RHYME AND STYLE FEATURES FOR MUSICAL GENRE CLASSIFICATION BY SONG LYRICS

Rudolf Mayer[1], Robert Neumayer[1,2], and Andreas Rauber[1]

[1]Department of Software Technology and Interactive Systems,
Vienna University of Technology, Vienna, Austria

[2]Department of Computer and Information Science,
Norwegian University of Science and Technology, Trondheim, Norway

ABSTRACT

How individuals perceive music is influenced by many different factors. The audible part of a piece of music, its sound, does for sure contribute, but is only one aspect to be taken into account. Cultural information influences how we experience music, as does the songs' text and its sound. Next to symbolic and audio based music information retrieval, which focus on the sound of music, song lyrics, may thus be used to improve classification or similarity ranking of music. Song lyrics exhibit specific properties different from traditional text documents – many lyrics are for example composed in rhyming verses, and may have different frequencies for certain parts-of-speech when compared to other text documents. Further, lyrics may use 'slang' language or differ greatly in the length and complexity of the language used, which can be measured by some statistical features such as word / verse length, and the amount of repetative text. In this paper, we present a novel set of features developed for textual analysis of song lyrics, and combine them with and compare them to classical bag-of-words indexing approaches. We present results for musical genre classification on a test collection in order to demonstrate our analysis.

1 INTRODUCTION

The prevalent approach in music information retrieval is to analyse music on the symbolic or audio level. Songs are therein represented by low level features computed from the audio waveform or by transcriptions of the music. Additionally, all songs but instrumental tracks can be treated as textual content by means of song lyrics. This information can be exploited by methods of classical text information retrieval to provide an alternative to music processing based on audio alone. On the one hand, lyrics provide the means to identify specific genres such as 'love songs', or 'Christmas carols', which are not acoustic genres per se, but, to a large degree, defined by song lyrics [10]. Christmas songs, for instance may appear in a wide range of genres such as 'Punk Rock' or 'Pop' in addition to the classic 'Christmas carol'. On the other hand, song lyrics might sometimes be the only available option for extracting features, for example, when audio files are not available in an appropriate format, or only via a stream when bandwith is a limiting factor. Further, processing of audio tracks might be too time consuming compared to lyrics processing.

Song lyrics may differ to a great extent from the documents often dealt with in traditional text retrieval tasks such as searching the web or office documents. In addition to its plain text content, song lyrics exhibit a certain structure, as they are organised in blocks of choruses and verses. Also, lyrics might feature other specific properties, such as slang language in 'Hip-Hop' or 'Rap' music, or other statistical information such as (average) line lengths, or words per minute. Also special characters, e.g. the number of exclamation marks used, might be of interest.

In this paper, we present a set of features composed of these various textual properties. We then use them for genre classification, and compare and combine them with features resulting from standard bag-of-words indexing of the song texts. We aim to show the following: a) rhyme, part-of-speech, and simple text statistic features alone can be used for genre classification, and b) the combination of bag-of-words features and our feature sets is worthwile.

The remainder of this paper is structured as follows. We first give an overview of related work on music information retrieval focusing on lyrics processing in Section 2. Then, we give an introduction to lyrics analysis and explain our feature sets in detail in Section 3. In Section 4, we report from experiments performed on a manually compiled collection of song lyrics. Finally, we draw conclusions and give an overview of future work in Section 5.

2 RELATED WORK

In general, music information retrieval (MIR) is a broad and diverse area, including research on a magnitude of topics such as classic similarity retrieval, genre classification, visualisation of music collections, or user interfaces for accessing (digital) audio collections. Many of the sub-domains of MIR – particularly driven by the large-scale use of digital

audio over the last years – have been heavily researched.

Experiments on content based audio retrieval, i.e. based on signal processing techniques, were reported in [3] as well as [14], focusing on automatic genre classification. Several feature sets have since been devised to capture the acoustic characteristics of audio material, e.g. [12].

An investigation of the merits for musical genre classification, placing emphasis on the usefulness of both the concept of genre itself as well as the applicability and importance of genre classification, is conducted in [8].

A system integrating multi modal data sources, e.g. data in the form of artist or album reviews was presented in [1]. Cultural data is used to organise collections hierarchically on the artist level in [11]. The system describes artists by terms gathered from web search engine results. A promising technique for automatic lyrics alignment from web resources is given in [4].

