General Purpose Multimedia Dataset - GarageBand 2008

Anders Meng*

March 14, 2008

Abstract

This document describes a general purpose multimedia dataset to be used in cross-media machine learning problems. In more detail we describe the genre taxonomy applied at `http://www.garageband.com`, from where the dataset was collected, and how the taxonomy have been fused into a more human understandable taxonomy. Finally, a description of various features extracted from both the audio and text are presented.
Keywords: Garageband, MP3 music, Review information, Music Genre, Ranking information

1 Description

The Internet site `http://www.garageband.com` is an online music reviewing portal for music artists. It allows artists to review and evaluate other artists music. The feedback can provide valuable information to the artist on his/her musical performance and is an efficient place for detecting new musical trends and talents. An example of a review to the song "Happy Place" by the band "Desperate Cry" (in the genre Metal), is provided by the user "xxxxxx" (not the persons real nick) from St. Petersburg in Florida, quote:

> *Demon Vocals, Samples and Power Riffs*
> The song starts off with a spoken word sample played over power chords and what sounds like keyboard resonance swells. The main vocals are of the standard demonic variety, although cookie monster makes a backing vocal appearance at times. The guitars are well played and the drums are solid. There are a lot of subliminal things happening in the background that could be more samples or some kind of track masking. An interesting effort. - *xxxxxx from St. Petersburg, Florida on 16May2006*

The review consist of a "title" some bodytext and finally information on the user which provided the review, followed by his geographical position and the date on which the review was created.

*This work was supported in part by the Danish Technical Research Council under Project 26-04-0092 Intelligent Sound (www.intelligentsound.org)

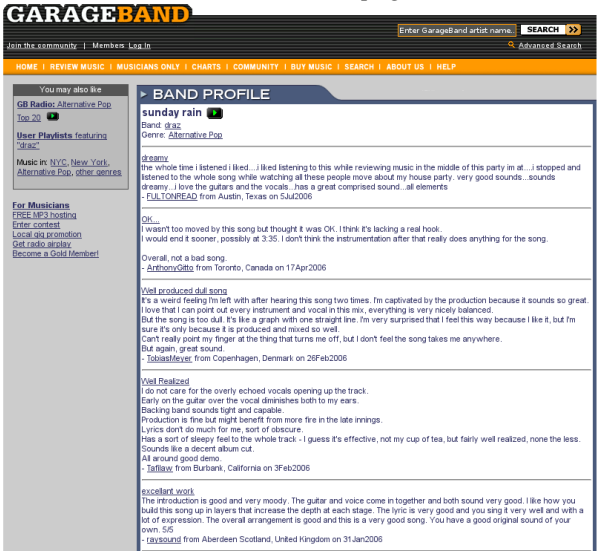
During January 2008 we downloaded a dataset consisting of 16706 song objects by analyzing the top 600 songs within each genre [1]. Each song object, see figure 6, consists of the following information:

- The initial 55 seconds of each downloadable MP3 file
- Review information for each song. Each song is on the average reviewed by 56 people. See example HTML-review page in figure 1.
- A front page with information about musical genre, a ranking within the genre (at the time of the download), top reviews, awards provided by the community and in some cases also extracts from the lyrics. See example in figure 1.

Each artist provide the MP3 file and attach two genres to the song. The artist select the "most correct" genre and a second genre, which can be considered more mainstream. They have done this to ensure a review of the song. The Garageband genre taxonomy consists of 47 musical genres, listed here in no specific order:
Acoustic, Alternative Metal, Alternative Pop, Alternative Rock, Ambient, Americana, Blues, Blues Rock, Classical, Comedy, Country, Dance Electronic, Electronica, Emo, Experimental Electronica, Experimental Rock, Folk, Folk Rock, Funk, Groove Rock, Hard Rock, Hardcore Metal, Hip Hop, Indie Rock, Industrial, Instrumental Rock, Jazz, Latin, Metal, Modern Rock, Pop, Pop Punk, Pop Rock, Power Pop, Progressive Rock, Punk, R&B, Rap, Reggae, Rock, Ska, Spoken Word, Techno, Trance, World, World Fusion.

