selected PDF. Fig. 2 shows the PDF that best fits the histogram showed in Fig. 1.

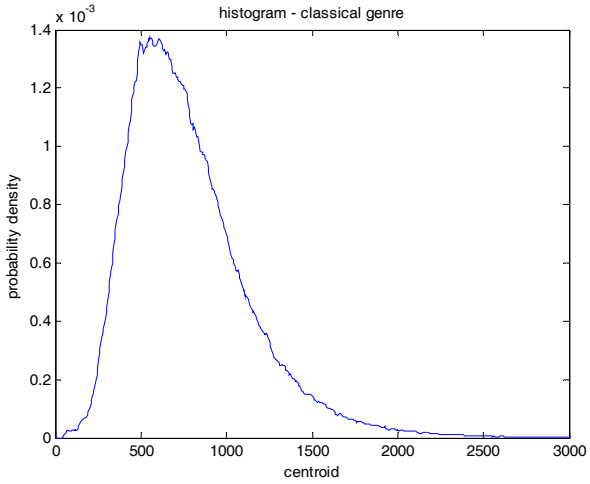

Figure 1. Example of histogram.

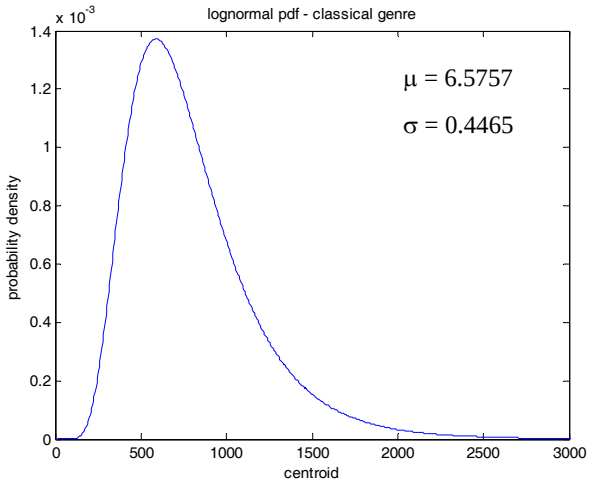

Figure 2. Example of PDF estimation.

The same process is repeated for all features and genres. Section 7 details the tests and achieved results.

It is important to note that, in theory, a better PDF candidate could emerge from the collected data if such information was summarized into a normalized histogram, which would then work as a non-standard PDF. In fact, this option was tested by dividing the audio material into two sets. The first set, corresponding to about 70% of the whole database, was used to determine such histogram-based PDF, and the remaining 30% of the audio files were used to test if such PDF would actually fit the characteristics of the corresponding genre and feature. The observed overall performance, using the same performance criteria presented in Section 7, was poorer than that achieved using classical PDFs because, although the histogram-based PDF fits quite well the training data, it does not have a good correspondence with the remaining data. This is so because the behavior of audio signals is very fuzzy, even considering only excerpts of a same genre. Such fuzzy behavior is more efficiently modeled by the traditional PDFs, which are more robust to deal with the inherent inconsistencies found among different audio excerpts.

7. TESTS AND RESULTS

The precision of the adjustment between an estimated PDF and the corresponding histogram is assessed by means of a Mean Square Error (MSE) multiplied by 1,000. However, such measure must be carefully used: if considered alone, the MSE can lead to misjudgments about the real performance of the estimated PDF. This occurs because, even after submitted to the moving average smoothing technique, some histograms still preserve an oscillatory behavior. In such cases, the estimated PDF can fit almost perfectly the envelope of the histogram, but not its ripple. An example of this situation, corresponding to the ZCR feature, is shown in Fig. 3. Despite the good agreement between the PDF and the histogram envelope, the MSE is relatively high. However, a visual inspection reveals that the estimated PDF curve reliably models the statistical characteristics of the feature, since the ripple is, in fact, an undesirable behavior for the purpose of signal classification.

