RAJIV GANDHI INSTITUTE OF TECHNOLOGY

GOVERNMENT ENGINEERING COLLEGE

KOTTAYAM - 686 501

DEPARTMENT OF
COMPUTER SCIENCE & ENGINEERING

CS 451-SEMINAR REPORT

DECEMBER 2022

Novel Audio Features For Music Emotion Recognition

Submitted by

V ADARSH (REG NO: KTE18CS057)

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY
THIRUVANANTHAPURAM

RAJIV GANDHI INSTITUTE OF TECHNOLOGY

GOVERNMENT ENGINEERING COLLEGE

KOTTAYAM – 686 051

DEPARTMENT OF
COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that this report entitled "Novel audio features for music emotion recognition" is an authentic report of the seminar done by V ADARSH (REG NO: KTE18CS057) during the academic year 2022-23, in partial fulfillment of the requirement for the award of B.Tech Degree in Computer Science and Engineering of Kerala Technological University, Thiruvananthapuram.

GUIDE

SEMINAR COORDINATOR

HEAD OF THE DEPARTMENT

ACKNOWLEDGEMENT

Behind every achievement lies and unfathomable sea of gratitude to the almighty, without whom it would ever have coming into existence I can barely find words to express all the wisdom, love and support from almighty.

I like to express my utmost gratitude to **Dr. C. Sathish Kumar,** Principal, Govt. Rajiv Gandhi Institute of Technology, Kottayam.

I am fortunate to be blessed with the guidance and encouragement of **Dr. Madhu K P**, Head of Department, Computer Science and Engineering, R.I.T Kottayam.

It is my pleasure to acknowledge the help, which I had received from different individuals. My fist sincere appreciation and gratitude goes to **Prof. Jeena Joy** for her guidance, valuable suggestion and inspirations during the entire period.

To derive benefits of their innumerous experience, it is a matter of great privilege for me. I also take this opportunity to express my sincere thanks to all the staff in the computer science department, who extended their wholehearted cooperation, moral support and rendering ungrudging assistance.

V ADARSH

ABSTRACT

The advanced music emotion recognition is a state-of-the-art proposing novel emotionally-relevant audio features. Reviewed all the existing audio features implemented in well-known frameworks and their relationships with the eight commonly defined musical concepts. This knowledge helped uncover musical concepts lacking computational extractors, namely related with musical texture and expressive techniques. To evaluate , a public dataset of 900 audio clips, with subjective annotations following Russell's emotion quadrants were used. The existent audio features (baseline) and the proposed features (novel) were tested using 20 repetitions of 10-fold cross-validation. Adding the proposed features improved the F1-score to 76.4% (by 9%), and results uncovered interesting relations, namely the weight of specific features and musical concepts to each emotion quadrant, and warrant promising new directions for future research in the field of music emotion recognition, interactive media, and novel music interfaces.

CONTENTS

Chapter 1

INTRODUCTION

In recent years, Music Emotion Recognition (MER) has attracted increasing attention from the Music Information Retrieval (MIR) research community.

However, several limitations still persist, namely, the lack of a consensual and public dataset and the need to further exploit emotionally-relevant acoustic features. Particularly, the belief that features specifically suited to emotion detection are needed to narrow the so-called semantic gap and their absence hinders the progress of research on MER. Moreover, existing system implementation shows that the state-of-the-art solutions are still unable to accurately solve simple problems, such as classification with few emotion classes

Several factors contribute to this glass ceiling of MER systems. To begin with, the perception of emotion is inherently subjective: different people may perceive different, even opposite, emotions when listening to the same song. Even when there is an agreement between listeners, there is often ambiguity in the terms used regarding emotion description and classification . It is not well-understood how and why some musical elements elicit specific emotional responses in listeners.

Second, creating robust algorithms to accurately capture these music-emotion relations is a complex problem, involving, among others, tasks such as tempo and melody estimation, which still have much room for improvement.

Third, as opposed to other information retrieval problems, there are no public, widely accepted and adequately validated, benchmarks to compare works. Typically, researchers use private datasets or provide only audio features. Even though the MIREX AMC task has contributed with one dataset to alleviate this problem, several major issues have been identified in the literature. Namely, the defined taxonomy lacks support from music psychology and some of the clusters show semantic and acoustic overlap

Finally, and most importantly, many of the audio features applied in MER were created for other audio recognition applications and often lack emotional relevance. Hence, the main working hypothesis is that, to further advance the audio MER field, research needs to focus on what people believe is its main, crucial, and current problem: to capture the emotional content conveyed in music through better designed audio features.

This classification experiments showed an improvement of 9% in F1-Score when using the top 100 baseline and novel features, while compared to the top 100 baseline features only. Moreover, even when the top 800 baseline features is employed, the result is 4.7% below the one obtained with the top100 baseline and novel features set.

Chapter 2
LITERATURE SURVEY

The main goal of this study is to analyse the relations between musical dimensions and emotions. To further clarify the focus of this survey, it is important to mention that approaches based on deep learning techniques

Regarding symbolic features, since some current approaches establish a bridge between the audio and the symbolic MER domains by integrating an audio transcription stage into the feature extraction stage e.g., possible research directions on the exploitation of symbolic features on MER.

The relations between music and emotion, which are detailed .There, describes the specific associations between each of the eight musical dimensions and different emotions ,to the existing emotionally relevant computational audio features, organizing them by musical dimension.

Emotion in music can be studied as: perceived, as in the emotion an individual identifies when listening; felt, regarding the emotional response a user feels when listening, which can be different from the perceived one or transmitted, representing the emotion that the performer or composer aimed to convey.

Regarding the relations between emotions and specific musical attributes, several studies uncovered interesting associations. As an example the major modes are frequently related to emotional states such as happiness or solemnity, whereas minor modes are often associated with sadness or anger and those related with simple, consonant, harmonies are usually happy, pleasant or relaxed. On the contrary, complex, dissonant, harmonies relate to emotions such as excitement, tension or sadness, as they create instability in a musical motion . Moreover, researchers identified many musical features related to emotion, namely: timing, dynamics, articulation, timbre, pitch, interval, melody, harmony, tonality, rhythm, mode, loudness, vibrato and some other musical form .

Despite the identification of these relations, many of them are not fully understood, still requiring further musicological and psychological studies, while others are difficult to extract from audio signals. Nevertheless, several computational audio features have been proposed over the years. While the number of existent audio features is high, many were developed to solve other problems (e.g., Mel Frequency Cepstral Coefficients (MFCCs) for speech recognition) and may not be directly relevant to MER .

