Music Recommendation System based on Facial Expression and Speech

Mrunmayee Shewale[1], Sahil Sinha[2], Prof. Satyajit Sirsat[3] Department of Computer Engineering,
Nutan Maharashtra Institute of Engineering and Technology,
Talegaon Dabhade, Pune

Abstract: We propose a new approach for playing music automatically using facial emotion Current methods often involve manually recording music, using computer-based tools, or classifying sounds. Instead, we recommend manually changing the way you rank and play. We use convolutional neural networks for emotion recognition. Pygame and Tkinter are available for down load. Our proposed method will reduce the calculation time as well as reduce the total cost of obtaining results and building the system, thus improving the overall accuracy of the system. The testing was done on the FER2013 dataset. Capture face with built in camera. Feature extraction is performed on facial images to direct emotions such as happiness, anger, sadness, surprise and neutrality

Keywords: Face Recognition, Feature extraction, Emotion detection, Convolutional Neural Network, Pygame

I.Introducton:

In recent years, the field of music recommendation systems has witnessed significant advancements driven by the convergence of artificial intelligence, machine learning, and human- computer interaction technologies. However, a burgeoning area of research explores novel approaches that incorporate multimodal signals, including facial expressions and speech, to enhance the accuracy and conceptuality of music recommendations. The integration of facial expression and speech analysis into music recommendation systems represents a paradigm shift towards more nuanced and emotionally intelligent computing. Facial expressions are powerful indicators of emotional states, offering valuable insights into an individual mood, engagement level, and affective response to stimuli, including music. Similarly, speech carries rich emotional and contextual cues that can significantly inform the music recommendation process. By leveraging these non-verbal and verbal signals, researchers are exploring innovative ways to tailor music suggestions that resonate with emotional and cognitive states in real-time.

II. Literature Review

Recognizing human emotion is considered a fascinating task for data scientists. Studying human emotion needs appropriate sensors to deploy for collecting the right data. These sensors are generally used for automatic emotion
collection, recognition, and making
intelligent decisions. The key central
nervous system (CNS) emotional-affective processes are, (i) Primary-process,
(ii) secondary-process, and (iii) tertiary-
process.[1] The music module has three modules: theory module, music classification module, and agreement module. The theory uses the user facial image as input and uses deep learning to identify the user emotions with 90.23% accuracy. The music classification module uses audio to classify music into 4 different types of emotions and achieves a significant result of 97.69%[2].For music recommendations,
Pygame &amp; Tkinter are
used. Our proposed method will reduce the computational time re quired to obtain the results and the total cost of building the system, thus improving the overall accuracy of the system. The testing was done on the FER2013 dataset. Capture face with builtin camera. Feature extraction is performed on facia I images to detect emotions such as happiness, anger, sadnes s, surprise and neutrality [3]. The meter, timbre, rhythm, and pitch of musicare
managed in areas of the brain that affects emotions and mood interaction between individuals may be a major aspect of lifestyle. It reveals perfect details and much of data among humans,whether they are in the form of body language, speech, facial expression, or emotions. Nowadays, emotion detection is considered the most important technique used in many applications such as smart card applications, surveillance, image database investigation, criminal, video indexing, civilian applications, security, and adaptive human-computer interface
with multimedia environments [4].

Black box learning competition, face recognition competition and multidisciplinary learning competition. We describe the materials created for these challenges and summarize the challenges. We offer advice for those working on future challenges and provide some advice on what insights can be gained from the machine learning
Challenge [5].The application works Unlike regular software, it scans the audio files present on the device and categorizes them according to parameters (audio properties) predefined in the app, creating a series of mood playlists. The graphical input given to the application is sorted (face recognition) to create a sense that is used to select the desired playlist from the previously written ones. The main goal of this article is to create a useful and accurate algorithm for creating playlists based on the user current mood and behavior [6]. Sapkota The face recognition is considered as one of the best way to determine a person mood. this image processing system is used for reducing the face space dimensions using the principal component analysis(PCA) method and then it applies fishers linear discriminant(FDL) or the LDA method to obtain the feature of the image characteristics, we especially use this because it maximizes the training process classification in between classes. This algorithm helps to process for image recognition is done in fisher face while, matching faces algorithm we use minimum Euclidean it helps us to classify the expression that implies the emotion of the user[7]. The music player that we are using it can be used locally and nowadays everything became portable and efficient to carry but it the emotion of a person can be taken by different of wearable sensors and easy to use rather than the whole manual work it would be possible using GSR(galvanic skin response) and PPG(plethysmography physiological sensors) that would give us enough data to predict the mood of the customer accurately. This system with enhanced will be able to benefit and the system with advanced features and needs to be constantly upgraded [8].

