Multimodal Music Mood Classification Using Audio and Lyrics

Cyril Laurier Jens Grivolla* Perfecto Herrera

Music Technology Group
Universitat Pompeu Fabra

Fundaci Barcelona Media
Av. Diagonal 177, 08018 Barcelona

Music Technology Group
Universitat Pompeu Fabra

Md . Nahiyan Uddin Md. Mirajul Islam

May 14, 2017

Outline

Outline

Introduction

What is multimodal music classification?

We listen to a lot of varieties of musics in our daily life. Our music preferences change according to our moods.
The term "Multimodal music classification" refers to classify musics we hear according to different modes of human psychology.

Outline

Related Works

- There is some existing work dealing with audio mood classification.

Related Works

- There is some existing work dealing with audio mood classification.
- Also some recent literature about mood detection in text

Related Works

- Neumayer and Rauber have shown the complementarity of audio and lyrics in the context of genre classification

Related Works

- Neumayer and Rauber have shown the complementarity of audio and lyrics in the context of genre classification
- Logan et al. have investigated the properties of lyrics using Latent Semantic Analysis.

Related Works

- Neumayer and Rauber have shown the complementarity of audio and lyrics in the context of genre classification
- Logan et al. have investigated the properties of lyrics using Latent Semantic Analysis.
- Natural genre clusters were discovered .Their conclusion was also that lyrics are useful for artist similarity searches.

Related Works

- Neumayer and Rauber have shown the complementarity of audio and lyrics in the context of genre classification
- Logan et al. have investigated the properties of lyrics using Latent Semantic Analysis.
- Natural genre clusters were discovered .Their conclusion was also that lyrics are useful for artist similarity searches.
- Studies in cognitive neuropsychology also demonstrated the independence of both sources of information and so the potential complementarity of both melody and lyrics in the case of emotional expression.

Related Works

- But very little has been done so far to address the automatic classification of lyrics according to their mood.
- We have found no prior articles studying the combination of lyrics and acoustic information for this particular classification purpose.

Outline

Database

According to the moods the songs are classified into four catagories.

- Happy

Database

According to the moods the songs are classified into four catagories.

- Happy
- Relaxed

Database

According to the moods the songs are classified into four catagories.

- Happy
- Relaxed
- Sad

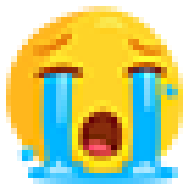


Database

According to the moods the songs are classified into four catagories.

- Happy
- Relaxed
- Sad
- Angry



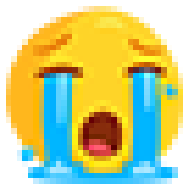



Database

According to the moods the songs are classified into four catagories.

- Happy
- Relaxed
- Sad
- Angry






Database

For choosing songs , the folloowing criterias are maintained in order to get a more accurate result

- Songs were collected from last.fm with their maintained tags

Database

For choosing songs , the folloowing criterias are maintained in order to get a more accurate result

- Songs were collected from last.fm with their maintained tags
- Only songs having English lyrics and an entry in LyricWiki were selected

Database

For choosing songs , the folloowing criterias are maintained in order to get a more accurate result

- Songs were collected from last.fm with their maintained tags
- Only songs having English lyrics and an entry in LyricWiki were selected
- The tags of the songs were once more evaluated by listeners to ensure more accuracy

Database

For choosing songs , the folloowing criterias are maintained in order to get a more accurate result

- Songs were collected from last.fm with their maintained tags
- Only songs having English lyrics and an entry in LyricWiki were selected
- The tags of the songs were once more evaluated by listeners to ensure more accuracy
- In total, 17 different evaluators participated and an average of 71.3% of the songs originally selected from last.fm were validated

Database

For choosing songs , the folloowing criterias are maintained in order to get a more accurate result

- Songs were collected from last.fm with their maintained tags
- Only songs having English lyrics and an entry in LyricWiki were selected
- The tags of the songs were once more evaluated by listeners to ensure more accuracy
- In total, 17 different evaluators participated and an average of 71.3% of the songs originally selected from last.fm were validated
- The database is composed of 1000 songs divided between 4 categories. An equal distribution of these binary classes was used.

Outline

Outline

Audio Classification

- To classify music by mood a state-of-the-art audio classification algorithm was used in a supervised learning approach.
- The features and the classifier were selected according to current litterature
- The results from the Audio Mood Classification evaluation task held by the Music Information Retrieval Evaluation eXchange (MIREX)

Audio Classification

- In order to classify the music from acoustical information, we first extracted audio features of different kinds:
  - Timbral (for instance MFCC, spectral centroid)
  - Rhythmic(for example tempo, onset rate)
  - Tonal (like Harmonic Pitch Class Profiles)
  - Temporal descriptors

Audio Classification

- Classifiers were used like Support Vector Machines(SVM) , Random Forest or Logistic Regres

	SVM	Logistic	RandForest
Angry	**98.1% (3.8)**	95.9%(5.0)	95.4%(4.7)
Happy	**81.5% (11.5)**	74.8%(11.3)	77.7%(12.0)
Sad	**87.7% (11.0)**	85.9%(10.8)	86.2%(10.5)
Relaxed	**91.4% (7.3)**	80.9%(7.0)	91.2%(6.7)
Mean	**89.8% (8.4)**	84.4%(8.5)	87.6%(8.5)

Table 1. Classification accuracy using audio features, for each category against its complementary (with standard deviation)

Among them SVM gives us the best result.

