Cochlea-based Features for Music Emotion Classification

Luka Kraljević, Mladen Russo, Mia Mlikota and Matko Šarić
University of Split, FESB, Laboratory for Smart Environment Technologies, Split, Croatia

Keywords: Music, Emotion Detection, Cochlea, Gammatone Filterbank.

Abstract: Listening to music often evokes strong emotions. With the rapid growth of easily-accessible digital music libraries there is an increasing need in reliable music emotion recognition systems. Common musical features like tempo, mode, pitch, clarity, etc. which can be easily calculated from audio signal are associated with particular emotions and are often used in emotion detection systems. Based on the idea that humans don't detect emotions from pure audio signal but from a signal that had been previously processed by the cochlea, in this work we propose new cochlear based features for music emotion recognition. Features are calculated from the gammatone filterbank model output and emotion classification is then performed using Support Vector Machine (SVM) and TreeBagger classifiers. Proposed features are evaluated on publicly available 1000 songs database and compared to other commonly used features. Results show that our approach is effective and outperforms other commonly used features. In the combined features set we achieved accuracy of 83.88% and 75.12% for arousal and valence.

1 INTRODUCTION

Music is an essential element of our life, and mostly everyone listens to some kind of music. It is well known that music has affective characteristics which are used for mood and emotion regulation of listeners. With the rapid growth of easily-accessible digital music libraries comes a request for a new type of music labels. Standard music classification by author, genre and artist becomes insufficient because it lacks recommendation power.

To address this problem, music emotions recognition (MER) systems use emotions as a subjective criterion for a music search and organization. However, MER implementation does not come without challenging tasks.

Music emotion recognition is a personal experience and it can be viewed through two phases. In the first phase raw audio signal is processed by human ear and transformed in the form acceptable by person's brain (auditory model). Note that this step is common to all healthy listeners while the second step is more subjective and is related with signal processing inside person's brain, often depending not only on demographic characteristics of the listener like age, gender or culture but also on the current emotional state and previous experiences. Even when the same music is played, people can perceive different emotions. The words used to describe emotions are ambiguous and there isn't universal way to quantify emotions.

Most MER related studies are based on two popular approaches: categorical (Hevner, 1936) and dimensional models (Russell, 1980). The categorical approaches involve finding and organizing a limited number of universal categories such as happy, sad, anger and peaceful. As an alternative to categorical approach, dimensional model can be characterized by two affective dimensions called valence and arousal, as depicted in Figure 1. The arousal emotion ranges from calm to excited whereas the valence emotion ranges from negative to positive.

In recent years many research on emotion recognition from music were conducted using either categorical or dimensional model. As an artefact of this studies many publicly available database emerged, like the one we are using in this research.

Lu, Liu and Zhang (2006), extracted three types of features form western classic music clips. They used Gaussian Mixture Models as classification to distinguish moods namely Contentment, Depression, Exuberance and Anxious. Authors performed hierarchical classification by classifying moods based on intensity and then on rhythm and timbre.

Another research on the same mood set was conducted by Hampiholi (2012). In this work RMS

Kraljević, L., Russo, M., Mlikota, M. and Šarić, M.
Cochlea-based Features for Music Emotion Classification.
DOI: 10.5220/0006466900640068
In *Proceedings of the 14th International Joint Conference on e-Business and Telecommunications (ICETE 2017) - Volume 5: SIGMAP*, pages 64-68
ISBN: 978-989-758-260-8

Energy, Low Energy Frames, Brightness, Zero Crossing, Bandwidth and Roll-Off are used as features extracted from a set of 122 samples. The accuracy obtained by C4.5 decision tree classifier was 60% for Bollywood songs, and 40% for western music.

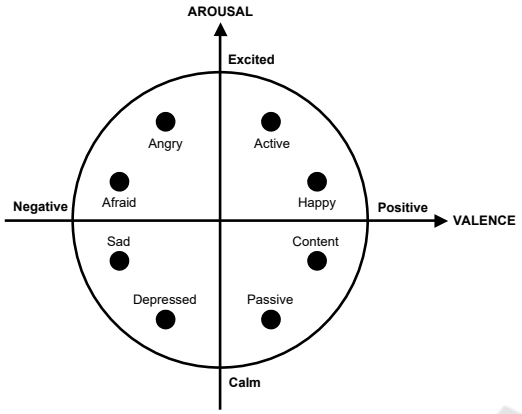

Figure 1: Valence-arousal dimensional model.

Wood and Semwal, (2016) proposed representing current emotion as blend of moods. Authors used set of algorithms to map feature data into complex combination of emotions.

Use of fuzzy classifiers for emotions classification divided into the four quadrants was proposed by Yang, Liu and Chen (2006). They used fuzzy vector to assign features as indication of relative strength for each class. Then they transformed fuzzy vector to AV space by considering the geometric relationship of the four emotion classes.

