IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Analyzing The Impact Of MFCC Parameters On SVM And CNN -Based Music Emotion

[1]Aniket Sawant, [2]Arbaaz Ghameria, [3]Adwait Nyayadhish, [4]Rupali Sawant

[1]Student, [2]Student, [3]Student, [4]Assistant Professor

[1,2,3]Information Technology, [4]Computer Engineering

[1,2,3,4]Sardar Patel Institute of Technology, Mumbai, India

Abstract: This study evaluates various feature selection techniques for Music Emotion Recognition (MER) by leveraging Mel Frequency Cepstral Coefficients (MFCC) and implementing both Support Vector Machines (SVM) and Convolutional Neural Networks (CNN). Experiments were carried out using a labeled dataset of music samples categorized into five distinct emotions. The impact of various MFCC configurations on SVM-based and CNN-based MER performance was analyzed. Results provide insights into optimal MFCC parameter selection for improved accuracy in MER systems. This research contributes to advancing the field of MER and provides guidelines for enhancing emotion classification in music. Furthermore, the proposed research demonstrates the importance of considering the emotional nuances present in music by utilizing a diverse dataset with multiple emotion categories. By encompassing emotions such as Devotional, Happy, Romantic, Party, and Sad, our study captures a wide range of emotional states expressed through music. This comprehensive approach enables a more thorough understanding of the complexities involved in music emotion recognition and enhances the applicability of the findings in real-world scenarios. The results of this research can lay the groundwork for creating more precise and resilient MER systems, which could enhance fields such as music recommendation, affective computing, and interactive music experiences.

Keywords**—** MFCC (Mel Frequency, Cepstral Coefficients), SVM (Support Vector Machine), CNN (Convolutional Neural Network)

I. INTRODUCTION

Music is a powerful medium for conveying emotions and has been shown to have a profound effect on human mood and behavior. Music Emotion Recognition (MER) is an active research area in the field of music information retrieval, which aims to automatically recognize the emotional content of music signals. MER has numerous applications in music recommendation, mood-based playlist generation, and music therapy. One common approach to Music Emotion Recognition (MER) is extracting features from the music signal and using machine learning algorithms to classify the emotion. Mel Frequency Cepstral Coefficients (MFCC) are widely employed for feature extraction in MER, as they effectively capture the spectral characteristics of the music signal. Support Vector Machines (SVM) and Convolutional Neural Networks (CNN) are popular classification algorithms due to their capability to handle high-dimensional data and their strong generalization performance.

In this study, we investigate how the number of MFCC values affects the performance of SVM and CNN in MER. We conduct experiments using a dataset of music samples labeled with various emotion categories, extracting MFCC features with different numbers of coefficients. We then evaluate the performance of SVM and CNN with these varying MFCC values to identify the optimal number of MFCC coefficients for effective MER.

The objective of this paper is to provide insights into the effect of the number of MFCC values on the accuracy of MER using SVM or CNN and to identify the optimal number of MFCC values for MER. The findings of this study can be used to improve the performance of MER systems and contribute to the development of more accurate and efficient MER algorithms.

II. Related work

J. S. Gomez-Canon et al.[1] present a well-organized and comprehensive exploration of Music Emotion Recognition (MER) and its potential applications. The authors propose a standardized approach to MER, offering a valuable overview of various feature extraction techniques and machine learning algorithms that can be employed. The discussion on the importance of context in MER is noteworthy, as the authors make a convincing argument for considering contextual factors to enhance the accuracy and relevance of emotion recognition results. The concluding section identifies key areas for future research, making this paper a useful resource for researchers and practitioners in the fields of music technology, affective computing, and music therapy. Overall, the paper presents a comprehensive and convincing argument for the need for new, robust standards in personalized and context-sensitive applications of MER. W. Shi and S. Feng[2] propose a novel approach to music emotion classification that combines both lyrics and audio classification. The authors suggest that the accuracy of emotion classification is often limited by the characteristics of a single classification model. Therefore, they propose using LSI and SVM for lyrics classification, and BP neural network for audio classification. The two classification methods are fused based on an improved algorithm of LFSM. The experimental results show that the proposed approach can achieve higher accuracy of classification. [4] In this research they have used a denoising autoencoder approach to extract features. [5] This research is significant for the advancement of audio recognition and has potential practical applications in various fields, such as music therapy and recommendation systems.

