A COMPARISON OF HUMAN AND AUTOMATIC MUSICAL GENRE CLASSIFICATION

S. Lippens, J.P Martens, T. De Mulder *

Ghent University
Department of Electronics and Information Systems
Sint-Pietersnieuwstraat 41
B9000 Gent, Belgium

G. Tzanetakis †

Carnegie Mellon University
Computer Science Department
5000 Forbes Avenue, Pittsburgh
PA 15218 USA

ABSTRACT

Recently there has been an increasing amount of work in the area of automatic genre classification of music in audio format. In addition to automatically structuring large music collections such classification can be used as a way to evaluate features for describing musical content. However the evaluation and comparison of genre classification systems is hindered by the subjective perception of genre definitions by users. In this work we describe a set of experiments in automatic musical genre classification. An important contribution of this work is the comparison of the automatic results with human genre classifications on the same dataset. The results show that, although there is room for improvement, genre classification is inherently subjective and therefore perfect results can not be expected neither from automatic nor human classification. The experiments also show that features derived from an auditory model have similar performance with features based on Mel-Frequency Cepstral Coefficients (MFCC).

1. INTRODUCTION

Musical genres are categorical labels that are created by humans in order to organize the vast universe of music. They arise through a complex interplay of culture, art, and market forces. The boundaries between different genres are not well defined and therefore it is difficult to find precise definitions and mathematical formulas that can automatically identify the genre of a piece of music. A listener judges the genre of a piece of music on the basis of objective and subjective measures. Table 1 shows the number of genres used by different musical content providers and illustrates the lack of consensus about musical genres. This variation is caused by differences about the existence, importance, boundaries and hierarchy of genres between different groups of people. A further complication arises from sales related influences (e.g. targeting a specific audience) which result in pseudo-genres such as "compilations", "box sets" and "children".

Based on these observations the task of automatic genre classification directly from audio signals is not trivial. Automatic music genre classification allows the automatic structuring and organization of large archives of music and also provides a good way to compare and evaluate feature sets that attempt to represent musical content. Recently a number of automatic musical genre classification algorithms that combine signal processing and statistical pattern classification have been proposed. However, to the best of our knowledge there has not been any attempt to compare the outputs of automatic genre classification algorithms with human annotations by average listeners. In this paper, we describe a set of experiments that compare automatically computed labels with human annotations for the task of musical genre classification using the same dataset. In addition we show that the use of a computationally demanding but psychoacoustically more accurate auditory model as a feature front end does not seem to provide any significant advantage for musical genre classification compared to the use of the standard Mel Frequency Cepstral Coefficients.

All Music	691	radio.real.com	18
mp3.com	288	BPI	17
Yahoo!	116	Warner Bros	16
Indiana University	31	Atlantis Records	13
Amazon	25	Virgin	13
Sony	25	IFPI Belgium	11
Altavista	25	MTV	5

Table 1. A snapshot of some music related organizations and the number of genres they use on their website (February 2003).

1.1. Related Work

Previous work in the area of automatic musical genre classification includes: features computed based on wavelet analysis [1], visual texture features of spectrograms [2], and a specialized architecture called "Explicit Time Modeling Neural Networks" [3]. A comparison of audio features with features extracted from the analysis of cultural meta-data such as download usage patterns is presented in [4]. A detailed study of automatic musical genre classification is presented in [5] and the proposed features have been used in the experiments presented in this paper. A more detailed description of the experiments presented in this paper can be found in [6]. A recent review of representing musical genre in digital music distribution is provided in [7] which covers manual annotation, automatic methods and usage-based methods such as collaborative filtering. To the best of our knowledge, the only result in the performance of humans for the task of musical genre classification is [8] which reports classification accuracy of approximately 70% using 10 genres. No attempt for comparison with automatic algorithms is made. The more general problem of organizing digital collections of music for searching and browsing is the topic of the emerging research area of Music Information Retrieval (two good recent overviews of MIR are [9, 10]).

