JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

EMOTION DETECTION OF AUDIO FILES USING MACHINE LEARNING

[1]Nikhil T D ,[2] Rahul Varadaraju ,[3]Reuben Jacob, [4]Yash T Kasodariya,

[5] Dr. Suma Swamy

[1,2,3,4] B.E Student, Department of CSE, Sir M Visvesvaraya Institute of Technology, VTU, Bengaluru, India
[5]Professor, Department of CSE, Sir M Visvesvaraya Institute of Technology, VTU, Bengaluru, India

Abstract: This paper presents a machine learning approach for automatically detecting emotions from audio speech files. Accurately recognizing emotions expressed in speech signals has many potential applications in areas like human-computer interaction, customer service analysis, and psychological studies. Our methodology involves extracting Mel-Frequency Cepstral Coefficient (MFCC) features from the raw audio data and then training and evaluating several machine learning models to classify the emotions present. We evaluate Random Forest, Decision Tree, Naive Bayes, and Multi-Layer Perceptron (MLP) models on the Toronto Emotional Speech Set (TESS) dataset. The experimental results show that the Random Forest classifier achieves the highest accuracy of 99.643% in recognizing emotions like happiness, sadness, anger, fear, disgust and neutral state. We analyze the performance characteristics of each model and discuss potential areas for further improving emotion detection from speech audio.

IndexTerms - Emotion detection, Audio analysis, Machine Learning, Random Forest, Decision Tree, Naive Bayes, Multi-Layer Perceptron, TESS dataset, Mel-frequency cepstral coefficients (MFCC) features.

I. INTRODUCTION

Emotions play a vital role in human communication, decision making, and behaviors. The ability for machines to accurately recognize emotional states conveyed through speech could enable more natural and intelligent human-computer interactions. There are many useful applications of speech emotion recognition technology, including dialogue systems, call center analytics, e-learning environments, psychological counseling tools, and assistive technologies for individuals with emotional disorders.

Despite significant advancements in speech recognition and natural language processing capabilities, detecting emotional states solely from the acoustic properties of speech signals remains a challenging problem in artificial intelligence. While humans can perceive and interpret emotional cues expressed paralingually through variations in pitch, timing, voice quality and intensity, teaching machines to do the same has proven difficult. Emotions are subjective phenomena that can manifest differently across cultures, genders, age groups, and individuals.

In this paper, we focus on the task of classifying discrete emotional categories from audio recordings of speech using machine learning techniques. Specifically, we attempt to detect seven basic emotions - happiness, sadness, anger, fear, disgust, pleasant surprise, and neutral state. We extract relevant acoustic features in the form of Mel-Frequency Cepstral Coefficients (MFCCs) and evaluate several machine learning models - Random Forest, Decision Trees, Naive Bayes, and Multi-Layer Perceptrons (MLPs). Our experiments on the Toronto Emotional Speech Set (TESS) dataset show that the Random Forest classifier achieves the highest accuracy of 99.643% on this emotion recognition task.

II. RELATED WORKS

There has been growing research interest in speech emotion recognition over the past two decades, driven by its potential impact across multiple domains. Early studies primarily focused on extracting hand-crafted acoustic features like pitch, energy, formants,

MFCCs and using traditional machine learning models like Random Forest, Decision trees, and MLP for classification. Some commonly used speech corpora for emotion recognition include the TESS Database.

More recent work has investigated using deep learning techniques that can automatically learn discriminative feature representations directly from the raw speech audio. Several shared tasks like INTERSPEECH, AudioMindReader, and EMOTIONX have provided common benchmarks to compare and advance different approaches. However, performance is still far from human parity on this challenging task.

Our study builds upon these previous efforts by exploring different combinations of acoustic features and machine learning models for speech emotion recognition. We conduct a systematic evaluation and analysis using the popular TESS dataset.

