Multi-modal Analysis of Music: A large-scale Evaluation

Rudolf Mayer
Institute of Software Technology and Interactive Systems
Vienna University of Technology
Vienna, Austria
mayer@ifs.tuwien.ac.at

Robert Neumayer
Department of Computer and Information Science
Norwegian University of Science and Technology
Trondheim, Norway
neumayer@idi.ntnu.no

***Abstract*—Multimedia data by definition comprises several different types of content. Music specifically inherits *audio* at its core, *text* in the form of lyrics, *images* by means of album covers, and *video* in the form of music videos. Yet, in many Music Information Retrieval applications, only the audio content is utilised. A few recent studies have however shown the usefulness of incorporating also other modalities; in most of these studies, textual information in the form of song lyrics or also artist biographies, were employed. Following this direction, the contribution of this paper is a large-scale evaluation of the combination of audio and text (lyrics) features for genre classification, on a database comprising over 20.000 songs. We briefly present the audio and lyrics features employed, and provide an in-depth discussion of the experimental results.**

I. Introduction

With the ever-growing spread of music available in digital formats – be it in online music stores or on consumers' computer or mobile music players – Music Information Retrieval (MIR) as a research area dealing with ways to organise, structure and retrieve such music, is of increasing importance. Many of its typical task such as genre classification or similarity retrieval / recommendation often rely on only one of the many modalities of music, namely the audio content itself. However, music comprises many more different modalities. Text is present in the form of song lyrics, as well as artist biographies or album reviews, etc. Many artists and publishers put emphasis on carefully designing an album cover to transmit a message coherent with the music it represents. Similar arguments also hold true for music videos.

Recent research has to some extent acknowledged the multi-modality of music, with most research studies focusing on lyrics for e.g. emotion, mood or topic detection. In this work, we apply our previous work on extracting rhyme and style features from song lyrics, with the goal of improving genre classification. Our main contribution is a large-scale evaluation on a database comprising over 20.000 songs from various different genres. Our goal in this paper is to show the applicability of our techniques to, and the potential of lyrics-based features on a larger test collection.

The remainder of this paper is structured as follows. In Section II, we briefly review related work in the field of multi-modal music information retrieval. Section III will outline the audio and lyrics features employed in our study. In Section IV, we describe our test collection, and outline its significant properties, while in Section V we discuss the results on genre classification on this collection. Finally, Section VI will give conclusions and an outlook on future work.

II. Related Work

Music Information Retrieval is a sub-area of information retrieval concerned with adequately accessing (digital) audio. Important research directions include, but are not limited to similarity retrieval, musical genre classification, or music analysis and knowledge representation. Comprehensive overviews of the research field are given in [1], [2].

The still dominant method of processing audio files in music information retrieval is by analysis of the audio signal, which is computed from plain wave files or via a preceding decoding step from other wide-spread audio formats such as MP3 or the (lossless) Flac format. A wealth of different descriptive features for the abstract representation of audio content have been presented. An early overview on content-based music information retrieval and experiments is given in [3] and [4], focussing mainly on automatic genre classification of music.

In this work, we employ mainly the Rhythm Patterns, Rhythm Histograms and Statistical Spectrum Descriptors [5], which we will discuss in more detail in Section III. Other feature sets may include for example MPEG-7 audio descriptors, MARSYAS or the Chroma feature set [6], which attempt to represent the harmonic content (e.g, keys, chords).

Several research teams have further begun working on adding textual information to the retrieval process, predominantly in the form of song lyrics and an abstract vector representation of the term information contained in text documents. A semantic and structural analysis of song lyrics is conducted in [7]. The definition of artist similarity via song lyrics is given in [8]. It is pointed out that acoustic similarity is superior to textual similarity yet a combination of both approaches might lead to better results.