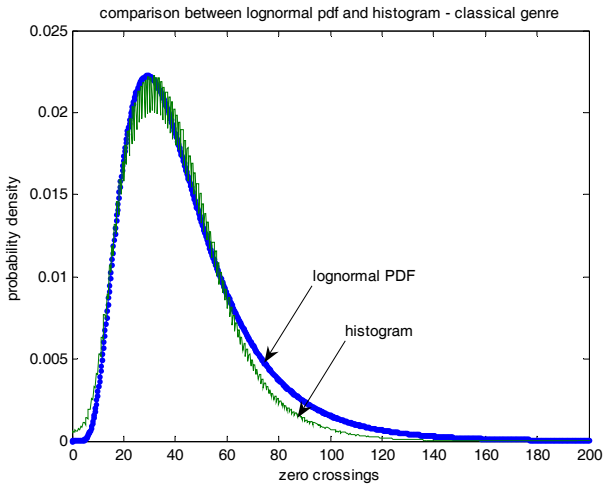

Figure 3. Comparison between histogram and estimated PDF.

The results obtained with the PDF estimation for all situations being considered are summarized in Table 2. Both the MSE value and the visual inspection criteria will be taken into account in the analysis of the results shown in Table 2.

Genre/Feature	*Spectral Centroid*	*ZCR*	*Spectral Roll-off*	*Spectral Flux*	*Bandwidth*	*Loudness*	*HZCRR*	*Low Energy*
Classical	0.2321	2.5044	1.8977	0.4346	2.7603	1.5819	1.9819	0.8581
Light Country	1.8865	1.5590	18.2395	4.1798	1.4842	0.8820	2.8994	2.0226
Dancing Country	2.3554	4.0611	5.5790	6.4158	1.0198	1.7222	3.4051	1.7892
Heavy Metal	4.3000	6.8973	1.4699	7.3621	4.0170	0.7493	0.6507	0.6209
Jazz	2.0323	2.7547	22.9105	2.1643	0.6027	0.7828	5.4653	1.0938
Latin	1.2075	5.8349	2.0290	6.3031	1.8327	0.8666	5.6007	1.6735
New Age	2.3221	2.5271	18.6730	0.5633	1.2961	3.2549	1.1131	0.7895
Opera	0.8887	1.5261	12.5870	3.3081	11.9496	13.0910	1.5964	2.8531
Pop	1.5791	11.6286	7.2360	3.0250	4.4235	1.8193	4.2826	1.0053
Rap	14.2140	9.4122	2.5336	6.4414	3.5193	1.8892	1.7313	3.3288
Reggae	10.1643	13.2286	1.3769	5.2521	3.6414	0.7562	0.8026	1.0781
Rock	6.3961	7.2369	3.9049	3.6824	2.8611	0.7933	1.1657	2.5163
Soft Rock	2.8582	4.3437	1.7702	3.8098	1.8435	1.3270	2.9094	3.3393
Soft	2.9635	3.3335	30.9373	1.7685	0.3242	2.5996	4.7577	2.4687
Techno	8.0124	5.1276	6.1967	7.9484	3.1741	1.5463	3.2628	7.5947
Female Speech	11.7352	2.1033	21.3518	5.6259	1.7249	28.4179	1.2066	4.7291
Male Speech	7.3253	2.2135	32.1390	3.6902	1.7342	17.2550	0.9007	1.1177

Table 2. Individual feature results.

The numbers in Table 2 represent the MSE multiplied by 1,000. The values in italics indicate the use of normal PDF. The other values were obtained using the lognormal PDF.

The analysis of the results for each feature is presented in the following. Due to space limitations, only the most representative figures are presented. The remaining figures can be found in www.decom.fee.unicamp.br/~amauri.

In order to compare the results of Table 2, it is necessary to define a maximum tolerable MSE value. In fact, such a value should be defined taking into account the sensibility of the results produced by estimated PDFs in a practical application, such as a signal classifier. Since such application is still under development, for the purposes of this paper, the maximum tolerable MSE value will be chosen taking into account the unitary area under each PDF curve. It seems that a maximum MSE of 0.007 (or 7 in Table 2) is an adequate reference.

1) *Spectral Centroid*: the MSE values are low for several genres, indicating that the PDF models adequately the feature characteristics. This is particularly true for classical and opera genres. However, for rap, reggae, rock, techno, and male and female speech, further analysis is necessary.