Nowadays, most proposed audio features are implemented and available in audio frameworks. Some of the current state-of-the-art (hereafter termed standard) audio features, available in widely adopted frameworks, namely, the MIR Toolbox, Marsyas and Psysound.

The conclusions derived from these current methods are related with tone color (63.7%). Also, many of these features are abstract and very low level, capturing statistics about the waveform signal or the spectrum. These are not directly related with the higher level musical concepts. As an example, MFCCs belong to tone color but do not give explicit information about the source or material of the sound. Nonetheless , they can implicitly help to distinguish between these features. This is an example of the mentioned semantic gap, where high level concepts are not being captured explicitly with the existent low level features. Also a few features are mainly related with expressive techniques, musical texture (which has none) or musical form. Thus, there is a need for audio features estimating higher-level concepts, e.g., expressive techniques and ornamentations like vibratos, tremolos or staccatos (articulation), texture information such as the number of musical lines or repetition and complexity in musical form.

Chapter 3
Architecture and Method

The proposed novel audio features describes the emotional classification experiments carried out. Since the available datasets have certain limitations, the process starts by creating a new set of datasets that has a feature to keep adding new tags to it, as progress is being obtained.

The new dataset is created in such a way that a simple and universal taxonomy is used and perform semi-automatic construction to limit the resource requirement and to scale the dataset.

The new dataset created requires a proper validation before it can be used to process data. Here a manual blind inspection of the candidate set was conducted to annotate the given audio clips into the Russell's quadrants.

This process also includes the process of marking the unclear clips and bad(noise) audio files to avoid confusion. To construct the final dataset, the audios marked unclear or bad where and those clips that are not matching the Russell's quadrants were excluded. The main classification in the Russell's quadrant is based on the positive or negative audio (arousal or valence).

Next is the process of adding the audio frameworks like MIR Toolbox, Marsyas and Psysound which contains a large number of computational audio features. Then a specific set of audio features are extracted from these three frameworks and a feature reduction stage is carried out to discard the redundant features obtained by similar algorithms across selected audio frameworks. These standard audio features serves to build a baseline models for the new approaches to reference.

To approach the problem of music emotion recognition with more detailing, the novel audio features are used. It includes the explicit determination of musical notes, frequency and intensity contours. This includes the following preliminary steps:

3.1 From the audio signal to MIDI notes

The estimation of predominant melody lines even though imperfect is one of the important steps. This phase starts by estimating the predominant fundamental frequencies and saliences which is the process of identifying which frequencies are present in the signal at each point of time. Next harmonic summation is used to estimate the pitches in these instants and how salient they are, and the series of consecutive pitches which are continuous in the frequency are used to form the pitch contours.

3.2 Melodic features

The melodic features are the key concepts of music and they are defined as the horizontal succession of pitches. They contain a set of metrics obtained features from the notes of melodic trajectory. The melodic features are mail grouped into (1) MIDI Note Number statistics, (2) Note Space Length and Chroma NSL, (3) Register Distribution, (4) Ratio of Pitch Transitions and (5) Note Smoothness statistics.

3.3 Dynamic features

The pitch salience of each note compares with neighbour notes in the scores to get the information about their individual intensities. These notes are classified as high(strong), medium and low(smooth) intensity based on the mean and standard deviation of all notes.

3.4 Rhythmic features

These can be defined as changing sequences of notes with specific time interval, and to calculate the dynamics of these time intervals and their changes, three possible categories are considered, short, medium and long notes. These ranges are also calculated according to the mean and standard deviation of all notes.

With the high number of features, the ReliefF feature selection algorithms were used to select the better suited ones for each classification problem. The output of the ReliefF algorithm is a weight between -1 and 1for each attribute, with more positive weights indicating more predictive attributes. Then the weights are averaged and a ranking of top N features are deduced for classification testing. The best performing N determines, how many features are required to obtain the best results.

Chapter 4

EXPERIMENTAL EVALUATIONS

First , the standard features were ranked with the ReliefF , were used to obtain a baseline results ,following the novel audio features were combined with the baseline and also tested. As a summary of the attained classification result. The baseline features attained 67.5% F1Score (macro weighted) with SVM and 70 standard features. The same solution achieved a maximum of 71.7% with a very high number of features. Adding the novel features (i.e., standard + novel features) increased the maximum result of the classifier to 76.4% (0.04 standard deviation).

The best result (76.4%) was obtained with 29 novel and 71 baseline features, which demonstrates the relevance of adding novel features to MER.

This allows to understand which emotions are more difficult to classify and what is the influence of the standard and novel features in this process.

This seems to indicate that emotions with higher arousal are easier to differentiate with the selected features. For the same number of features , the experiment using novel features shows an improvement of 9% in F1-Score when compared to the one using only the baseline features. Regarding the misclassified songs, analyzing the confusion matrix shows that the classifier is slightly biased towards positive valence, predicting more frequently songs from quadrants 1 and 4 than from 2 and 3.

Chapter 5
CONCLUSION

This shows the influence of musical audio features in MER applications. The standard audio features available in known frameworks were studied and organized into eight musical categories. Additional experiments were carried out to uncover the importance of specific features and musical concepts to discriminate specific emotional quadrants. In the future, a further exploration of the relationship between the voice signal and lyrics by experimenting with multi-modal MER approaches. Studying the emotion variation detection model helps to build sets of interpretable rules providing a more readable characterization of how musical features influence emotions. To evaluate, a new dataset was built semi-automatically, containing 900 song entries and respective metadata (e.g., title, artist, genre and mood tags), annotated according to the Russell's emotion model quadrants. Classification results show that the addition of the novel features improves the results from 67.4% to 76.4% when using a similar number of features (100), or from 71.7% when 800 baseline features were used.

Chapter 6

REFERENCES

[1] R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. P. Paiva, (**2013**)"*Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis," in 10th International Symposium on Computer Music Multidisciplinary Research –* CMMR'.

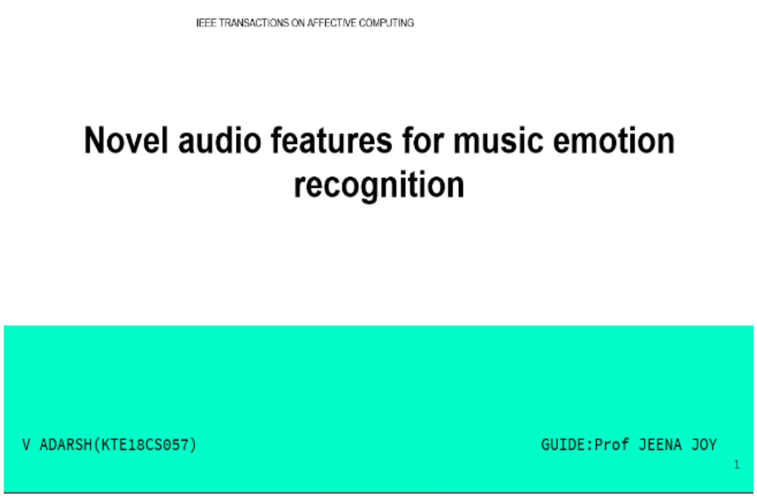
[2] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P.Richardson, J. Scott, J. A. Speck, and D. Turnbull\ ,(**2010**) "*Music Emotion Recognition: A State of the Art Review," in Proc. of the11th Int. Society for Music Information Retrieval Conf*. (ISMIR **2010**).