Also, the analysis of karaoke music is an interesting new research area. A multi-modal lyrics extraction technique for tracking and extracting karaoke text from video frames is presented in [9]. Some effort has also been spent on the automatic synchronisation of lyrics and audio tracks at a syllabic level [10]. A multi-modal approach to query music,

text, and images with a special focus on album covers is presented in [11]. Other cultural data is included in the retrieval process e.g. in the form of textual artist or album reviews [12].

Another area were lyrics have been employed is the field of emotion detection and classification, for example [13], which aims at disambiguating music emotion with lyrics and social context features. More recent work combined both audio and lyrics-based feature for mood classification [14].

First results for genre classification using the rhyme features used later in this paper are reported in [15]; these results particularly showed that simple lyrics features may well be worthwhile. This approach has further been extended on two bigger test collections, and to combining and comparing the lyrics features with audio features in [16].

III. Audio and Lyrics Features

In this section we describe the set of audio and lyrics features we employed for our experiments. The audio feature sets comprise Rhythm Patterns, Statistical Spectrum Descriptors, and Rhythm Histograms. The latter two are based on the Rhythm Patterns features, and skip or alter some of the processing steps, resulting in a different feature dimensionality. The lyrics features are bag-of-words features computed from tokens or terms occurring in documents, rhyme features taking into account the rhyming structure of lyrics, features considering the distribution of certain parts-of-speech, and text statistics features covering average numbers of words and particular characters.

A. Rhythm Patterns

Rhythm Patterns (RP) are a feature set for handling audio data based on analysis of the spectral audio data and psycho-acoustic transformations [17], [5].

In a pre-processing stage, music in different file formats is converted to raw digital audio, and multiple channels are averaged to one. Then, the audio is split into segments of six seconds, possibly leaving out lead-in and fade-out segments, and further skipping other segments, e.g. out of the remaining segments every third one may be processed.

The feature extraction process for a Rhythm Pattern is then composed of two stages. For each segment, the spectrogram of the audio is computed using the short time Fast Fourier Transform (STFT). The window size is set to 23 ms (1024 samples) and a Hanning window is applied using 50 % overlap between the windows. The Bark scale, a perceptual scale which groups frequencies to critical bands according to perceptive pitch regions [18], is applied to the spectrogram, aggregating it to 24 frequency bands. Then, the Bark scale spectrogram is transformed into the decibel scale, and further psycho-acoustic transformations are applied: computation of the Phon scale incorporates equal loudness curves, which account for the different perception of loudness at different frequencies [18]. Subsequently, the values are transformed into the unit Sone. The Sone scale relates to the Phon scale in the way that a doubling on the Sone scale sounds to the human ear like a doubling of the loudness. This results in a psycho-acoustically modified Sonogram representation that reflects human loudness sensation.

In the second step, a discrete Fourier transform is applied to this Sonogram, resulting in a (time-invariant) spectrum of loudness amplitude modulation per modulation frequency for each individual critical band. After additional weighting and smoothing steps, a Rhythm Pattern exhibits magnitude of modulation for 60 modulation frequencies (between 0.17 and 10 Hz) on 24 bands, and has thus 1440 dimensions.

In order to summarise the characteristics of an entire piece of music, the feature vectors derived from its segments are simply averaged by computing the median. This approach extracts suitable characteristics of semantic structure for a given piece of music to be used for music similarity tasks.

B. Statistical Spectrum Descriptors

Computing Statistical Spectrum Descriptors (SSD) features relies on the first part of the algorithm for computing RP features. Statistical Spectrum Descriptors are based on the Bark-scale representation of the frequency spectrum. From this representation of perceived loudness, seven statistical measures are computed for each of the 24 critical band, in order to describe fluctuations within the critical bands. The statistical measures comprise mean, median, variance, skewness, kurtosis, min- and max-value. A Statistical Spectrum Descriptor is extracted for each selected segment. The SSD feature vector for a piece of audio is then calculated as the median of the descriptors of its segments.