The cases of rap and reggae are similar. Fig 4 illustrates the curves for rap. In this case, the PDF is not able to reliably fit the entire histogram. It can be observed that the decreasing slopes of the curves are quite matched. However, the increasing slopes are not matched at all. If the PDF were moved right in order to fit the increasing slope of the histogram, this would cause a major unfit between the decreasing slopes.

A visual inspection reveals that the error is more intense for centroid values smaller than 200, so this is the area of the plot that must be carefully analyzed. For centroid values smaller than 140 (smaller than the peak), the error is high, but the histogram values associated to such range are small. Therefore, such centroid values have little importance. For centroid values greater than 140, the PDF reproduces with reasonable accuracy the behavior of the histogram

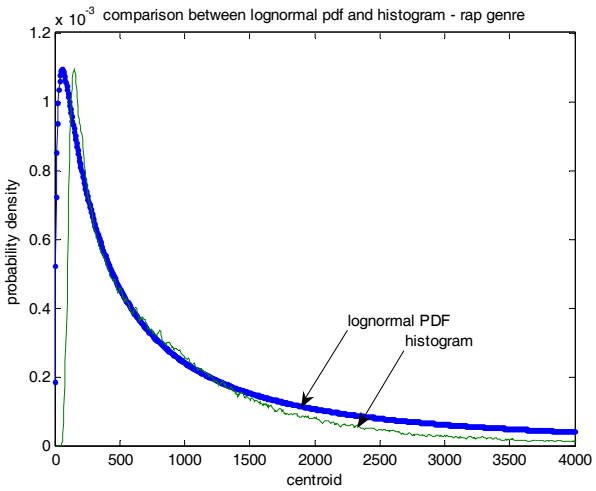

Figure 4. Comparison between centroid curves for rap genre.

Moreover, it must be taken into account that, despite all selection efforts, the group of signals that represent each genre still presents a mild degree of heterogeneity among them.

Therefore, considering that the estimated PDFs worked well for most genres, they may have the ability to compensate possible histogram inconsistencies resulting from such heterogeneity. To test this possibility, further tests were carried out using a new group of rap signals, in order to determine if the discrepancy between the curves is due to a failure of the standard PDF in modeling the data or to the heterogeneity of the data. The results confirmed that the PDF models better the characteristics of the signals than a histogram-based PDF, even in such a case where the match between the curves is not good. This also supports the observations made in Section 6. This observation is also valid for the other 3 genres with higher MSE values. In those further cases, a visual inspection indicates that the

errors are mostly concentrated at low probability regions of the graphics, as can be seen in Fig. 5.

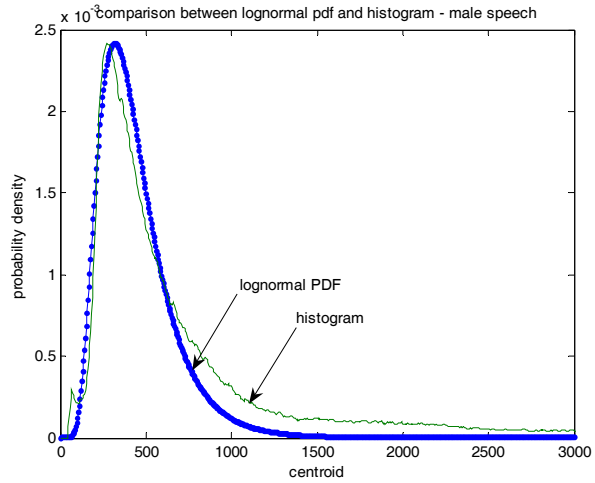

Figure 5. Comparison between centroid curves for male speech.

2) *Zero-Crossing Rate*: the PDFs estimated for this feature also have achieved good fit with most histograms. Only for 4 genres the MSE has indicated possible significant disagreement between the curves. This feature is one of the most affected by the ripple present in the histogram. A visual inspection reveals that all 4 genres with higher MSE values present only mild disagreement between the curves. In this case, the relatively high error values are mostly due to oscillations in the histogram, as can be seen in Fig. 6. Such oscillations seem to be generated by the process of averaging along the texture windows and by the construction of the histogram itself and, therefore, would not be a natural characteristic of the ZCR data. A more detailed analysis revealed that the envelope of the histogram describes very accurately the data behavior, meaning that the oscillations can be treated as a mild side effect of the histogram construction process, which occurs only for ZCR. Under this point of view, the match between the curves can be considered adequate.

