BRAC
UNIVERSITY

Inspiring Excellence

CSE424
Pattern Recognition

Mini Research Project Partial Report

Title: Emotion Recognition From Human Speeches: Introducing the Three-Stage Pipeline

Submitted By:
20301305 MD Rishat Sheakh
20301326 Sartaj Emon Prattoy
20301367 Awon Bin Kamrul

Group: **26**
Section: 1, CSE424

INTRODUCTION

Speech emotion recognition (SER) has been increasingly popular among researchers for several decades and shows promise in various fields such as Human-Computer Interaction, Multimedia, and Biomedicine. Speech is a significant means of communication therefore a detailed analysis of speech signals is necessary. Emotion, apart from verbal content, is crucial in speech signals as it can significantly alter the interpretation of a sentence. SER holds potential in speech-enabled interfaces like AI voice assistants, which can monitor emotions to predict psychological changes or signs of mental stress and depression. Its applications extend to medical fields for detecting conditions like Autism and Parkinson's Disease, aiding in educational software for student mental health detection, and enhancing safety in automated vehicles by assessing driver speech for signs of impairment.

SER has a wide range of potential applications, such as helping AI voice assistants monitor emotions for predicting psychological shifts, aiding in medical diagnoses like detecting Autism and Parkinson's Disease, identifying mental health concerns in educational environments, and improving safety in automated vehicles by evaluating the emotional state of drivers.

The process of SER involves two essential phases: feature extraction and classification. Feature extraction includes temporal features like signal energy and spectral features obtained through Fourier Transforms. Deep learning, particularly 2D convolutional neural networks (CNN), has been successful in image classification tasks, leading to the proposal of using Mel Spectrogram features for SER.

MOTIVATION AND CONTRIBUTION

As modern technology is evolving day by day with all the automatons with gestures and voice, Speech Emotion Recognition (SER) plays a crucial role in the development of cutting-edge Human-Computer Interaction interfaces. From voice-enabled security devices and authentication systems to Automated Vehicle Environments, emotion can be analyzed to prevent identity mismatches or accidents. Many different fields can benefit from the classification of emotion from its audience, for example, the medical sciences field too can benefit from the classification of emotion from patients' speech for treating Parkinson’s disease and Autism to name a few.

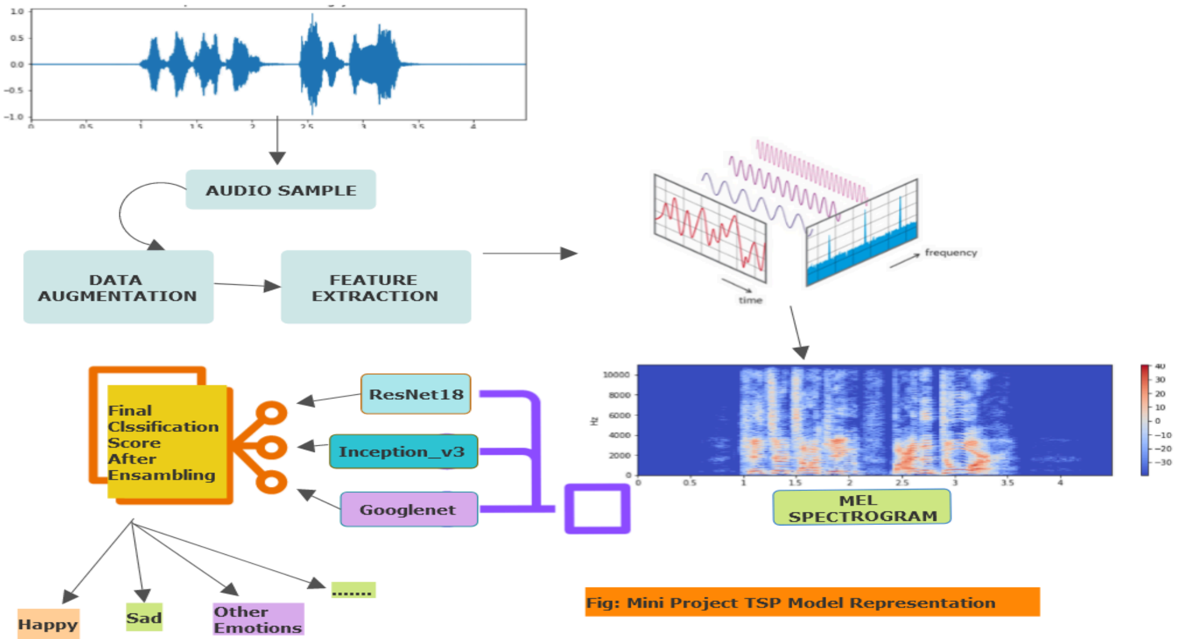


Fig: Mini Project TSP Model Representation

Key points of our proposed model:

Building an end-to-end SER model requires a substantial amount of data. Due to limited labeled audio data availability, we opted to incorporate three pre-trained transfer learning models—Inception_V3, GoogLeNet, and ResNet18—into our ensemble pipeline.

Training audio samples on a single classification model can lead to imbalance issues. The Ensemble approach gives an aggregate opinion to all the individual models thereby decreasing noise and giving better and unbiased prediction scores. This makes it a novel approach for SER.

A modified Gompertz function is employed to assign fuzzy ranks to the decision scores of the individual models. This approach ensures prediction scores rarely reach zero, as the Gompertz function exponentially saturates to an asymptote. This method, different from traditional ensemble pipelines, efficiently assigns adaptive priority weights to individual model prediction scores.

Comparative analysis demonstrates that our Three-Stage-model model outperforms other contemporary SER approaches, validating the novelty and effectiveness of our framework.

The Three-Stage-model model is trained and evaluated on three open-source datasets—SAVEE, CREMA-D, and RAVDESS— to achieve state-of-the-art accuracy compared to current methodologies.

LITERATURE REVIEW

Abdul-hadi et al. (2020) conducted a study on emotion recognition techniques using SVM classification based on facial expressions and speech signals. The goal was to Enhance emotion recognition in human-computer interaction systems by integrating information from both facial expressions and speech signals effectively. He used preprocessed video data, extracting features using histograms of oriented gradients (HOG) from facial images and speech signals, and training the SVM model to get the outcome with an accuracy was 85.72 % of e the emotion of crying and laughing by speaking.

singh et al. (2021) proposed a hierarchical deep learning approach for SER, integrating acoustic features and hierarchical DNNs to enhance emotion classification accuracy. The framework utilizes 33 acoustic features extracted from audio data, enabling comprehensive representation. By employing hierarchical DNNs, emotions are processed and classified hierarchically, potentially capturing intricate patterns. The superiority of this approach is evidenced by its outperformance of recent SER techniques. Experimental results demonstrate promising accuracies, with reported rates of 81.2%, 81.7%, and 74.5% on SAVEE, RAVDESS, and IEMOCAP datasets, respectively, showcasing the framework's robustness and applicability across domains.

The study on "Automatic Speech Emotion Recognition Using Support Vector Machine" highlights the importance of SER in Human-Computer Interaction. By analyzing the Berlin Emotional Speech Database, the research extracts features like energy, pitch, LPCC, MFCC, and LPCMCC, employing SVM-based classification. Combining these features significantly boosts recognition accuracy, with the best-achieved accuracy of 82.5% using energy, pitch, and LPCMCC. Future work aims to improve feature extraction techniques and explore real-time applicability for enhanced practical use of SER systems.

Luna-Jiménez et al. (2021) investigate the efficacy of multimodal emotion recognition systems by integrating speech and facial information. Methodologies involve employing transfer learning techniques, fine-tuning CNN-14 for speech emotion recognition, and utilizing STN for facial emotion recognition. The study achieves an accuracy of 80.08% on the RAVDESS dataset. However, limitations are noted, including potential challenges in generalizability across diverse demographic groups and conditions, emphasizing the necessity for further research to address these issues.