In contrast to the Rhythm Patterns feature set, the dimensionality of the feature space is much lower – SSDs have 168 instead of 1440 dimensions, still at matching performance in terms of genre classification accuracies [5].

C. Rhythm Histogram Features

The Rhythm Histogram features are a descriptor for the rhythmic characteristics in a piece of audio. Contrary to the Rhythm Patterns and the Statistical Spectrum Descriptor, information is not stored per critical band. Rather, the magnitudes of each modulation frequency bin (at the end of the second phase of the RP calculation process) of all 24 critical bands are summed up, to form a histogram of 'rhythmic energy' per modulation frequency. The histogram contains 60 bins which reflect modulation frequency between 0.168 and 10 Hz. For a given piece of audio, the Rhythm Histogram feature set is calculated by taking the median of the histograms of every 6 second segment processed.

The dimensionality of Rhythm Histograms is with 168 features as well much lower than with the Rhythm Patterns.

D. Bag-Of-Words

Classical bag-of-words indexing at first tokenises all text documents in a collection, most commonly resulting in a set of words representing each document. Let the number of documents in a collection be denoted by N, each single document by d, and a term or token by t. Accordingly, the

TABLE I
RHYME FEATURES FOR LYRICS ANALYSIS

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 II
OVERVIEW OF TEXT STATISTIC FEATURES

Feature Name	Description
exclamation_mark, colon, single_quote, comma, question_mark, dot, hyphen, semicolon	simple counts of occurrences
d0 - d9	occurrences 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

term frequency $tf(t,d)$ is the number of occurrences of term t in document d and the *document frequency* $df(t)$ the number of documents term t appears in.

The process of assigning weights to terms according to their importance or significance for a document is called 'term-weighing'. The basic assumptions are that terms occurring very often in a document are more important for classification, whereas terms that occur in a high fraction of all documents are less important. The weighing we rely on is the most common model, namely the *term frequency times inverse document frequency* [19]. These weights are computed as:

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

This results in vectors of weight values for each document d in the collection, i.e. each lyrics document. This representation also introduces a concept of distance, as lyrics that contain a similar vocabulary are likely to be semantically related. We did not perform stemming in this setup, earlier experiments showed only negligible differences for stemmed and non-stemmed features [15]; the rationale behind using non-stemmed terms is the occurrence of slang language in some genres.

E. Rhyme Features

Rhyme denotes the consonance or similar sound of two or more syllables or whole words. This linguistic style is most commonly used in poetry and songs. The rationale behind the development of rhyme features is that different genres of music should exhibit different styles of lyrics. We assume the rhyming characteristics of a song to be given 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 their characteristic sound. To automatically identify such patterns we introduce several descriptors from the song lyrics to represent different types of rhymes.

For the analysis of rhyme structures we do not rely on lexical word endings, but rather apply a more correct approach based on phonemes – the sounds or groups thereof in a language. Hence, we first need to transcribe the lyrics to a phonetic representation. The words 'sky' and 'lie', for instance, both end with the same phoneme /ai/. Phonetic transcription is language dependent; however, as our test collection is composed of tracks predominantly in English language, we exclusively use English phonemes.

After transcribing the lyrics into a phoneme representation, we distinguish two patterns of subsequent lines in a song text: *AA* and *AB*. The former represents two rhyming lines, while the latter denotes non-rhyming. Based on these basic patterns, we extract the features described in Table I.

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 such as *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. Besides, we define the unique rhyme words as the fraction of unique terms used to build rhymes 'UniqueRhymeWords', which describes whether rhymes are frequently formed using the same word pairs, or a wide variety of words is used for the rhymes.

In order to initially investigate the usefulness of rhyming at all, we do not take into account rhyming schemes based on assonance, semirhymes, alliterations, amongst others. We also did not yet incorporate more elaborate rhyme patterns, especially not the less obvious ones, such as the 'Ottava Rhyme' of the form *ABABABCC*, and others. Also, we assign to all the rhyme forms the same weights, i.e. we do for example not give more importance to complex rhyme schemes. Experimental results lead to the conclusion that some of these patterns may well be worth studying. An experimental study on the frequency of occurrences might be a good starting point first, as modern popular music does not seem to contain many of these patterns.