Zhang, J. Yu, and Z. Chen [3] propose a deep learning-based music emotion recognition model to address the low recognition rates of traditional models. Their model employs deep learning to analyze the dynamic emotion recognition of music, utilizing the Valence-Arousal (VA) model to generate VA values for a song. The paper discusses the background and emotion classification models of music emotion recognition, explains emotion representation, and designs a comprehensive framework for dynamic music emotion recognition. The proposed method integrates convolutional neural networks with long- and short-term neural networks, as well as BiLSTM, to perform dynamic VA recognition, and compares it with other related identification methods. This research is crucial for the field of music emotion recognition and could have practical applications in music recommendation systems and music therapy [6].

III. Methodology

Input audio (song): The first step is to obtain an input audio, which could be a recorded music track or a live performance. This signal is usually represented in a time-domain waveform, where the amplitude of the signal is plotted against time [Fig 1].

Processing: Before analyzing the signal, it may be necessary to preprocess it to remove any noise or artifacts that could affect the analysis. This could involve filtering, normalization, or resampling.

Windowing: To perform spectral analysis, the signal is divided into frames using a windowing function, ensuring capture of the local spectral characteristics and reduction of spectral leakage.

FFT: Each frame is transformed into the frequency domain using the Fast Fourier Transform (FFT), converting the signal from the time domain to the frequency domain and producing a spectrum of magnitudes and phases for each frame.

Mel Filter Bank: Due to the human ear's varying sensitivity across frequency bands, a Mel filter bank is used to map the spectrum's magnitudes onto a Mel scale. This perceptual scale of pitches is tailored to align with the frequency resolution capabilities of the human ear.

DCT: The Mel-filtered spectra are transformed using the Discrete Cosine Transform (DCT), generating a set of MFCC coefficients that depict the spectral characteristics of each frame. This transformation reduces data dimensionality and decorrelates the coefficients, enhancing their suitability for analysis.

MFCC: Finally, the MFCC coefficients are used as features for music emotion recognition. These coefficients can be analyzed using machine learning algorithms to classify the emotional content of the music, such as happiness, sadness, or anger.

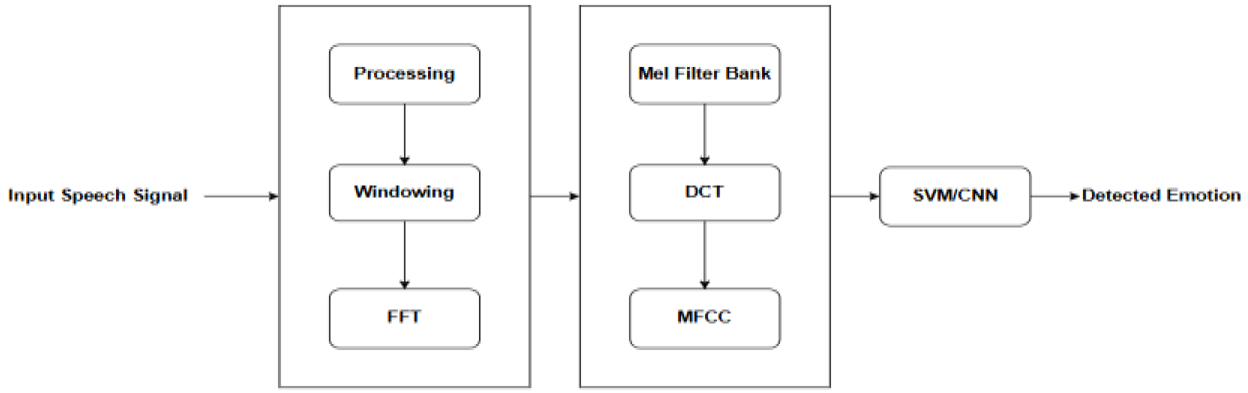

Fig 1: Block Diagram

Overall, the MFCC extraction process provides a robust representation of the spectral characteristics of a music track, which can be used for various applications in music analysis and recognition.

