International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 02 Feb 2024, pp: 451-455 www.ijaem.net ISSN: 2395-5252

A Systematic Review on Auditory Parameters Analysis and Recommendation for Songs/Podcasts

Prof.Srushti Gunthe, Prasad Nalawade, Suchita Chavan, Rutuja Jadhav, Prajwal Korade

Department of Information Technology, SITS,Pune

Date of Submission: 20-02-2024 Date of Acceptance: 04-03-2024

ABSTRACT-
In recent times almost, everything around us is being automated in some way just to make life easier for us. With the use of ma-chine learning and deep learning we can help, respond and ease the tasks of human. The re-search generally focuses on recognition of au-dio parameters and recommending the best audio signals for songs and podcasts. The re-search genarally focuses on recognition of au-dio parameters and recommending the best audio signals for songs and podcasts. The voice recognition component utilizes MFCC to extract distinctive features from vocal in-puts, enabling precise identification and dif-ferentiation. KBFE, through K-means clus-tering, refines the feature set, optimizing the model's ability of noticing small differences in what users like. Our recommendation sys-tem, focused on smart learning like RNN, learns from how each user acts and changes over time. KNN helps by finding users who are alike, making the system better at guess-ing what users might like. GMM and HMM are like the brain of the system, figuring out patterns in lots of audio data. They help smoothly put together the recommendation system. These fancy algorithms make sure the system not only hears voices well but also gives suggestions that really match what users like. This project is a mix of high-tech learn-ing methods, making it easy for users to enjoy audio content that fits them. By combining MFCC, KBFE (knowledge based front end), RNN, KNN, GMM, HMM, and more, our project will recognize the audio parameters and will recommend which audio signals are suitable for songs and podcasts.).
Keywords-MFCC (Mel-Frequency Cepstral Coefficient), KNN, GMM, HMM, KBFE(Knowledge based front end)

I. INTRODUCTION

Emotion recognition in speech, speech separation, music recommendations, and machine learning al-gorithm comparisons are pivotal research areas with widespread applications in enhancing human-computer interactions, voice processing systems, and recommendation systems [9]. Understanding and processing emotions in speech is essential for improv-ing the responsiveness of computers to human feel-ings, thereby benefitting customer service, health-care, and more[1]. The use of Mel Frequency Cep-stral Coefficients (MFCCs) and deep learning mod-els like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have contributed to more accurate emotion recognition [8][3]. To further advance these systems, access to diverse databases of real-life emotional speech is cru-cial, and researchers are exploring the amalgamation of various emotion classifiers for superior results[2]. The future holds the promise of real-time applica-tions, such as empathetic chatbots that can sense and respond to human emotions naturally.

Speech separation through Sparse Non-Negative Ma-trix Factorization (SNMF) addresses the challenge of isolating multiple voices in a single recording, with potential implications for voice recognition sys-tems [10]. Researchers are utilizing techniques that dissect audio into individual components, comple-mented by deep learning methods like CNNs and Recurrent Neural Networks (RNNs) to enhance the clarity and naturalness of separated voices. These advancements are augmenting voice processing sys-tems across various domains [6]. In the realm of music recommendations, sophisticated

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 02 Feb 2024, pp: 451-455 www.ijaem.net ISSN: 2395-5252

methodolo-gies are employed to tailor song suggestions to in-dividual preferences. Music properties are analyzed using techniques like Shortest Time Fourier Trans-form (STFT), and dynamic K-means clustering is utilized to discern users' music tastes [3]. With the increasing availability of music online, real-time rec-ommendations have gained significance, particularly in the context of music streaming. The future en-visions systems that can amalgamate multiple rec-ommendation approaches to provide users with the ultimate musical experience [4].

In the context of machine learning algorithms, this survey delves into a comparative analysis of K-Nearest Neighbor (KNN), Genetic Algorithm (GA), Support Vector Machine (SVM), Decision Tree (DT), and Long Short Term Memory (LSTM) [12]. These algorithms function as diverse tools applica-ble across various domains, including fraud detection and image recognition. The objective is to ascertain the most suitable algorithm for specific tasks, of-ten relying on performance metrics[11]. Findings of-ten highlight the superior performance of algorithms such as LSTM and SVM. The future is expected to bring about smarter systems with the integration of AI into various tasks, optimizing their functionality and efficiency[7].

This survey paper comprehensively reviews these re-search areas, offering a detailed examination of the methodologies, techniques, and potential future de-velopments in emotion recognition, speech separa-tion, music recommendations, and machine learn-ing algorithm selection. Each section will provide an in-depth discussion, examining the significance of these domains and the research progress made within them.

