Integration of Text and Audio Features for Genre Classification in Music Information Retrieval

Robert Neumayer and Andreas Rauber

Vienna University of Technology
Institute for Software Technology and Interactive Systems
{neumayer,rauber}@ifs.tuwien.ac.at

Abstract. Multimedia content can be described in versatile ways as its essence is not limited to one view. For music data these multiple views could be a song's audio features as well as its lyrics. Both of these modalities have their advantages as text may be easier to search in and could cover more of the 'content semantics' of a song, while omitting other types of semantic categorisation. (Psycho)acoustic feature sets, on the other hand, provide the means to identify tracks that 'sound similar' while less supporting other kinds of semantic categorisation. Those discerning characteristics of different feature sets meet users' differing information needs. We will explain the nature of text and audio feature sets which describe the same audio tracks. Moreover, we will propose the use of textual data on top of low level audio features for music genre classification. Further, we will show the impact of different combinations of audio features and textual features based on content words.

1 Introduction

The large-scale adoption of new business models for digital audio material is happening already. What many content providers and online music vendors are still missing are appropriate means of presenting their media to their users according to different information needs. Amazon[1] or last.fm[2] have shown the potential of recommendation engines based on the mining of transactional data.

It further is an intrinsic need for every Music Information Retrieval system to include not only recommendation or playlist generation engines, but also possibilities for searching and browsing. Music Information Retrieval has made huge progress in terms of devising sophisticated descriptors for the acoustic content of music. Research in this direction facilitates numerous content-based search scenarios, such as query by humming, or organisation tasks, such as genre classification, playlist generation, and browsing access by perceived sound similarity.

Song lyrics cover semantic information about a song's contents on a level that could never be covered by audio features only. Many users may rather

[1] http://www.amazon.com

[2] http://www.last.fm

G. Amati, C. Carpineto, and G. Romano (Eds.): ECIR 2007, LNCS 4425, pp. 724–727, 2007.

be interested in songs that cover similar topics than sound alike. Some musical genres, such as e.g. Christmas candles, can only be detected by including textual features, as they occur across many different musical genres and the definition of the genre itself is rather done on a topic level.

We thus combine both textual as well as audio information for music genre classification, i.e. automatically assigning musical genres to tracks based on audio features as well as content words in song lyrics.

The remainder of this paper is organised as follows. Section 2 provides a brief overview of related work. This is followed by a description of our classification experiments in Section 3, as well as outlook on current work in Section 4.

2 Related Work

The area of Music Information Retrieval has been heavily researched, particularly focusing on audio feature extraction. First experiments with content-based Music Information Retrieval were reported in [1] as well as [6], the focus being on automatic genre classification of music. In this paper we use a modified version of the Rhythm Patterns features, previously used within the SOMeJB system [5]. Based on that feature set, [2] shows that the Statistical Spectrum Descriptors yield relatively good results at a manageable dimensionality.

A sophisticated semantic and structural analysis including language identification of songs based on lyrics is conducted in [4]. Artist similarity based on song lyrics is performed in [3], pointing out that similarity retrieval using lyrics is inferior to acoustic similarity. It is also suggested that a combination of lyrics and acoustic similarity could improve results, further motivating the research reported in this paper.

3 Experiments

Due to the lack of public benchmark corpora, we created a parallel corpus of audio and song lyrics files of a music collection of 9.758 titles organised into 41 genres. Class sizes ranged from only a few songs for the 'Classical' genre to about 1.900 songs for 'Punk Rock'. In order to utilise the information contained in music for genre classification, we describe sets of audio features derived from the waveform of audio tracks as well as the bag-of-word features for song lyrics.

3.1 Audio Features

Three features were computed from audio tracks in standard PCM format with 44.1 kHz sampling frequency (e.g. decoded MP3 files). *Rythm Patterns* (RP) [5], also called Fluctuation Patterns, denote a matrix representation of fluctuations on critical bands (parts of it describe rhythm in the narrow sense), resulting in a 1.440 dimensional feature space. *Statistical Spectrum Descriptors* (SSDs, 168 dimensions) are statistical moments derived from a psycho-acoustically transformed spectrogram [2]. *Rhythm Histograms* (RH, 60 dimensions) are calculated as the sums of the magnitudes of each modulation frequency bin of all 24 critical bands.