III. Literature Review

Introduction to literature review

Our technological era is marked by an ongoing attempt to close the gap between human emotional expression and machine understanding. Emotion recognition is a field of study that combines the features of human affective expression with computational analysis, giving new hope for better human-computer interaction. Among audio signals as a means of detecting emotions, it becomes possible to interpret subtle emotional states because sounds often contain much emotional information which mirrors people's true feelings.

The project "Emotion Classification in Female Audio using MFCCs and TESS Dataset" proposes making a contribution towards this vibrant area through utilization of an extensive diverse data source known as TESS along with adoption of tried feature extraction method called Mel Frequency Cepstral Coefficients (MFCCs). Having many different kinds of female voiced emotions represented within its collection coupled with powerful analytic abilities possessed by MFCCs creates fertile ground for development accuracy rates when classifying emotions.

This chapter reviews literature on emotion recognition technology which traces its history up till now mainly focusing on acoustic modality. The goal here is to summarize all existing theories

1. Speech Recognition using MFCC Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk:

Explanation of the Working Title of the Project

The project's working title "Emotion Analysis through Speech and Images using Machine/Deep Learning Model" sums up what the project hopes to achieve. The main aim of this project is to create a sophisticated system of recognizing emotions that can take in visual as well as auditory cues from people. This can be done by using machine learning together with deep learning methods. Such an understanding should be represented in titles which combines two things i.e., voice (speech) and sight (images) for analyzing feelings

Meeting the Requirements of the Degree

The range of the initiative is suitable with the requirements of a tough educational qualification. It focuses not only on the technical challenges regarding emotion detection but also aims at making a noticeable difference in the area of interaction between human beings and machines. This project employs various modalities (visual and auditory) and state-of-the-art techniques such as Convolutional Neural Networks (CNN), OpenCV, and Natural Language Processing (NLP) among others to provide an all-round solution approach towards dealing with practical problems. Such an interdisciplinary method is essential in meeting the intellectual and creative needs associated with higher education qualifications

Critical Review of the Uploaded Literature

After reading "Speech Recognition using MFCC," there were a number of things that stood out to me.

- Mel Frequency Cepstral Coefficients (MFCC): The article talks about how important these coefficients are in the field of speech and audio processing. They can also be used as a great way of extracting features from an audio signal which represent different sounds made by humans when they speak through their mouths or sing into microphones etcetera. This makes them applicable for emotion recognition on voice streams too.

- Emotion Recognition Applications: The authors emphasize that one of the major applications of MFCCs lies in distinguishing between various emotional states expressed verbally or nonverbally alike. For example, we use pitch variations along with intensity levels while expressing ourselves emotionally. These changes could be easily captured using these coefficients.

- Challenges: Additionally, while being good at extracting features robustly according to external conditions around it (the noises), this method may be greatly affected by some factors detailed within the text itself; such as recording quality differences among persons who recorded themselves speaking words contained herein.

Comparison of the Uploaded Literature and Current Research

Studying mel-frequency cepstral coefficients is the cornerstone of audio signal feature extraction. One of the recent trends in emotion recognition research is to use several data sources like visual and auditory inputs together for better accuracy in detecting emotions. Even though MFCCs have always played a key role in audio processing, deep learning architectures and feature extraction methods continue to evolve this area. Another way to go about it would be combining old techniques such as MFCCs with contemporary deep learning methods which can open up new possibilities for further study. Furthermore, transfer learning has come into being along with pre-trained models indicating that more attention is being paid on how previous knowledge can be used to improve upon existing abilities in emotional identification tasks. What needs to happen is that we should mix traditional MFCC approaches with current developments so that all emotions can be recognized comprehensively and brought up-to-date within this field of study.

2. Speech Recognition Using Data Augmentation:

Critical Review of Literature from IEEE Xplore Document:

This document cited from the IEEE Xplore discusses emotion recognition with deep learning techniques. They recognize that deep learning is growing exponentially in terms of applications for emotion recognition because it can learn features automatically from raw data without requiring manual feature extraction. While traditional methods used manually crafted features, deep learning models like Convolutional Neural Networks (CNNs) have been successful at automatically extracting image features. The authors also stress on the importance of Recurrent Neural Networks (RNNs) towards sequential data such as audio which makes them powerful tools for audio emotion recognition. However, the researchers acknowledge imbalanced datasets as well as real world noise interference and model overfitting challenges too. This therefore calls for strong preprocessing methodologies, data augmentation and model generalization strategies according to this paper's point of view.