The rest of the paper is organized as follows:Section 2 reviews literature survey of different research pa-pers, Section 3 provides an overview of methodolo-gies, Section 4 discuss challenges related to technolo-gies, Section 4 provides an overview of methodolo-gies, Section 5 reviews research directions, Section 6 concludes the paper.

II. LITERATURE REVIEW

Recognizing emotions in speech is crucial for appli-cations like making computers more responsive to our feelings, improving customer service, and health-care. Researchers often use a technique called Mel Frequency Cepstral Coefficients (MFCCs) to analyze speech features and deep learning methods, like Con-volutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, to make emotion recognition more accurate. To build better systems, they also need larger and more diverse databases of real-life emotional speech . A clever approach is to combine different emotion classifiers for better re-sults. The future looks promising for real-time appli-cations like chatbots that can sense emotions, mak-ing human-computer interactions more natural and empathetic. [2] [3].

Time-Frequency Matrix (TFM) presents a novel ap-proach for audio feature extraction and classifica-tion, with a particular focus on environ mental au-dio signals. These signals are challenging due to their non-stationarity and discontinuities. The pri-mary objective of the research is to develop a fea-ture extraction technique that effectively quanti-fies the non-stationarities in audio signals and im-proves the accuracy of audio classification. [4]. For speech separation using Sparse Non-Negative Matrix Factorization (SNMF), it's about separating multi-ple voices when there's only one recording. This can help improve things like voice recognition sys-tems. The researchers had used techniques that break down audio into its individual [10]. They had also added deep learning methods like Convolutional Neural Networks (CNNs) and Recurrent Neural Net-works (RNNs) to make the separated voices sound clearer and more natural. These advances are mak-ing voice processing systems better, which is help-ful in many areas [6]. An effective pitch estimation method using Canonical Correlation Analysis (CCA) for speech processing. The proposed method com-bines various techniques, including Dominant Har-monic Model (DHM), Empirical Mode Decomposi-tion (EMD), and Normalized Autocorrelation Func-tion (NACF) for accurate pitch estimation. It uti-lizes CCA to select Intrinsic Mode Function (IMF) components from reference signals, which are then used to reconstruct the signal, resulting in improved pitch estimation accuracy [1].

In the world of music recommendations, researchers are using fancy methods to suggest songs you might like. They analyze music properties using things like Shortest Time Fourier Transform (STFT) and figure out your music tastes with dynamic K-means clus-tering. It's all about making music recommendations more tailored to you. As more and more music is available online, real-time recommendations are be-coming important, especially when you're streaming music. The future might hold systems that can mix different recommendation methods to give you the best musical experience [9]. In the paper comparing machine learning algorithms like K-Nearest Neigh-

bor (KNN), Genetic Algorithm (GA), Support Vector Machine (SVM), Decision Tree (DT), and Long Short Term Memory (LSTM), it's about finding out which algorithm is the smartest for different jobs. These algorithms are like different tools, and they're used in various areas, from catching fraud to recog-nizing images [8]. Researchers want to know which one works best, and they used measurements to fig-ure it out. They find that algorithms like LSTM and SVM often give the best results. The future is all about making things smarter and using AI to do a lot of different tasks [12].

The middle-high frequency range, particularly be-tween 3,000 Hz and 4,000 Hz, contains key compo-nents that contribute to the glossiness of a profes-sional singer's voice. Through neural networks, the system can analyze the difference between an ama-teur singer's voice and a professional opera singer's voice. Parallel data is created by aligning the ama-teur and professional singing voices using phoneme information from lyrics [7]. When it comes to rec-ognizing emotions in speech, the review focuses on identifying different feelings people express when they speak. It's essential for making devices like virtual assistants more aware of how you feel. Researchers use features from speech, like Mel Frequency Cepstral Coefficients (MFCCs), and machines like Gaussian Mixture Models (GMM) and k-Nearest Neighbour (k-NN) to figure out these emo-tions [5]. They use fancy measurements like Gross Pitch Error to check how well the system does. They're always looking for more real-life emotional data to make their systems work better. The fu-ture could bring real-time apps that sense your emo-tions and make your interactions with computers and robots more human-like and caring [11].

The survey paper presents various methodologies and approaches employed in the reviewed literature, focusing on four key areas: Emotion Recognition in Speech, Speech Separation, Music Recommenda-tions, and Machine Learning Algorithm Comparison. Each section highlights the methods used to collect and analyze data in these domains.

