**EMOTION RECOGNITION USING SPEECH**

**-ANANYA CHAKRABORTY**

**Collab link -** [**https://colab.research.google.com/drive/199ZLjBkipsLLFxZn1UMP0i3nbVW3xHXL?usp=sharing**](https://colab.research.google.com/drive/199ZLjBkipsLLFxZn1UMP0i3nbVW3xHXL?usp=sharing)

**Video link - https://drive.google.com/file/d/1AYnzoiC1Rs9fGKvYFSArE6Y-2KUYWOSf/view?usp=sharing**

**INTRODUCTION**

Speech emotion recognition is a technique that uses a computer to extract emotional features from speech signals and compares and analyzes the obtained characteristic parameters and emotional change. Finally, the speech and emotion legislation was completed, and speech emotional states were rated in accordance with the law. Speech emotion detection is now a popular research area in signal processing and pattern recognition, as well as a developing branch of artificial intelligence and artificial psychology. Human-computer interaction, interactive teaching, entertainment, security, and other sectors have all benefited from the study.

Voice signal acquisition, feature extraction, and emotion recognition were the three aspects of a speech emotion processing and recognition system. Figure 1 depicts the system framework.

Diagram

Description automatically generated

As voice user interfaces increase user experience and engagement, emotion recognition is becoming more important (VUIs). Speech-based emotion recognition systems are being developed for practical use. When the system is employed in actual applications, however, real-world background noise affects speech-based emotion recognition ability, offsetting some of these advantages.

SER (Voice Emotion Recognition) is a study topic that entails inferring emotion from voice data. It is one of the most challenging challenges in the field of speech signal analysis.

**What can we do with it?**

SER has extended into a multitude of disciplines throughout the years, despite its lack of recognition, including:

The capacity of a medical practitioner to discern what a patient is genuinely feeling when they are evaluated using mobile platforms in the medical industry may be beneficial in the healing process.

Customer service: In a call center, discussion may be used to evaluate call attendants' interactions with customers, which can assist improve service quality.

Recommender systems: These systems might be useful for recommending products to customers depending on their preferences.

SER has extended into a multitude of disciplines throughout the years, despite its lack of recognition, including:

When patients are examined via mobile platforms in the medical field, the ability of a medical specialist to determine what the patient is truly feeling might be helpful in the healing process.

Customer service: In a call center, discussion may be used to evaluate call attendants' interactions with customers, which can assist improve service quality.

Recommender systems: These systems might be useful for recommending products to customers depending on their preferences.

**ABOUT THE PROJECT**

**Dataset**

We utilized four datasets (including the custom dataset from this repository) that were previously downloaded and prepared in the data folder:

● RAVDESS is a 24 actor (12 male, 12 female) Ryson Audio-Visual Database of Emotional Speech and Song that vocalizes two lexically-matched phrases in a neutral North American accent.

● TESS stands for Toronto Emotional Speech Set, which was inspired by Northwestern University's Auditory Test No. 6. (NU-6; Tillman & Carhart, 1966). Two actors spoke a set of 200 target words in the carrier phrase "Say the word . (aged 26 and 64 years).

In 1997 and 1999, we recorded a database of emotional utterances said by actors as part of the DFG-funded research project SE462/3-1. The recordings were made in the Technical University Berlin's Technical Acoustics department's anechoic room. Prof. Dr. W. Sendlmeier of the Technical University of Berlin's Institute of Speech and Communication's department of communication science was the project's director. Felix Burkhardt, Miriam Kienast, Astrid Paeschke, and Benjamin Weiss were the key members of the project.

**Emotions available**

* "Neutral," "calm," "happy," "sad," "angry," "fear," "disgust," "ps" (pleasant surprise), and "boredom" are the nine emotions offered.
* Extraction of Features
* The essential component of a spoken emotion identification system is feature extraction. It is performed mostly by converting the voice waveform to a parametric representation at a lower data rate.
* We used the most often used aspects of the librosa library in this repository, including:
* MFCC\sChromagram\sMEL Frequency Spectrogram (mel)
* Tonnetz in contrast (tonal centroid features)

**Algorithms Used**

This repository can be used to build machine learning classifiers as well as regressors for the case of 3 emotions {'sad': 0, 'neutral': 1, 'happy': 2} and the case of 5 emotions {'angry': 1, 'sad': 2, 'neutral': 3, 'ps': 4, 'happy': 5}

**Classifiers**

* SVC
* RandomForestClassifier
* GradientBoostingClassifier
* KNeighborsClassifier
* MLPClassifier
* BaggingClassifier
* Recurrent Neural Networks (Keras)

**Regressors**

* SVR
* RandomForestRegressor
* GradientBoostingRegressor
* KNeighborsRegressor
* MLPRegressor
* BaggingRegressor
* Recurrent Neural Networks (Keras)

This is a histogram-

Chart, bar chart

Description automatically generated

**ABOUT SPEECH EMOTION RECOGNITION**

Speech Emotion Recognition is the task of recognizing the emotional aspects of speech regardless of its semantic content (SER). While individuals can do this well as a natural part of speech communication, the ability to do it automatically via programmable devices is still a work in progress.