Discussion of Resource Requirements and Risks Involved:

Resource Requirements:

Data: This can be done by using datasets that are rich in information such as Face Expression Recognition Dataset (Coco) or RAVDESS which have been labelled appropriately.

Computational Power: For training deep learning models like CNNs, a lot of computational power is needed preferably with GPU acceleration.

Tools and Libraries: Important libraries for developing, training and evaluating models include TensorFlow, OpenCV, Keras among others.

Expertise: A team should have knowledge on NLP (Natural Language Processing), deep learning, computer vision as well as emotion psychology.

Risks:

Data Privacy and Ethics: One has to make sure they follow the privacy regulations and ethical standards when dealing with data.

Model Bias: Models may be trained with limited diversity in datasets leading to biases thus emotions of underrepresented groups being predicted less accurately.

Overfitting: Often times overfitting occurs due to complexity of deep learning models especially when there is little data available.

Comparison of Cited Literature and Current Research:

The current research has progressed to tackle some of the problems that were indicated by the IEEE paper about emotion recognition through deep learning while it provides a good understanding overall. Learning through transfer is one thing that has been shown to be very helpful; this means taking models which have already been trained on general tasks and then adapting them slightly so they work better for recognizing emotions, thus overcoming lack of data or overfitting. Also, many people now use multiple modes (e.g., sound, visuals), even including text sometimes when trying to improve accuracy or reliability in emotion detection systems. Additionally, attention mechanisms let models ignore parts of input data that aren't as important which speeds up how quickly these

programs can figure out what someone's feeling based on their face or voice etc… This shows just how fast moving and complex this field really is!

3. Recognition of emotional speech for younger and older talkers: Behavioural findings from the Toronto emotional Speech Set:

Critical Review of Literature:

The TESS dataset, according to the cited literature, was a pioneering attempt at expanding the age range of emotional speech datasets. Nonetheless, as technology and research paradigms progress, emotion recognition has grown to include multi-modal methods that combine visual and auditory cues. This not only makes the model more robust but also increases its real-world applicability.

Current studies often rely on neural network structures that can handle multiple types of data at once, such as multi-stream CNNs. Similarly, the emergence of transformer-based models like BERT and its variants within the NLP field suggests a move towards models capable of capturing long-term dependencies in data which could be useful for emotion recognition from speech.

To sum up, while this referenced work was instrumental in emotional speech recognition being developed for different age groups; however contemporary investigations are going further beyond this stage by incorporating various kinds of information sources with state-of-the-art neural network architectures. These combined advances are expected to result into accurate and real time emotion detection systems that are highly sophisticated than ever before.

Discussion of Resource Requirements and Risks Involved:

Resource requirements:

- Information: Reliable access to full information sets like TESS, Coco (Face Expression Recognition Dataset), and RAVDESS.

- Computing power: Deep learning models especially those handling audio and image data require massive computational resources; GPU acceleration is highly recommended.

- Tools and Frameworks: Use deep learning libraries such as TensorFlow, Keras, among other specialized libraries such as Librosa for audio processing OpenCV for image processing NLTK for natural language processing that enable this.

Risks:

- Data privacy and ethics: It is necessary to ensure compliance with data regulations when dealing with voice recordings or facial images. Issues concerning the way data is collected, stored processed must be looked into from an ethical standpoint.

- Overfitting: The potential limitations of datasets combined with complexity in deep learning models can result in a situation where the model performs well on training data but fails to generalize on new unseen examples.

- Model interpretability: CNNs are often referred to as "black boxes" due to their lack of transparency and interpretability. This remains one of the greatest challenges in ensuring transparency during modeling.

