SPEECH EMOTION RECOGNITION USING MACHINE LEARNING

[1]Arjun Gireesh, [2]Farhan Naseer Ahmed, [3]Rashad KP , [4]Jithin Ralf Justin, [5]Amuthavalli M

[1,2,3,4] Department of Computer Science and Engineering, Hindusthan College of Engineering and Technology Coimbatore, India,

20104011@hicet.ac.in, 20104024@hicet.ac.in, 20104048@hicet.ac.in, 20104033@hicet.ac.in

[5]Assistant Professor, Department of Computer Science and Engineering, Hindusthan College of Engineering and Technology Coimbatore, India,amuthavalli.cse@hicet.ac.in

Abstract

Emotion is a complex state of human being that depicts the physical, physiological or mental condition of a person. According to the human sciences, emotion is a mental process of neural mechanisms and disorders. Digital processing of speech signal is very important for high-speed and precise automatic voice recognition technology. Nowadays it is being used for health care, telephony military and people with disabilities therefore the digital signal processes such as Feature Extraction and Feature Matching are the latest issues for study of voice signal. In order to extract valuable information from the speech signal, make decisions on the process, and obtain results, the data needs to be manipulated and analyzed. Basic method used for extracting the features of the voice signal is to find the Mel frequency cepstral coefficients. Mel-frequency cepstral coefficients (MFCCs) is the coefficients that collectively represent the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. After calculating feature, neural networks are used to model the speech recognition. Based on the speech model the system decides whether or not the uttered speech matches what was prompted to utter

I. INTRODUCTION

Emotion is a complex state of human being that depicts the physical, physiological or mental condition of a person. According to the human sciences, emotion is a mental process of neural mechanisms and disorders. The expression of emotion may vary in linguistics as per the meaning of sounds. Numerous research activities have been carried out by researchers during the last few decades in understanding human emotion and for its effective recognition. It may help the society in many ways The speech emotion recognition (SER) system can serve as a tool for assessing the efficiency of a call center executive dealing with potential customers Intelligent on-line tutors can take clues from the emotional state of the students in delivering fruitful lectures by finding new methods of teaching or motivation Designing personal robots for entertainment or education purpose is another such significant application. The recognition system can alert
the security organization by identifying the stress from a phone conversation of security guards on duty. Interactive movies, and bio-medical engineering, etc. are few other vivid application domains that will be benefited from this type of study. A categorization of emotions into three broad areas such as primary, secondary and tertiary However, automatic SER is a complex research area because of the dynamism involved due to a varied degree of psychological perception to the speech emotion, which is further categorized into dimensional(continuous) or categorical.

II . LITERATURE REVIEW

i. Wu et al. proposed a new modulation spectral features (MSFs) human speech emotion recognition. Appropriate feature extracted from an auditory- spired long-term spectro-temporal by utilizing a modulation filterbank and an auditory filterbank for speech decomposition. This method obtained acoustic frequency and temporal modulation frequency components for convey important data which is missing from traditional short-term spectral features. For classification process, SVM with radial basis function (RBF) are adopted. Berlin and Vera am Mittag (VAM) are employed to evaluate MSFs. In experimental result, the MSFs display capable performance in comparison with MFCC and perceptual linear prediction coefficients (PLPC).

ii. Albornoz et al. investigate a new spectral feature in order to determine emotions and to characterize groups. Different classifier such as HMM, GMM and MLP have been evaluated with distinct configuration and input features to design a novel hierarchical techniques for classification of emotions. The innovation of the proposed method is two things, first the election of foremost performing features and second is employing of foremost class-wise classification performance of total features same as the classifier. Experimental result in Berlin dataset demonstrates the hierarchical approach achieves the better performance compare to best stand.

iii. Nicholson et al. (2000) explore the use of neural networks for emotion recognition in speech. Their paper likely details the network architecture, training process, and evaluation methods employed in their system. The research might focus on factors like accuracy, specific emotions recognized, and comparisons to existing techniques. While the paper predates modern deep learning advancements, it's valuable for understanding how researchers approached speech emotion recognition in the early 2000s.

iv. Nwe et al. (2003) investigate speech emotion recognition using Hidden Markov Models (HMMs) in their paper published in Speech Communication. HMMs are statistical models suited for sequential data, potentially making them effective for analyzing the flow of emotions in speech. The research likely explores feature extraction techniques for speech signals, along with the design of HMMs to represent different emotional states. The paper's focus might be on the evaluation methods used to assess the accuracy of this HMM-based emotion recognition system.