A study focusing solely on the semantic and structural analysis of song lyrics including language identification of songs based on lyrics is conducted in [6]. Song lyrics can also be feasible input for artist similarity computations as shown in [5]. It is pointed out that similarity retrieval using lyrics, i.e. finding similar songs to a given query song, is inferior to acoustic similarity. However, it is also suggested that a combination of lyrics and acoustic similarity could improve results, which motivates future research. Genre classification based on lyrics data as well as its combination with audio features is presented in [9].

Further, the combination of a range of feature sets for audio clustering based on the Self-Organising Map (SOM) is presented in [7]. It is shown that the combination of heterogeneous features improves clustering quality. Visualisation techniques for multi-modal clustering based on Self-Organising maps are given in [10], demonstrating the potential of lyrics analysis for clustering collections of digital audio. Similarity of songs is defined according to both modalities to compute quality measures with respect to the differences in distributions across clusterings in order to identify interesting genres and artists.

3 LYRICS FEATURES

In this section we present the feature set computed from the song lyrics, namely bag-of-words, rhyme, part-of-speech, and statistical features.

3.1 Bag-of-Words Features

A common approach in text retrieval is to index documents with the bag-of-words method. Here, each unique term occurring in any of the documents of the collection is regarded a feature. To populate the feature vectors, information about the frequency of occurrences of the terms in the collection is gathered. A simple approach is the Boolean Model, which only considers whether a term is present in a document or not. More sophisticated, one can apply a term weighting scheme based on the importance of a term to describe and discrimante between documents, such as the popular $tf \times idf$ (term frequency $\times$ inverse document frequency) weighting scheme [13]. In this model, a document is denoted by d, a term (token) by t, and the number of documents in a corpus by N. The *term frequency* $tf(t,d)$ denotes the number of times term t appears in document d. The number of documents in the collection that term t occurs in is denoted as *document frequency* $df(d)$. The process of assigning weights to terms according to their importance for the classification is called 'term-weighing', the $tf \times idf$ weight of a term in a document is computed as:

$$tf \times idf(t,d) = tf(t,d) \cdot ln(N/df(t)) \qquad (1)$$

This weighting scheme is based on the assumption that terms are of importance when they occur more frequently in one document, and at the same time less frequently in the rest of the document collection. The bag-of-words approach focuses on grasping the topical content of documents, and does not consider any structural information about the texts. This method tends to, already with a low number of documents, result in high-dimensional feature vectors. Thus, often a feature space reduction is required for further processing, which is often achieved by cutting of words which occur either too often, or too rarely.

3.2 Rhyme Features

A rhyme is a linguistic style, based on consonance or similar sound of two or more syllables or whole words at the end of one line; rhymes are most commonly used in poetry and songs. We assume that different genres of music will exhibit different styles of lyrics, which will also be characterised by the degree and form of the rhymes used. 'Hip-Hop' or 'Rap' music, for instance, makes heavy use of rhymes, which (along with a dominant bass) leads to its characteristic sound. We thus extract several descriptors from the song lyrics that shall represent different types of rhymes.

One important notion is that consonance or similarity of sound of syllables or words is not necessarily bound to the lexical word endings, but rather to identical or similar phonemes. For example, the words 'sky' and 'lie' both end with the same phoneme /ai/. Before detecting rhyme information, we therefore first transcribe our lyrics to a phonetic representation. Phonetic transcription is language dependent, thus the language of song lyrics would first need to be identified, using e.g. TextCat [2] to determine the correct transcriptor. However, as our test collection presented in this paper features only English songs and we therefore use English phonemes only, we omit details on this step.

We distinguish two patterns of subsequent lines in a song text: *AA* and *AB*. The former represents two rhyming lines,

Feature Name	Description
Rhymes-AA	A sequence of two (or more) rhyming lines ('Couplet')
Rhymes-AABB	A block of two rhyming sequences of two lines ('Clerihew)
Rhymes-ABAB	A block of alternating rhymes
Rhymes-ABBA	A sequence of rhymes with a nested sequence ('Enclosing rhyme')
RhymePercent	The percentage of blocks that rhyme
UniqueRhymeWords	The fraction of unique terms used to build the rhymes

Table 1. Rhyme features for lyrics analysis

Feature Name	Description
exclamation_mark, colon, single_quote, comma, question_mark, dot, hyphen, semicolon	simple count
d0 - d9	counts of digits
WordsPerLine	words / number of lines
UniqueWordsPerLine	unique words / number of lines
UniqueWordsRatio	unique words / words
CharsPerWord	number of chars / number of words
WordsPerMinute	the number of words / length of the song

Table 2. Overview of text statistic features

while the latter denotes non-rhyming. Based on these basic patterns, we extract the features described in Table 1.