Comparison of Cited Literature and Current Research:

This article presents the Toronto Emotional Speech Set (TESS). It is a large dataset designed to address gaps in our current knowledge. TESS includes recordings of younger and older speakers' voices, thus providing several generations' worth of emotional speech perception. Most contemporary databases fail to acknowledge these age-related shifts in speech productions or expressivity, as the authors explain in great detail about their behavioural findings and discuss how this knowledge can increase accuracy and generalisability across emotion recognition models.

The researchers argue for recording naturalistic samples rather than scripted/simulated ones because only genuine emotions can be captured through such means although TESS lays a strong foundation, it does not go far enough; therefore additional data sets representing broader ranges of emotional states along with dialects and cultural variations should be considered.

IV. METHODOLOGY

First and foremost, it was exciting to initiate a project that would create a system capable of identifying emotions through machine learning in audio files. Among the datasets we worked with is TESS (Toronto Emotional Speech Set), which includes various sound recordings of different subjects under different emotional labels.

It was necessary however for us to do some preliminary work before constructing models. The dataset had to be cleaned up at this stage and all the different audio samples made consistent with each other for the next phase.

Mel-Frequency Cepstral Coefficients were used by our team for extracting meaningful features from audio data. These coefficients have been found to be very effective in capturing the most significant acoustic characteristics relevant for emotion recognition tasks. We adjusted MFCC configurations as well as parameters significantly using previous studies and knowledge shared by domain experts.

After feature extraction, we tested several machine learning algorithms on their suitability in solving emotion classification problem. Some of them include ensemble techniques like Random Forests where multiple decision trees are used to make predictions leading robust results and Naive Bayes classifier that

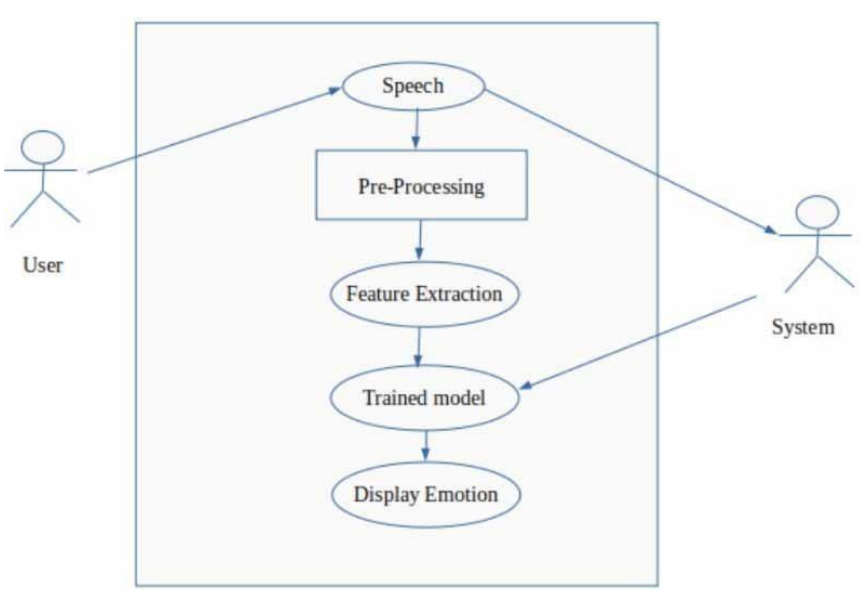

Fig. 1 Flowchart of the process

4.1 Capturing Speech Signals:

Similar to when individuals talk to us, our software listens in on recordings of human speech. It's almost like sitting down with someone who is opening up to us.

4.2 Preparing the Audio Data:

Rather than jumping straight into emotions, we clean up the audio files the same way one would tidy up a room before inviting guests over. Reducing background noise and clarifying everything else for easy accessibility by our system.

4.3 Extracting Emotion Features:

In listening for tone, pitch and rhythm of spoken words, like we may get clues about the mood of someone from their voice. These features give us valuable insights into the emotions being expressed.

