Integration of Text and Audio Features

for Genre Classification in Music Information Retrieval

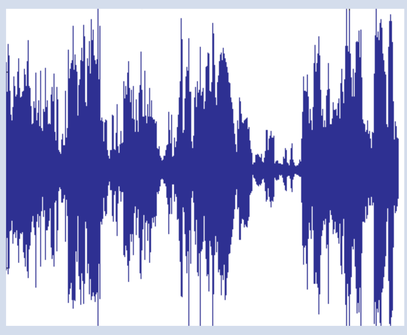
Robert Neumayer and Andreas Rauber
Vienna University of Technology
Institute for Software Technology and Interactive Systems
www.ifs.tuwien.ac.at/mir

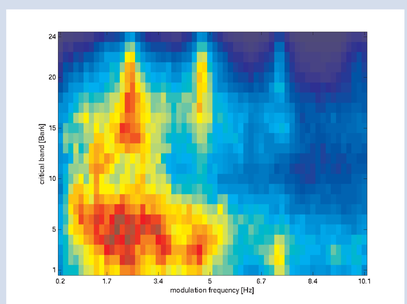
Multimedia content can be described in versatile ways as its essence is not limited to one view. For music data these multiple views include a song's audio features as well as its lyrics. Both of these modalities have their advantages as text may be easier to search in and could cover more of the `content 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 analyse the nature of text and audio feature sets which describe the same audio tracks. Moreover, we propose the use of textual data on top of low level audio features for music genre classification. Further, we show the impact of different combinations of audio features and textual features based on song lyrics.

Audio Feature Extraction

Three features were computed from audio tracks in standard PCM format with 44.1 kHz sampling frequency (e.g. decoded MP3 files).

Rhythm Patterns (RP) 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.
Rhythm Histograms (RH, 60 dimensions) are calculated as the sums of the magnitudes of each modulation frequency bin of all 24 critical bands.





Lyrics Space

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.

This results in tfidf vectors of variable dimensionality suitable for classification experiments.

Genre Classification Experiments

Audio Test Collection

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.

Lyrics Fetching

For every piece of music, three lyrics portals were accessed, using artist name and track title as queries. 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. Then the bag-of-word features for song lyrics are extracted.

Musical Genre Classification

Classification accuracies are shown 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.

Selected Accuracies Based on Audio Features by Genre (RP, 1440 dim)

Selected Accuracies Based on Lyrics Features by Genre (1440 dim)

Audio Accuracies	Genre	Lyrics Accuracies
.2333	BritPop	.2833
.3099	Reggae	.3662
.6169	Punk Rock	.5611
.7844	Hip-Hop	.7490
.4128	Overall	.4445

Combined Genre Classification Results

Experiment	Feature Types	Dimensionality	Classification Accuracy
A1	Rhythm Histograms (RH)	60	.3100
A2	Statistical Spectrum Descriptors (SSD)	168	.4166
A3	Rhythm Patterns (RP)	1440	.4128
L1	Text Features Only	60	.2451
L2	Text Features Only	168	.3204
L3	Text Features Only	1440	.4445
C1	Concatenation of Text and Rhythm Histograms (RH)	120	.3268
C2	Concatenation of Text and Statistical Spectrum Descriptors (SSD)	336	.4817
C3	Concatenation of Text and Rhythm Patterns (RP)	2880	.4841

Conclusion and Outlook

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). For combination experiments (C1 - C3) we use balanced combinations of features, i.e. same dimensionality for audio and lyrics vectors. C2 shows similar accuracy at much lower dimensionality compared to audio and lyrics only features.

Presented at the European Conference on Information Retrieval (ECIR'07), April 2-5 2007, Rome, Italy

Multimedia Understanding through Semantics, Computation and Learning

Part of this work was supported by the European Union through the MUSCLE NoE.