In the work done by Kartikay, Ganesan and Ladwani (2016) V-A ratings from 1000 songs database where used together with several common music features to classify emotions into four categories: Happy, Sad, Angry and Peaceful. They used Support Vector Machine (SVM), Naive Bayes, Linear Discriminant Analysis and Decision Trees as classification algorithms and they obtained accuracy of 59.8%, 75.4%, 78.5% and 71.4% respectively.

Interesting approach was proposed by Bai et al. (2016). Authors used regression approach modelling emotions like continues variable in valence-arousal space. As a performance measure they reported R^2 statistic as 29.3% and 62.5% respectively for Random Forest Regression and Support Vector Regression.

Some of the above studies concern on creating better algorithms for the prediction using only common acoustic features. On the other hand, others investigate combination of multiple acoustic features trying to find most informative features. But only a small amount of studies are dedicated to finding new type of features like Kumar et al (2016).

In their work they proposed two affective features namely compressibility and sparse spectral components. Authors reported that compressibility performance (53% and 73% accuracy for valence and arousal) outperforms other features while SSC (47% for valence 68% for arousal) is comparable to Mel-frequency cepstral coefficients (MFCC) (52% valence and 66% arousal).

The main idea behind our paper is that humans don't detect emotions from pure audio signal but from signal that had been previously processed by the cochlea, e.g. affective experience is integral to auditory perception. In our approach, we use a Gammatone filterbank auditory model of the cochlea to calculate novel features that could be used as state-of-the-art music emotion recognition features.

2 PROPOSED APPROACH

In this section we describe the proposed features together with brief description of other typical features used for accuracy evaluation. To ensure verifiability and comparability of our approach, we used publicly available 1000 Song Database (Soleymani et. al., 2013), as our benchmark database. For comparison we used some common acoustic features (tempo, pitch, RMS…) which are often used in literature, e.g. (Kartikay, Ganesan and Ladwani, 2016) (Kim et.al 2010). Also to ensure high-quality test, instead of using "hand-crafted" approach for feature calculation these common features were extracted using the Music Information Retival (MIR) toolbox (Lartillot and Toiviainen, 2007).

Our results and analysis are based on valence-arousal dimensional model, Figure 1. Valence represents natural attractiveness or awareness of any emotion, e.g., how positive/negative emotion is. Arousal can be viewed as the strength of emotion, e.g. how exciting or calming emotion is.

2.1 1000Song Database

Original database consists of 1000 music excerpts provided by Free Music Archive (FMA). Authors found that database contains some duplicates and after removing redundant files they provided a reduced set of 744 music clips that was used as our benchmark database. Each music excerpt is 45s long

and sampled with frequency rate of 44100Hz. For each music clip authors provided a file with continuous annotation rating on scale between -1 and 1 collected with 2Hz sampling rate. Authors also provided a file with arousal or valence rating of the hole clip on a nine-point scale collected with use of self-assessment manikins (SAM) which is a tool that uses pictorial scale, displaying comics characters expressing emotions.

In this work, those ratings have been thresholded to form binary classification problem. Emotions rated above five have been classified as high otherwise as low placing emotional state in one of four quadrants of V-A space.

2.2 Gammatone Filter Bank

Modelling the natural response of the human auditory system we can simulate machines to act as the human ear. Gammatone filterbank is a widely used model of auditory filters. (Patterson et al., 1992) It performs spectral analysis and outputs sound signal into channels where each channel represents motion of basilar membrane. The impulse response of a Gammatone filter is given as a product of a Gamma distribution function and a sinusoidal tone at central frequency f_c

$$g(t) = Kt^{(n-1)}e^{-2\pi Bt}\,cos(2\pi f_c + \varphi) \quad t > 0 \quad (1)$$

where K is the amplitude factor, n is the filter order, f_c is the central frequency in Hz, φ is phase shift and B is duration of impulse response. Scaling of Gammatone filterbank is determined by B which is related to equivalent rectangular bandwidth (ERB). ERB represents a psychoacoustic measure for width of auditory filter at each point along the cochlea.

$$B = 1.019 \times 2\pi \times ERB \quad (2)$$

$$ERB = \left[\left(\frac{f_c}{EarQ}\right)^p + (minBW)^p\right]^{1/p} \quad (3)$$

where $EarQ$ is the asymptotic filter quality at high frequencies and $minBW$ is min bandwidth at low frequency. In this work we used bank of 20 Gammatone filters with $EarQ = 9.26449$ and $minBW = 24.7$ as parameters proposed by Glasberg and Moore (1990).