A 'Couplet' *AA* describes the rhyming of two or more subsequent pairs of lines. It usually occurs in the form of a 'Clerihew', i.e. several blocks of Couplets *AABBCC...* *ABBA*, or *enclosing rhyme* denotes the rhyming of the first and fourth, as well as the second and third lines (out of four lines). We further measure 'RhymePercent', the percentage of rhyming blocks, and define the unique rhyme words as the fraction of unique terms used to build rhymes 'UniqueRhymeWords'.

Of course, more elaborate rhyming patterns, especially less obvious forms of rhymes, could be taken into account. In order to initially investigate the usefulness of rhyming at all, we do not take into account rhyming schemes based on assonance, semirhymes, or alliterations, amongst others. Initial experimental results lead to the conclusions that some of these patterns may well be worth studying. However, the frequency of such patterns in popular music is doubtful.

3.3 Part-of-Speech Features

Part-of-speech tagging is a lexical categorisation or grammatical tagging of words according to their definition and the textual context they appear in. Different part-of-speech categories are for example nouns, verbs, articles or adjectives. We presume that different genres will differ also in the category of words they are using, and therefore we additionally extract several part of speech descriptors from the lyrics. We count the numbers of: *nouns*, *verbs*, *pronouns*, *relational pronouns* (such as 'that' or 'which'), *prepositions*, *adverbs*, *articles*, *modals*, and *adjectives*. To account for different document lengths, all of these values are normalised by the number of words of the respective lyrics document.

3.4 Text Statistic Features

Text documents can also be described by simple statistical measures based on word or character frequencies. Measures such as the average length of words or the ratio of unique words in the vocabulary might give an indication of the complexity of the texts, and are expected to vary over different genres. The usage of punctuation marks such as exclamation or question marks may be specific for some genres. We further expect some genres to make increased use of apostrophes when omitting the correct spelling of word endings. The list of extracted features is given in Table 2.

All features that simply count character occurrences are normalised by the number of words of the song text to accommodate for different lyrics lengths. 'WordsPerLine' and 'UniqueWordsPerLine' describe the words per line and the unique number of words per line. The 'UniqueWordsRatio' is the ratio of the number of unique words and the total number of words. 'CharsPerWord' denotes the simple average number of characters per word. The last feature, 'WordsPerMinute' (WPM), is computed analogously to the well-known beats-per-minute (BPM) value [1] .

4 EXPERIMENTS

In this section we report on experiments performed for a test collection of 397 song lyrics. We used classical bag-of-words indexing as well as the feature sets introduced in Section 3. Genre classification was done for a range of setups with Naïve Bayes, k-Nearest Neighbour with different values for k, SVMs with linear and polynomial kernels (more complex kernels did not improve accuracy), and a Decision

[1] Actually we use the ratio of the number of words and the song length in seconds to keep feature values in the same range. Hence, the correct name would be 'WordsPerSecond', or WPS.

Genre	Songs	Genre	Songs
Country	29	Pop	42
Folk	44	Punk Rock	40
Grunge	40	R&B	40
Hip-Hop	41	Reggae	33
Metal	46	Slow Rock	42

Table 3. Composition of the test collection

Tree (J48). We used a slightyly modified version of the Weka toolset for running and evaluating our experiments [2] .

4.1 Test Collection

The task of compiling a feasible test collection for genre classification by lyrics is tedious for the preprocessing task must provide correct lyrics both in terms of structure and content. Therefore, all lyrics were manually preprocessed in order to remove additional markup like '[2x]' or '[chorus]', and to include the unabridged lyrics for all songs. We paid special attention to completeness in terms of the resultant text documents being an as adequate and proper transcription of the songs' lyrics as possible.

Starting from a private collection of about 12.000 songs, we selected a random sample of 30-45 songs from each of the ten genres listed below. We removed songs that seemed not feasible because of their length (extremely short or overlength tracks) and removed songs in languages other than English to avoid the complexity that comes with phoneme transcription to other languages. This resulted in a collection of 397 remaining songs.