F. 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*,

TABLE III
COMPOSITION OF THE TEST COLLECTION

Genre	Artists	Albums	Songs
Pop	1.150	1.730	6.156
Alternative	457	780	3.699
Rock	386	871	3.666
Hip-Hop	308	537	2.234
Country	223	453	1.990
R&B	233	349	1.107
Christian	101	170	490
Comedy	20	44	206
Reggae	30	48	121
Dance / Electronic	48	59	112
Blues	23	39	99
Jazz	38	49	97
Scores / Soundtrack	46	24	70
Classical	21	21	62
Total	3.084	5.074	20.109

and *adjectives*. To account for different document lengths, all of these values are normalised by the number of words of the respective lyrics document.

G. 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. Further, 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 II.

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].

IV. TEST COLLECTION

The collection we used in the following set of experiments was introduced in [20]. It is a subset of the collection marketed through Verisign Austria's[2] content download platform, and comprises 60.223 of the most popular audio tracks by more than 7.500 artists. The collection contained several duplicate songs, which were removed for our experiments.

For 41.679 of these, song lyrics have been automatically downloaded from lyrics portals on the Web. We considered only songs that have lyrics with at descent length, to remove lyrics that are most probably not correctly downloaded, but just contain HTML fragments.

The tracks are manually assigned by experts to one or more of a total of 34 different genres. 9977 songs did not receive any ratings at all, and were thus removed from the databases for our purpose of genre classification. Further, we only considered songs that have a rather clear assignment to one genre, thus of those tracks that have received more than one voting, we only considered those that have at least $2/3$ of the experts agreeing on the same genre, skipping another 12572 songs. Also, genres that had less than 60 songs were not considered.

Finally, after all the removal steps, and thus considering only tracks that have both a clear genre assignment and lyrics in proper quality available, we obtain a collection of 20.109 songs, categorised into 14 genres. Details on the number of songs per genre in this collection can be found in Table III. It is noticeable that the different genres vary a lot in size. As such, the smallest class is "Classical", with just 62 songs, or 0.29%. Also, Scores / Soundtrack, Jazz, Blues, Dance / Electronic, Reggae and Comedy comprise less or just about 1% of the whole collection. Contrarily to this, the largest class, Pop, holds 6.156 songs, or 30.6%, followed by two almost equally big classes, Alternative and Rock, each accounting for almost 3.700 songs or 18.4% of the collection.

While this collection clearly is imbalanced towards the Pop and Alternative Rock and Rock genres, together accounting for more than $2/3$ of the collection, it can surely be regarded as a lifelike collection. For the experimental results, the class distribution implies a baseline result of the size of the biggest class, thus 30.6%.

V. EXPERIMENTS

In this section we present the results of our experiments, where we will compare the performance of audio features and text features using various classifiers. To this end, we first extracted the audio and lyrics feature sets described in Section III. We then build several combinations of these different feature sets, both separately within the lyrics modality, as well as combinations of audio and lyrics feature sets. This results in several dozens of different feature set combinations, out of which the most interesting ones are presented here. Most combinations are done with the SSD features, as those are the best performing ones.

For all our experiments, we employed the WEKA machine learning toolkit [3], and unless otherwise noted used the default settings for the classifiers and tests. We used k-Nearest-Neighbour, Naïve Bayes, J48 Decion Trees, and Support Vector Machines, and performed the experiments based on a ten-fold cross-validation. All results given in this sections are micro-averaged classification accuracies. Statistical significance testing is performed per column, using a paired t-test with an α value of 0.05. In the result tables, plus signs $(+)$ denote a significant improvement, whereas

[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.