Corresponding impulse response is show in Fig. 2; top plot represents response in frequency domain, bottom plot represents time domain response for each of 20 gammatone filters.

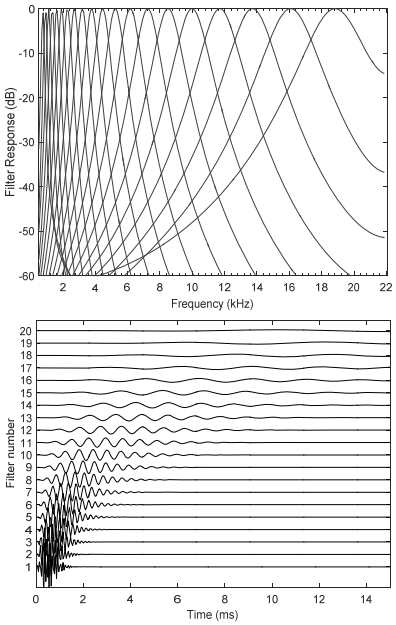

Figure 2: Impulse response of 20-channel Gammatone filter bank.

2.3 Feature Extraction

Feature extraction is an important step in obtaining affective information from the audio signal. In our approach, first we design 20 filters needed to implement the ERB cochlear model using default values of the algorithm as stated above. Then we filtered an audio signal with the bank of Gammatone filters resulting with 20 channels output, each representing the output of the basilar membrane at particular location corresponding to filter's central frequency. After that, average power of each filter's output is calculated to be input feature, forming the feature vector of 20 elements.

All features are extracted on clip level since we observe static clip annotations. In order to compare our approach to other state of the art features, we used MIR Toolbox and extracted the following features:

Energy: the root mean square represents the global energy of the entire audio signal. It also represents sensitivity to loudness which is proportional to the average value of the various peak levels.

Tempo: represents the speed of the song. It is measured by detecting periodicities in a range of beats per minute (BPM).

Event Density: number of similar events in a clip, the average frequency of events during specified time (number of notes per second).

Mode: indicates tonalities that can be used and indicates tonalities with special importance i.e. major vs. minor, returned as a numerical value between -1 and +1.

Pitch: is responsible for discerning sounds as lower or higher. Represents average frequency of the song.

Key Clarity: represents key strength of the best keys

Pulse Clarity: represents strength of the beats, the ease with which one can perceive the underlying rhythmic or metrical pulsation

Roughness: or sensory dissonance, represents average of all the dissonance between all possible pairs of peaks.

MFCC: coefficients representing the spectral shape of the sound. It approximates the human auditory response taking perceptual consideration.

Sometimes it is hard to directly compare features' performance between other researches mainly because accuracy interpretation may differ in terms of the emotion model (categorical or dimensional), difference in duration of music piece (full song, verse, fixed-length clip or segment e.g., 500ms), quality of an audio database (how was ground truth collected). In order to evaluate recognition results, proposed feature will be compared with two sets of features: Basic set contains RMS, Tempo, Event Density, Mode, Pitch, Key Clarity, Pulse Clarity and Roughness yielding with the 8-dimensional feature vector and MFCC set with 13 MFCC coefficients. Comparison with MFCC is particularly important because both MFCC and proposed features are perceptually-motivated. That's why we also compared the overall accuracy gain combining one or the other with the base set.

3 CLASSIFICATION AND PERFORMANCE EVALUTION

The proposed approach uses two different classifiers to perform two class classification on valence and arousal emotional dimensions. Support Vector Machine is a learning algorithm for pattern classification with good generalization properties, insensitive to overtraining and the curse of dimensionality. TreeBagger is ensemble of decision trees.

Forming ensembles gives a boost in accuracy on dataset. It combines the results of many decision trees improving generalization by reducing effects of overfitting.

To correctly evaluate classification performance, 10% of complete dataset was randomly taken as a test set leaving rest for training phase. Training was performed using the standard 10-fold cross-validation protocol in an effort to minimize the potential of overfitting. Results were calculated as average of 100 iterations. The comparison of performance between the proposed feature, basic set and MFCC coefficients is given in the Table 1 regarding SVM and Table 2 when TreeBagger is used as classifier.

Table 1: Accuracy comparison using SVM.

	Arousal	Valence
Basic	75.13	73.10
Proposed	77.39	68.08
MFCC	74.17	67.02
Basic + Proposed	80.97	74.45
Basic + MFCC	79.76	72.72

Table 1 and Table 2 also summarize comparison of overall accuracy gain when proposed features or MFCC features are combined with the basic set.

Table 2: Accuracy comparison using TreeBagger.