Outline

Outline

Experiment 1

Classification based on similarity using Lucene

- The representation of the songs is reduced to a bag of words i.e. the set of words or terms used in a song as well as their frequency
- This is then used, with the help of the Lucene document retrieval system [4] , to rank documents by their similarity
- More importance was attributed to those terms that are frequent in the given song, but less frequent overall in the collection

Experiment 1

Result

We conducted experiments with varying numbers of similar documents (k) to be taken into account.

- A low k provides less stability, as the predicted label depends strongly on individual examples from the collection
- Large ks on the other hand can mean that examples are taken into account that are not actually very similar (and thus representative) of the one that is to be classified

The optimum depends on the application and the distribution of the datapoints and can not be easily predicted a-priori

Experiment 1

Result

As we are seeing , the prediction power of the similarity-based approach for lyrics remains limited, with averaged accuracy around 60%. The most predictable catagory is "angry" and the least predictable is "sad".

	k=3(%)	k=5(%)	k=7(%)	k=9(%)	k=11(%)
Angry	69.5	67.5	69.0	68.5	67.0
Happy	55.9	57.4	60.9	64.5	64.1
Sad	55.0	52.8	58.9	54.5	55.0
Relaxed	61.8	65.8	61.0	59.8	59.1
Mean	60.5	60.9	62.5	61.8	61.3

Table: Classification accuracies using K - NN with a if.idf based distance on lyrics for different values of k

Experiment 1

Limitations

But there are some limitations :

- It is difficult to directly integrate the results
- While the audio side, the featured vectors can be used with different clasification algorithms , this is not as easy case for the lyrics
- Vocabulary size already reached 7000 words while more complete collections reach vocabulary sizes of several hundred thousand distinct words

Outline

Experiment 2

Classification using Latent Semantic Analysis (LSA)

One approach to deal with the dimensionality problem is to

- project the lyrics into a lower-dimensional space that is manageable by generic classifiers
- project the data into a space of a given dimensionality using Latent Semantic Analysis (LSA)
- maintain a good approximation of the distances between data points.
  Experiments were conducted to determine the impact of the number of dimensions used in the LCA . As could be expected, performance (using lyrics alone) is very low for extremely low dimensionality and tends to improve with a greater number of dimensions.

Experiment 2

Result

The use of LSA does not dramatically improve performance compared to our first experiment, depending on the category it can even be worse. The reduction in dimensionality does, however, provide more flexibility, as different types of classifiers can be used on the resulting representation. The results shown here use a reduction to 30 dimensions.

	SVM(%)	Logistic(%)	RandForest(%)
Angry	62.1(9.1)	62.0(10.2)	61.3(11.5)
Happy	55.2(10.3)	54.1(12.5)	54.8(10.7)
Sad	66.4(9.7)	65.3(11.0)	56.7(12.1)
Relaxed	57.5(8.2)	57.3(9.1)	56.8(9.79)
Mean	61.3(9.3)	59.7(10.7)	57.4(11.0)

Table: Classification accuracies using LSA

Experiment 2

Limitations

But there are some limitations :

- If our mood categories do not relate to clusters of songs that would be considered similar according to the metrics used in document retrieval

Experiment 2

Limitations

But there are some limitations :

- If our mood categories do not relate to clusters of songs that would be considered similar according to the metrics used in document retrieval
- This severally limits the potential of any approaches that are based on document distances with tf.idf weighting

Experiment 2

Limitations

But there are some limitations :

- If our mood categories do not relate to clusters of songs that would be considered similar according to the metrics used in document retrieval
- This severally limits the potential of any approaches that are based on document distances with tf.idf weighting
- LSA doesn't overcome this problem as the distances between the data points in the projected space directly reflect their tf.idf based distance used as a basic for the transformation

Outline

Experiment 3

Classification using Language Model Differences(LMD)

- Distances between songs based on lyrics cannot separate our mood categories very well
- But lyrics convey other types of information to be exploited in persuing their separation according to mood.
- In order to assess their potential the language models corresponding to the different categories were analyzed

Experiment 3

Classification using Language Model Differences(LMD)

Figure 1 shows document frequencies for the 200 most frequent terms in the angry category compared to the frequencies to the not angry category

