A HIERARCHICAL APPROACH TO AUTOMATIC MUSICAL GENRE CLASSIFICATION

Juan José Burred — Communication Systems Group, Technical University Berlin, Germany, burred@zplane.de

Alexander Lerch — zplane.development, Berlin, Germany, lerch@zplane.de

ABSTRACT

A system for the automatic classification of audio signals according to audio category is presented. The signals are recognized as speech, background noise and one of 13 musical genres. A large number of audio features are evaluated for their suitability in such a classification task, including well-known physical and perceptual features, audio descriptors defined in the MPEG-7 standard, as well as new features proposed in this work. These are selected with regard to their ability to distinguish between a given set of audio types and to their robustness to noise and bandwidth changes. In contrast to previous systems, the feature selection and the classification process itself are carried out in a hierarchical way. This is motivated by the numerous advantages of such a tree-like structure, which include easy expansion capabilities, flexibility in the design of genre-dependent features and the ability to reduce the probability of costly errors. The resulting application is evaluated with respect to classification accuracy and computational costs.

1. INTRODUCTION

In a musical context, audio data is organized mostly according to the musical genre. In recent years, many different approaches have been proposed to perform genre extraction from raw audio data, ranging from music/speech discriminators [1, 2] to systems based on elaborate musical and non-musical taxonomies [3, 4]. All of these systems rely on pattern recognition techniques, in which each signal is represented by a set of features that are used to train a statistical or neural classifier.

Although many combinations of features and classifiers have been evaluated in these works, little attention has been paid to the following issues:

- *Genre dependency of the features:* Clearly, some features will be more suitable than others when classifying into a given set of subgenres. For example, features describing beat strength are more likely to perform better in separating classical from pop music than in classifying into chamber music subgenres. This suggests a hierarchical classification scheme.
- *Problems of dimensionality:* In pattern recognition applications, adding new features (i.e., adding new dimensions in the feature space) does not necessarily result in a higher classification rate, especially when few training samples are available for each class. Reducing the number of features allows to reduce computational costs while maintaining a similar classification rate. In some cases, the classification rate can even benefit from the reduction in dimensionality.
- *Inappropriate taxonomies:* Many proposed taxonomies are too simple or musicologically inconsistent.

2. AUDIO TAXONOMY

A special effort was made in defining the audio taxonomy to be implemented. The class structure was chosen to be simple enough to allow classification with feasible features, complete enough to allow an acceptable classification of as much input signal types as possible, and musicologically consistent. The obtained taxonomy contains a total number of 17 classes (3 speech classes, 13 music classes and 1 background noise class) and is shown in Figure 1.

3. FEATURE EXTRACTION

The system presented here is intended to operate on audio signals stored as audio files. Furthermore, the files are supposed to be homogeneous, i.e., to contain only one type of audio. Thus, a single class decision is met for each file. For each underlying frame-based feature, the four following file-based *subfeatures* are computed: mean (M), standard deviation (S), mean of the derivative (DM) and standard deviation of the derivative (DS) over all analysis frames in the files.

3.1. Timbral features

The following well-known features [1, 3] describing timbral content have been implemented and evaluated :

- *Zero crossings:* An approximate measure of noisiness.
- *Centroid:* Models the sound *sharpness.*
- *Rolloff:* A measure of spectral shape.
- *Flux:* A measure of the spectral rate of change.
- *Mel Frequency Cepstral Coefficients (MFCC):* MFCCs are a compact representation of the spectrum of an audio signal that takes into account the nonlinear human perception of pitch, as described by the *mel scale*.

3.2. MPEG-7 features

The new MPEG-7 standard [5] deals with the content-based description of raw data. In the context of MPEG-7, a feature is called a Low-Level Descriptor (LLD). In the present work, four audio LLDs have been selected for the implementation and evaluation in the classification task. Only descriptors that are applicable to any audio type have been taken into consideration. There are other LLDs intended for single-voiced, quasi-periodic audio