[3] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, Feb. (**2008**)"*A Regression Approach to Music Emotion Recognition," IEEE Trans. Audio.Speech. Lang. Processing*, vol. 16, no. 2, pp. 448–457,.

[4] A. B. Warriner, V. Kuperman, and M. Brysbaert, Dec. (**2013**)"*Norms of valence, arousal, and dominance for 13,915 English lemmas,"Behav. Res. Methods,* vol. 45, no. 4, pp. 1191–1207.

[5] M. M. Bradley and P. J. Lang, **(1999)**"*Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings, "Psychology, vol. Technical*.

APPENDIX A: PRESENTATION SLIDES



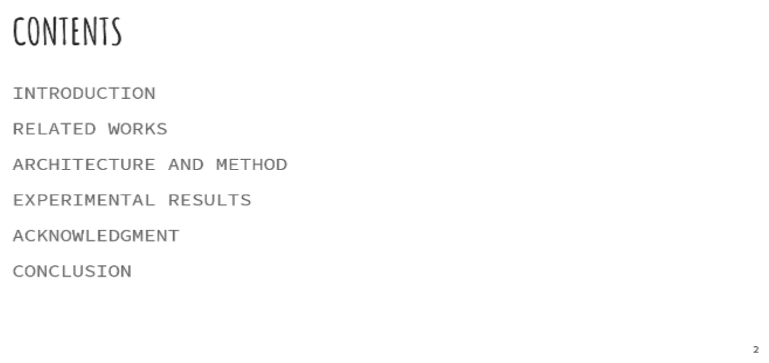
IEEE TRANSACTIONS ON AFFECTIVE COMPUTING

Novel audio features for music emotion recognition

V ADARSH(KTE18CS057)

GUIDE:Prof JEENA JOY

1



CONTENTS

INTRODUCTION

RELATED WORKS

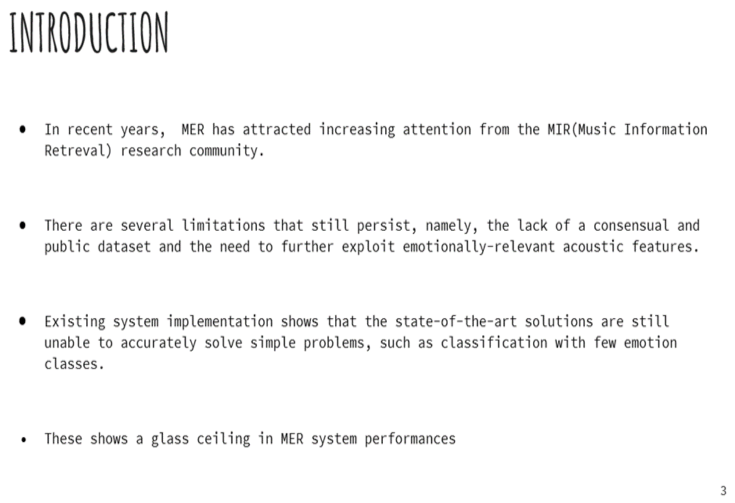
ARCHITECTURE AND METHOD

EXPERIMENTAL RESULTS

ACKNOWLEDGMENT

CONCLUSION

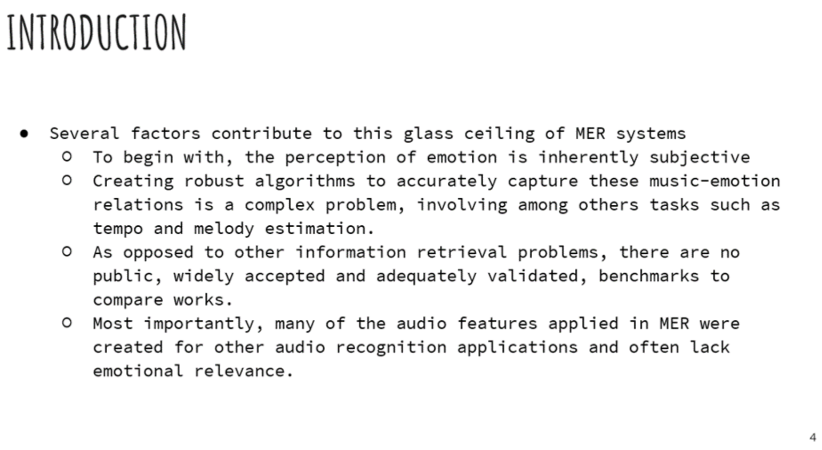
2



INTRODUCTION

- In recent years, MER has attracted increasing attention from the MIR(Music Information Retreval) research community.
- There are several limitations that still persist, namely, the lack of a consensual and public dataset and the need to further exploit emotionally-relevant acoustic features.
- Existing system implementation shows that the state-of-the-art solutions are still unable to accurately solve simple problems, such as classification with few emotion classes.
- These shows a glass ceiling in MER system performances

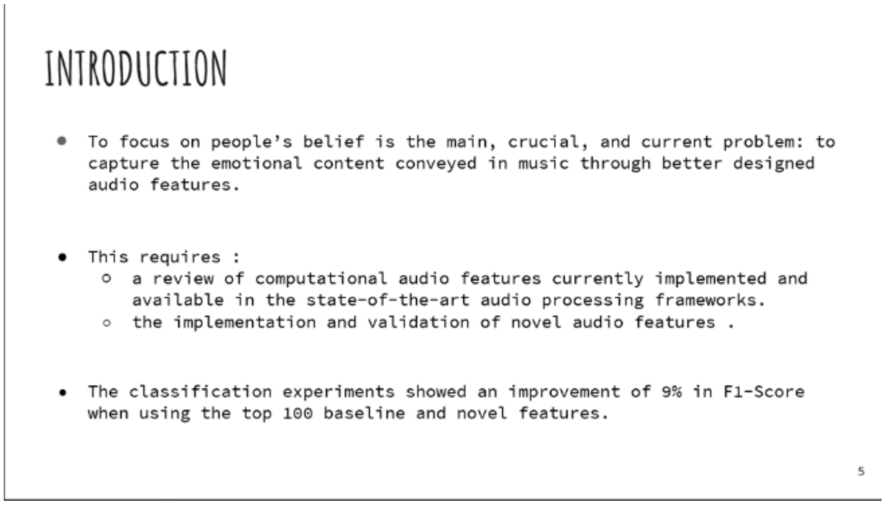
3



INTRODUCTION

- Several factors contribute to this glass ceiling of MER systems
  - To begin with, the perception of emotion is inherently subjective
  - Creating robust algorithms to accurately capture these music-emotion relations is a complex problem, involving among others tasks such as tempo and melody estimation.
  - As opposed to other information retrieval problems, there are no public, widely accepted and adequately validated, benchmarks to compare works.
  - Most importantly, many of the audio features applied in MER were created for other audio recognition applications and often lack emotional relevance.