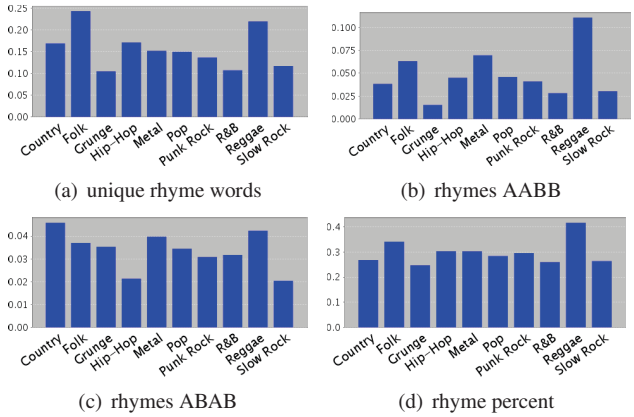
We aimed at having a high number of different artists in order to prevent biased results by too many songs from the same artist; our collection comprises 143 different artists. Table 3 shows the numbers of songs per genre.

The lyrics for all songs were automatically retrieved from the Internet. All songs were manually filtered, e.g. we removed instrumental songs. Annotations like [chorus] or [verse] were substituted by the respective parts of the text, i.e. the full texts of all choruses and verses were placed correctly.

4.2 Feature Analysis

To illustrate the discriminative power of the new feature set, we present how values vary across different genres. Due to space limitations, we selected a subset of four exemplary features from each type of rhyme, part-of-speech, and text statistic features.

In Figure 1, a selected set of Rhyme features is illustrated. The first subplot shows the number of unique words



(a) unique rhyme words (b) rhymes AABB

(c) rhymes ABAB (d) rhyme percent

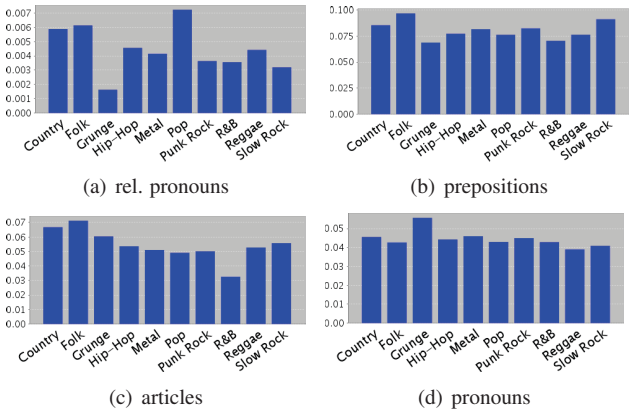
Figure 1. Average values for rhyme features

used to build the rhyme patterns. It can be observed that 'Folk' and 'Reggae' have by far the highest value, while other genres like 'R&B', 'Slow Rock' and 'Grunge' are obviously using less rhyming language. 'Reggae' and 'Grunge' can also be well distinguished by the *AABB* pattern, as well as 'R&B' and 'Slow Rock'. The latter can also be well discriminated against other genres regarding the usage of the *ABAB* pattern, while 'Reggae' is particularly intriguing when it comes to the total percentage of rhyme patterns in its song lyrics. Other rhyme features have similar characteristics, with *AA* patterns having being the least discriminative between the genres.

Figure 2 gives an overview of part-of-speech features, namely relational pronouns, prepositions, articles, and pronouns. Relational pronouns show a pretty strong fluctuation acrros all genres, with 'Grunge' notably having the lowest value, while 'Pop', 'Country' and 'Folk' exhibit the highest values. Articles help to distinguish especially 'R&B' from the other features by having a much lower average than the other genres. To some extent, also 'Folk' can be discriminated by having the highest value. Pronoun usage is mostly equally distributed, but 'Grunge' stands out with a much higher value. Preposition also do not show a huge variation across the different classes. For the other features, nouns, verbs and adjectives have very similar values, while both adverbs and modals do show some fluctuations but do not significantly differ across genres.