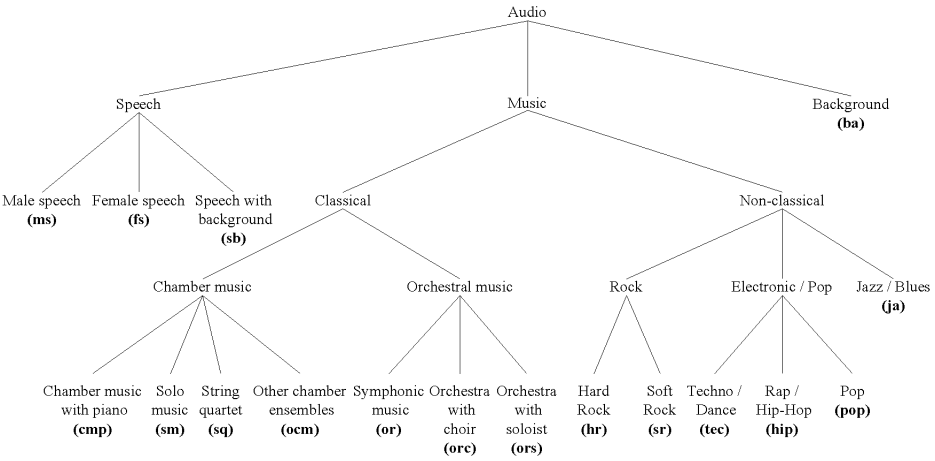

Figure 1: *Audio taxonomy.*

segments, which have not been used here, since they would yield unpredictable results with general, complex signals. The implemented descriptors are:

- *Audio Spectrum Centroid:* A perceptually adapted definition of the centroid.
- *Audio Spectrum Spread:* It describes how the spectrum is concentrated around the centroid.
- *Audio Spectrum Flatness:* A measure of the deviation of the spectral form from that of a flat spectrum.
- *Harmonic Ratio:* A measure of the proportion of harmonic components within the spectrum. It is defined as the maximum value of the autocorrelation of each frame.
- *Modified Harmonic Ratio:* The definition of the harmonic ratio provided within the standard was modified in such a way that the first peak of the autocorrelation was skipped when searching for the maximum. This resulted in a larger value range and has proven to work better than the standard version in the classification task.

3.3. Rhythm features

Instead of just measuring the tempo, it is more interesting for classification purposes to extract information about rhythmical structure and beat strength. A *beat histogram* is a curve describing beat strength as a function of a range of tempo values, and allows the extraction of the mentioned properties. Peaks on the histogram correspond to the main beat and other subbeats. Several methods have been proposed for its computation [3, 6]. In this work, an implementation similar to the one presented in [6] has been used. All the rhythm features were extracted from the beat histograms, as follows:

- *Beat strength:* To obtain an overall measure of beat strength, the following statistical measures of the histogram have been evaluated: mean, standard deviation, mean of the derivative, standard deviation of the derivative, skewness, kurtosis and entropy. These measures are computed in the "beat domain", and should not be confused with the time-based statistical subfeatures mentioned earlier.
- *Rhythmic regularity:* A beat histogram in which the peaks are spaced periodically denotes high rhythmic regularity. This can be measured by the normalized autocorrelation function of the beat histogram. It will contain clear peaks for rhythmically regular music examples, and will be the more linear the weaker the regularity is.

3.4. Other features

- *Root Mean Square:* RMS energy of each signal frame.
- *Time envelope:* maxima of each frame's absolute amplitude.
- *Low energy rate:* Percentage of frames within a file that have an RMS energy lower than the mean RMS energy across that file.
- *Loudness:* A basic exponential model of loudness of the form $L = E^{0.23}$ is used, where E is the energy of the current frame. This model has proven to be highly effective in spite of its simplicity.
- *Central moments:* The third and fourth order central moments of the time-domain audio signal, i.e., its *skewness* and its *kurtosis* are evaluated here as possible audio features.
- *Predictivity ratio:* ratio of the energy of the linearly predicted signal to the energy of the original signal.

4. FEATURE SELECTION

Altogether, 20 frame-based features, plus one file-based feature (low energy rate) and 9 beat-histogram-based features have been evaluated. Furthermore, the 4 subfeatures, which are applicable to all of the 20 frame-based features, make a total number of 90 different available features to implement the classification system. The so-called *curse of dimensionality* is a well-known phenomenon that appears in many pattern recognition applications. It implies that it is advantageous to reduce the number of features in order to reduce computational costs while keeping similar levels of performance, and in some cases even to improve the classification rate. In general, a well-designed feature should be invariant to irrelevant transformations of the signal, it should have good discriminative power between classes, and it should be uncorrelated to other features.

	Best features		Worst features
1.	Low energy rate	90.	1st MFCC / DM
2.	Beat histogram entropy	89.	Audio Spectrum Centroid / M
3.	Root mean square / DS	88.	Kurtosis / DM

Table 1: *Results of the noise test. 3 features most robust (left) and most susceptible to the addition of white noise.*

	Best features		Worst features
1.	2nd MFCC / DS	90.	Predictivity ratio / S
2.	1st MFCC / DS	89.	Predictivity ratio / M
3.	5th MFCC / DS	88.	Predictivity ratio / DM

Table 2: *Results of the bandwidth test. 3 features most robust (left) and most susceptible to lowpass filtering.*

Features were selected in a completely systematical way. We proceed in two steps: the first step corresponds to the above criterium of invariancy, the second to the criteria of discriminative power and uncorrelation.