DATASET

Speech Emotion Recognition is a classification problem. The easiest sentiment analysis definition for speech recognition is that it is a classification task. Therefore, it can only be repeated with the quality and quantity of correctly labeled data. Emotion perception is subjective as different people can hear a sense of the same speech as different ones. Consequently, when we choose a dataset, interest is taken in those labeled by people of the same demographic as the base creators. We use simulated databases to mitigate the ways made dialect, allowing some actors and speakers to reproduce the expected feelings by performing prepared lines. In our Research, we have utilized the public SAVEE[] and RAVDESS[] datasets. These datasets are free of noise and have correctly tagged audio samples. The datasets comprise thorough tables on which data is tagged and how audio samples are particularly distributed across several emotional streams. The detailed description of these datasets is as follows:

RAVDESS

Ryerson Audio-Visual Database of Emotional Speech and Song(RAVDESS) contains 1440 files. There are 60 trials per actor and it has 24 actors. 12 of them are male and 12 of them are female. So 1440 files vocalize two lexically matched statements in a neutral North American accent. Speech emotions include calm, happy, sad, angry, fearful, surprised, and disgusted expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. The ratings were provided by 247 individuals for reliability assessment.[]

SAVEE

The Surrey Audio-Visual Expressed Emotion (SAVEE) database was recorded from 4 native male English speakers. The text material consisted of 15 TIMIT sentences per emotion. These include 3 common, 2 emotion-specific and 10 generic sentences that were different for each emotion and they were phonetically-balanced. There were seven emotions that were anger, fear, disgust, surprise, sad, happy, and neutral. The file in the dataset includes 480 audio samples in .wav format.

CREMA-D

Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) is a Diverse Dataset that includes 7,442 original clips from 91 actors. Varied demographics: 48 male, 43 female, spanning ages 20-74, multiple races/ethnicities.Emotion Representation: 12 sentences expressing 6 emotions at 4 intensity levels. Emotions include Anger, Disgust, Fear, Happy, Neutral, Sad. []

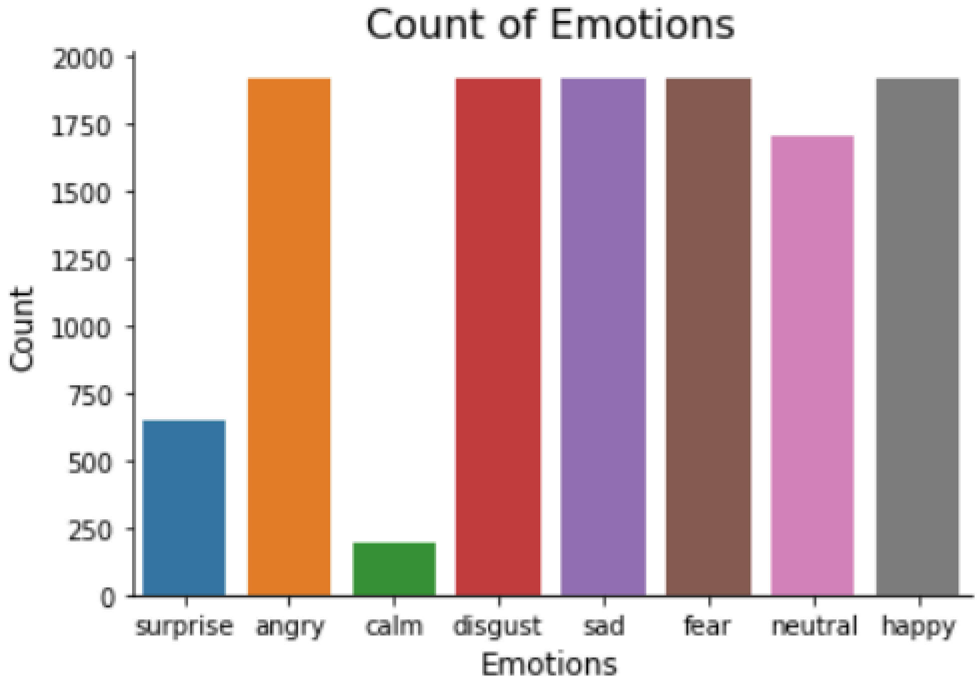
METHODOLOGY

In this project, we followed a certain framework using the datasets mentioned in the previous section. At the first stage, we began with data preparation. The later segments include data augmentation, followed by feature extraction. The extracted data then goes through our selected pre-trained model through a transferred learning process model.

DATA PREPARATION

The two different datasets that we are working with require a data frame for storing all the emotions of the data in a data frame with their paths. The data frame is used in feature extraction for model training.

The Data from all these datasets did go through Visualization and Exploration where each emotion count was plotted and wave plots and spectrograms for audio signals were shown.



DATA AUGMENTATION

Data augmentation is the process by which we create new synthetic data samples by adding small perturbations to our initial training set. For Audio data we can apply noise injection, shifting time, changing pitch and speed to generate syntactic data.