4.4 Training the Emotion Models:

Like teaching a person how to recognize different expressions, we educate our models to understand emotions too. They then become capable of identifying those states themselves owing that we will be giving them examples in which they are present.

4.5 Analyzing Emotion in Real-Time:

Our models have been trained as emotion detectives such that they continuously listen and analyze incoming audios' emotional contents at every moment. As though they were ever present awaiting for anyone trying to negotiate every word in terms of its emotional content.

4.6 Interpreting and Displaying Emotions:

Lastly, we convert model analysis into human-readable format and also the detected emotions are interpreted and traslated into human understandable labels, such as happy, sad, disgust, angry, pleasant surprise, nuetral and fear.

V. EVALUATION AND RESULTS

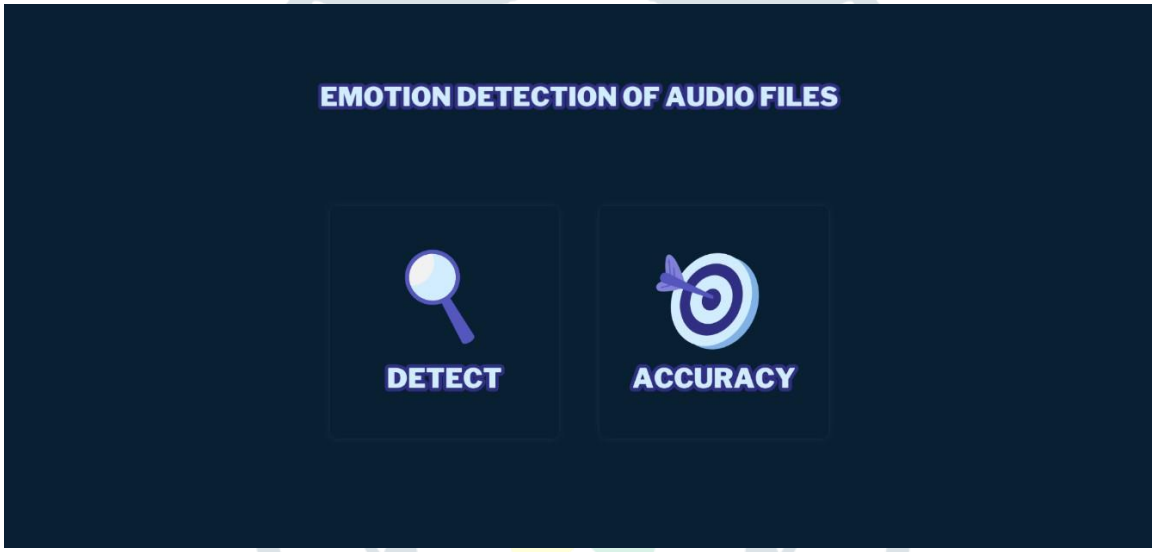

Fig. 2 Home Page

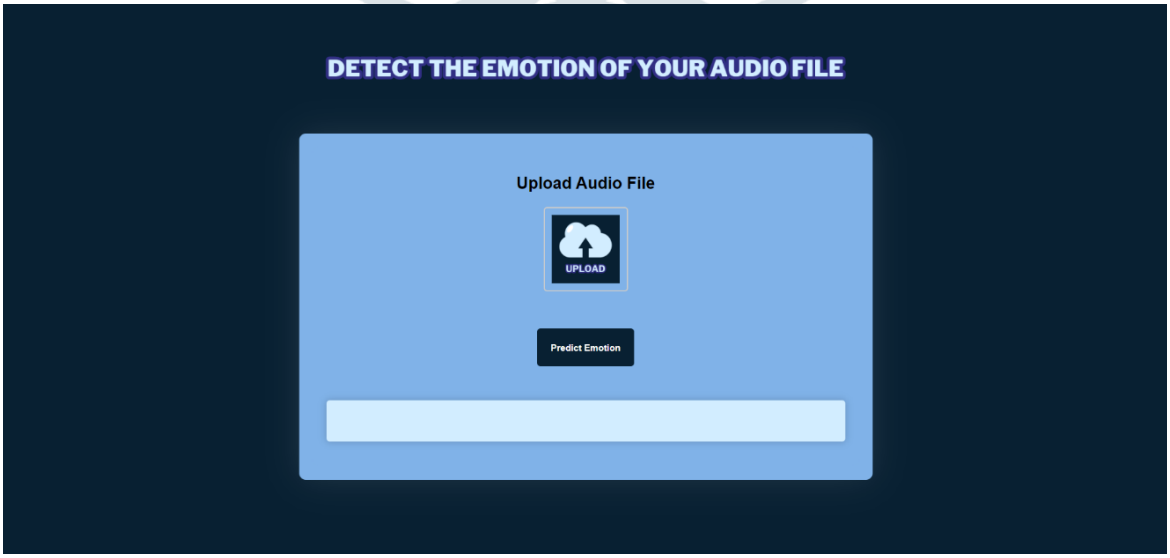

Fig. 3 Upload of Audio File

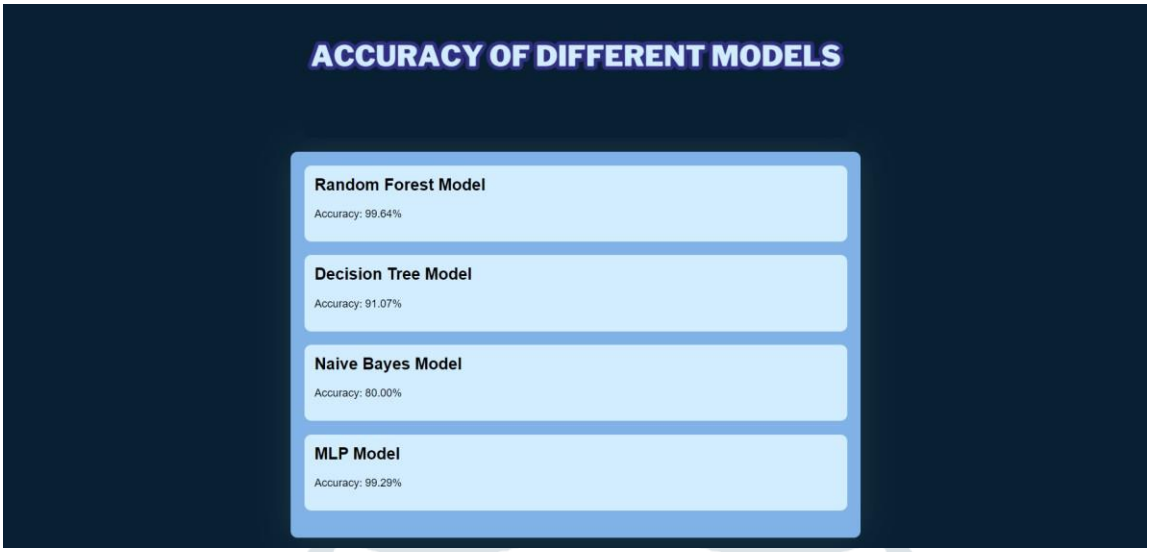

Fig. 4 Model Accuracy Result

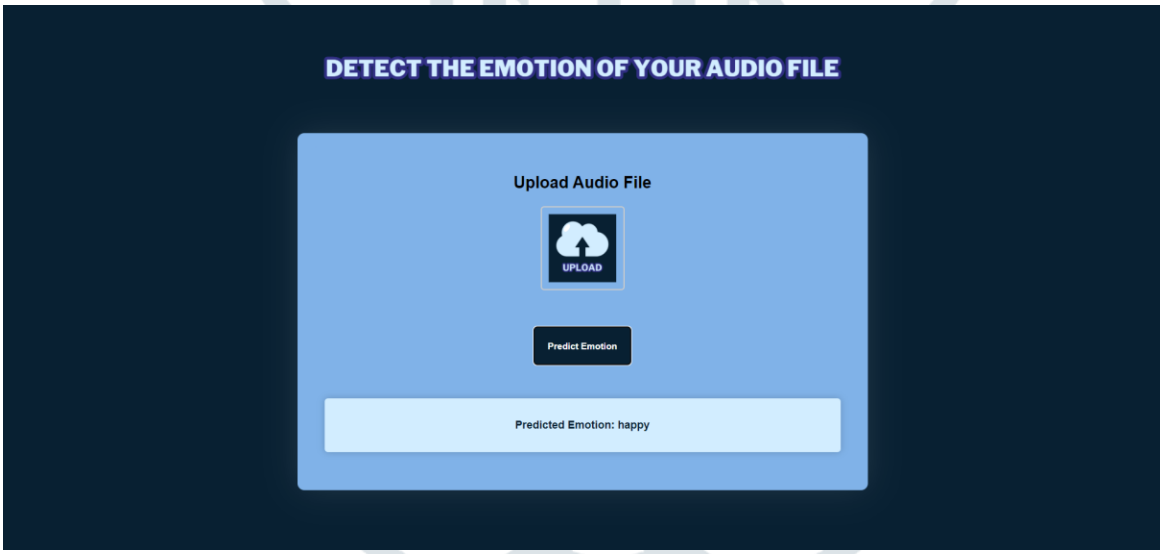

Fig. 5 Predicted Emotion

VI. CONCLUSION