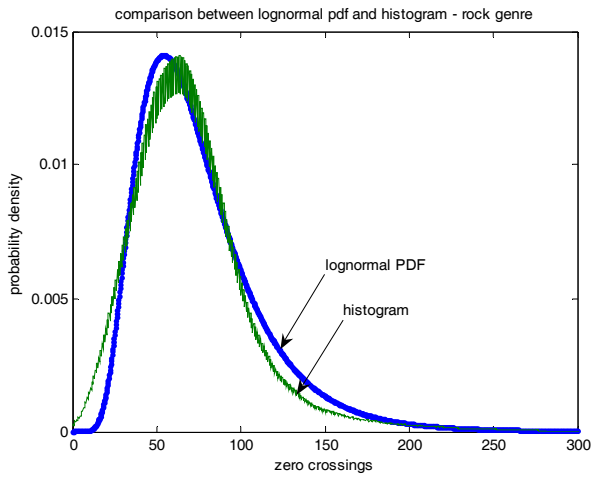

Figure 6. Comparison between ZCR curves for rock genre.

3) *Spectral Roll-off*: about half the PDFs estimated for this feature has achieved good performance. The other half has resulted in very high MSE values. All genres that generated impaired PDFs have shown similar behavior, which is exemplified in Fig. 7.

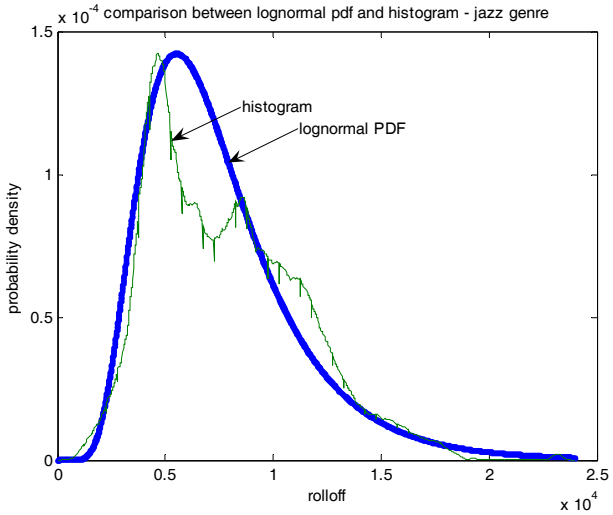

Figure 7. Comparison between roll-off curves for jazz genre.

As can be seen in Fig. 7, the behavior of the histogram for roll-off values between 4,000 and 12,000 is not adequately approximated by the selected PDF function. To investigate the reasons for such behavior, further tests were carried out using a new set of audio files. A non-standard PDF, based on the actual histogram presented in Fig. 7, was determined to be compared with the lognormal PDF of Fig. 7. Both curves have showed practically the same performance in modeling the data. This means that the lognormal PDF is indeed compensating, at least partially, the heterogeneity present in the data used to determine its shape. However, it is important to emphasize that the results obtained for this feature are poorer than expected and further studies will be necessary to find a better solution for its modeling.

4) *Spectral Flux*: almost all PDFs obtained for this feature fit well with the respective histograms. Only heavy metal and techno presented high values. In both cases, the curves have a moderate disagreement for high values of spectral flux. It is a situation quite similar to that one shown in Fig. 5, but in this case there is practically no disagreement for middle and low values, as can be seen in Fig. 8.

5) *Bandwidth*: only opera PDF resulted in high MSE value. This is due to the steepness of the descent slope, as can be seen in Fig. 9. Since the disagreement is located in a region of low probability, the effect of the error is not significant.

6) *Loudness*: the PDFs obtained for this feature have achieved excellent results for all genres but three: opera, male speech and female speech. All three genres have one point in common - a strong predominance of human voices. In the case of speech, the loudness varies considerably depending on the type of phoneme - fricative, voiced, unvoiced - that is being pronounced. Besides, there are

always several intervals of silence. Therefore, it is very difficult to detect some kind of pattern or tendency for the loudness of this kind of signals. In the case of opera signals, a similar analysis can be performed. Additionally, the singer often varies greatly the intonation of its performance, making even more difficult to identify a pattern. Those observations lead to the conclusion that some caution is needed when applying this feature to signals with strong voice elements.