III. METHODOLOGY

1. Mel-Frequency Cepstral Coefficients (MFCCs) Mel-Frequency Cepstral Coefficients (MFCCs) are a critical feature extraction technique used in voice recognition and automatic speech recognition (ASR) systems. MFCCs are particularly effective in repre-senting the spectral characteristics of audio signals, making them a valuable tool in preprocessing and analyzing spoken language.
2. Recurrent Neural Networks (RNNs) Recurrent Neural Networks (RNNs) are a class of deep learn-ing models commonly used in voice recognition sys-tems, particularly in the automatic speech recogni-tion (ASR) component. RNNs are well-suited for tasks that involve sequential data, making them a natural choice for processing audio signals, which are essentially sequential in nature. Here's how RNNs can be employed in a voice recognition system.
3. K-Nearest Neighbors (KNN) K-Nearest Neigh-bors (KNN) is not typically used as a primary al-gorithm for voice recognition systems, especially in automatic speech recognition (ASR) tasks. KNN is more commonly applied in classification and cluster-ing tasks. However, it can still play a role in certain aspects of voice recognition systems.
4. Hidden Markov Models (HMMs) Hidden Markov Models (HMMs) have historically played a crucial role in voice recognition systems, particularly in the domain of automatic speech recognition (ASR). HMMs are a statistical modeling technique that can be used to model both the acoustic and linguistic aspects of speech.

IV. CHALLENGES

The challenges in auditory parameters analysis and recommendation for songs and podcasts are multi-faceted. One significant challenge lies in personal-ization, as creating recommendation systems that truly understand individual tastes and moods re-mains complex. Maintaining user privacy and data ethics is another hurdle, ensuring that recommen-dations are accurate without compromising personal information. Content licensing and copyright issues add a layer of complexity, as respecting intellectual property rights while recommending content is essential. Scalability is also a challenge, especially for streaming platforms with vast libraries and millions of users. Integrating auditory analysis with emerging technologies like virtual and augmented reality re-quires addressing compatibility and user experience concerns. Lastly, providing adaptive recommenda-tions that evolve with changing user preferences in real-time remains an ongoing challenge.

Voice recognition and recommendation systems en-counter their own set of challenges. These in-clude improving accuracy in recognizing

emotions in speech and integrating real-life emotional data into systems. Striking the right balance between pro-viding personalized recommendations and safeguard-ing user data and privacy is a constant challenge. Overcoming copyright and licensing constraints in audio content poses a challenge, particularly when recommending songs and podcasts. The scalability of such systems to accommodate a large user base and extensive content libraries is also a significant obstacle. Ensuring compatibility and seamless in-tegration with various devices and platforms adds to the complexity. Lastly, keeping recommendations adaptive and responsive to evolving user preferences presents an ongoing challenge, demanding continu-ous research and innovation in the field.

V. RESEARCH DIRECTIONS

In the field of auditory parameter analysis and recommendation for songs and podcasts, several promising research directions are emerging.

1. Enhanced Personalization: Investigating more ad-vanced methods to tailor recommendations to indi-vidual preferences, considering factors beyond mu-sic genre and exploring mood-based or context-aware recommendations.
2. Data Privacy and Ethics: Research into safe-guarding user data and privacy while developing rec-ommendation systems, addressing ethical concerns, and ensuring responsible data usage.
3. Content Licensing and Copyright: Exploring solu-tions for content recommendation systems to respect copyright and intellectual property rights while pro-viding valuable recommendations.
4. Scalability: Developing methods to handle the vast and ever-growing volume of audio content avail-able online and making recommendations scalable for millions of users.
5. Integration with Emerging Technologies: Explor-ing how auditory analysis and recommendations can be integrated into emerging technologies like virtual and augmented reality for more immersive audio ex-periences.
6. Adaptive Recommendations: Researching ways to adapt recommendations in real-time as user prefer- ences evolve, offering more dynamic and up-to-date suggestions.

VI. CONCLUSION

The research we've explored shows great progress in how computers understand and interact with our voices. They're becoming better at recognizing our emotions, which can make computer interfaces more responsive and improve areas like customer service and healthcare. But to make these systems even better, we need bigger and more diverse databases of real-life emotional speech. We're also using advanced techniques like deep learning to make these systems more accurate and empathetic. In the field of mu-sic recommendations, we're personalizing music sug-gestions using techniques that analyze your music taste. This makes your music experience more tai-lored, especially when you're using online streaming services. Moreover, we're working on finding the best algorithms for various tasks. Some, like Long Short-Term Memory (LSTM) and Support Vector Machine (SVM), are proving to be very effective, promising a future where intelligent algorithms make various pro-cesses smarter and more efficient. All in all, these studies are leading us toward a future where tech-nology understands our emotions and preferences, making our interactions with computers more nat-ural and caring.