An overview of the most interesting statistic features is presented in Figure 3. Genres like 'Reggae', 'Punk Rock' and 'Hip-hop' use an expressive language that employs a lot of exclamation marks. Similar, 'Hip-Hop', 'Punk Rock' and 'Metal' seem to use a lot of digits in their lyrics. This can be explained partly with the names of other artists mentioned in the songs, as in many 'Hip-Hop' pieces of '50 Cent' or '2 Pac'. Also, 911 is an often mentioned number. 'Folk', 'Hip-Hop', and to a lesser extent also 'Country' seem to be characterised by more creative lyrics, which

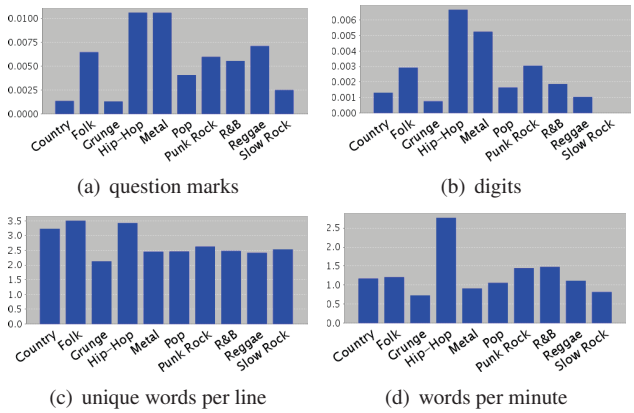
[2] http://http://www.cs.waikato.ac.nz/ml/weka/



(a) rel. pronouns (b) prepositions

(c) articles (d) pronouns

Figure 2. Average values for part-of-speech features



(a) question marks (b) digits

(c) unique words per line (d) words per minute

Figure 3. Average values for statistic features

manifests in a bigger word pool used in these genres. Last, 'Words per Minute' is the most discriminative feature for 'Hip-Hop', which indeed is characterised by fast language, and in most of the cases has very short lead-in and fade-out sequences. Also 'R&B' and 'Punk Rock' have these characteristics, while 'Grunge', 'Slow Rock' and 'Metal' feature long instrumental parts or a more moderate singing tempo.

4.3 To Stem or Not to Stem

We used Porter's stemming algorithm to remove word endings. Its impact on classification accuracies varies; in some cases, it lowers classification accuracy. We performed experiments for a standard Naïve Bayes classifier, and Support Vector Machines (SVM) with linear kernel and Decision Trees (J48) for three different dimensionalities computed by document frequency thresholding feature selection. For the Bayes classifier and SVMs, stemming improved the results for all three dimensionalities (full set, 3000, 300). For the Decision Tree setup, it did not yield improvements. Overall, we did not find significant differences in accuracies, yet, we had expected setups without stemming to be superior caused by word endings in 'slang' language.

4.4 Classification Results

We performed a set of classification experiments, a summarisation of the results is given in Table 4. It shows the experiments we performed on the test collection, omitting results on different values for k for the k-Nearest Neighbour and kernels other than the linear one for SVMs, as they showed the same trends, but had a worse overall performance. For a set of ten genres, we assume the absolute baseline for classification experiments to be the relative number of tracks in the largest class, i.e. $46/397 = 11.59\%$. However, we rather consider the results obtained with the bag-of-words approach as the baseline to compare to. The values given are macro-averaged classification accuracies for the correctly classified tracks computed with ten-fold cross validation. We want to point out that classification accuracies are lower than comparable results based on audio content, yet we show that genre classification by lyrics features can be complementary for they might classify different genres better than audio only.

For all three classifiers, feature sets including our proposed features achieved the best results (experiments 6, 4, and 3). The addition to the bag-of-words approach (with stemming) always yields better results, this becomes even clearer in the case of the SVM (experiments 6, 8, and 10). The best result of all experiments was achieved with a dimensionality of 38, which is sa significantly lower dimensionality than the one of the best performing bag-of-words approach (experiment 6, dimensionality: 3336). We want to point out the improvement in performance of about 7.5% absolute between 24.83 and 33.47 (relative increase of about a third). We also performed the same set of experiments with non-stemmed bag-of-words features and combinations thereof; they revealed the same trends but are omitted due to space limitations.

5 CONCLUSIONS

In this paper, we introduced style and rhyme features for lyrics processing. We showed that all our feature sets outperform the baseline in terms of classification accuracies in a set of experiments. Further, we showed that the combination of style features and classical bag-of-words features outperforms the bag-of-words only approach and in some cases reduces the needed dimensionality.