v. Cao, Verma, and Nenkova (2015) propose a novel approach to speech emotion recognition in their paper published in Computer Speech and Language. pen_spark expand_more Their method focuses on ranking emotions rather than directly classifying them.expand_more This ranking approach might be particularly useful for speaker-sensitive recognition, where the system adapts to individual vocal characteristics.expand_more

vi. Huang et al. (2019, in press) present a novel approach to depression detection in speech. Their work, likely published in the IEEE Journal of Selected Topics in Signal Processing, explores how natural language processing (NLP) techniques can be used to analyze not only the acoustic properties of speech but also specific speech events. This focus on landmark events within speech, combined with NLP, might offer new insights into identifying speech patterns indicative of depression.

vii. Xu et al. (2018) propose a method for speech emotion recognition that combines two machine learning techniques in their IEEE published paper. The approach leverages Subspace Learning, which aims to find a lower-dimensional representation of the speech data, and Extreme Learning Machines (ELMs), a type of artificial neural network known for its efficiency. This combination might help improve the recognition of emotions in speech compared to traditional methods.

III. PROBLEM AND EXISTING SYSTEM

- Accurately recognizing emotions in spoken language poses a significant challenge. Speaker variations, cultural influences on expression, and ever-present background noise make analysis difficult. Existing limitations also include a narrow range of detectable emotions and a lack of transparency in how these systems reach conclusions. Overcoming these hurdles is crucial for creating robust speech emotion recognition systems that can unlock a deeper understanding of human communication and foster innovative applications across various fields..

- **EXISTING SYSTEM**
    - **Gender Recognition (Preprocessing):**
        a. The system first identifies the speaker's gender using a reference database.
        b. It likely uses a statistical approach based on pitch to determine if the voice is male or female.
        c. This step might help improve the accuracy of emotion recognition since speech patterns differ between genders.
    - **Feature Extraction:**
        a. The speech sample is divided into small frames (16 milliseconds each).
        b. For each frame, Mel-Frequency Cepstral Coefficients (MFCCs) are calculated. MFCCs are a common feature extraction technique that captures the spectral characteristics of speech relevant to emotion recognition.
    - **Emotion Recognition:**
        a. The system maintains a reference database containing MFCCs of pre-recorded speech samples labeled with emotions (sad, anger, neutral, happy).

b. It compares the MFCCs of each analyzed frame with the MFCCs in the reference database.
c. Based on the distance between the analyzed frame's MFCCs and the reference database entries, the system classifies the frame's emotion (anger, happy, or neutral).

Overall, this system uses a two-step approach: identifying the speaker's gender and then recognizing emotions based on MFCC features and comparison with a reference database

IV. SYSTEM ARCHITECTURE

- **Five-Part System:** The human emotion recognition system comprises five components:

  - Input speech signal (raw speech data)
  - Pre-processing (cleansing and preparing speech data)
  - Feature extraction & selection (extracting relevant features like MFCCs)
  - Classification (categorizing emotions using features)
  - Emotion recognition (outputs identified emotions)

- **Data Collection:** A large speech dataset with emotional labels (anger, sadness, joy, etc.) needs to be gathered, ideally from real-world scenarios for better generalizability.

- **Pre-processing Techniques:** Software modules are developed to handle tasks like:

  - Noise reduction (removing background noise)
  - Silence removal (eliminating silent sections)

- **Feature Extraction with MFCCs:** Mel-Frequency Cepstral Coefficients (MFCCs) are a common technique used to extract relevant features from the pre-processed speech data. These features capture characteristics of the speech signal that are informative for emotion recognition, such as pitch variations and spectral energy distribution.

- **Machine Learning Model Selection:** A machine learning algorithm Neural Networks is chosen for classification. This model will be trained on the extracted features and corresponding emotion labels from the dataset.

- **Model Training and Optimization:** The chosen machine learning model is trained on the prepared data. This involves feeding the data into the model and adjusting its internal parameters to learn the relationship between features and emotions. Iterative optimization based on evaluation results might involve refining pre-processing steps, feature extraction algorithms, or the model itself to improve accuracy and robustness.

- **Model Evaluation with Metrics:** The performance of the trained model is assessed using a separate evaluation dataset. Metrics like accuracy, precision (identifying true positives), and recall (capturing all relevant emotions) are used to gauge the model's effectiveness.

- **Deployment Considerations:** Depending on the application, the emotion recognition system might need to be integrated with other software components (e.g., call center system).

- **User Interface and Maintenance:** A user interface (UI) is developed to upload the file and display the recognized emotions.