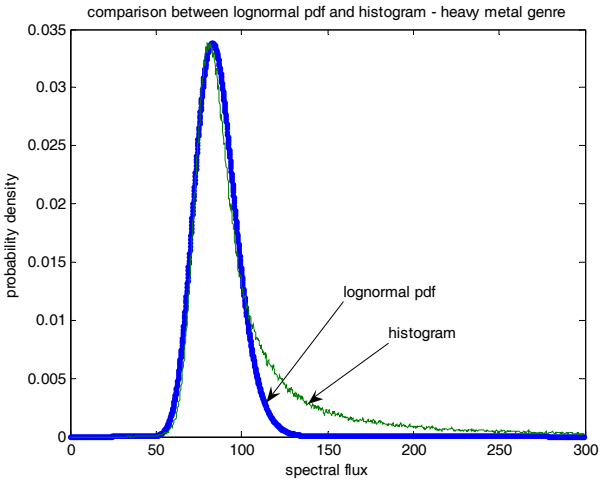

Figure 8. Comparison between spectral flux curves for techno genre.

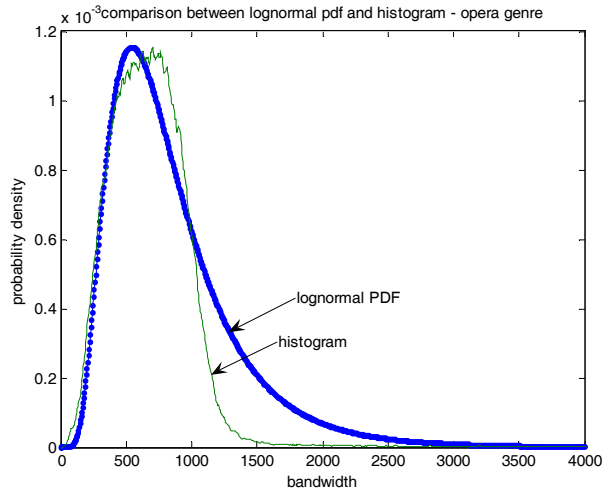

Figure 9. Comparison between bandwidth curves for opera genre.

7) *High Zero-Crossing Rate Ratio*: all PDFs generated for this feature match quite well the respective histograms. No important disagreements have been observed.

8) *Low Energy Ratio*: almost all PDFs have shown good performance. Only techno genre has presented a MSE value a little above the threshold, but most of the error is due to an expressive ripple present in the histogram.

An important point to be analyzed is if the PDFs obtained for a given feature are diverse enough among them to provide a good distinction between the genres. It has been observed that there are genres with very particular PDF shapes, but, on the other hand, there are also genres with very similar PDFs. Therefore, each feature is capable to successfully distinguish between some genres, but not all of them. This indicates that, to work properly, several features must be taken into account at same time. The way to properly combine such features is currently being conducted, and the first results are presented in Section 7.1. An example of a fuzzy classifier and respective results is presented in Section 7.2.

7.1 STRATEGY 1

This strategy is quite simple and direct, consisting of the following basic steps:

1. The signal is divided into 21.3 ms frames, with an overlap of 50% between consecutive frames.
2. All short-term features are extracted for the 96 first frames (first texture window), and their values are averaged over the segment.
3. The medium-term features are extracted for the first texture window.
4. The resulting value of each feature over the texture window, together with the corresponding PDF, generates a probability value for each feature. Since there are 8 features, there will be 8 probability values associated to each genre.
5. A single probability value is determined for each genre by multiplying the respective probabilities.
6. The genre with greater probability value is taken as winner for that texture window.
7. Steps 2 to 6 are repeated for the remaining texture windows, generating a winner genre for each 1 s segment.
8. The genre that appears more times as winner is taken as definitive classification for the signal. If there is a draw, the probabilities obtained for the drawn genres are all summed, and the greater value determines the final winner.

Therefore, there are two families of random variables in the model:

- Genre G, which is discrete and assumes values g_j, being described by an individual probability;
- Features H_i, which are continuous and are described by probability densities.