* This work was partly funded by the Flemish Institute for Promotion of Scientifi c and Technical Research in Industry (010035-GBOU)

† This work was partially supported by NSF Award 0085945

2. FEATURE EXTRACTION AND CLASSIFICATION

Most of the features investigated in this paper were proposed in [5]. These features attempt to represent timbral texture, rhythmic content and pitch content information. The features used to represent timbral texture are based on standard features proposed for music-speech discrimination and speech recognition. They consist of a set of 4 features derived from the Short Time Fourier Transform (STFT) magnitude spectrum such as the Spectral Centroid (defined as the first moment of the magnitude spectrum) as well as the first 5 Mel-Frequency Cepstral Coefficients (MFCC) [11]. These features are computed using an analysis window of 20 milliseconds. Means and variances of the features over a larger texture window (1 second) with a hop size of 20 milliseconds are computed resulting in a set of 18 features. An additional feature (the percentage of low energy frames over the texture window) results in a timbral texture feature vector of 19 dimensions.

The basis of representing rhythmic content is the calculation of a Beat Histogram (BH) that shows the distribution of various beat periodicities of the signal. For example a piece with tempo 60 Beats-per-Minute (BPM) would exhibit BH peaks at 60 and 120 BPM. The BH is calculated using periodicity detection in multiple octave channels that are computed using a Discrete Wavelet Transform. In [5], six numerical features that attempt to summarize the BH are computed and used for classification.

In addition to the feature set proposed in [5] we explored the use of an auditory model as a front-end for feature calculation. The model is described in [12] and attempts to represent in more detail the physiology of the human ear. More specifically it implements the transition from waveform in the air to activity pattern in the auditory nerves with the following stages: 1) a low-pass filter (2nd order, 10 dB boost at 4kHz) mimics the propagation in the outer and middle ear; 2) 40 band-pass filters model the mechanical filtering in the cochlea (the central frequencies are uniformly spaced on a critical band frequency scale and have 3dB bandwidths of one critical band); 3) 40 hair cell models convert the filter outputs to neural signals (this incorporates compression of dynamic range, half-wave rectification, short term adaptation and coding of temporal information as found in physiological measurements); 4) 40 low-pass filters finally extract neural signal intensities which are sampled every 10 milliseconds. Further calculation of the timbral features is very similar to the calculation of the MFCC based features. Instead of using the logarithm of the STFT coefficients as in the case of the standard MFCC calculation, the 40 channel values were used to calculate the DCT coefficients. Means and variances of these DCT coefficients over a larger window (typically 30 seconds) result in the auditory model based feature vector. We experimented with feature sets derived from the first 5, 7, 9, 11 and 13 DCT coefficients of the auditory model's output.

For classification, a number of standard statistical pattern recognition classifiers were used. The simple Gaussian (GS) classifier, modeling each class probability density function (pdf) as a multidimensional Gaussian distribution whose parameters are estimated on the training set. In the Gaussian Mixture Model (GMM), each class pdf is assumed to be a mixture of K weighted multidimensional Gaussian distributions. The iterative Expectation Maximization (EM) algorithm can be used to estimate the parameters of each Gaussian component and the mixture weights. The K-nearest neighbor classifier (KNN) is an example of a non-parametric classifier where each sample is labeled according to the majority of its K nearest neighbors. More information about these classifiers and statistical pattern recognition in general can be found in [13].

3. EVALUATION

The typical approach in the published literature on genre classification has been to use the labels provided by some authority, train classifiers and present classification accuracy results by using cross-validation. These results are difficult to interpret because they are not directly comparable with the performance of humans for the same task. In order to put our classification results in context we establish both lower (random classification) and upper bounds (human classification) on the classification accuracy.

3.1. Establishing bounds for the classification results

In this section we will try to sketch a profile of the dataset we used both in a qualitative and quantitative way. This is important because the used dataset has a huge influence on the achievable results and this fact hinders the comparison among different experiments with different datasets.

The MAMI dataset is a collection of 160 full length tracks of music. This dataset is used for a variety of research in the area of content based musical audio mining, with a focus on 'query by humming'. The construction of the set is aimed at giving a representative view on the western music consumption today, based on the sales figures from IFPI (the International Federation of Phonographic Industry) in Belgium for the year 2000. Originally, the tracks were annotated with 11 musical genres, according to the classification of IFPI. Initial experiments made clear that this labeling was not appropriate for training genre classifiers. For example some genres had few examples or were very heterogeneous. In order to address this issue, a set of 6 basic genres was defined and user experiments were conducted to confirm that their definitions are consistent among different subjects.