	Arousal	Valence
Basic	75.52	72.83
Proposed	81.02	68.40
MFCC	71.45	64.43
Basic + Proposed	83.88	75.12
Basic + MFCC	79.64	72.72

MFCC achieved the best validation score when SVM was used, and that's why we first compared our proposed features to MFCC with SVM learning algorithm. It is visible from the results that proposed features outperform MFCC features in direct comparison as well as in accuracy gain when combined with the base set. In order to improve learning efficiency, and possibly improve prediction performance of our proposed features, we also used TreeBagger as a classifier since it reported best validation accuracy using 10-fold cross-validation. Our results show that TreeBagger classifier enables performance boost of the proposed approach for up to 4%.

4 CONCLUSIONS

Generally speaking, one cannot listen to music without affection involvement. Emotion-detecting is a perceptive task and nature has developed an efficient strategy to accomplishing it. Based on the

idea that humans don't detect emotions from pure audio signal but from signal that had been previously processed by the cochlea, in this work we proposed a new feature set for music emotion recognition.

An audio signal was filtered with a bank of Gammatone filters resulting with 20 channels each representing output of the basilar membrane at particular location corresponding to filter's central frequencies. During automated process proposed features were extracted as average power of each filter's output and compared with other state of the art features. Support vector machine and TreeBagger classifiers were used for performance evaluation. Experimental results on 1000 Songs Database showed that the proposed feature vector outperforms basic set as well as MFCC features. Proposed features also performed better in terms of accuracy gain when combined with the base set. Comparison with MFCC coefficient is particularly relevant because it gives us the real insight on how well our proposed feature performs over other perceptually-motivated feature.

Although the results are good they still need to be improved for real-life applications where emotional changes should be tracked continuously during audio clip. In our future work, we will focus our research in direction of developing improved features based on auditory perception.

Extracting features from a more complex model of auditory processing, thus simulating cochlea in more detail could bring us further in improving music emotion recognition.

ACKNOWLEDGEMENTS

This work has been fully supported by the Croatian Science Foundation under the project number UIP-2014-09-3875.

REFERENCES

Bai, J., Peng, J., Shi, J., Tang, D., Wu, Y., Li, J., Luo, K., 2016. Dimensional music emotion recognition by valence-arousal regression, *2016 IEEE 15th International Conference on Cognitive Informatics & Cognitive Computing* (ICCI*CC), Palo Alto, CA, pp. 42-49.

Glasberg, B.R., Moore, B. C. J., 1990. Derivation of auditory filter shapes from notched-noise data, *Hearing Res.*, vol. 47, pp. 103–108.

Hampiholi, V., 2012. A method for Music Classification based on Perceived Mood Detection for Indian Bollywood Music, *World Academy of Science, Engineering and Technology*. vol.6 December.

Hevner, K., 1936. Experimental studies of the elements of expression in music. *American Journal of Psychology*,vol. 48, pp. 246–268.

Kartikay, A., Ganesan, H., Ladwani, V. M, 2016. Classification of music into moods using musical features, *International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, pp. 1-5.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton B. G., Richardson P., Scott, J., Speck, J. A., Turnbull, D., 2010. Emotion Recognition: a State of the Art Review., *11th International Society for Music Information and Retrieval Conferenc.*

Kumar, N., Guha, T., Huang, C. W., Vaz, C., Narayanan, S. S., 2016. Novel affective features for multiscale prediction of emotion in music, *IEEE 18th International Workshop on Multimedia Signal Processing (MMSP)*, Montreal, QC, pp. 1-5.

Lartillot, O., Toiviainen, P., 2007. A Matlab Toolbox for Musical Feature Extraction From Audio, *International Conference on Digital Audio Effects*, Bordeaux.

Lu, L., Liu, D., Zhang, H.-Y. 2006. Automatic Mood Detection and Tracking of Music Audio Signals, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 1, January.

Patterson, R. D., Robinson K., Holdsworth, J., McKeown, D., Zhang, C., Allerhand, M. H., 1992. Complex sounds and auditory images *In Auditory Physiology and Perception, (Eds.)*, Oxford, pp. 429-446.

Russell, J. A, 1980. A circumplex model of affect. *J. Personality Social Psychology*, vol. 39, no. 6, pp. 1161–1178.

Soleymani, M., Caro, M. N., Schmidt, E. M., Sha, C.-Y., Yang, Y.-H., 2013. 1000 Songs for Emotional Analysis of Music, *Proc. of the 2Nd ACM International Workshop on Crowdsourcing for Multimedia.*, Barcelona, Spain.

Wood, P. A., Semwal, S. K., 2016. An algorithmic approach to music retrieval by emotion based on feature data, *2016 Future Technologies Conference (FTC)*, San Francisco, CA, pp. 140-144.

Yang, Y.-H., Liu C.-C., Chen, H.-H., 2006. Music emotion classification: A fuzzy approach, *ACMMM*, pp. 81-84.