4

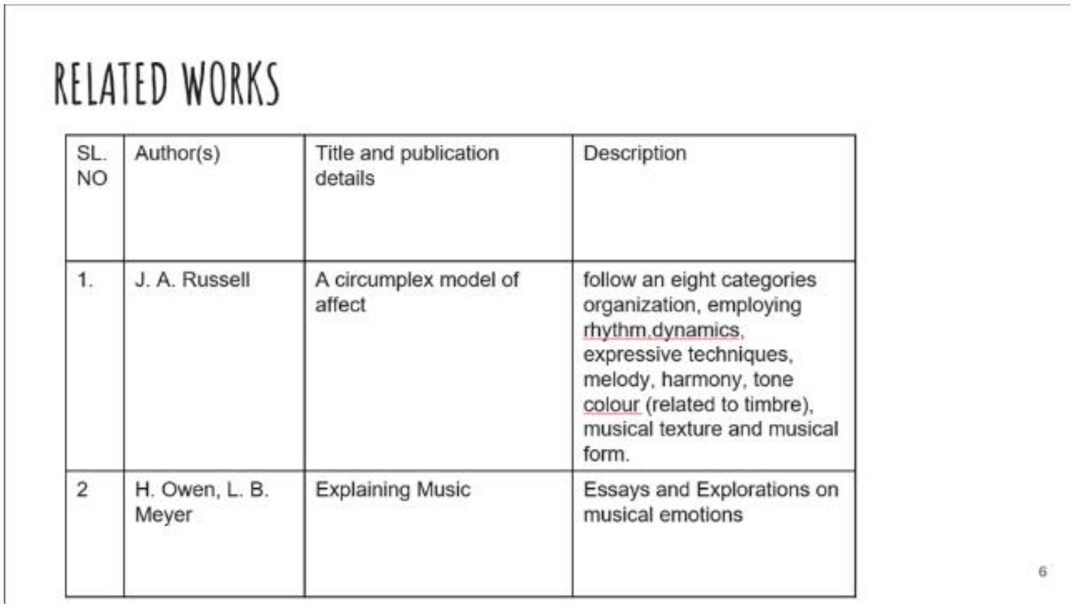


INTRODUCTION

- To focus on people's belief is the main, crucial, and current problem: to capture the emotional content conveyed in music through better designed audio features.

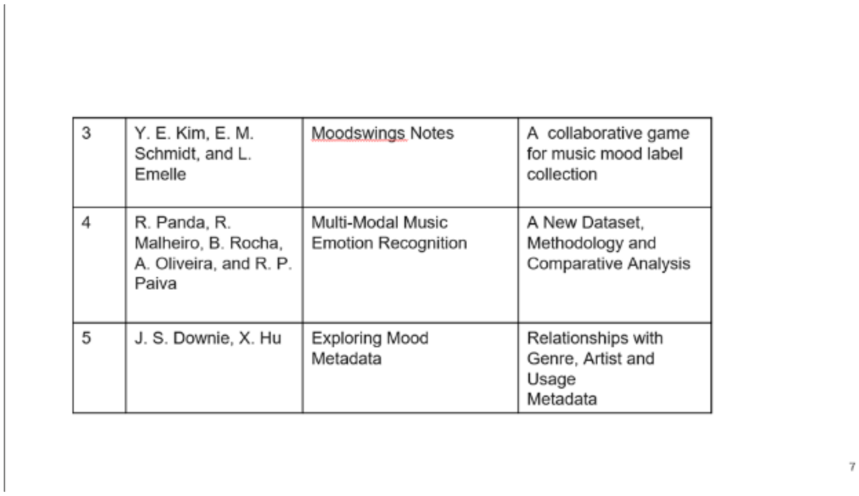
- This requires :
  - a review of computational audio features currently implemented and available in the state-of-the-art audio processing frameworks.
  - the implementation and validation of novel audio features .

- The classification experiments showed an improvement of 9% in F1-Score when using the top 100 baseline and novel features.

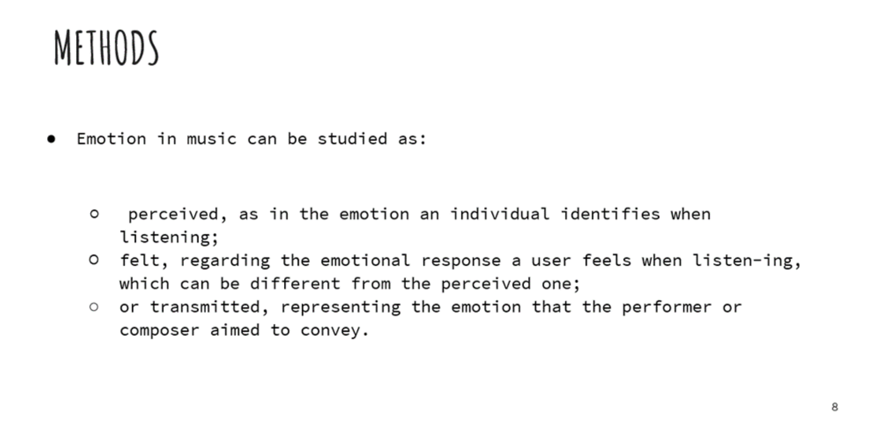


RELATED WORKS

SL. NO	Author(s)	Title and publication details	Description
1.	J. A. Russell	A circumplex model of affect	follow an eight categories organization, employing rhythm.dynamics, expressive techniques, melody, harmony, tone colour (related to timbre), musical texture and musical form.
2	H. Owen, L. B. Meyer	Explaining Music	Essays and Explorations on musical emotions