V. ARCHITECTURE DIAGRAM

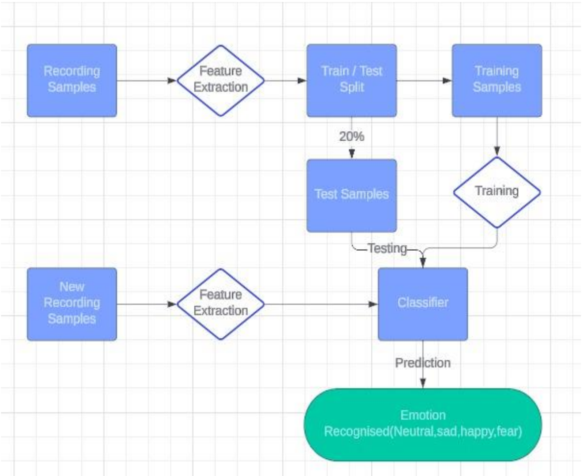


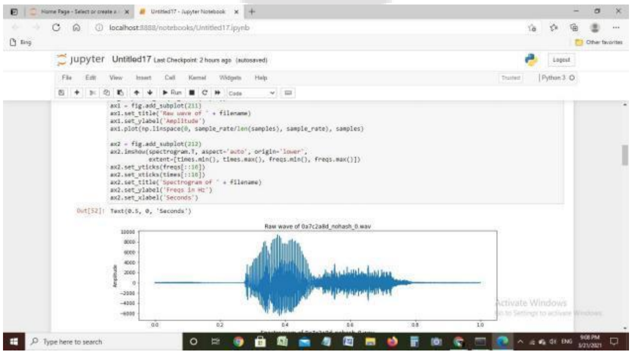
Fig. 2. Architecture Diagram of system.

VI. IMPLEMENTATION AND DEPLOYMENT

Bringing this speech emotion recognition system to life involves several steps. First, a large collection of real-world speech data labeled with emotions is required. This data undergoes pre-processing to remove noise and prepare it for analysis. Then, features like MFCCs are extracted to capture the emotional nuances within the speech. A machine learning model, like SVM or a Neural Network, is chosen and trained on this data, learning to associate features with specific emotions. The model's performance is evaluated to ensure accuracy, and any necessary adjustments are made to the pre-processing, feature extraction, or the model itself. Finally, for deployment, the system might be integrated with other software depending on the application (e.g., call center system). Real-time processing might be crucial, requiring optimization for speed. A user interface is developed to display the recognized emotions. Throughout the system's life cycle, performance monitoring and potential retraining with new data are essential for maintaining accuracy. Additionally, addressing privacy concerns and ensuring the explainability of the model's decisions are important for responsible deployment.

VII. RESULT AND DISCUSSIONS

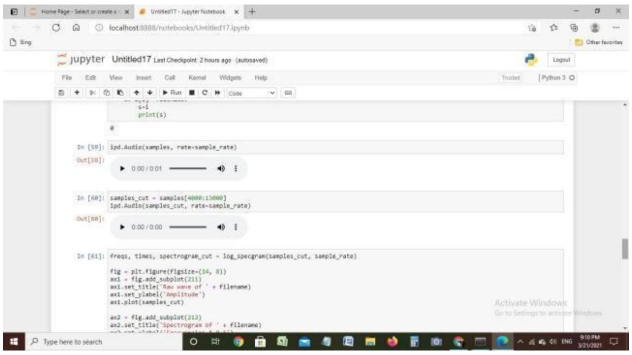
Advancements in machine learning have opened doors for machines to understand not just user commands but also the emotions behind them. This speech emotion detection system, leveraging machine learning to analyze speech audio, offers a range of applications. It can be used in call centers, virtual assistants, and linguistic research. Future enhancements can focus on improving efficiency, accuracy, and recognizing a wider range of emotions like depression and sarcasm detection, potentially creating valuable tools for therapy and human-computer interaction.


Home Page - Select or create a
Untitled17 - Jupyter Notebook
localhost:8888/notebooks/Untitled17.ipynb
jupyter Untitled17 Last Checkpoint: 2 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
Raw wave of 0a7c2a8d_nohash_0.wav
Amplitude
10000
8000
6000
4000
2000
0
−2000
−4000
−6000
0.0
0.2
0.4
0.6
0.8
1.0
Activate Windows
Go to Settings to activate Windows.
Type here to search
ENG
9:08 PM
3/21/2021

```
ax1 = fig.add_subplot(211)
ax1.set_title('Raw wave of ' + filename)
ax1.set_ylabel('Amplitude')
ax1.plot(np.linspace(0, sample_rate/len(samples), sample_rate), samples)

ax2 = fig.add_subplot(212)
ax2.imshow(spectrogram.T, aspect='auto', origin='lower',
           extent=[times.min(), times.max(), freqs.min(), freqs.max()])
ax2.set_yticks(freqs[::16])
ax2.set_xticks(times[::16])
ax2.set_title('Spectrogram of ' + filename)
ax2.set_ylabel('Freqs in Hz')
ax2.set_xlabel('Seconds')
```

```
Out[52]: Text(0.5, 0, 'Seconds')
```