The strategy tries to choose j that maximizes $P\left(G = g_j \mid H_1 = c_1\right) \cdot \mathrm{L} \cdot P\left(G = g_j \mid H_8 = c_8\right)$. It is assumed that the probabilities of each genre are equal, and that the features are mutually independent. The first assumption holds for the dataset used in this work, and is also reasonable in the real world. The second assumption, on the other hand, is not entirely true, since features like ZCR and spectral roll-off are slightly related. However, this fact does not significantly harm the effectiveness of the strategy. The same observations are valid for strategy 2 (Section 7.2).

The procedure, although simple, has led to quite good results, as shown in Table 3. The overall accuracy is above 60 %, performance that is compatible with the best proposals found in the literature [3, 7, 11].

Genre	Accuracy
Classical	87.3 %
Light Country	45.4 %
Dancing Country	31.2 %
Heavy Metal	67.8 %
Jazz	51.2 %
Latin	52.3 %
New Age	50.1 %
Opera	73.2 %
Pop	55.7 %
Rap	62.9 %
Reggae	62.3 %
Rock	59.8 %
Soft Rock	59.7 %
Soft	61.1 %
Techno	55.5 %
Female Speech	70.2 %
Male Speech	72.6 %
Total	60.5 %

Table 3. Classification accuracy for strategy 1.

7.2 STRATEGY 2

The strategy presented in this section explores the fact that several audio signals have characteristics that fit more than one genre, making it difficult to find a single correct classification. In such cases, it makes sense to classify the signal according to its multiple genre characteristics.

The initial part of the strategy is similar to the 5 first steps presented in Section 7.1. After that, the probability values determined for each genre are multiplied along all texture windows, resulting in a single general probability value for each genre. Then, the following rules are applied:

1. The greater probability value is determined and used to normalize all probabilities.

2. The 3 greater normalized probability values (NPV) and respective genres are determined, and the remaining probability values are discarded.

3. Genres whose NPV are greater than 0.85 are considered as having strong influence over the signal. If such value lies between 0.7 and 0.85, the genre is considered as having weak influence over the signal. Finally, if the NPV is smaller than 0.7, the respective genre is considered as having no influence.

Since the first genre NPV will be always equal to 1, such a genre will always have strong influence. The second and third place genres may have any NPV value. In cases where the signal is clearly influenced by only one genre, it is expected that the second and third NPV values be smaller than 0.7, and therefore the signal is classified according to only one genre. On the other hand, if the signal is composed by elements coming from three different genres, it is possible that three strong genres be assigned to such signal. Since signals influenced by 4 or more genres are quite rare, such situation has not been considered here.

Tests have revealed that in 66 % of the cases, the target label (label manually assigned to each signal during the database assembly) of a signal was among the genres considered having strong influence over the signal; in 14 % of the cases, the target genre was among the genres considered as having weak influence; finally, in only 20 % of the cases, the target genre was not identified as having influence over the signal. In other words, in only 20 % of the cases the procedure has completely misclassified the signals. It was also observed that the strategy is quite effective in correctly identifying signals that are weakly or strongly influenced by more than one genre. Such results are very good, especially if one takes into account that this is a very simple strategy. It is expected that improved strategies be able to achieve better results, making the approach used in this strategy an excellent alternative to the traditional techniques.

8. CONCLUSION

This paper has presented statistical models for several features commonly used in the classification of audio signals. An individual Probability Density Function has been determined for each feature and audio genre. The PDFs were adjusted minimizing the Mean Square Error between each PDF function and the respective feature histogram.

The objective is to use the PDF models to generate audio classifier systems capable to outperform the classifiers presented in the literature. Although such new classifiers are still under development, the paper presented two simple proposals to illustrate the application. Such experiments have revealed that the PDF approach has several desirable characteristics, such as flexibility and the possibility to manipulate the data to be classified in a finer way. They also revealed that the accuracy attained by the proposed statistical models is adequate for this application.

ACKNOWLEDGEMENTS

This work was supported by Fapesp under grants 04/08281-0 and 03/09858-6.

REFERENCES

[1] E. Wold, T. Blum, D. Keislar, J. Wheaton, “Content-Based Classification, Search, and Retrieval of Audio”, *IEEE MultiMedia*, vol. 3, no. 3, pp. 27-36, September 1996.