4.1. Tests on Robustness to Irrelevancies

Noise content and signal bandwidth are regarded here as irrelevant, since they should not influence the classification. Discarding the features that are more susceptible to moderate noise and bandwidth changes allows to ensure similar classification rates for a wide range of audio qualities.

In order to test the features for robustness against the addition of noise, four representative training samples belonging to the speech, classical music, popular music, and background noise classes were chosen. Each example was normalized and mixed with white gaussian noise of -25 dB RMS power and subjected to file-based feature extraction. The resulting features of the four noisy signals were compared with the ones extracted from the original signals. The variations were averaged across the four samples. Table 1 shows the 3 features that were least susceptible to noise (3 best features), and the 3 most susceptible features (3 worst features). The 20 worst features in the ranking were discarded.

To test robustness to bandwidth changes, a low-pass filtering with a cut-off frequency of 11025 Hz was applied to the same four examples and their extracted features were compared with the original ones in the same way. The 20 worst features were also discarded. Table 2 shows the selected results.

Some general conclusions can be drawn from the results of the tests. The M and DM subfeatures are in most cases highly sensitive to noise and to lowpass filtering. The DS subfeatures are especially robust to noisy changes in the signal. MFCCs are highly robust to lowpass filtering, except for their DM variants. Predictivity ratio features are extremely sensitive to lowpass filtering. As a result of both tests, a total number of 32 features (the ones that appear at least once in both bottom-ranking lists) were discarded.

4.2. Feature subset selection

The remaining 58 features were subjected to a feature selection algorithm that selects a subset of features containing the highest class discriminating power. A vectorial *sequential feed forward* algorithm was used, which searches for a feature subset that maximizes a measure of class separability based on the scatter matrices of the training set [7].

	speech/music/background		classical/non-classical
1.	2nd MFCC / S	1.	Zero Crossings / DS
2.	4th MFCC / DS	2.	Loudness / M
3.	Rhythmic regularity	3.	Rhythmic regularity

Table 3: *Results of the feature subset selection. 3 best features for the speech/music/background and classical/non-classical splits.*

The algorithm yields a list in which the 58 features are ordered according to their ability to separate classes. It also ensures that features selected consecutively are as uncorrelated as possible.

In this work, a hierarchical classification approach has been used. Rather than making a single decision to classify into one of the 17 classes (direct approach), the hierarchical approach consists of successive classification decisions with a number of classes ranging from 2 to 4, with the hierarchy corresponding to the audio taxonomy tree depicted in Figure 1. This has a parallel approach in the context of feature selection. Instead of using the whole training database to obtain a single list of selected features, a genre-dependent feature selection was used, in which only the training samples belonging to the current branch in the classification tree are used to evaluate the separability of the current 2, 3 or 4 classes.

As a result, a set of 9 feature lists was obtained, one for each split in the tree. Each list shows which features are most appropriate for distinguishing between a given set of music or audio subgenres. As an example, Table 3 shows the 3 best features for the speech/music/background and classical/non-classical splits.

The following general conclusions could be drawn from the results: The DM is not an effective subfeature for classification. This can be explained by the fact that the mean of the derivative is very likely to take values close to zero. The proposed rhythmic regularity feature has excellent separating performance. It belongs to the top-3 of the lists in all cases except for the speech and the classical subsplits. The zero crossings / DS feature is an excellent separator. It tops the list in 4 occasions: classical/non-classical, and the 3 classical subsplits.

5. CLASSIFICATION

A 3-component Gaussian Mixture Model was used as classifier. At a number of around 20 features, performance stopped growing, and in some evaluation experiments it even decreased. For these reasons, the number of features was fixed at 20. That is, at each classification step, the system extracts the 20 best features from the signal, as indicated by the feature selection list corresponding to the current tree split.

6. EVALUATION

50 audio examples were collected for each of the 17 classes, resulting in an 850 file database. Each sample is approximately 30 seconds long, resulting in over 7 hours of audio.

The evaluation was carried out using a *stratified 10-fold cross-validation*. Table 4 shows the percentages of correct classifications (mean plus standard deviation across the cross-validation experiments) for the hierarchical approach and for each of the tree splits, as well as an all-class classification rate which takes into account all 17 classes. The *accumulative performance* is the percentage of samples of the test set correctly classified. The *independent performance* is the percentage of samples correctly classified at level i that have been correctly classified at level $i + 1$. Figure 2