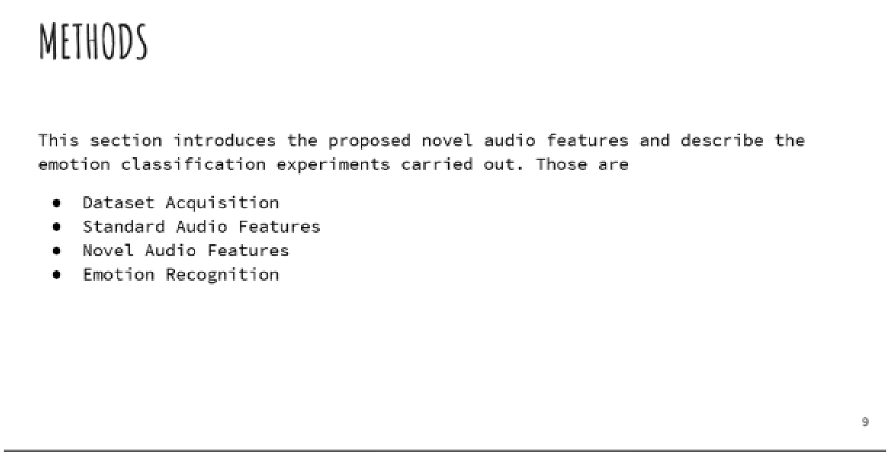


3	Y. E. Kim, E. M. Schmidt, and L. Emelle	Moodswings Notes	A collaborative game for music mood label collection
4	R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. P. Paiva	Multi-Modal Music Emotion Recognition	A New Dataset, Methodology and Comparative Analysis
5	J. S. Downie, X. Hu	Exploring Mood Metadata	Relationships with Genre, Artist and Usage Metadata



METHODS

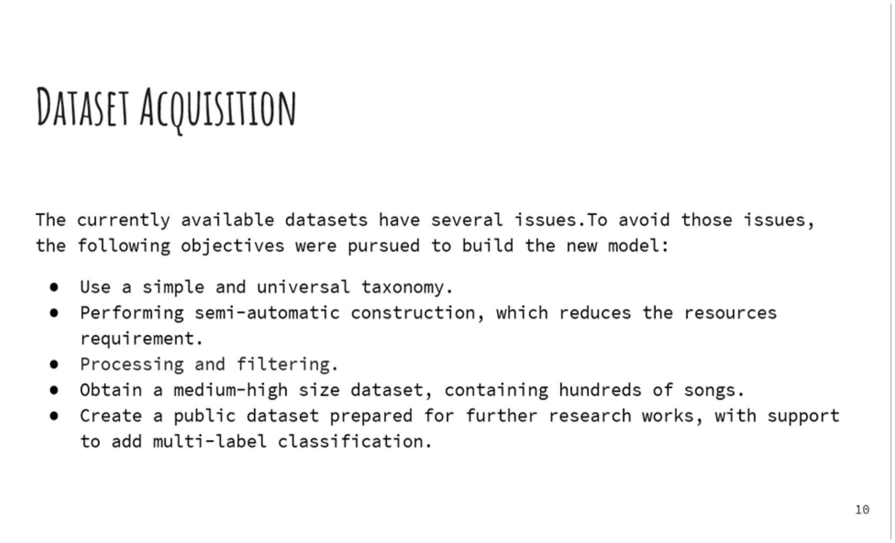
- Emotion in music can be studied as:
  - perceived, as in the emotion an individual identifies when listening;
  - felt, regarding the emotional response a user feels when listen-ing, which can be different from the perceived one;
  - or transmitted, representing the emotion that the performer or composer aimed to convey.



METHODS

This section introduces the proposed novel audio features and describe the emotion classification experiments carried out. Those are

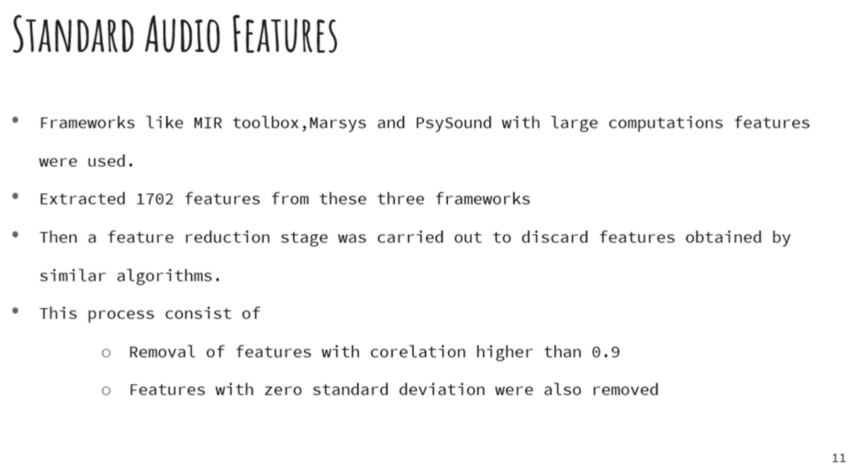
- Dataset Acquisition
- Standard Audio Features
- Novel Audio Features
- Emotion Recognition



Dataset Acquisition

The currently available datasets have several issues.To avoid those issues, the following objectives were pursued to build the new model:

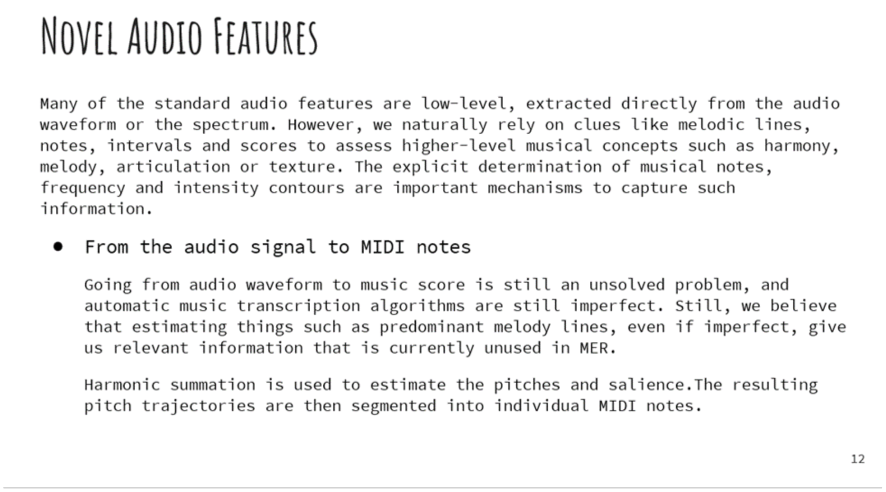
- Use a simple and universal taxonomy.
- Performing semi-automatic construction, which reduces the resources requirement.
- Processing and filtering.
- Obtain a medium-high size dataset, containing hundreds of songs.
- Create a public dataset prepared for further research works, with support to add multi-label classification.