[1] Not every band allows the publicity to download their music. We have only downloaded music snippets from the public available ones.

(a) HTML Front page



(b) Review information page

Figure 1: Example of (a) HTML front page, and (b) Review information page

2 Fusing the genre Taxonomy

Running through the genre taxonomy there exists quite a few genres where the users (and artists) will be confused. One such example is Electronic and Electronica. It is not very obvious how one can differentiate between these two genres. This would require some very detailed description on how one can differentiate between the two. This information, though, is not present on their site. To minimize the confusion between genres we have created a fused genre taxonomy. The fusion process have been based on information's found in the textual reviews. Figure 2 shows a diagram illustrating the steps towards the fused taxonomy. The individual steps in the diagram are explained in more detail below:

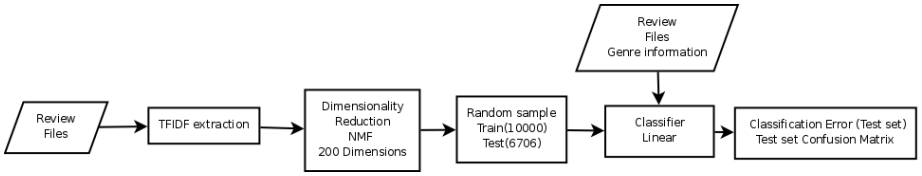


Figure 2: Overview of the genre fusion process.

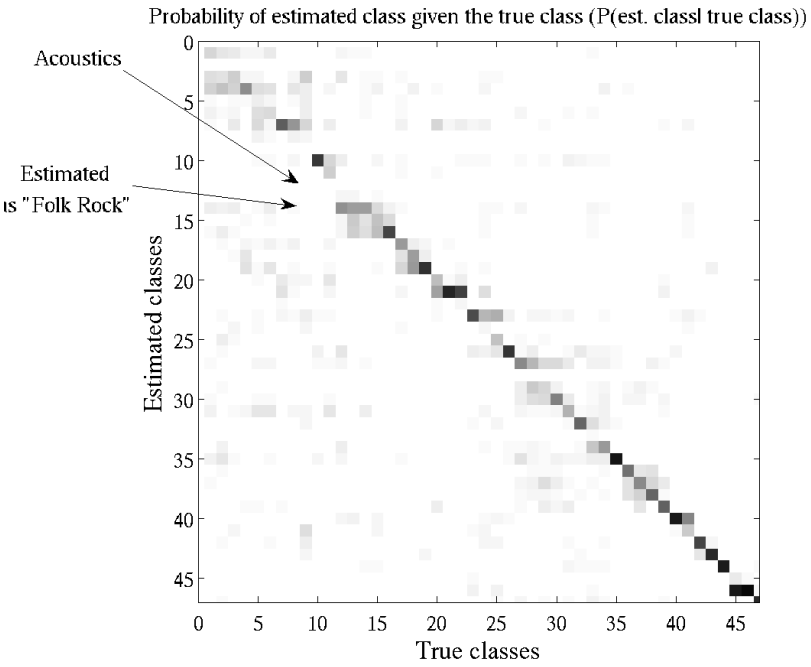


Figure 3: Confusion matrix created from test data. Example of human confusion is the genre Acoustics, which is confused mostly with folk rock and folk. Complete black refers to $P(C_{estimated}|C_{true}) = 1$ and white to $P(C_{estimated}|C_{true}) = 1$.