Face detection is one of the applications which is considered under computer vision technology. This is the process in which algorithms are developed and trained to properly locate faces or objects in object detection or related system in mages. This detection can be real-time from a video frame or images. Face detection uses such classifiers, which are algorithms that detect what either a face (1) or not a face (0) in an image. Classifiers are trained to detect faces using numbers of images to get more accuracy. OpenCV uses two sorts of classifiers, LBP (Local Binary Pattern) and Hair Cascades [9]. This paper proposes a completely unique framework for expression recognition by victimization look options of elite facial patches, a number of outstanding facial patches, reckoning on the position of facial landmarks, area unit extracted that area unit active throughout feeling stimulus [10]. These active patches area unit more processed to get the salient patches that obtain discriminative options for classification of every try of expressions, thereby choosing completely different facial patches
as salient for various try of expression categories. One against one classification methodology is
adopted victimization these options [11]. This Paper proposes the robust, discriminative features
learned by the Convolution neural network The outputs of the model are going to be feature
maps, which are an intermediate representation for all layers after the very first layer. Load the
input image for which we want to view the Feature map to know which features were promi nent
to classify the image [12]. Computer interfaces
(BCIs) aim at providing ano muscular channel for causing commands to the external world
exploitation the medical instrument activity
or alternative
electrophysiological measures of the brain
perform [13]. This literature review surveys key
studies and developments in this area, focusing on
the integration of multimodal signals to enhance
music recommendation accuracy and user
experience.

III. Methodology

The methodology employed in the development of a music recommendation system integrating facial expression and speech analysis is multifaceted, encompassing data collection, feature extraction, model implementation, and evaluation. This section outlines the step-by-step approach adopted to design and implement the proposed system. Data Collection: The first phase involves acquiring multimodal data consisting of facial expression recordings and speech samples from a diverse set of users. This may be done using webcams for capturing facial
expressions during music listening sessions and microphone recordings for extracting speech features. Its essential to ensure a diverse

dataset that encompasses various emotional states and user preferences to train and validate the recommendation model effectively. Facial Expression Analysis: Facial expression analysis is performed using computer vision techniques to extract meaningful features from facial images. This includes detecting key facial landmarks, such as eyes, nose, and mouth, and analyzing changes in expressions over time.

Techniques like facial landmark detection, facial action unit recognition, and emotion classification are employee to quantify emotional responses during music listening. Speech Analysis: Speech samples are preprocessed to extract relevant features that convey emotional and contextual information. This involves techniques such as speech signal processing, including feature extraction (e.g., MFCC Mel- Frequency Cepstral Coefficients), and

sentiment analysis to derive emotional cues from spoken language. Natural language processing tools may be utilized to analyze speech transcripts and identify sentiment, topic, and user preferences related to music. Develop machine learning or deep learning models to learn the relationship between multiple inputs (facial expressions, speech) and musical preferences Commonly used models include convolutional neural networks (CNN) for image based processing (face analysis), recurrent neural networks (RNN) for data processing (speech analysis), and hybrid are hitectures that combine both methods. Transfer learning using predefined models (e.g. ResNet, LSTM) can increase training and improve performance.

Evaluation:

IV. Emotion Detection module:

Facial expression analysis is a critical component of music recommendation systems that aim to incorporate users emotional states into personalized recommendations. This process involves capturing and interpreting facial movements and expressions using computer vision techniques. Here are the key aspects and methodologies involved in facial expression analysis for music recommendation: Data Acquisition: Facial expression analysis begins with the acquisition of facial data using cameras or webcams capable of capturing high-resolution images or videos of faces. It essential to gather a diverse dataset that includes a wide range of facial expressions corresponding to different emotional states (e.g., happiness, sadness, surprise, disgust) and cultural backgrounds.

Preprocessing: Preprocessing steps involve preparing facial images for analysis by removing noise, standardizing lighting conditions, and aligning faces to a common coordinate system. Techniques like face detection and face alignment are employed to localize and normalize facial regions of interest (e.g., eyes, nose, mouth) within the images. Feature Extraction: Facial features are extracted from preprocessed images to quantify facial expressions. This typically involves detecting and analyzing key facial landmarks) Common techniques for feature extraction include: Facial Landmark Detection: Identifying specific points on the face (e.g., eyes, nose, mouth) using algorithms like the Active Shape.