Standard Audio Features

- Frameworks like MIR toolbox,Marsys and PsySound with large computations features were used.
- Extracted 1702 features from these three frameworks
- Then a feature reduction stage was carried out to discard features obtained by similar algorithms.
- This process consist of
    - Removal of features with corelation higher than 0.9
    - Features with zero standard deviation were also removed

11



Novel Audio Features

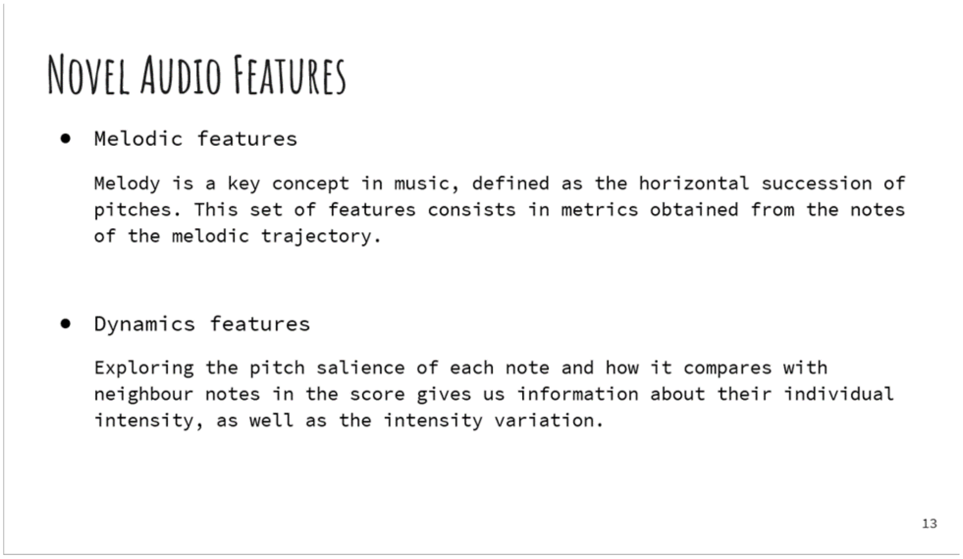
Many of the standard audio features are low-level, extracted directly from the audio waveform or the spectrum. However, we naturally rely on clues like melodic lines, notes, intervals and scores to assess higher-level musical concepts such as harmony, melody, articulation or texture. The explicit determination of musical notes, frequency and intensity contours are important mechanisms to capture such information.

- From the audio signal to MIDI notes

  Going from audio waveform to music score is still an unsolved problem, and automatic music transcription algorithms are still imperfect. Still, we believe that estimating things such as predominant melody lines, even if imperfect, give us relevant information that is currently unused in MER.

  Harmonic summation is used to estimate the pitches and salience.The resulting pitch trajectories are then segmented into individual MIDI notes.

12



Novel Audio Features

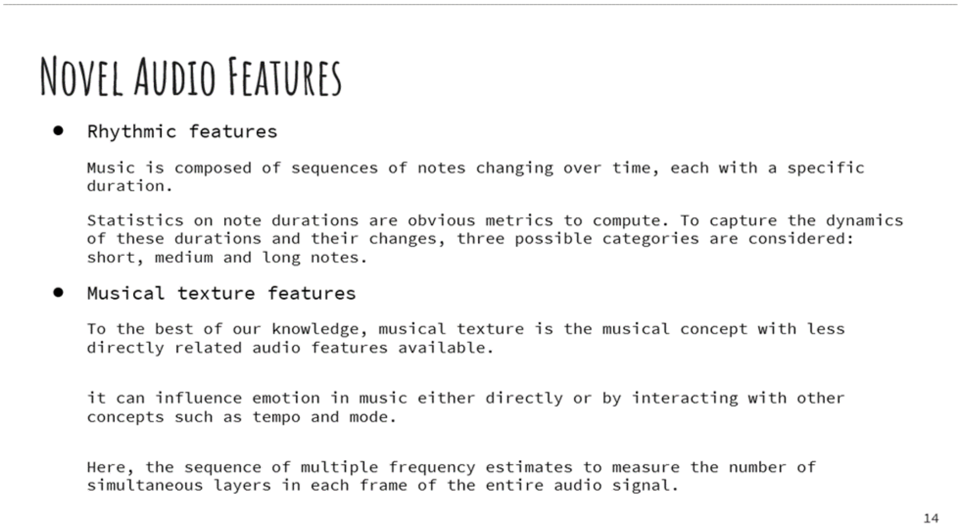
- Melodic features

  Melody is a key concept in music, defined as the horizontal succession of pitches. This set of features consists in metrics obtained from the notes of the melodic trajectory.

- Dynamics features

  Exploring the pitch salience of each note and how it compares with neighbour notes in the score gives us information about their individual intensity, as well as the intensity variation.

13



Novel Audio Features

- Rhythmic features

  Music is composed of sequences of notes changing over time, each with a specific duration.

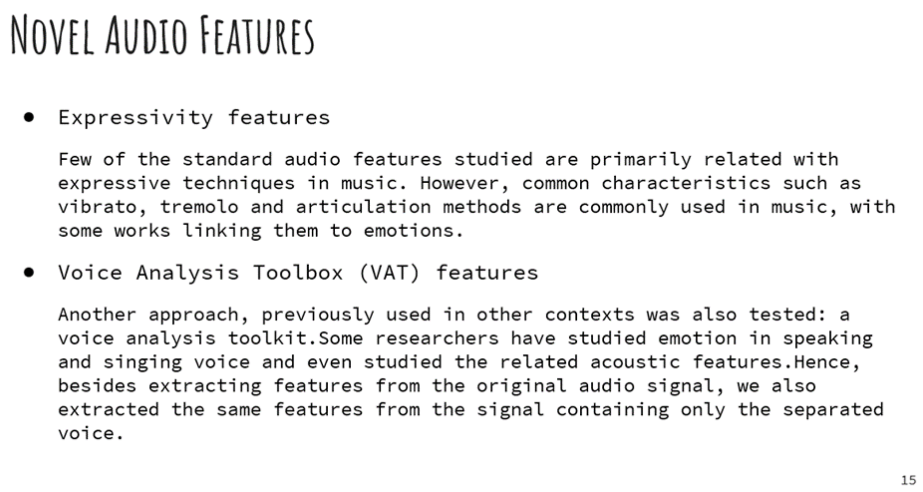
  Statistics on note durations are obvious metrics to compute. To capture the dynamics of these durations and their changes, three possible categories are considered: short, medium and long notes.