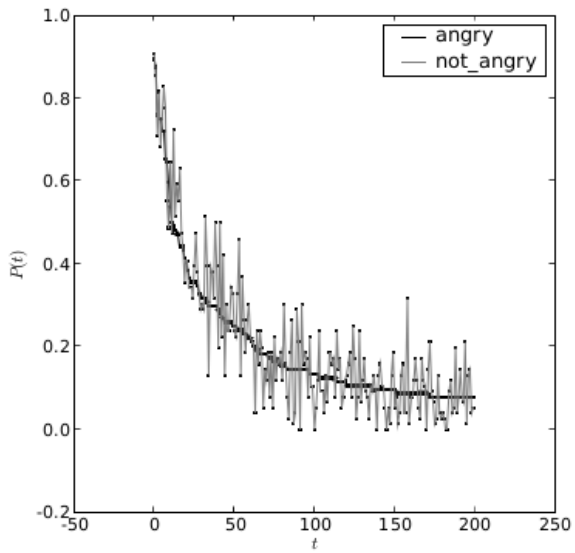


Figure 1. Document frequencies ($P(t)$) of terms in "angry" and "not angry" category where t is the term id.

Experiment 3

Classification using Language Model Differences(LMD)

- When comparing two language models, the simplest approach is to calculate the difference in document frequency for all terms.
- This can be computed either as an abosolute difference in document frequency for all terms. Example of terms ranked by absloute difference:
  - *angry:* world,die,death,control,...
  - *not angry:* me,love,i'm,can,could,so,but,...

  Examples of terms ranked by relative difference:
  - *angry:* realms,bear,four,...
  - *not angry:* chillin,nursery,hanging,...

Experiment 3

Classification using Language Model Differences(LMD)

Here are some necessary equaions to evaluate the absolute difference and relative

$$\Delta_{abs}(t) = abs(P(t|LM_1) - P(t|LM_2))$$

$$\Delta_{rel}(t) = \frac{abs(P(t|LM_1) - P(t|LM_2))}{max(P(t|LM_1), P(t|LM_2))}$$

$$\Delta_{mixed}(t) = \frac{abs(P(t|LM_1) - P(t|LM_2))}{\sqrt{(max(P(t|LM_1), P(t|LM_2)))}}$$

Experiment 3

Result

For each category, we selected the n terms with the highest mixed . We obtained a vector representation with n dimensions that can be used with different classifiers. We made 10 runs of 10-fold cross-validation (this includes the term selection, of course) and tried different values n.

	SVM(%)	Logistic(%)	RandForest(%)
Angry	77.9(10.3)	60.6(12.0)	71.0(11.5)
Happy	80.8(12.1)	67.5(13.3)	70.8(11.4)
Sad	84.4(11.2)	83.9(7.0)	75.1(12.9)
Relaxed	79.7(9.5)	71.3(10.5)	78.0(9.5)
Mean	80.7(10.8)	70.8(10.7)	73.7(11.3)

Table: Classification performances using the 100 most discriminant terms

Outline

Combining Audio and Lyric Information

We used two approaches to integrate these two information sources.

- The first one used separate predictions for audio and lyrics and combined them through voting.
- The second approach was to combine all features in the same space, having a vector composed of both audio and lyrics features. This allowed to use audio and lyrics information within one classifier
  Only the second one is reported here.

Outline

Combining Audio and Lyric Information

Having audio and lyrics information in the same vector allows to exploit interdependencies between aspects from both modalities

- the combination of the language model differences with the audio descriptors yielded to rel- atively good results.
- This combination gives significant improvements over both individual approaches, leveraging the complementary information available from audio and lyrics, at least for two of the four categories: happy and sad with both a sig- nificant (p less than 0.05 using a Paired T-Test) overall increase around 5% for both
- For the angry and relaxed categories there is also a slight increase in classification performance

Combining Audio and Lyric Information

- However, the extremely high baseline of over 98% accuracy on audio alone for the angry category, as well as the large difference in performance between lyrics and audio for relax limits the benefits of using a hybrid method
  Only the second one is reported here.

Experiment 3

Result

For each category we show the accuracy of the SVM classifier for the audio analysis, for the lyrics analysis, and for the multimodal approach combining both. As in the previous experiments, the accuracies shown Table 38 are averages over the 10 runs of 10-fold cross-validation.

	Audio(%)	Lyrics(%)	Mixed(%)
Angry	98.1(3.8)	77.9(10.3	98.39(3.7))
Happy	81.5(11.5)	80.8(11.2)	86.8(10.6)
Sad	87.7(11.0)	84.4(11.2)	92.8(8.7)
Relaxed	91.4(7.30)	79.7(9.5)	91.7(7.1)

Table: Classification accuracies using audio features, lyrics with language model differences and finally a mixed feature space both.

Outline

Summary

- This multimodal approach increases the performances for all the mood categories.
- There are more work to be done with lyric modal approach to classify the mood of the songs.