- Extract the genre information from each song (we restrict this to a single genre. Most of the music snippets have sub-genre information).
- Strip genre information from each review file (header information).
- Create a term-document matrix (counts) where terms corresponds to words and the documents consists (on the average) of 56 reviews. Terms which does not occur in 10 or more documents are pruned away. Furthermore, terms which does not appear at least twice in a document are pruned away. This removes spelling errors from the corpus. Stemming is applied. The resulting term-document matrix is of dimension 30566×16706. The "libbow" software [5] was used to create the term document matrix.
- The term document matrix was column normalized, hence, not differentiating between the lengths of the documents. After normalization the rows were weighted using the IDF scheme (inverse document frequency), thus, penalizing non discriminative terms.
- Dimensionality reduction was performed using the non-negative matrix factorization (NMF) utilizing a squared error function. An analysis of variability in the number of dimensions was carried out by sweeping the number of dimensions from 10 to 200 and observe the cross validation error on the training set (consisting of 10000 randomly selected songs of the complete set of 16706). The analysis showed that 200 dimensions were appropriate, obtaining a cross validation error of $\approx 50\%$.
- A linear classifier was optimized to the training set of 10000 songs randomly selected from the dataset. Each of the remaining 6706 songs in the test set was used to create a confusion matrix between the predicted genre and "true" genre. The ordered and normalized confusion matrix can be seen from figure 3. Here black indicate 1 (100%) and white 0. An example of the deficiency in the original garageband taxonomy is illustrated with the arrows. The arrow showing the genre "acoustics" are newer predicted from the users choice of words. The "acoustics" genre is typically confused with that of "folk" and "folk rock".
- From the normalized confusion matrix we created a dendrogram by measuring a distance between the "true genres" (garageband ones), given all the predicted genres. The distance measure can be expressed as:

$$D(i,j) = \text{Dist}\left\{\text{P}(\text{C}_\text{e} = \text{k}|\text{C}_\text{t} = \text{i}), \text{P}(\text{C}_\text{e} = \text{k}|\text{C}_\text{t} = \text{j})\right\}, \quad \text{for k} = \{1,\ldots,18\} \text{ and i,j} = \{1,\ldots,47\} \tag{1}$$

  where C_e refers to the estimated genre and C_t refers to the true genre. The distance ($D(i,j)$) indicates a distance between the true genres based on the observed confusions. For the dendrogram creation we used a cosine distance measure, simply extracting the angle between the distributions. This cosine distance measure has the property that the dendrogram values lies in the interval $0-1$, where 1 refers to complete similarity.
- Figure 4 shows the created dendrogram. For our fused genre taxonomy we chose 0.7 as a threshold value, since, it seems that the taxonomy here is the most stable (due to a larger distance to all the branching/merging of genres).

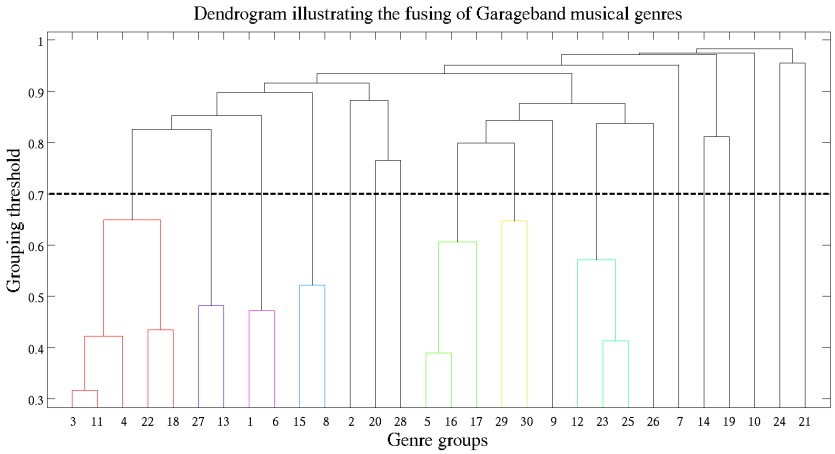


Figure 4: Dendrogram illustrating the groupings of genres determined from the confusion matrix

The leaf nodes in the dendrogram of figure 4 and fused genre names (and id's) are shown in table 1. There is several observations which illustrate the natural confusion of genres by humans. A very straightforward example is the fusion of electronica and electronic. This confusion is pretty obvious, since in order to discriminate between these two a very detailed taxonomy description would be needed (and there is none). Another interesting fusing is that of Emo, Punk and Pop Punk. Basically, Emo is a derived genre from "Hardcore Punk", see e.g. `http://en.wikipedia.org/wiki/Emo`.