- Musical texture features

  To the best of our knowledge, musical texture is the musical concept with less directly related audio features available.

  it can influence emotion in music either directly or by interacting with other concepts such as tempo and mode.

  Here, the sequence of multiple frequency estimates to measure the number of simultaneous layers in each frame of the entire audio signal.



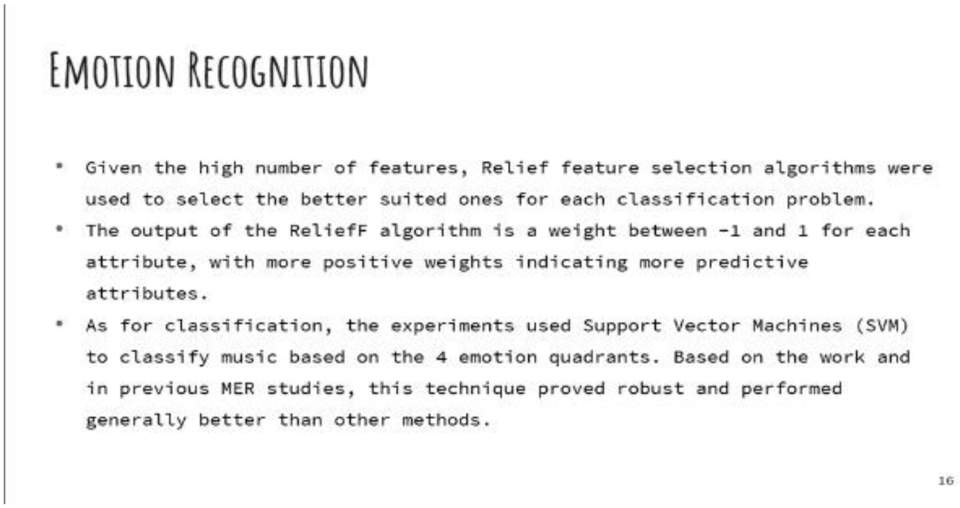
Novel Audio Features

- Expressivity features

  Few of the standard audio features studied are primarily related with expressive techniques in music. However, common characteristics such as vibrato, tremolo and articulation methods are commonly used in music, with some works linking them to emotions.

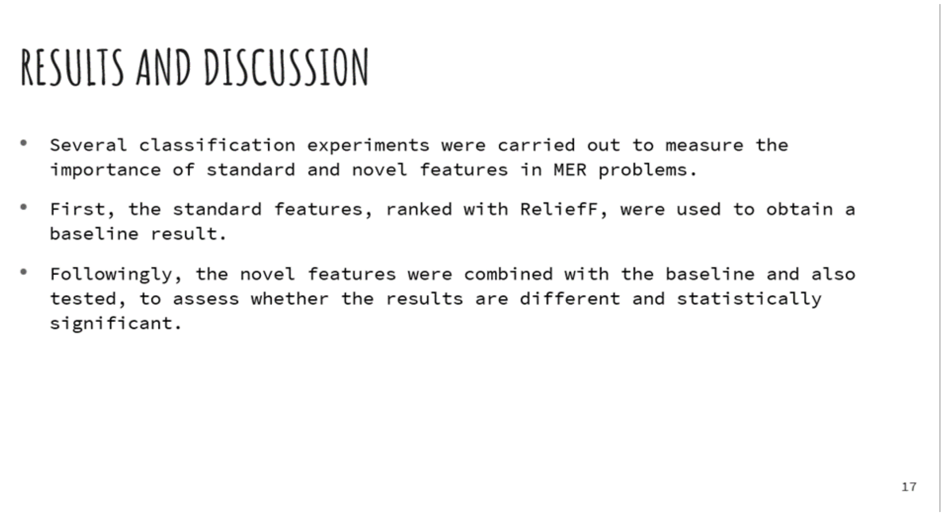
- Voice Analysis Toolbox (VAT) features

  Another approach, previously used in other contexts was also tested: a voice analysis toolkit.Some researchers have studied emotion in speaking and singing voice and even studied the related acoustic features.Hence, besides extracting features from the original audio signal, we also extracted the same features from the signal containing only the separated voice.



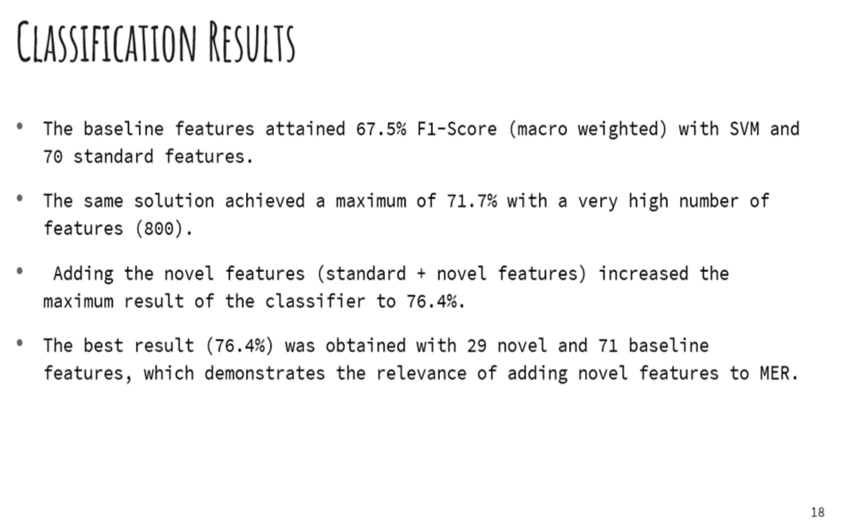
Emotion Recognition

- Given the high number of features, Relief feature selection algorithms were used to select the better suited ones for each classification problem.
- The output of the ReliefF algorithm is a weight between -1 and 1 for each attribute, with more positive weights indicating more predictive attributes.
- As for classification, the experiments used Support Vector Machines (SVM) to classify music based on the 4 emotion quadrants. Based on the work and in previous MER studies, this technique proved robust and performed generally better than other methods.