These results indicate that style and rhyme features can contribute to better music classification or a reduction in dimensionality for similar classification accuracies. Hence, we plan to investigate more sophisticated rhyme patterns. As the classification experiments reported in this paper were based on a simple concatenation approach, more elaborate

Exp.	Feature Combination	Dim	5-NN		Naïve Bayes	SVM		Decision Tree	
1	rhyme	6	13.17		13.67	− 12.83	−	13.82	-
2	part-of-speech (pos)	9	15.99		20.79	16.04	−	15.89	-
3	text statistic	23	**28.53**	+	20.6	− 30.12		**26.72**	
4	rhyme / pos / text statistic	38	25.33	+	23.37	**33.47**	+	24.60	
5	bag-of-words #1	3298	13.63		25.89	24.83		23.63	
6	bag-of-words / rhyme / pos / text statistic #1	3336	13.88		**27.58**	29.44	+	24.39	
7	bag-of-words #2	992	12.34		25.13	21.96		23.11	
8	bag-of-words / rhyme / pos / text statistic #2	1030	16.56		25.78	27.37		23.77	
9	bag-of-words #3	382	14.06		22.74	22.27		22.17	
10	bag-of-words / rhyme / pos / text statistic #3	420	15.06		24.50	29.36		24.05	

Table 4. Classification results for different feature combinations. The highest accuracies per column are printed in bold face; statistically significant improvement or degradation over the base line experiment (5, column-wise) is indicated by (+) or (−), respectively

methods such as ensemble learning will be taken into consideration. Finally, we plan on improve lyrics preprocessing by heuristics for rhyme detection and automatic alignment thereof.

References

[1] Stephan Baumann, Tim Pohle, and Shankar Vembu. Towards a socio-cultural compatibility of mir systems. In *Proc. of the 5th Int. Conf. of Music Information Retrieval (ISMIR'04)*, pages 460–465, Barcelona, Spain, October 10-14 2004.

[2] William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *Proc. of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, USA, 1994.

[3] Jonathan Foote. An overview of audio information retrieval. *Multimedia Systems*, 7(1):2–10, 1999.

[4] Peter Knees, Markus Schedl, and Gerhard Widmer. Multiple lyrics alignment: Automatic retrieval of song lyrics. In *Proc. of 6th Int. Conf. on Music Information Retrieval (ISMIR'05)*, pages 564–569, London, UK, September 11-15 2005.

[5] Beth Logan, Andrew Kositsky, and Pedro Moreno. Semantic analysis of song lyrics. In *Proc. of the 2004 IEEE Int. Conf. on Multimedia and Expo (ICME'04)*, pages 827–830, Taipei, Taiwan, June 27-30 2004.

[6] Jose P. G. Mahedero, Álvaro Martínez, Pedro Cano, Markus Koppenberger, and Fabien Gouyon. Natural language processing of lyrics. In *Proc. of the 13th annual ACM Int. Conf. on Multimedia (ACMMM'05)*, pages 475–478, Singapore, 2005.

[7] Tamsin Maxwell. Exploring the music genome: Lyric clustering with heterogeneous features. Master's thesis, University of Edinburgh, 2007.

[8] Cory McKay and Ichiro Fujinaga. Musical genre classification: Is it worth pursuing and how can it be improved? In *Proc. of the 7th Int. Conf. on Music Information Retrieval (ISMIR'06)*, pages 101–106, Victoria, BC, Canada, October 8-12 2006.

[9] Robert Neumayer and Andreas Rauber. Integration of text and audio features for genre classification in music information retrieval. In *Proc. of the 29th European Conf. on Information Retrieval (ECIR'07)*, pages 724–727, Rome, Italy, April 2-5 2007.

[10] Robert Neumayer and Andreas Rauber. Multi-modal music information retrieval - visualisation and evaluation of clusterings by both audio and lyrics. In *Proc. of the 8th RIAO Conf. (RIAO'07)*, Pittsburgh, PA, USA, May 29th - June 1 2007.

[11] Elias Pampalk, Arthur Flexer, and Gerhard Widmer. Hierarchical organization and description of music collections at the artist level. In *Research and Advanced Technology for Digital Libraries ECDL'05*, pages 37–48, 2005.

[12] Elias Pampalk, Arthur Flexer, and Gerhard Widmer. Improvements of audio-based music similarity and genre classificaton. In *Proc. of the 6th Int. Conf. on Music Information Retrieval (ISMIR'05)*, pages 628–633, London, UK, September 11-15 2005.

[13] Gerald Salton. *Automatic text processing – The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc., 1989.

[14] George Tzanetakis and Perry Cook. Marsyas: A framework for audio analysis. *Organized Sound*, 4(30):169–175, 2000.