The prior distribution of the 16706 songs in the fused 18-genre taxonomy is shown in figure 5. The prior is not uniform. Jazz has the smallest size ≈ 250 song objects, while the largest group is the rock group with approximately 3000 songs. Many of the genres are well represented with 1000 song objects or more.

Dendrogram id	Garageband genres	New Id	Fused taxonomy
3 11 4 22 18	Alternative pop, Pop, Pop rock Power pop Alternative Rock, Indie Rock Hard Rock, Modern Rock Rock	1	Rock
27 13	Instrumental Rock Progressive Rock	2	Progressive Rock
1 6	Acoustic, Folk, Folk Rock Americana, Country	3	Folk/Country
15 8	Emo Pop Punk, Punk	4	Punk
2	Alternative Metal, Hardcore Metal, Metal	5	Heavy Metal
20	Funk, Groove Rock, R&B	6	Funk
28	Jazz	7	Jazz
5 16 17	Ambient, Electronica, Electronic Experimental Electronica Experimental Rock	8	Electronica
29 30	Latin World, World Fusion	9	Latin
9	Classical	10	Classical
12 23 25	Dance Techno Trance	11	Techno
26	Industrial	12	Industrial
7	Blues, Blues Rock	13	Blues
14	Reggae	14	Reggae
19	Ska	15	Ska
10	Comedy	16	Comedy
24	Hip-Hop, Rap	17	Rap
21	Spoken Word	18	Spoken Word

Table 1: Overview of the Garageband 47 genre-taxonomy and the fused 18 genre-taxonomy generated from an analysis of the reviewer information.

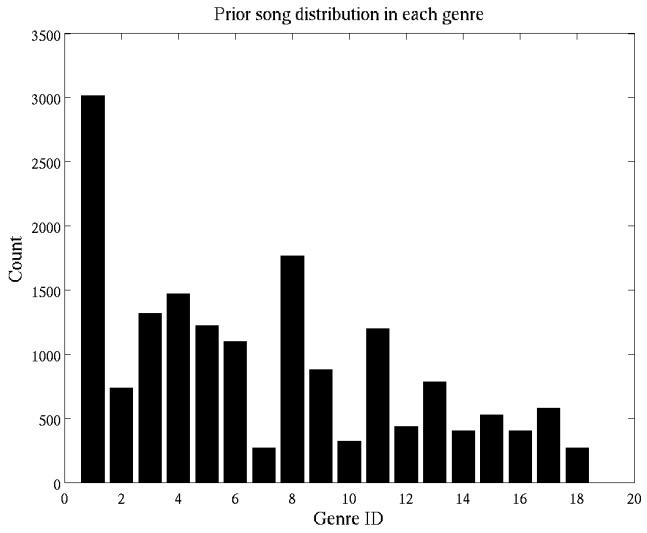


Figure 5: The prior distribution of the fused genre taxonomy.

3 Feature extraction

Features were extracted from both the audio and the available text. The features extracted are explained in more detail in the following subsection. Figure 6 illustrates the information typically present in each song object.

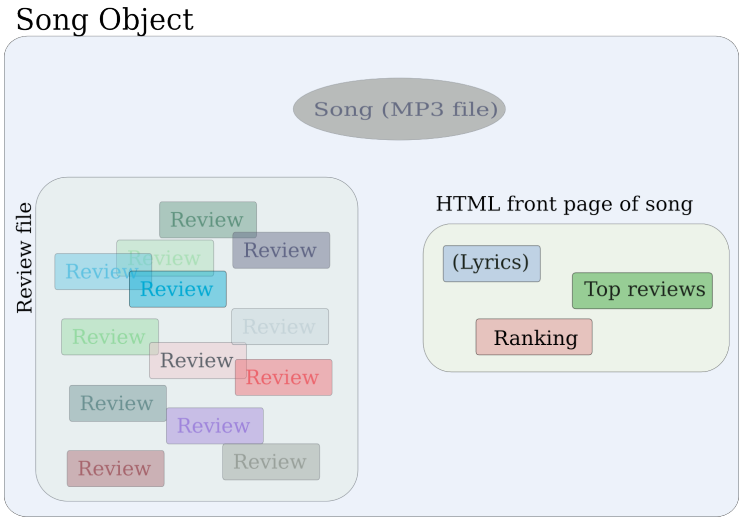


Figure 6: A song object. Not all song objects have lyrics. On the average each song title have been reviewed by 56 different persons.