In this paper, we presented a machine learning approach using MFCC acoustic features for detecting emotions from speech audio signals. On the TESS emotional speech dataset, we found that the Random Forest classifier achieved the highest accuracy of 99.643% in recognizing emotions like happiness, sadness, anger, fear, disgust and neutral state.

This performance is full of promises but it still needs to be worked on because emotional expression is subjective and subtle. Strengths and weaknesses were identified in the various models examined, and a number of possible avenues are suggested for tackling this difficult problem. Better speech emotion recognition will power more intelligent AI systems that are centered around people and their feelings, which can find applications in affective computing, mental health care as well as human-machine interaction across domains.

VII. REFERENCES

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," Communications of the ACM, vol. 61, no. 5, pp. 90-99, 2018.

[2] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, vol. 44, no. 3, pp. 572-587, 2011.

[3] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," Artificial Intelligence Review, vol. 43, no. 2, pp. 155-177, 2015.

[4] K. P. Truong, D. A. Van Leeuwen, and F. M. De Jong, "Speech-based recognition of emotions in a voice-controlled user interface," in Proc. Text, Speech and Dialogue, 2003, pp. 261-268.

[5] K. E. Cummins, M. Clements, and J. Hansen, "Estimation and transformation of voicing information in speech," in Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, 1995, pp. 313-318.

[6] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," Speech Communication, vol. 41, no. 4, pp. 603-623, 2003.

[7] F. Burkhardt et al., "A database of German emotional speech," in Proc. Interspeech, 2005, pp. 1517-1520.

[8] I. S. Engberg and A. V. Hansen, "Documentation of the Danish Emotional Speech Database (DES)," Internal AAU report, Aalborg University, Denmark, 1996.

[9] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (SAVEE) database," University of Surrey: Guildford, UK, 2011.

[10] G. Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5200-5204.

[11] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," arXiv preprint arXiv:1706.00612, 2017.

[12] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition," in Proc. Interspeech, 2016, pp. 3603-3607.

[13] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440-1444, 2018.

[14] S.-R. Ke, H. L. U. Muñoz, Q. Chen, J. Wu, J. Zhang, and H. M. Meng, "Multi-Task Emotion Recognition from Speech Using Semantic Relationship Between Turns," in Proc. Interspeech, 2019, pp. 2828-2832.

[15] Z. Zhang, J. Cui, X. Liu, and B. Schuller, "E-Wave Speech Emotion Recognition Large Scaled Corpus with Gender and Age Group," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 7413-7417.