REFERENCES

[1]. Subrata Kumer Paul, Rakhi Rani Paule "Effective pitch estimation using Canonical Correlation Analysis " In 2020 Dept. of Computer Science and Engineering Univer-sity of Rajshahi, Rajshahi, Bangladesh, doi: 110.1109/ICAICT51780.2020.9333460. "Effec-tive pitch estimation using Canonical Correla-tion Analysis"

[2]. Zhandos Yessenbayev "Robust Segmentation of Speech Signal Using MFCC and Acoustic Parameters" In 2012 Department of Informa-tion Technology, L.N. Gumilev Eurasian Na-tional University, Astana, Kazakhstan DOI: 10.1109/AMS.2012.26. "Robust Segmentation of Speech Signal Using MFCC and Acoustic Pa-rameters "

[3]. Om Deshmukh, Carol Y. Espy-Wilson and Amit Juneja "ACOUS TIC - PHONETIC SPEECH PARAMETERS FOR SPEAKERINDE-PENDENT SPEECH RECOGNITION" In 2002 University of Maryland, Department of Electrical and Computer Engineering, A. V. Williams Bldg., College Park, MD 20752 DOI: 10.1109/ICASSP.2002.5743787.

”ACOUS- TIC -PHONETIC SPEECH PARAMETERS FOR SPEAKER-INDEPENDENT SPEECH RECOGNITION ”

[4]. Behnaz Ghoraani,Sridhar Krishnan “Time–Frequency Matrix FeatureExtrac-tion and Classification of Environmental Audio Signals” In 2011 Dept. of Elec-trical Computer and Biomedical Engi-

[5]. neering, Toronto Metropolitan University, TorontoDOI:10.1109/TASL.2011.2118753. ” Time–requency Matrix Feature Extraction and Classification of Environmental Audio Signals”

[6]. Hyan-Soo BaeHo-Jin Lee; Suk-Gyu Lee “Voice recognition based on adaptive MFCC and deep learning” In 2017 2016 IEEE 11th, fer-ence on Industrial Electronics and Applications (ICIEA) DOI: 10.1109/ICIEA.2016.7603830. ” Voice Recognition Based on Adaptive MFCC and Deep Learning ”

[7]. Ashesh Jain1,2, Amir R. Zamir2, Silvio Savarese2, and Ashutosh Saxena3 “Structural-RNN: Deep Learning on Spatio-Temporal Graphs” In 2016 Cornell University1, Stanford University2, Brain Of Things Inc.3 DOI: 10.1109/CVPR.2016.573 . “ Structural-RNN: Deep Learning on Spatio-Temporal Graphs”

[8]. Ryuka Nanzaka, Tsuyoshi Kitamura, Yuji Adachi, Kiyoto Tai, Tetsuya Takiguchi “Spec-trum Enhancement of Singing Voice Using Deep Learning” In 2018 IEEE International Symposium on Multimedia (ISM) DOI: 10.1109/ISM.2018.00-18. ”Spectrum Enhance-ment of Singing Voice Using Deep Learning”

[9]. Shiqing Zhang, Xiaoming Zhao and Qi Tian “Spontaneous Speech Emotion Recognition Using Multiscale Deep Convolutional LSTM” In 2019 nstitute of Intelligent Information Processing Taizhou University Taizhou, China. DOI: 10.1109/TAFFC.2019.2947464. “Spon-taneous Speech Emotion Recognition Using Multiscale Deep Convolutional LSTM”

[10]. Dong-Moon Kim, Kun-su Kim, Kyo-Hyun Park , Jee-Hyong Lee and Keon Myung Lee “A Music Recommendation System with a Dynamic K-means Clustering Algorithm” In 2007 Department of Electrical and Electronic Engineering, SungkyunKwan University, South Korea. DOI: 10.1109/ICMLA.2007.97. ”A Mu-sic Recommendation System with a Dynamic K-means Clustering Algorithm ”

[11]. Mikkel N. Schmidt and Rasmus K. Ols-son “Single-Channel Speech Separation us-ing Sparse Non-Negative Matrix Factoriza-tion ” In 2006 Informatics and Mathematical Modelling, Technical University of Denmark, doi: 10.21437/Interspeech.2006-655. ” Single-Channel Speech Separation using Sparse Non-Negative Matrix Factorization”

[12]. Rahul B. Lanjewar, Swarup Mathurkar, Nilesh Patel “Implementation and Comparison of Speech Emotion Recognition System using Gaussian Mixture Model(GMM) and K-Nearest Neighbor(KNN) techniques.” ”Implementation and Comparison of Speech Emotion Recognition System using Gaussian Mixture Model(GMM) and K-Nearest Neighbor(KNN) techniques. ”

[13]. Malti Bansal, Apoorva Goyal, Apoorva Choud-hari “A comparative Analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning” ”A compara-tive Analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in ma-chine learning. ”