Speech Analysis:

Speech analysis is a fundamental component of music recommendation systems that leverage spoken language to infer emotional states, preferences, and contextual information. This section outlines key methodologies and techniques involved in speech analysis for enhancing music recommendations: Speech Data Collection: Speech analysis begins with the collection of spoken language data, typically obtained through microphone recordings during user interactions with music content. Diverse datasets encompassing various languages, accents, and emotional contexts are essential to train robust speech analysis models. Preprocessing: Preprocessing steps involve preparing raw speech data for analysis by removing noise, normalizing audio levels, and segmenting speech into meaningful units (eg. phonemes, words, sentences). Techniques such as audio filtering, normalization, and silence removal enhance the quality of speech data for subsequent analysis. Feature Extraction: Relevant features are

extracted from preprocessed spocch data to capture acoustic and linguistic characteristics indicative of emotional states and preferences. Commonly used features in speech analysis include Mel-Frequency Cepstral Coefficients (MFCCs): Represent spectral characteristics of speech signals, capturing information about vocal timbre and pitch. Pitch and Intensity Quantify fundamental frequency (pitch) and energy (loudness) variations in speech, which are correlated with emotional expression.

Prosodic Features: Capture rhythmic and intonational aspects of speech, including speech rate, pauses, and pitch contour, which convey emotional nuances.

V. Applications in Music Recommendation:

Speech analysis provides valuable insights into user's emotional responses, preferences, and contextual cues related to music. In music recommendation systems, speech-derived emotional states and preferences can be integrated with music features (eg, genre, tempo, lyrics) to personalize music playlists. By leveraging speech analysis, recommendation algorithms can dynamically adapt music suggestions based on expressed emotions and contextual information conveyed through spoken language.

VI. Future Scope:

Despite advancements, challenges remain in facial expression analysis, including robustness to variations in facial appearance, facial occlusions, and generalization across diverse populations. Future research directions may focus on multimodal fusion of facial expression data with other modalities (e.g., speech, physiological signals) to enhance the richness and accuracy of emotion-aware music recommendation systems.

In summary, facial expression analysis plays a pivotal role in emotion-aware music recommendation systems by enabling the capture and interpretation; emotional responses through facial movements and expressions. Leveraging computer vision techniques and machine learning algorithms, researchers can unlock new possibilities for personalized music experiences that remote authentically with users emotional and cognitive states.

VII. Implementation model:

Implementing a music recommendation system that integrates facial expression and speech analysis involves designing and deploying machine learning or deep learning models capable of leveraging multimodal inputs to generate personalized music recommendations aligned with emotional states and preferences. Here a comprehensive guide to model implementation for such a system: Data Representation: Represent facial expression and speech data in a format suitable for model training. This may involve encoding facial features (e.g., landmarks, emotion labels) and speech features (e.g., MFCCs, sentiment scores) into numerical representations. Ensure compatibility and consistency between data modalities (facial expression, speech) to facilitate multimodal fusion during model training. Model Architecture: Design a multimodal neural network architecture capable of processing both facial expression and speech inputs. Common architectures include: Early Fusion Models: Concatenate or combine feature vectors from different modalities (facial expression, speech) into a single input representation. Late Fusion Models: Process each modality separately using dedicated subnetworks and fuse their outputs at a higher level representation. Hybrid Models: Employ attention mechanisms, recurrent connections, or transformer architectures to learn complex interactions between multimodal inputs. Training Strategy: Define a training strategy to optimize the model parameters using labeled multimodal data (e.g., paired facial expression-speech- music data). Utilize appropriate loss functions (e.g, regression loss, classification loss) tailored to the specific recommendation task (e.g., emotion- aware music recommendation, preference prediction). Regularize the model using techniques like dropout, batch normalization, or L2 regularization to prevent overfitting and enhance generalization. Model Integration with Music Features:Integrate the multimodal model with music-related features (e.g., audio features, genre labels, artist metadata) to generate comprehensive user profiles. Use integration or content based filtering to combine user data with music products (e. g. songs, albums) based on user preferences and interests.