Automatic emotion detection systems are being studied in order to create efficient, real-time algorithms for recognizing the emotions of mobile phone users, contact center operators and consumers, automotive drivers, pilots, and other human-machine interface users. Emotions have been highlighted as a crucial stage in obtaining a human-like look and behavior in robots.

Emotion-aware robots may be able to express emotional personalities and offer appropriate emotional replies. Computer-generated characters capable of delivering very authentic and compelling discussions by appealing to human emotions may be used to replace humans in specific scenarios. Machines must be able to understand speech-based emotions. Only with this ability can a completely meaningful conversation be held based on mutual human-machine trust and comprehension.

Traditionally, machine learning (ML) comprises extracting and computing feature parameters from raw data (e.g., speech, images, video, ECG, EEG). The attributes are utilized to train a model that learns to deliver the desired output labels. With this method, choosing attributes is a common problem.

It's not apparent which features contribute to the best successful data categorization into several categories in general (or classes). Testing a large number of different features, combining numerous features into a single feature vector, or experimenting with different feature selection methods may provide some results. The quality of the resulting hand-crafted features can have a significant influence on categorization performance.

By avoiding the obstacle of optimum feature selection, deep neural networks (DNN) classifiers have presented an elegant answer. The idea is to use an end-to-end network that takes raw data as input and returns a class label as output. It is not necessary to calculate handcrafted attributes or determine whether parameters are acceptable for classification. The network is in charge of everything.

In many cases, such as with SER, just a little quantity of data is available for training. The problem of insufficient training data may be handled to a large part utilizing a technique known as transfer learning, as described in this paper. It uses a pre-trained network on a large amount of data to answer a broad categorization issue. This network is then subsequently trained (fine-tuned) to do a more particular job using a small set of accessible data.

**SER in Real-Time**

Real-time speech processing demands a continuously flowing input signal, rapid processing, and consistent data output all within milliseconds of the time the processed data samples were formed.

The length of time it takes to generate the feature parameters impacts whether a SER technique can be implemented in real time. While the system training process is time-consuming, it is a one-time task that is usually conducted offline in order to produce a set of class models. These models can be kept and utilized at any time to perform classification on incoming voice sequences. The classification procedure includes the calculation of feature parameters as well as model-based inference of emotional class labels..

The accuracy of speech emotion recognition was directly influenced by the quality of feature extraction in this system. In the feature extraction method, the entire emotion phrase was usually used as a unit for feature extraction, and the extraction contents were four components of emotion speech: time construction, amplitude construction, fundamental frequency construction, and formant construction. Then, using these four elements to compare emotion speech to no emotion speech, obtain the law of emotional signal distribution, and classify emotion speech according to the law.

**Emotion Speech Database**

We must first input emotional speech signal before extracting speech emotion features. The Emotion Speech Library, which supplies standard speech for speech Speech emotion recognition is built on the foundation of emotion recognition.

There is a lot of literature on this issue right now, and single language emotion speech databases have been built all around the world in English, German, Spanish, and Chinese. A few speech libraries offer a wide range of languages. The goal of this paper is to discern speech emotion using Chinese language.

When picking experimental sentences, we must follow the following rules in order to create a flawless speech data sample library. (i)The sentence chosen must not contain any emotional tendencies, ensuring that the recorded statements will not influence the experimenter's decision. (ii)The sentence chosen has a lot of emotional latitude. That is, the sentence can communicate multiple emotions rather than just one; otherwise, we won't be able to compare the emotional speech parameters in the same emotional sentence across emotional states.

The construction of speech signal amplitude and the emotional condition of speech are also linked. The volume of speech is usually loud when the speaker is furious or happy. The level of speech is often low when the speaker is sad or unhappy. As a result, amplitude construction analysis of speech emotion features is more useful.

Figure 2 depicts a comparison of emotional and calm speech, with the average amplitude difference shown. Figure 2 illustrates that the amplitude of joy, rage, and surprise, three types of emotional speech, is larger than the amplitude of calm voice signal, however the amplitude of sad speech is smaller.