Split	Accumulative performance	Independent Performance
speech/music/background	94.59 ± 1.77	94.59 ± 1.77
male/female/speech+background	76.67 ± 8.46	82.31 ± 8.63
classical/non-classical	91.08 ± 3.68	96.08 ± 2.02
chamber music/orchestral music	74.29 ± 7.25	81.52 ± 7.88
rock/pop/jazz	63.67 ± 6.17	70.33 ± 8.65
chamber subgenres	42.50 ± 12.08	54.67 ± 13.92
orchestral subgenres	52.67 ± 10.63	75.21 ± 11.83
hard rock/soft rock	55.00 ± 16.50	79.52 ± 20.18
pop subgenres	62.00 ± 9.96	76.15 ± 9.55
All classes	58.71 ± 2.85	

Table 4: *Classification performance using the hierarchical approach.*

Split	Accumulative performance	Independent Performance
speech/music/background	96.35 ± 1.70	96.35 ± 1.70
male/female/speech+background	76.67 ± 9.03	81.00 ± 9.19
classical/non-classical	94.31 ± 3.48	96.67 ± 2.45
chamber music/orchestral music	75.43 ± 7.15	78.06 ± 6.81
rock/pop/jazz	65.33 ± 5.49	71.83 ± 9.38
chamber subgenres	50.50 ± 9.26	63.05 ± 13.24
orchestral subgenres	52.00 ± 15.65	75.86 ± 18.26
hard rock/soft rock	59.00 ± 19.69	78.91 ± 21.09
pop subgenres	58.67 ± 16.87	71.57 ± 16.04
All classes	59.76 ± 5.23	

Table 5: *Classification performance using the direct approach.*

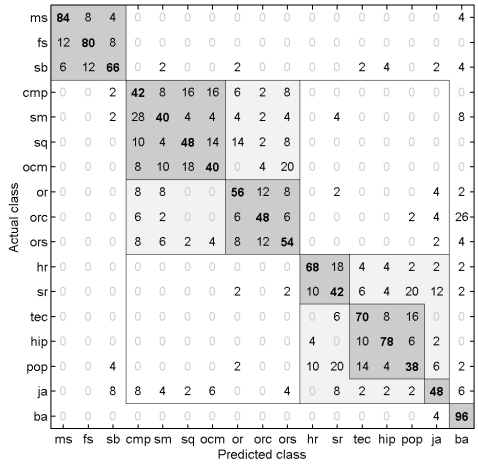

Figure 2: *Confusion matrix. Shaded boxes correspond to the consecutive splits in the hierarchy. For the class codes, see Figure 1.*

shows the corresponding confusion matrix. Numbers indicate the percentages of test samples belonging to each actual class. The evaluation was repeated for the direct approach. The results are shown on Table 5.

7. CONCLUSIONS

As can be seen, the classification rates are very similar for both direct and hierarchical approaches. Considering that the difference in performance is small, it has been opted for the hierarchical approach for the final implementation of the system, since it features the following additional advantages [8]:

- It allows the errors to be more acceptable than in the case of a direct classification. For example, it is more acceptable to wrongly classify a symphonic music sample as orchestral music with soloist, than as Hip-Hop. Dividing the decision in subdecisions makes the errors concentrate within the given subgenre.
- A hierarchical approach takes into consideration the class dependency of the features.
- It closely reflects the underlying audio taxonomy, thus allowing to evaluate the separability of commonly used genres and their suitability for automatic classification.
- It provides the framework for the future design of more sophisticated genre-dependent features.
- It makes future expansions of the taxonomy easier. Only the genre branch to which a new class is added should be modified with respect to feature selection and training, the rest of the models remaining unchanged.

The classification rates differ substantially across levels of the tree, showing the different grades of difficulty in separating each corresponding set of classes. The best independent performances were achieved at the highest levels in the tree, for example achieving 94.59% accuracy in differentiating between speech, background and music, 96.08% in separating classical from non-classical music and 81,52% in separating chamber music from orchestral music.

In contrast, the main difficulties arise in the most specific genres at the lowest levels of the tree, especially in the case of the four chamber music subgenres, where the total classification accuracy is of 54.67%. The lower levels of the tree, as well as the high number of classes considered, make the all-class classification rate drop to 58.71%. To achieve higher rates at these levels, more sophisticated, genre-specific features are needed.

Concerning computational costs, it was measured that the implemented prototype application operates faster than real-time on a 1 GHz processor.

8. REFERENCES

[1] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," in *Proc. ICASSP*, 1997.

[2] T. Zhang and J. Kuo, "Hierarchical system for content-based audio classification and retrieval," in *Proc. ICASSP*, 1998.

[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] M. Casey, "General sound classification and similarity in mpeg-7," *Organized Sound*, vol. 6:2, 2002.

[5] ISO/IEC FDIS 15938-4, "Information technology - multimedia content description interface - part 4: Audio," 2001.

[6] E.D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Am.*, vol. 103 (1), January 1998.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.

[8] J.-J. Aucouturier and F. Pachet, "Representing musical genre: A state of art," *Journal of New Music Research*, vol. 32(1), 2003.