[2] http://www.verisign.at/

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

TABLE IV
CLASSIFICATION ACCURACIES FOR SINGLE AUDIO AND TEXT FEATURE SETS

Dataset	Dim.	1-NN	3-NN	NB	DT	SVM
RH	60	34.93	35.12	20.50	30.32	40.75
RP	1440	38.69	39.41	17.25	30.35	49.47
SSD	168	**46.26**	**46.87**	**26.45**	**39.42**	**55.77**
Rhyme	6	25.29	24.92	22.75	28.13	30.61
POS	9	27.96	26.61	**32.33**	27.81	30.61
TextStat	23	**29.40**	**28.32**	2.13	**33.62**	**32.39**
BOW	20					33.11
BOW	50					36.47
BOW	100					39.31
BOW	300					45.14
BOW	698					48.35
BOW	997					49.42
BOW	1500					50.16
POS, Rhyme	15	27.91	26.87	**29.00**	27.97	30.59
POS, TextStat	32	**30.67**	29.66	3.48	**34.19**	35.22
Rhyme, TextStat	29	27.82	26.46	2.33	33.94	32.54
POS, Rhyme, TextStat	38	30.13	**29.68**	3.82	33.73	**36.09**

minus signs (−) denote significant degradation.

A. Single Feature Sets

Table IV gives a first overview on the results of the classification. Regarding the audio features, shown in the first section of the table, for all classifiers tested, the highest classification accuracies were always achieved with SSD features; all other features in all classifiers were significantly inferior. The results achieved with Naïve Bayes are extremely poor, and are below the above mentioned baseline of the percentage of the largest class, 30.61%. Also, the Decision Tree on RP and RH features fails to beat that baseline. SSD being the highest performing features set, we will use it as the baseline we want to improve on in the subsequent experiments with feature combinations.

As to the lyrics-based rhyme and style features shown in the second section of Table IV, the overall performance is not satisfying. Generally, the text-statistics features are performing best, with the exception of Naïve Bayes, which seems to have some problem with the data set and ranks at an all-time low of 2.13%. Rhyme and POS features on SVMs achieve exactly the baseline, where the confusion matrix reveals that simply all instances have been classified into the biggest genre, thus rendering the classifier useless. Only the part-of-speech features with Naïve Bayes, and the text-statistics on the Decision Tree and SVMs manage to marginally outperform the baseline, of which only text-statistics on Decision Trees have a statistically significant improvement.

The third section in Table IV gives the results with the bag-of-words features, with different numbers of features selected via frequency thresholding, as described in Section III-D. Compared to the audio-only features, the results are really promising, clearly out-performing the RH features, and with a high dimensionality of 1.500 terms even outperforming the RP.

Combining the rhyme and style features, slight improvements can be gained in most cases, as seen in the last section of Table IV. Besides with Naïve Bayes, the best combination always includes the text statistics and part-of-speech features, and two out of four cases also the rhyme features. However, the results are still far away from the lyrics features, and not that much better than the baseline. In comparison to the results on the lyrics-based features reported on the database used in [16], it can be noted that the absolute numbers of classification accuracies for the best combinations are not that much higher in this set of experiments. The relative improvement over the baseline (10% in [16]) can not be matched.

B. Feature Set Combinations

Even though the classification results of the lyrics-based features fall short of the SSD features as the assumed baseline, they can still be utilised to achieve (statistically significant) improvement over the audio-only baseline. Genre classification accuracies for selected feature set combinations are given in Table V.

Due to constraints in computational resources, caused by the big size of the dataset, for this final set of experiments we had to limit the number of classifiers and feature combinations. Thus, we trained only Support Vector Machines, as they have shown to be the by far most performing classifier in the experiments on the single feature sets, as well as in our previous work. To ensure that this is also the case with this data set, we still ran a few selected feature set combinations also with k-NN, Naïve Bayes and Decision Trees, which indeed were clearly outperformed by Support Vector Machines. Further, we focused on combining the lyrics feature sets with SSDs, again as SSD have clearly outperformed the other audio feature sets on the first set of experiments and in our previous work.