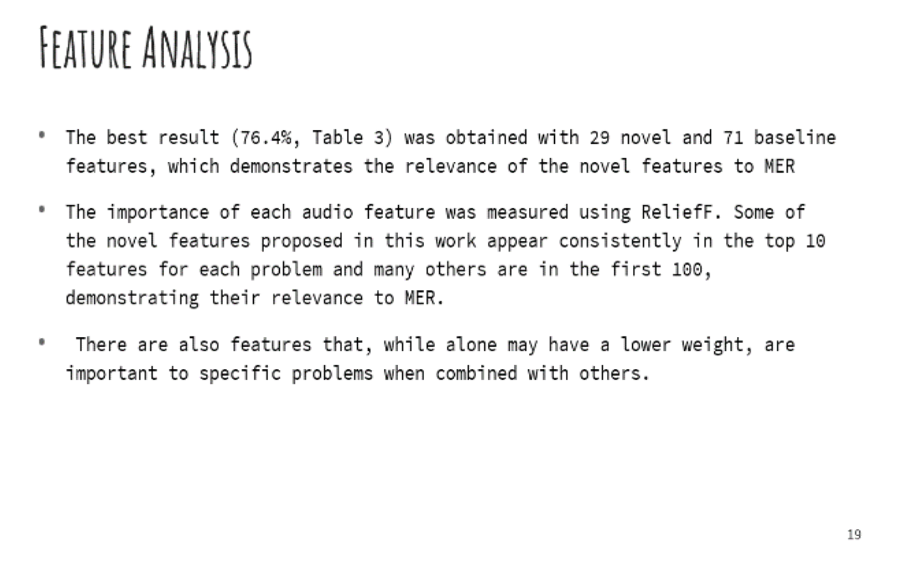
RESULTS AND DISCUSSION

- Several classification experiments were carried out to measure the importance of standard and novel features in MER problems.
- First, the standard features, ranked with ReliefF, were used to obtain a baseline result.
- Followingly, the novel features were combined with the baseline and also tested, to assess whether the results are different and statistically significant.



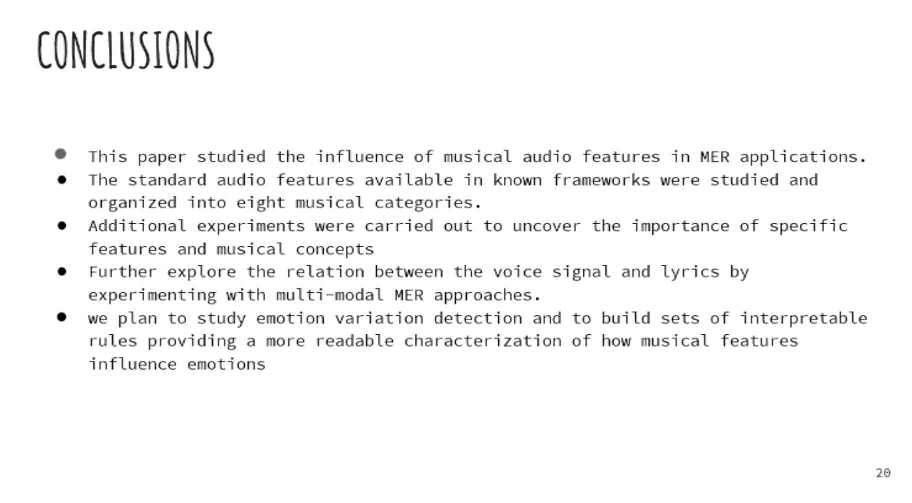
Classification Results

- The baseline features attained 67.5% F1-Score (macro weighted) with SVM and 70 standard features.
- The same solution achieved a maximum of 71.7% with a very high number of features (800).
- Adding the novel features (standard + novel features) increased the maximum result of the classifier to 76.4%.
- The best result (76.4%) was obtained with 29 novel and 71 baseline features, which demonstrates the relevance of adding novel features to MER.



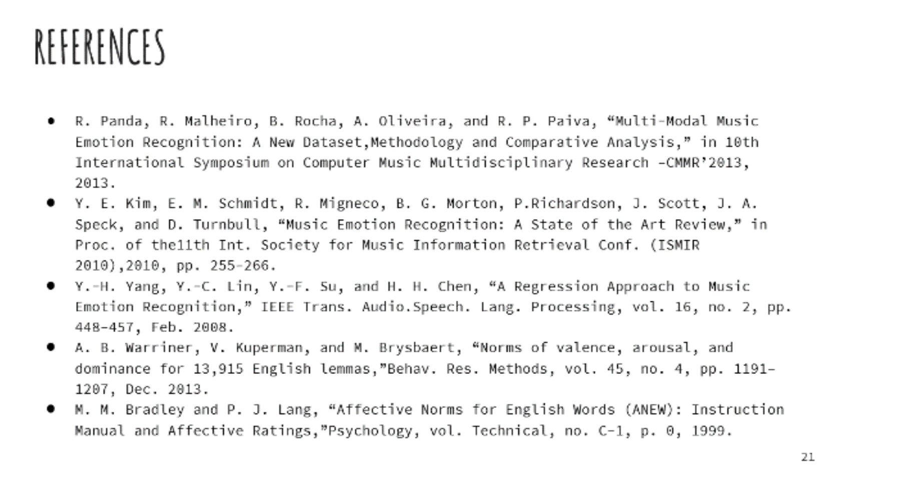
Feature Analysis

- The best result (76.4%, Table 3) was obtained with 29 novel and 71 baseline features, which demonstrates the relevance of the novel features to MER
- The importance of each audio feature was measured using ReliefF. Some of the novel features proposed in this work appear consistently in the top 10 features for each problem and many others are in the first 100, demonstrating their relevance to MER.
- There are also features that, while alone may have a lower weight, are important to specific problems when combined with others.



CONCLUSIONS

- This paper studied the influence of musical audio features in MER applications.
- The standard audio features available in known frameworks were studied and organized into eight musical categories.
- Additional experiments were carried out to uncover the importance of specific features and musical concepts
- Further explore the relation between the voice signal and lyrics by experimenting with multi-modal MER approaches.
- we plan to study emotion variation detection and to build sets of interpretable rules providing a more readable characterization of how musical features influence emotions



REFERENCES

- R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. P. Paiva, “Multi-Modal Music Emotion Recognition: A New Dataset,Methodology and Comparative Analysis,” in 10th International Symposium on Computer Music Multidisciplinary Research -CMMR’2013, 2013.
- Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P.Richardson, J. Scott, J. A. Speck, and D. Turnbull, “Music Emotion Recognition: A State of the Art Review,” in Proc. of the11th Int. Society for Music Information Retrieval Conf. (ISMIR 2010),2010, pp. 255-266.
- Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, “A Regression Approach to Music Emotion Recognition,” IEEE Trans. Audio.Speech. Lang. Processing, vol. 16, no. 2, pp. 448–457, Feb. 2008.
- A. B. Warriner, V. Kuperman, and M. Brysbaert, “Norms of valence, arousal, and dominance for 13,915 English lemmas,”Behav. Res. Methods, vol. 45, no. 4, pp. 1191-1207, Dec. 2013.
- M. M. Bradley and P. J. Lang, “Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings,”Psychology, vol. Technical, no. C-1, p. 0, 1999.