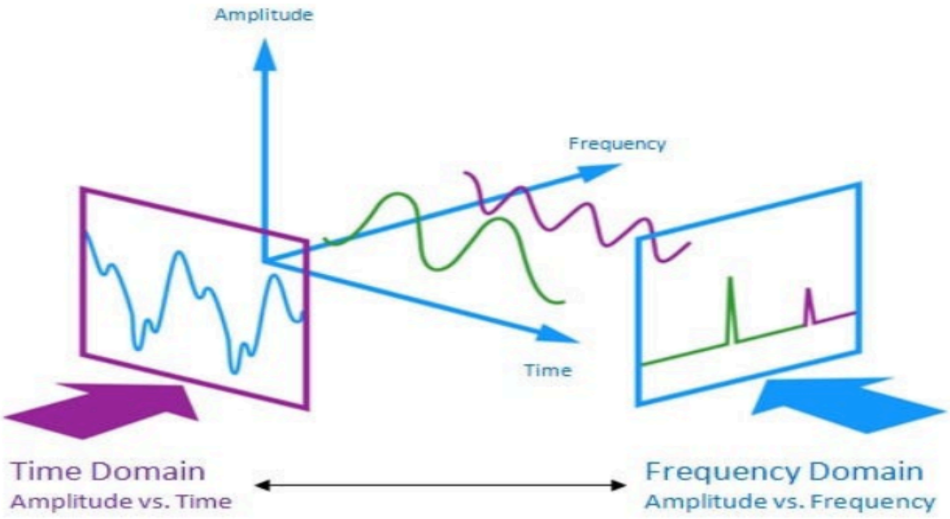
Various disproportionate data can cause problems in the prediction. Thus, to overcome these problems we have to use augmentation on the datasets that we are using. From the various augmentation techniques mentioned above, we are using shifting time with the equation:

$$Xnew = X [n \pm s]$$

Here, Xn is the audio signal and s is the number of samples shifted.

FEATURE EXTRACTION

A crucial step in analyzing and determining relationships between various things is feature extraction. Since the models cannot directly understand the audio data that is provided, we must transform it into a format that the models can understand. To do this, we employ feature extraction.



The audio signal is a three-dimensional signal in which three axes represent time, amplitude and frequency. There are various ways to do feature extraction some of which are Chroma Deviation, Spectral Centroid,Entropy of Energy, Zero Crossing Rate, Mel Spectrogram and so on.

In this project for feature extraction Mel Spectrogram was used as we are using the CNN model to train our datasets. Spectrograms are generated with the help of Fourier Transforms on sound signals. The sound signal is divided into small segments of time to which FT is applied individually. Thus, we get a frequency versus time graph. Human perception of frequency, however, is logarithmic rather than linear. This issue is resolved by the Mel scale, which maps a tone's perceived frequency to its measured frequency.

TRANSFER LEARNING

The three databases are divided in an 8:1:1 ratio into training, validation, and testing sets. We use the training and validation datasets to optimize our Mel spectrogram data transfer learning models. For the use in ensemble stage, the confidence scores for the testing split dataset are created for each of the three models and saved as a CSV file. We used three transfer learning models Namely GoogleNet, Inception_v3, ResNet18

Googlenet:

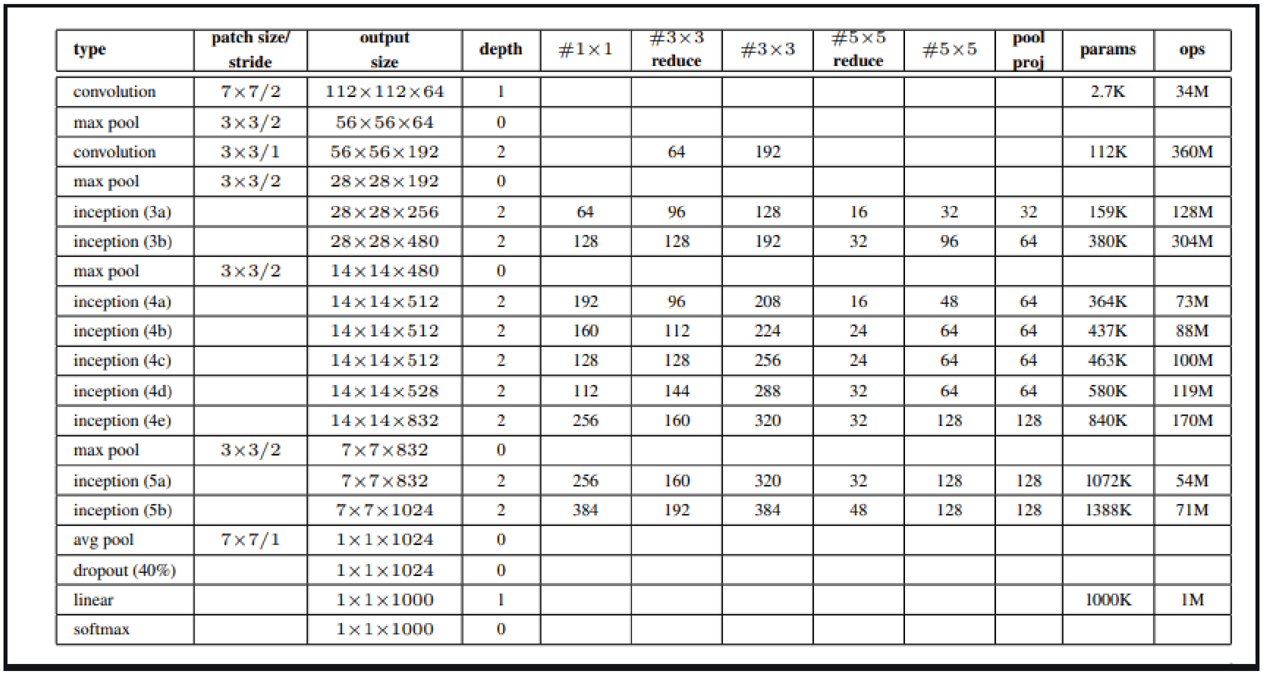
Google Net was proposed by research at Google (with the collaboration of various universities) in 2014 in the research paper titled "Going Deeper with Convolutions". The GoogLeNet architecture is very different from previous state-of-the-art architectures such as AlexNet and ZF-Net. It uses many different kinds of methods such as 1×1 convolution and global average pooling that enable it to create deeper architecture.

In Google architecture, there is a method called global average pooling is used at the end of the network. This layer takes a feature map of 7×7 and averages it to 1×1. This also decreases the number of trainable parameters to 0 and improves the top-1 accuracy by 0.6%

The inception module is different from previous architectures such as AlexNet, and ZF-Net. In this architecture, there is a fixed convolution size for each layer.

In the Inception, module 1×1, 3×3, 5×5 convolution and 3×3 max pooling are performed in a parallel way at the input and the output of these are stacked together to generate final output. The idea behind that convolution filters of different sizes will handle objects at multiple scales better.

Below is the layer by Layer architectural details of GoogLeNet.



type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

The overall architecture is 22 layers deep. The architecture was designed to keep computational efficiency in mind. The idea behind that the architecture can be run on individual devices even with low computational resources.

Inception_v3

Inception_v3 is a widely used 2D CNN model known for its strong classification performance through transfer learning. It's an extension of the GoogLeNet model, incorporating Batchnorm extensively in activation layers. Operating on input sizes of 299 × 299 × 3, it employs various convolution layers for feature extraction. The model's inception blocks enable diverse filter computation by concatenating them into a single feature map, reducing computational complexity by minimizing parameters.

ResNet18

ResNet, also known as Residual Network, was introduced in 2015 to tackle the issue of the "degradation problem" in deep networks by utilizing residual mapping. This architectural approach significantly improves the optimization process of CNN models. Similar to many other popular CNN models for image recognition, ResNet-18 is pre-trained on the ImageNet dataset, accepting input images of size 3 × 224 × 224, which is smaller than that of the Inception_v3 model. The performance of ResNet improves with increasing network depth. Variants of ResNet with different depths include ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-110. Among these, ResNet-18, utilized in our proposed three-stage-pipeline framework, strikes a balance between computational complexity and accuracy.

CONCLUSION

In conclusion, we expect that our paper will successfully evaluate the input data and based on that the different models such as GoogleNet, Inception_v3 and ResNet18 will provide output that will efficiently recognize the human spoken speech data.