The second part of Table V shows the results on combining SSD features with the rhyme and style features. It can be seen that each combination performs better than the SSD features only, and that the improvement is statistically significant. The best result is achieved with combining SSD with all rhyme and style features, i.e. rhyme, part-of-speech and text statistics. This combination achieves 58.09%, an improvement of 2.3 percent points over the baseline, with only a minor increase in the dimensionality.

Combining SSD with the bag-of-words features, as seen in the third part of Table V, also leads to statistically significant improvements, already with only 10 terms used. The best result is achieved when using around 800 keyword dimensions, for which we achieved 60.49% classification accuracy, which is in turn statistically significantly better than the SSD combination with the rhyme and style features.

The last two parts of Table V finally present the results of combining all SSD, rhyme and style and bag-of-words features. One of these combinations also achieves the best result in this experiment series, namely the last combination presented in the table, with 704 dimensions, achieving 60.79%.

TABLE V
CLASSIFICATION ACCURACIES AND RESULTS OF SIGNIFICANCE TESTING FOR COMBINED AUDIO AND TEXT FEATURE SETS. STATISTICALLY SIGNIFICANT IMPROVEMENT OR DEGRADATION OVER DATASETS (COLUMN-WISE) IS INDICATED BY (+) OR (−), RESPECTIVELY

Dataset	Dim.	SVM	
SSD	168	55.77	
SSD / POS	177	56.52	+
SSD / TextStat	191	57.35	+
SSD / POS / Rhyme	183	56.87	+
SSD / Rhyme / TextStat	197	57.42	+
SSD / POS / TextStat	200	57.92	+
SSD / POS / Rhyme / TextStat	206	58.09	+
BOW / SSD	178	56.32	+
BOW / SSD	188	57.12	+
BOW / SSD	198	57.89	+
BOW / SSD	218	58.46	+
BOW / SSD	243	58.84	+
BOW / SSD	318	59.17	+
BOW / SSD	468	59.98	+
BOW / SSD	966	60.49	+
BOW / SSD / TextStat	201	57.59	+
BOW / SSD / TextStat	211	58.14	+
BOW / SSD / TextStat	231	58.74	+
BOW / SSD / TextStat	266	59.22	+
BOW / SSD / TextStat	341	59.95	+
BOW / SSD / TextStat	391	60.17	+
BOW / SSD / TextStat	989	60.68	+
BOW / SSD / TextStat / POS / Rhyme	216	58.39	+
BOW / SSD / TextStat / POS / Rhyme	226	58.90	+
BOW / SSD / TextStat / POS / Rhyme	246	59.20	+
BOW / SSD / TextStat / POS / Rhyme	281	59.70	+
BOW / SSD / TextStat / POS / Rhyme	356	60.07	+
BOW / SSD / TextStat / POS / Rhyme	506	60.57	+
BOW / SSD / TextStat / POS / Rhyme	704	60.79	+

VI. CONCLUSION

In this paper, we presented a large-scale evaluation on multimodal features for automatic musical genre classification. Besides using features based on the audio signal, we utilised a set of features on the song lyrics as an additional, partly orthogonal dimension. Next to measuring the performance of single features sets, we in detail studied the power of combining audio with lyrics-based features.

The main contribution is the large-scale evaluation of these features and their combination, on a database of over 20.000 songs. We showed that similar effects as for the smaller, carefully assembled databases of 600 and over 3.000 songs presented in earlier work hold true as well for this larger database. Besides being a large database, it is also one taken from a real-world scenario, exhibiting potentially more challenging conditions, such as having a very imbalanced distribution of genres.

One surprising observation is that the bag-of-words features alone already achieve very good results, even outperforming Rhythm Patterns features. This, and the improved classification performance achieved on the combination of lyrics and audio feature sets, are promising results for future work in this area. Increased performance gains might be achieved by combining the different feature sets in a more sophisticated approach, e.g. by applying weighting schemes or ensemble classifiers.