Chart, line chart

Description automatically generated

**Methodology**

Dataset

One of the key goals of this research is to create a real-time SER system that can cover a wide range of emotion recognition applications. There have been several models offered to categorize various sorts of emotions; one of the most extensively utilized in emotion recognition research is Paul Ekman's model, which divides emotions into six main categories: happiness, sorrow, fear, anger, disgust, and surprise. Because it is one of the only publicly available SER datasets annotated with the six emotion classes indicated above, our tests were based on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset.This dataset has been extensively utilized in comparable studies focused on emotion recognition from speech, allowing us to compare our findings to those of earlier studies. The RAVDESS dataset includes recordings of 24 professional actors (12 females and 12 males) vocalizing each emotion type at two different levels of emotional intensity. The dataset's emotional content was evaluated by 247 people, who scored each statement ten times by ten different raters.We used 1880 audio-only recordings (16 bit, 48 kHz.wav) to cover the six basic emotions in our tests. We looked at 376 occurrences of happiness, sorrow, wrath, and fear, respectively, and 192 instances of disgust and astonishment.

**Proposed Method**

Figure 1 depicts the general framework of the suggested technique. The feature extraction technique is divided into two parts: the first creates the MFCC feature set from the raw speech signals, and the second creates image features from the spectrograms. Mel-frequency cepstrum (MFC) is a type of cepstrum analysis that was created to replicate how the human auditory system analyzes sounds. Unlike traditional cepstrum analysis, which gives equal weight to all frequency ranges, MFC gives lower frequencies more weight.This is consistent with the essential bandwidths of the human ear's well-known reliance on frequency. Human hearing ranges from 20 Hz (lowest pitch) to 20 kHz (highest pitch), but its capacity to discern individual tones is significantly better at low frequencies. As a result, a human perceives the distance between 100 Hz and 200 Hz to be far more important than the gap between 10 kHz and 10.1 kHz, despite the fact that their pairwise lengths are the same.

Diagram, schematic

Description automatically generated

Graphical user interface, website

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Chart, timeline

Description automatically generated

**Conclusion**

In recent years, researchers in the United States and overseas have been paying careful attention to SER technology due to its ability to consistently recognize emotions and so improve the quality of human-computer interaction.

We present a deep learning approach with fused features for SER in this study. In terms of data processing, we used additive Gaussian white noise to quadruple-process the RAVDESS dataset containing 5760 audio samples (AWGN). To classify emotions into one of the eight categories, we built two concurrent convolutional neural networks (CNNs) to extract spatial features and a transform encoder network to extract temporal features. On the RAVDESS dataset's holdout test set, we achieved an accuracy of 80.46 percent using CNNs in spatial feature representation and sequence coding modification. The recognition of emotions by transforming speech into text paired with semantics is explored based on the outcomes.

Chart

Description automatically generated

Graphical user interface, website

Description automatically generated

**References**

[1] Wu S, Li F, Zhang P. Weighted Feature Fusion Based Emotional Recognition for Variable-length Speech using DNN[C]//2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC). IEEE, 2019: 674-679. (in USA).

[2] Anagnostopoulos C N, Iliou T, Giannoukos I. Features and classifiers for emotion recognition from speech: a survey from 2000 to 201 1[J]. Artificial Intelligence Review, 2015, 43(2): 155-177. (in NETHERLANDS).

[3] Nwe T L, Foo S W, De Silva L C. Speech emotion recognition using hidden Markov models[J]. Speech communication, 2003, 41(4): 603-623. (in NETHERLANDS).

[4] Bourlard H，Konig Y, Morgan N, et al. A new training algorithm for hybrid HMM/ ANN speech recognition systems[ C]/ / 1996 8th European Signal Processing Conference. Trieste, Italy: IEEE, 1996:1-4. (in Italy).

[5] Kipyatkova I. LSTM-based language models for very large vocabulary continuous Russian speech recogni tion [ M ]/ /Speech and Computer. Cham: Springer International Publishing, 2019: 219-226. (in Germany).

[6] Zhang Y Y，Du J，Wang Z R，et al. Attention based recognition [ C ]//2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Honolulu, USA: IEEE, 2018: 1771-1775. (in USA).

[7] Bystrom M, Modestino J W. Combined source-channel coding schemes for video transmission over an additive white Gaussian noise channel[J]. IEEE Journal on Selected Areas in Communications, 2000, 18(6): 880-890.(in USA).