A new labeling of the dataset was obtained by surveying 27 human listeners. We let them listen to the central 30 seconds of each track m (m = 1 to 160) and asked them independently to choose a musical genre s out of 6 possibilities: *classical*, *dance*, *pop*, *rap*, *rock* or *other* (the latter was for the case none of the previous was really applicable). For each track m we define $Q_s^{(m)}$ as the number of votes for genre s. The maximum number of votes among the genres, called $Q_{max}^{(m)}$, indicates the elected genre $G^{(m)}$, which is used as the new label for each track. This results in the following structure of the MAMI dataset: 24 classical, 18 dance, 69 pop, 8 rap, 25 rock and 16 other tracks. As an extra, we have $Q_{max}^{(m)}$ at our disposal as a measure of unanimity about the musical genre. Figure 1 shows the histogram of $Q_{max}^{(m)}$ for the whole dataset. Besides the peak around 27 votes, there is a considerable second peak around 15 votes. These are mainly tracks with many votes for 'other' and little consensus among the human listeners. The average of $Q_{max}^{(m)}$ is 20.6 votes.

Another application of the survey is the evaluation of the human classification. For each human respondent we compared their selected genres with $G^{(m)}$, leading to a *percentage corresponding classification*. The 27 listeners achieve 76% corresponding classification on average (further referred to as C_h), with individual results ranging from minimum 57% to maximum 86%. All these presented measures indicate that there is a high degree of subjectivity involved with genres in the MAMI dataset. This situation is clearly not optimal for an unambiguous training of musical gen-

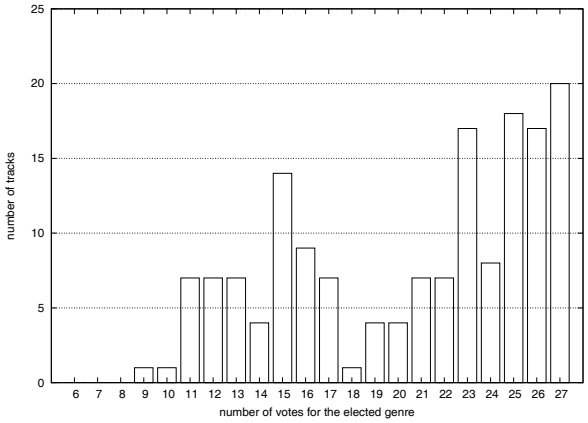

Fig. 1. histogram of the number of votes for the elected genre

res. On the other hand, the dataset is more representative for real life applications than a dataset with a well thought-out selection of genres and examples. To examine the influence of this fact we selected the 98 tracks with the following two properties: 1) the genre $G^{(m)}$ can be all but 'other' and 2) $Q^{(m)}_{max}$ is 18 or more. We call this constructed subset of the MAMI dataset the 'MAMI2' dataset. On this dataset the human listeners achieve 90% average corresponding classification (C_h).

Besides the human performance on the dataset, which we shall refer to as an upper bound (C_h), we can also define lower bounds to the performance of automatic classification. Those bounds give us a so called *reference framework* for putting the achieved results of automatic classification in perspective.

A straightforward definition of a lower bound is random classification where every class (genre) has an equal probability of selection. If there are K classes ω_k, the expected correct classification is $1/K$. A more advanced scheme for random classification uses the *prior probabilities* of the classes. We define the number of instances in each class ω_k ($1 \leq k \leq K$) as N_k and the total number of instances as $N = \sum_k N_k$. The prior probability of an instance belonging to class ω_k is then $P(\omega_k) = N_k/N$. In the prior probability random classification scheme we select each class ω_k with its prior probability $P(\omega_k)$. The expected correct classification is in this case $C_r = \sum_k P^2(\omega_k)$. Because the bayesian decision theory employed for the automatic classification is based on prior probabilities, C_r is more appropriate to use as a lower bound in the reference framework.