Table 1. Classification accuracies for different types and combinations of audio features and features based on lyrics. The experiments A1 - A3 denote audio-only, L1 - L4 lyrics-only, and C1 - C3 features combined from audio and lyrics feature sets. The type column shows the types of feature sets used, dimensionality notes the resulting dimensionality of the data.

Name	Type	Dimensionality	Classification Accuracy
A1	RH.	60	.3100
A2	SSD.	168	.4168
A3	RP.	1440	.4128
L1	LYRICS	60	.2451
L2	LYRICS	168	.3204
L3	LYRICS	1440	.4445
L4	LYRICS	3000	.4708
C1	LYRICS + RH	120	.3268
C2	LYRICS + SSD	336	.4817
C3	LYRICS + RP	2880	.4841

3.2 Lyrics Features

For every piece of music, three lyrics portals were accessed, using artist name and track title as queryies. If the results from *lyrc.com.ar* were of reasonable size, these lyrics were assigned to the track. If *lyrc.com.ar* fails, *sing365lyrics.com* will be checked for validity, then *oldielyrics.com*.

All lyrics were processed using the bag-of-words model and weighted by *tfidf* information. Feature selection was done via document frequency thresholding, i.e. the omittance of terms that occur in a very high or very low number of documents. For the matrices used for the experiments terms occurring in more than half of the documents were omitted, the lower threshold was then adjusted to meet the desired dimensionality. Downscaling was performed to different dimensionalities matching the dimensionalities of the audio feature spaces.

3.3 Classification Results

Table 1 shows classification accuracies for a set of experiments based on audio and lyrics features as well as combinations thereof. Experiments were performed by Weka's implementation of Support Vector Machines for ten-fold stratified cross validation. Results shown are the macro averaged classification accuracies.

Results show that a combination of lyrics and audio features improves overall classification performance. The best results were achieved by the 'LYRICS + RP' setting (C3), closely followed by the 'LYRICS + SSD' experiment (C2). The higher-dimensional the data for the lyrics experiments is, the higher is its classification accuracy, implying that there is even more discriminating information contained in lyrics (see experiments L1 - L4), which is not covered in this context because of the limitations of the simple concatenation approach. For combination experiments (C1 - C3) we use balanced combinations of features, i.e. the dimensionality of the lyrics component always equals the dimensionality of the audio feature component.

For statistical significance testing we used a paired T-test for a significance level of $\alpha = .05$. Results showed that A2 performs better than A1 ($p = .0143$), but there is no significant difference between A2 and A3 ($p = .9353$). Further, it is shown that C3 performed better than both A2 ($p = .1934$) and L3 ($p = .0129$). However, the results are not significantly different from experiment L4 ($p = .1755$), leading to the conclusion that high-dimensional lyrics data only is a strong basis for a classifier. Hence a classifier based on differing numbers of lyrics than audio features, e.g. more dimensions in the lyrics than in the audio space, might further improve classification accuracy. Yet, by combining lyrics and audio (C2) the same performance was achieved at a fraction of the dimensionality.

4 Conclusions and Future Work

We showed that the combination of multi-modal features for information retrieval increases classification accuracy. Future work will deal with better means of combining classification results. Ensemble methods might prove useful, overcoming the limitation of implicit feature weighting encountered in the current setting. Additionally, stylistic features for text genre classification are currently being integrated.

References

1. Jonathan Foote. An overview of audio information retrieval. *Multimedia Systems*, 7(1):2–10, 1999.
2. Thomas Lidy and Andreas Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR 2005)*, pages 34–41, London, UK, September 11-15 2005.
3. Beth Logan, Andrew Kositsky, and Pedro Moreno. Semantic analysis of song lyrics. In *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo, ICME 2004, 27-30 June 2004, Taipei, Taiwan.* IEEE, 2004.
4. Jose P. G. Mahedero, Álvaro Martínez, Pedro Cano, Markus Koppenberger, and Fabien Gouyon. Natural language processing of lyrics. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 475–478, New York, NY, USA, 2005. ACM Press.
5. Andreas Rauber, Elias Pampalk, and Dieter Merkl. Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by musical styles. In *Proceedings of the 3rd International Symposium on Music Information Retrieval*, pages 71–80, Paris, France, October 13-17 2002.
6. George Tzanetakis and Perry Cook. Marsyas: A framework for audio analysis. *Organized Sound*, 4(30), 2000.