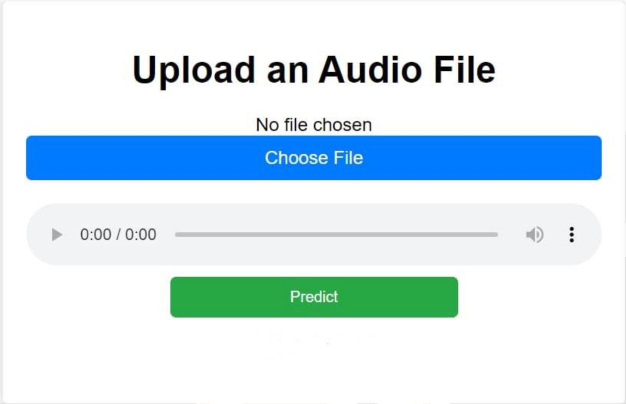
Fig. 1. RAW WAVE FORM



jupyter Untitled17 Last Checkpoint: 2 hours ago (autosaved)
localhost:8888/notebooks/Untitled17.ipynb
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code

```
        s=1
        print(s)
0

In [59]: ipd.Audio(samples, rate=sample_rate)
Out[59]:
         0:00 / 0:01

In [60]: samples_cut = samples[4000:13000]
         ipd.Audio(samples_cut, rate=sample_rate)
Out[60]:
         0:00 / 0:00

In [61]: freqs, times, spectrogram_cut = log_specgram(samples_cut, sample_rate)

         fig = plt.figure(figsize=(14, 8))
         ax1 = fig.add_subplot(211)
         ax1.set_title('Raw wave of ' + filename)
         ax1.set_ylabel('Amplitude')
         ax1.plot(samples_cut)

         ax2 = fig.add_subplot(212)
         ax2.set_title('Spectrogram of ' + filename)
```

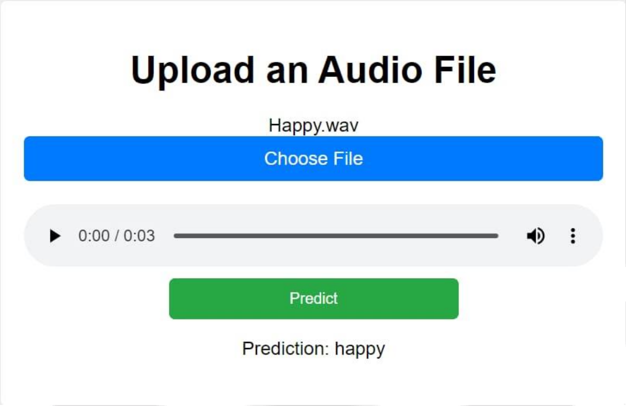
Fig. 3.SILENCE REMOVAL



Upload an Audio File

No file chosen

Fig. 4. UI TO UPLOAD THE FILE



Upload an Audio File

Happy.wav

Prediction: happy

Fig.5. PREDICTS THE EMOTION

VIII. CONCLUSIONS

The emerging growth and development in the field of AI and machine learning have led to the new era of automation. Most of these automated devices work based on voice commands from the user. Many advantages can be built over the existing systems f besides recognizing the words, the machines could comprehend the emotion of the speaker (user). Some applications of a speech emotion detection system are computer based tutorial applications, automated call center conversations, a diagnostic tool used for therapy and automatic translation system. we showed how we can leverage Machine learning to obtain the underlying emotion from speech audio data and some insights on the human expression of emotion through voice. This system can be employed in a variety of setups like Call Centre for complaints or marketing, in voice-based virtual assistants or chatbots, in linguistic research, etc.

IX. REFERENCES

[1]. Nicholson, J., Takahashi, K., & Nakatsu, R. (2000). Emotion recognition in speech using neural networks. Neural computing & applications, 9(4), 290-296.

[2]. Harár, P., Burget, R., &Dutta, M. K. (2017, February). Speech emotion recognition with deep learning. In Signal Processing and Integrated Networks (SPIN), 2017 4th International Conference on (pp. 137-140). IEEE.

[3]. Rawat, A., & Mishra, P. K. (2015). Emotion recognition through speech using neural network. Int. J, 5, 422-428.

[4]. Sun, W., Zhao, H., & Jin, Z. (2017).An efficient unconstrained facial expression recognition algorithm based on Stack Binarized Autoencoders and Binarized Neural Networks. Neuro computing, 267, 385-395

[5] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," Comput. Speech Lang., vol. 28, no. 1, pp. 186–202, Jan. 2015.

[6] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," Digit. Signal Process., vol. 22, no. 6, pp. 1154–1160, Dec. 2012.

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

[8] J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, May 2009.