3.2. Experiments and Results

A number of experiments were conducted in order to examine different choices in the feature extraction and classification system. The experiments were implemented using Marsyas (`http://marsyas.sourceforge.net`), a free software framework for rapid development and evaluation of Computer Audition applications. Feature extraction is a key operation because it has to capture precisely those components of the input signal that determine the genre. We took the central 30 seconds of each track of the MAMI dataset and applied all the possible combinations of feature sets and classification models to it. The leave-one-out evaluation method was used where each example is withheld for testing and the remaining examples are used for training. The best result we obtained was 58% correct classification with just the timbral texture features (STFT and MFCC) and the 3-nearest neighbor classifier. If we look at the top 10 results (ranging from 58% to 55%), the combined set of MFCC-based features and Beat Histogram features, hereafter called the MFCC+Rhythm feature set, appears 5 times, both in combination with nearest neighbor classifiers and parametric models (Single Gaussian and Gaussian Mixture Model). Other feature sets only seem to perform well in combination with nearest neighbor classifiers. Another good point for the MFCC+rhythm feature set is the fact that even its worst result is better than the average performance measured for the other sets. As an illustration, table 2 shows the confusion matrix for the experiment with this feature extractor and GS classifier.

classified	class					
as	classical	dance	pop	rap	rock	other
classical	**.75**	.00	.07	.00	.04	.25
dance	.04	**.83**	.17	.50	.12	.00
pop	.04	.17	**.48**	.25	.20	.13
rap	.00	.00	.00	**.13**	.00	.00
rock	.08	.00	.13	.00	**.60**	.00
other	.08	.00	.14	.13	.04	**.63**

Table 2. Confusion matrix for the MFCC+rhythm feature set in combination with a GS classifier.

To examine the auditory model based features, we used feature sets derived from the first 5, 7, 9, 11 and 13 DCT coefficients. This achieved no better results than the comparable MFCC-based feature extractor. This indicates that the use of more computationally demanding but psychoacoustically accurate auditory model as a feature front-end doesn't make a big difference for the task of automatic musical genre classification. The best results were acquired using the first 5 DCT coefficients. The use of more coefficients resulted in a lower performance. This is in accordance with [5].

In the previous section about the reference framework, we already introduced the use of different datasets. Figure 2 shows the results of automatic classification with the MFCC+rhythm features and the GS classifier for the two datasets MAMI and MAMI2. Also the reference framework is shown. The automatic classification clearly outperforms the random classification as expected, but there's still a gap of around 20% with the human classification.

We also focused on the influence of the choice of fragments in the tracks. Initially one fragment with varying length (1 to 30 seconds) was used per track. It always positioned in the middle of the track so no automatic segmentation was used for fragment selection. For features aimed to capture the musical texture, we did not experience significant improvements using fragments longer than 10 seconds, even 5 seconds were sufficient in most cases. The beat related features on the contrary showed a slight increase of performance with longer fragments.

The use of fragments of tracks made it possible to do a variation on the classification approach. We took as many as possible non overlapping fragments of 30 seconds per track. For each track we trained the system with all the fragments of the other tracks and then classified all the fragments of the track with the trained model. Finally we classified the whole track according to the most estimated genre among its fragments. We found no significant improvement of the classification accuracy of this multi-fragment track classification scheme over the standard classification scheme with only one central fragment per track. We think the use of extra

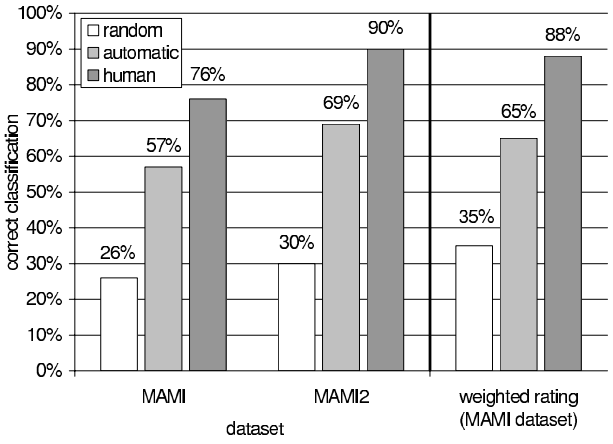

Fig. 2. The results of automatic classification with the MFCC+rhythm feature set and a GS classifier for different datasets and an alternative rating scheme. The reference framework (C_r and C_h) is also shown.

fragments per track in combination with the not optimal character of the MAMI dataset causes an extra blurring of the trained genre model, so there is no noticeable gain.