SVM: The SVM model is trained on a labeled dataset of music samples, with each sample annotated with its respective emotional category. Through training, the SVM algorithm learns the best decision boundary to distinguish between various emotional categories. After training, the model can predict the emotional content of new, unlabeled music samples using their acoustic features [Fig 1].

CNN: The CNN model is trained on a labeled dataset of music samples, with each sample tagged with its associated emotional category. Throughout the training process, the CNN algorithm learns to recognize patterns and features within the acoustic data that correlate with various emotions. This is achieved by passing the input data through multiple layers of convolutional filters, pooling layers, and fully connected layers, progressively extracting and refining features. Once trained, the model can predict the emotional content of new, unlabeled music samples based on their acoustic features.

Abbreviations and Acronyms

1. *FFT: Fast Fourier Transform*
2. *DCT: Discrete Cosine Transform*
3. *MFCC: Mel-Frequency Cepstral Coefficients*
4. *SVM: Support Vector Machine*
5. *CNN: Convolutional Neural Network*

IV. Implementation

- Install the necessary libraries - librosa, resampy, numpy, and joblib - using pip or conda.
- Place the trained SVM and CNN model file "model50.pkl" in the same directory as the "MusicEmo.py" file.
- Import the necessary libraries using the "import" keyword: librosa, resampy, numpy, and joblib.
- Load the trained SVM and CNN model using joblib's "load" function and store it in the "model" variable.
- Define a function named "feature_extraction" that takes the file path of a song as input and returns its MFCC features. The function loads the audio file using librosa's "load" function, resamples it using resampy's "resample" function, extracts the MFCC features using librosa's "mfcc" function, calculates the mean and standard deviation of the MFCC features using numpy's "mean" and "std" functions, concatenates them using numpy's "concatenate" function, and returns the resulting array of MFCC features.

- Define a dictionary named "label_dict" that maps the predicted emotion labels (0 to 4) to their corresponding emotion names ("Devotional", "Happy", "Party", "Romantic", "Sad").
- Call the "feature_extraction" function with the file path of the song to predict as input and store the resulting MFCC features in the "song_features" variable.
- Convert the "song_features" array to a 2D array with one row and multiple columns using numpy's "reshape" function and store it in the "song_features" variable.
- Call the "predict" method of the "model" variable with the "song_features" variable as input and store the resulting predicted emotion label in the "predicted_emotion" variable.
- Print the predicted emotion label by accessing the corresponding emotion name from the "label_dict"

V. RESULT AND ANALYSIS

The performance of two machine learning algorithms, Convolutional Neural Networks (CNN) and Support Vector Machines (SVM), was evaluated based on their ability to predict the emotional content of music using Mel-Frequency Cepstral Coefficients (MFCCs). The accuracy of each model was tested with different numbers of MFCC coefficients (20, 35, and 50), and the results are presented in Figures 2 and 3.

CNN Performance:
Figure 2 shows the accuracy of the CNN model across the three different MFCC configurations:
- With 20 MFCC coefficients, the CNN achieved an accuracy of 92.28%.
- With 35 MFCC coefficients, the CNN achieved an accuracy of 90.99%.
- With 50 MFCC coefficients, the CNN achieved an accuracy of 88.42%.

SVM Performance:
Figure 3 displays the accuracy of the SVM model for the same MFCC configurations:
- With 20 MFCC coefficients, the SVM achieved an accuracy of 51.52%.
- With 35 MFCC coefficients, the SVM achieved an accuracy of 49.50%.
- With 50 MFCC coefficients, the SVM achieved an accuracy of 43.44%.