3.1 Text

3.1.1 Vector Space model

The vector space model, see e.g. [2], have been used to model the reviewer information. Each document consists of several reviews by different people, however, they are all reviewing the same song title. Hence, in the vector space model we let each document refer to the collection of reviews. A term-document matrix, consisting of the raw word counts, was generated from the document files. Words which did not occur in 10 or more documents were pruned away. Furthermore, words which did not appear at least twice in a document were pruned away. This removes spelling errors from the corpus. Finally stemming was applied, thus, resulting in a term-document matrix of dimensions 30566×16706. The "libbow" software [5] was used. The sparse matrix together with relevant genre information, vocabulary etc. were saved in the Matlab 7.0 format ("tfmatrix.mat").

3.1.2 Timeinformation

Each review is tagged with geographical information as well as date information on when the review was provided. This information have not been extracted, but can easily be extracted from the reviews.

3.2 User co-writer matrix

Each review consists of approximately 56 songs on average. A matrix was generated recording whenever a user has created a song-review. The matrix (users-files) is of dimension $\approx 114000 \times 16706$. However, only around 25000 have commented 10 songs or more, and approximately 11000 users have commented 20 songs or more. A log-log plot of the amount of songs along the x-axis and the amount of users on the y-axis is shown in figure 7. If a user have reviewed the same song more than once, the count have just been incremented.

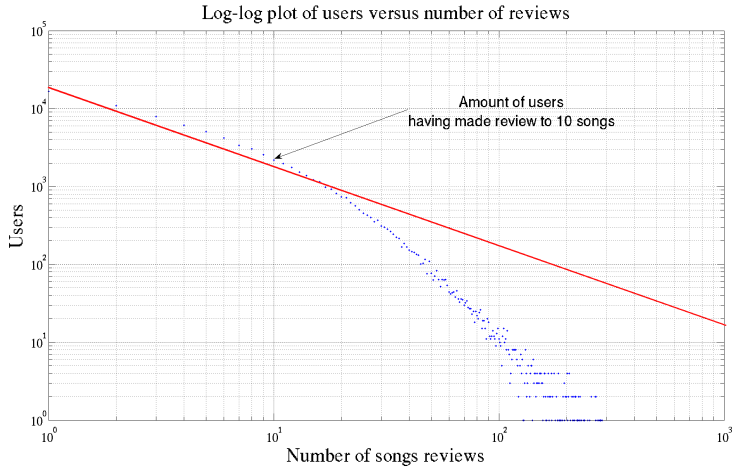


Figure 7: Log-log plot of the number of users having reviewed x number of songs.

3.3 Audio

Two sets of features have been extracted from the audio files, one set of "short-time" features; the MFCC's (Mel frequency cepstral coefficients) and one set of features at a longer time-scale, at the music snippet time-scale of 30s., the so called MAR (Multivariate Autoregressive Model) features. Both feature sets will be explained in little more detail in the following sub-sections. Figure 8 shows a block-diagram over the feature extraction process. Before extraction of the features we ensure that all the music samples have the same sample-frequency (to ensure an unified MFCC extraction). This is done by downsampling music snippets which had a samplefrequency of 44100. We used the "mpg123" decoder for both extraction and downsampling of the music snippets, see `http://www.mpg123.org` for further information.