In a last set of experiments we used an alternative ranking scheme based on the human votes. The purpose was to examine the presence of so called 'graceful errors' [7]. E.g., errors like a classification of a particular 'pop' track as 'soft rock' are subjectively more understandable and should be less punished than errors like a 'baroque' track being classified as 'hardcore punk'. Practically, for each track m we rated the classification in a genre s with a score $Q_s^{(m)}/Q_{max}^{(m)}$, with $Q_s^{(m)}$ the number of votes for genre s for that track m and $Q_{max}^{(m)}$ the number of votes for the elected genre, as previously defined. The right part of figure 2 shows the results of this experiment and should be compared with the first three bars of the left part of the figure (unweighted rating with the same dataset). The human classification clearly benefits of the milder validation, but the automatic classification shows less improvement in accuracy. It is even less than the improvement of the random classification, which indicates that no substantial presence of graceful errors can be identified.

4. CONCLUSIONS AND FUTURE WORK

A set of experiments comparing human and automatic musical genre classification was presented. The results indicate that there is significant subjectivity in genre annotations by humans, and that there is still a significant gap between automatic and human classifications. In addition it was found that the features emerging from a computationally intensive auditory model do not outperform the standard MFCC features for the presented task.

In the future, we plan to explore new feature sets as well as more classifiers, such as Support Vector Machines (SVM) [14], in the hope that by doing so the current gap between automatic algorithms and human annotations of musical genre classification will gradually disappear.

5. REFERENCES

[1] 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 *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1998.

[2] Hrishikesh Deshpande, Rohit Singh, and Unjung Nam, "Classification of Musical Signals in the Visual Domain," in *Proc. COST G-G Conf. on Digital Audio Effects (DAFX)*, Limerick, Ireland, Dec. 2001.

[3] H. Soltau, T. Schultz, M. Westphal, and A. Waibel, "Recognition of Music Types," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 1998, vol. 2, pp. 1137–1140.

[4] Brian Whitman and Paris Smaragdis, "Combining Musical and Cultural Features for Intelligent Style Detection," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, Paris, France, Oct. 2002, pp. 47–52.

[5] George Tzanetakis and Perry Cook, "Musical Genre Classification of Audio Signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, July 2002.

[6] Stefaan Lippens, "Automatische Genreclassificatie van Muziek," M.S. thesis, Department of Electronics and Information Systems, Ghent University, 2003.

[7] Jean Julien Aucouturier and Francois Pachet, "Musical Genre: a Survey," *Journal of New Music Research*, vol. 32, no. 1, 2003.

[8] D. Perrot and Robert Gjerdigen, "Scanning the dial: An exploration of factors in identification of musical style," in *Proc. Society for Music Perception and Cognition*, 1999, p. 88, (abstract).

[9] Joe Futrelle and Stephen J. Downie, "Interdisciplinary Communities and Research Issues in Music Infomration Retrieval," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, Paris, France, Oct. 2002, pp. 215–221.

[10] Francois Pachet, "Content Management for Electronic Music Distribution: The Real Issues," *Communications of the ACM*, vol. 46, no. 4, Apr. 2003.

[11] Steven Davis and Paul Mermelstein, "Experiments in syllable-based recognition of continuous speech," *IEEE Transcactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357–366, Aug. 1980.

[12] Luc M. Van Immerseel and Jean-Pierre Martens, "Pitch and voiced/unvoiced determination with an auditory model," *Acoustical Society of America*, vol. 91, no. 6, pp. 3511–3526, June 1992.

[13] Richard Duda, Peter Hart, and David Stork, *Pattern classification*, John Wiley & Sons, New York, 2000.

[14] Tao Li and George Tzanetakis, "Factors in automatic musical genre classification," in *Proc. Workshop on applications of signal processing to audio and acoustics WASPAA*, New Paltz, NY, 2003, IEEE.