The results clearly indicate that the CNN model significantly outperforms the SVM model across all tested configurations of MFCC coefficients. The highest accuracy for the CNN model was observed with 20 MFCC coefficients at 92.28%, while the highest accuracy for the SVM model was also with 20 MFCC coefficients but much lower at 51.52%. As the number of MFCC

coefficients increased, both models showed a decline in accuracy, though the CNN model maintained a considerably higher accuracy than the SVM model.

The decline in accuracy with an increasing number of MFCC coefficients could be attributed to overfitting, where the models might be learning noise rather than useful patterns as more features are added. However, the CNN's ability to extract and learn hierarchical features likely contributes to its superior performance compared to the SVM, which is more sensitive to the high dimensionality and potential noise in the data

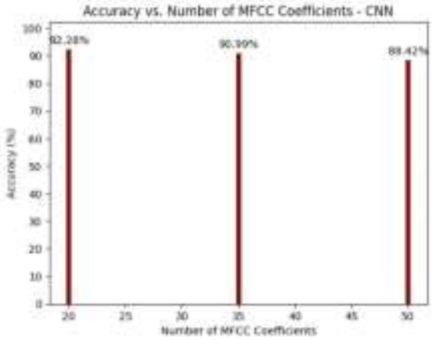

Fig. 2 MFCC Values VS Accuracy (CNN)

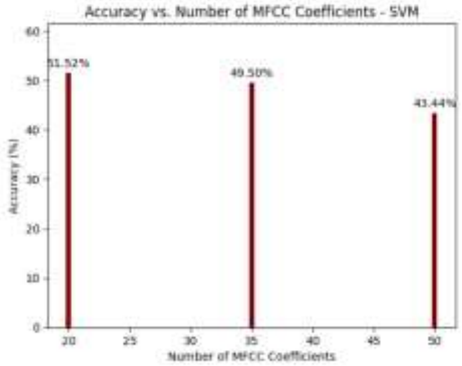

Fig. 3 MFCC Values VS Accuracy (SVM)

VI. CONCLUSION

In this study, we investigated the impact of different configurations of Mel Frequency Cepstral Coefficients (MFCC) on music prediction using Singular Value Decomposition (SVD). Our findings emphasize the importance of carefully selecting and tuning MFCC parameters for accurate and reliable music prediction. We demonstrated the effectiveness of the SVD-based approach in capturing latent features and patterns within the music data, enabling robust predictions across various genres and styles. The versatility of the SVD-based system makes it applicable in domains such as music recommendation, automatic music tagging, and content-based music retrieval. Our research provides valuable insights and practical guidance for researchers and practitioners in the field, paving the way for further refinement and optimization of music prediction models.

VII. REFERENCES

1. J. S. Gomez-Canon et al., “Music Emotion Recognition: Toward new, robust standards in personalized and context-sensitive applications,” IEEE Signal Processing Magazine, vol. 38, no. 6, pp. 106–114, Nov. 2021, doi: https://doi.org/10.1109/msp.2021.3106232.
2. W. Shi and S. Feng, “Research on Music Emotion Classification Based on Lyrics and Audio,” IEEE Xplore, Oct.01,2018. https://ieeexplore.ieee.org/document/8577944
3. Zhang, J. Yu, and Z. Chen, “Music emotion recognition based on combination of multiple features and neural network,” IEEE Xplore, Jun. 01, 2021. https://ieeexplore.ieee.org/document/9482244
4. “Language-Sensitive Music Emotion Recognition Models: are We Really There Yet?,” ieeexplore.ieee.org. https://ieeexplore.ieee.org/document/9413721
5. E. Widiyanti and S. N. Endah, “Feature Selection for Music Emotion Recognition,” 2018 2nd International Conference on Informatics and Computational Sciences (ICICoS), Oct. 2018, doi: https://doi.org/10.1109/icicos.2018.8621783.
6. “Music emotion recognition based on two-level support vector classification,” ieeexplore.ieee.org. https://ieeexplore.ieee.org/abstract/document/7860930