The features extracted are saved in Matlab 7.0 *format. There is a single file available for each song title.*

3.3.1 Mel Frequency Cepstral Coefficients (MFCC) features

The Mel Frequency Cepstral Coefficients, see e.g. [4], are frequency based "short-time" features. The MFCC's are in principle a compact representation of the general frequency characteristics important for human hearing. The coefficients are ranked in such a way that the lower coefficients contain information about the small variations of

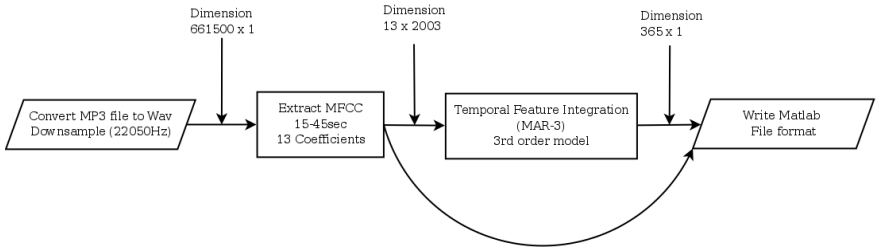


Figure 8: Audio feature extraction scheme applied to extract information at different time-scales. The multivariate 3rd order AR model create a single feature vector consisting of mean-value, covariance matrix(upper) and 3 matrices explaining the correlation structure at lag 1,2,3 in the MFCC's.

the spectral envelope. Hence, adding a single coefficient will increase the detail level of the envelope. For more detailed information about the MFCC's, see [6, 1].

We extracted the initial 13 MFCC's from the song-snippets ($15-45$ seconds of the MP3[2]) using the implementation of [3] with a hopsize and framesize corresponding to 15 and 30msec., respectively. The MFCC's have been applied in various audio applications during the last couple of years. For a more thorough explanation see e.g.

3.3.2 Multivariate AR (MAR) features

Temporal feature integration, see [7] for a more thorough description, was applied to the MFCC's. A multivariate autoregressive model of order 3 was applied in this stage. A model order of 3 was selected from previous experience within music genre classification, see e.g. [7]. Furthermore, the dimonsionality of $\mathbf{x}$ was 10, thus the initial 10 MFFC's were modelled with the multivariate AR model.

The multivariate AR model can be written as

$$\mathbf{x}_n = \sum_{p=1}^{3} \mathbf{A}_p \mathbf{x}_{n-p} + \mathbf{u}_n \tag{2}$$

where the noise term $\mathbf{u}_n$ is assumed i.i.d. with mean value $\mathbf{v}$ and finite covariance matrix $\mathbf{C}$. Note that the mean of the noise process $\mathbf{v}$ is related to the mean $\mathbf{m}$ of the time-series (series of MFCC's) by

$$\mathbf{m} = \left(\mathbf{I} - \sum_{p=1}^{P} \mathbf{A}_p \right)^{-1} \mathbf{v}. \tag{3}$$

The matrices $\mathbf{A}_p$ for $p = 1, 2, 3$ are the coefficient matrices of the 3rd order multivariate autoregressive model. The encode how much of the previous short-time features $\mathbf{x}_{n-1}, \mathbf{x}_{n-2}, \mathbf{x}_{n-3}$ that can be used to predict the short time feature $\mathbf{x}_n$.

4 Contact information

The dataset can be requested by contacting jl@imm.dtu.dk. You are welcome to use the dataset for research purposes, however, please acknowledge its use with a citation:

[2]The initial 15 seconds was skipped, since many music files have a quiet intro.

Anders Meng, "General Purpose Multimedia Dataset - GarageBand 2008",
`http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=5641`,
Technical University of Denmark, 2008.

References

[1] J.-J. Aucouturier, F. Pachet, and M. Sandler. The way it sounds : Timbre models for analysis and retrieval of polyphonic music signals. *IEEE Transactions on Multimedia*, 7(6):8, December 2005.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[3] M. Brookes. VOICEBOX (a MATLAB toolbox for speech processing), 1997.

[4] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-28(4):357–366, August 1980.

[5] Andrew Kachites McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow, 1996.

[6] A. Meng. *Temporal Feature Integration for Music Organisation*. PhD thesis, Technical University of Denmark, IMM, 2006.

[7] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen. Temporal Feature Integration for Music Genre Classification. *IEEE Transactions on Audio, Speech, and Language Processing.*, 15(5):1654–1664, July 2007.