Real-time Inference:

Spotify: Optimize the model for real-time inference to support interactive music recommendation applications. Employ efficient computation techniques (eg, GPU acceleration, model quantization) to reduce latency and ensure responsiveness during inference. Evaluate the performance of the implemented music recommendation system using appropriate metrics (eg, accuracy, precision, recall, user satisfaction). Conduct validation experiments on held-out datasets or through user studies to assess the system effectiveness in generating emotionally intelligent music recommendations. Deployment and Scaling: Use training models in product development using flexible tools (eg. cloud computing, packaging) to support user inters actions. Implementation of monitoring and interrupts to track performance patterns and user interactions to optimize and improve performance. Ethical Considerations: Ensure ethical compliance and user privacy protection throughout the model implementation and deployment process. Implement data anonymization techniques, obtain informed consent from users, and adhere to applicable regulations (e.g., GDPR, CCPA) governing personal data usage.

VIII. Conclusion:

In conclusion, the research paper has explored the innovative intersection of music recommendation systems with facial expression and speech analysis, aiming to enhance user experiences by delivering personalized music content aligned with individual emotional states and preferences. The integration of multimodal signals from facial expressions and speech redefined the landscape of music recommendation, enabling more nuanced and emotionally intelligent interactions between users and music content. Throughout this study, we have investigated various methodologies, including data collection, facial expression analysis, speech analysis, and model implementation, to develop an emotion-aware music recommendation system. By leveraging computer vision techniques, natural language processing algorithms, and advanced machine learning models, we have demonstrated the feasibility and effectiveness of incorporating non-verbal and verbal cues into the recommendation process.

IX. References:

[1]. R. Ramanathan, R. Kumaran, R. R. Rohan, et al., &quot;An intelligent music player based on emotion recognition,&quot; in 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), pp.1-5,2017.

[2]. S. Gilda, H. Zafar, C. Soni, et al., &quot;Smart music player integrating facial emotion recognition and music mood recommendation,&quot; in 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp.154-158, 2017.

[3]. D. Ayata, Y. Yaslan, and M. E. Kamasak, &quot;Emotion based music recommendation system using wearable physiological sensors,&quot; IEEE trans. consum. electron., vol. 64, no. 2, pp. 196-203, 2018 [4]A. Alrihaili, A. Alsaedi, K. Albalawi, et al., &quot;Music recommender system for users based on emotion detection through facial features,&quot; in 12th International Conference on Developments in eSystems Engineering (DeSE), pp.1014- 1019,2019.

[5]. 1. J. Goodfellow, D. Erhan, Y. Bengiou, et al., &quot;Challenges in representation learning: A report on three machine learning contests,&quot; in Neural Information Processing, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 117-124 2013.

[6]. J.S.Preema Rajashree, M.Sahana, et al.,&quot; Review on facial expressionbased music player,&quot; International Journal of Engineering Re search &amp; Technology (IJERT), vol. 6, no.15, 2018.

[7]. A. Guidel, B Sapkota, K. Sapkota, &quot;Music recommendation by facial analysis,&quot; 2020.

[8]. C.H. Sadhvika, G. Abigna, P. S. Reddy, &quot;Emotion-based music recommendation system, Sreenidhi Institute of Science and Technology,&quot; Yamnampet, Hyderabad; International Journal of Emerging Technologies and Innovative Research (JETIR), vol.7, no. 4, April 2020.

[9]. V. Tabora, Face detection using OpenCV with Haar Cascade Classifiers,&quot; Becominghuman.ai,2019.

[10]. Z. Qin, F. Yu, C. Liu,et al, &quot;How convolutional neural network see the world A survey of convolutional neural network visualization methods,&quot; 2018.

[11]. K. Chankuptarat, R. Sriwatanaworachai and S. Chotipant, &quot;EmotionBased Music Player,&quot; 5th International Conference on Engineering, Applied Sciences and Technology (ICEAST), pp. 1-4, 2019.

[12]. F. Norden and F.V.R. Marlevi, &quot;A Comparative Analysis of Machine Learning Algorithms in Binary Facial Expression Recognition, &quot;TRITAEECS-EX:143 pp.9.2019. [13]. P. Singhal, P. Singh, and A. Vidyarthi, &quot;Interpretation and localization of Thorax diseases using DCNN in Chest XRay, &quot;Journal of In formatics Electrical and Electronics Engineering, vol.1, no.1, pp.1-7,2020,

[14]. M. Vinny, P. Singh, &quot;Review on the Artificial Brain Technology: Blue Brain, &quot;Journal of Informatics Electrical and Electronics Engineer ing, vol.1, no.1, pp. 1-11,2020,