[2] J. Saunders, “Real-Time Discrimination of Broadcast Speech/Music”, Proceedings of the IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing, vol. 2, 1996, pp 993-996.

[3] G. Tzanetakis and P. Cook, “Musical Genre Classification of Audio Signals”, *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, July 2002.

[4] 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.

[5] G. Backfried, R. Rainoldi, J. Riedler, “Automatic language identification in broadcast news”, *Proceedings of the 2002 International Joint Conference on Neural Networks*, vol. 2, pp. 1406-1410, Honululu, USA, 2002.

[6] A. Berenzweig, B. Logan, D.P.W. Ellis, B. Whitman, A Large-Scale Evaluation of Acoustic and Subjective Music Similarity Measures, *Computer Music Journal*, vol. 28, no. 2, pp. 63-76, 2004.

[7] J. J. Burred and A. Lerch, "Hierarchical Approach to Automatic Musical Genre Classification", *Proceedings of the 6th International Conference on Digital Audio Effects DAFX03*, London, UK, September 2003.

[8] M. Casey, "General sound classification and similarity in MPEG-7", *Organized Sound*, vol. 6, no. 2, pp. 153-164, Aug. 2001.

[9] K. El-Maleh, A. Samouclian, P. Kabal, "Frame-Level Noise Classification in Mobile Environments", *Proc. IEEE Conf. Acoustics, Speech, Signal Proc.*, Phoenix, AZ, USA, March 1999.

[10] J. Makhoul, F. Kubla, T. Leek, D. Liu, L. Nguyen, R. Schwartz, A. Srivastava, "Speech and Language Technologies for Audio Indexing and Retrieval", *Proceedings of the IEEE*, pp. 1338-1353, August 2000.

[11] M. F. McKinney, J. Breebaart, "Features for Audio and Music Classification", *Proceedings of ISMIR*, Baltimore, USA, 2003.

[12] P. J. Moreno, R. Rifkin, "Using The Fisher Kernel Method for Web Audio Classification", *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 2417-2420, 2000.

[13] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi and T. Sorsa, "Computational auditory scene recognition", *Proceedings of IEEE Conf. Acoustics, Speech, Signal Proc.*, Florida, USA, May 2002.

[14] http://en.wikipedia.org/wiki/Jazz

[15] G. J. Lu, T. Hankinson, "A Technique Towards Automatic Audio Classification and Retrieval", *Proc. IEEE Intl. Conf. on Signal Processing*, vol. 2, pp. 1142--1145, 1998.

[16] G. Agostini, M. Longari and E. Pollastri, "Musical Instrument Timbres Classification with Spectral Features", *EURASIP Journal on Applied Signal Processing*, no. 1, pp. 1-11, 2003.

[17] S. Z. Li, "Content-based classification and retrieval of audio using the nearest feature line method", *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 619–625, 2000.

[18] L. Lu, H.-J. Zhang and Hao Jiang "Content Analysis for Audio Classification and Segmentation", *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 504-516, October 2002.

[19] T. Zhang and C.-C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification", *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 441-457, 2001.

[20] E. Scheirer and M. Slaney: "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator", *Proceedings of ICASSP97*, pp. 1331-1334, April 1997, Munich, Germany.

[21] G. J. Lu, T. Hankinson, "A Technique Towards Automatic Audio Classification and Retrieval", *Proc. IEEE Intl. Conf. on Signal Processing*, vol. 2, pp. 1142--1145, 1998.

[22] T. Zhang, C.-C.J. Kuo, "Hierarchical System for Content-based Audio Classification and Retrieval", *Proceedings of SPIE, Multimedia Storage and Archiving Systems III*, pp. 398-409, San Diego, USA, July 1998.

[23] E. Zwicker, H. Fastl, *Psychoacoustics, Facts and Models*, Berlin: Springer Verlag, 1990.

Jayme Garcia Arnal Barbedo received the B.S. degree in electrical engineering from the Federal University of Mato Grosso do Sul, Brazil, in 1998, and the 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 sound source separation. His interests also include audio and video encoding applied to digital television broadcasting and other digital signal processing areas.

Amauri Lopes received his 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 theDepartment 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.