REFERENCES

[1] J. Downie, *Annual Review of Information Science and Technology*. Medford, NJ: Information Today, 2003, vol. 37, ch. Music Information Retrieval, pp. 295–340.

[2] N. Orio, "Music retrieval: A tutorial and review," *Foundations and Trends in Information Retrieval*, vol. 1, no. 1, pp. 1–90, September 2006.

[3] J. Foote, "An overview of audio information retrieval," *Multimedia Systems*, vol. 7, no. 1, pp. 2–10, 1999.

[4] G. Tzanetakis and P. Cook, "Marsyas: A framework for audio analysis," *Organized Sound*, vol. 4, no. 30, pp. 169–175, 2000.

[5] T. Lidy and A. Rauber, "Evaluation of feature extractors and psycho-acoustic transformations for music genre classification," in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR'05)*, London, UK, September 11-15 2005, pp. 34–41.

[6] M. Goto, "A chorus section detection method for musical audio signals and its application to a music listening station," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 5, pp. 1783–1794, 2006.

[7] J. P. G. Mahedero, Á. Martínez, P. Cano, M. Koppenberger, and F. Gouyon, "Natural language processing of lyrics," in *Proceedings of the ACM 13th International Conference on Multimedia (MM'05)*, New York, NY, USA, 2005, pp. 475–478.

[8] B. Logan, A. Kositsky, and P. Moreno, "Semantic analysis of song lyrics," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'04)*, Taipei, Taiwan, June 27-30 2004, pp. 827–830.

[9] Y. Zhu, K. Chen, and Q. Sun, "Multimodal content-based structure analysis of karaoke music," in *Proceedings of the ACM 13th International Conference on Multimedia (MM'05)*, Singapore, 2005, pp. 638–647.

[10] D. Iskandar, Y. Wang, M.-Y. Kan, and H. Li, "Syllabic level automatic synchronization of music signals and text lyrics," in *Proceedings of the ACM 14th International Conference on Multimedia (MM'06)*, New York, NY, USA, 2006, pp. 659–662.

[11] E. Brochu, N. de Freitas, and K. Bao, "The sound of an album cover: Probabilistic multimedia and IR," in *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*, C. M. Bishop and B. J. Frey, Eds., Key West, FL, USA, January 3-6 2003.

[12] S. Baumann, T. Pohle, and S. Vembu, "Towards a socio-cultural compatibility of mir systems." in *Proceedings of the 5th International Conference of Music Information Retrieval (ISMIR'04)*, Barcelona, Spain, October 10-14 2004, pp. 460–465.

[13] D. Yang and W. Lee, "Disambiguating music emotion using software agents," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Barcelona, Spain, October 2004.

[14] C. Laurier, J. Grivolla, and P. Herrera, "Multimodal music mood classification using audio and lyrics," San Diego, CA, USA, December 11–13 2008, pp. 688–693.

[15] R. Mayer, R. Neumayer, and A. Rauber, "Rhyme and style features for musical genre classification by song lyrics," in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08)*, September 14-18 2008.

[16] ——, "Combination of audio and lyrics features for genre classification in digital audio collections," in *Proceedings of the ACM Multimedia 2008*, Vancouver, BC, Canada, October 27-31 2008, pp. 159–168.

[17] A. Rauber, E. Pampalk, and D. 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 (ISMIR'02)*, Paris, France, October 13-17 2002, pp. 71–80.

[18] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*, 2nd ed., ser. Series of Information Sciences. Berlin: Springer, 1999, vol. 22.

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

[20] F. Kleedorfer, P. Knees, and T. Pohle, "Oh oh oh whoah! towards automatic topic detection in song lyrics," in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, Philadelphia, PA, USA, September 14 – 18 2008, pp. 287 – 292.