SPEECH EMOTION RECOGNITION WITH GENDER PREDICTION USING CNN

BY:
Anil kumar Makkena

ABSTRACT

Emotions cannot be estimated with text alone as irony, humor, cadence etc. are important components of a speech. So, text has a limited capacity in conveying emotions. There is various work in sentiment analysis with text and speech but this project deals with the identification of mood in a non-discrete way. The traditional method of prediction emotion where labeling a certain number of discrete emotions is obviously a method of classification using various learning techniques. As an essential way of human emotional behavior understanding, speech emotion recognition (SER) has attracted a great deal of attention in human-centered signal processing. The key to speech emotion recognition is extraction of speech emotion features and the convolution neural network is used as the feature extractor to extract the speech emotion feature from the normalized spectrogram. Accuracy in SER heavily depends on finding good affect- related, discriminative features. In this paper, we propose to learn affect-salient features for SER using convolutional neural networks (CNN). The training of CNN involves two stages. In the first stage, unlabeled samples are used to learn local invariant features (LIF) using a variant of sparse auto-encoder (SAE) with reconstruction penalization. In the second step, LIF is used as the input to a feature extractor, salient discriminative feature analysis (SDFA), to learn affect-salient, discriminative features using a novel objective function that encourages feature saliency, orthogonality, and discrimination for SER. Our experimental results on benchmark datasets show that our approach leads to stable and robust recognition performance in complex scenes (e.g., with speaker and language variation, and environment distortion) and outperforms several well-established SER features.

CONTENTS:

1. Introduction...5

2. Literature...6

3. Methodology 7

4. Implementation 11

5. Experiment and Results 14

    5.1. Training Phase 14

    5.2. Test Phase 15

6. Final Web Application

7. Conclusion 16

        6.1 Applications and Future Work...17

8. REFERENCE...18

LIST OF FIGURES:

Figure 2: Prepared Dataset 7

Figure 3: CNN Summary 9

Figure 4: Feature Extraction 13

Figure 5: Input Frequency Spectrogram 14

Figure 6: Validating the Model 14

Figure 7: Web Application Results 16

1. INTRODUCTION:

In this field several systems are proposed for recognizing emotional state of human from speaker's voice or speech signal. Some universal emotions include anger, happiness, sadness, surprise, neutral, disgust, fearful, stressed etc. For the last two decades several intelligent systems are proposed by researchers. These different systems also differ by the nature of features used for classification of speech signals. Selection of features and size of the database plays important role for recognition scheme. The main challenge of emotion recognition from speech is that each speech is of different length, now MFCC feature extraction method works in a sliding window method that means it set a 25ms frame over the speech signal and compute 13 cepstral coefficient from each frame those are used as features. Now depending on various length MFCC return different number of frames. As a result, from each speech signal we have different number of features which is not acceptable. Therefore, we have done some preprocessing to make each speech signal of equal length. We have used CNN-LSTM architecture as our classification purpose, basically CNN is used for 2-dimensional input space there are some work where a spectrogram image generated from audio signal is used as input for CNN in our work, we have used one dimensional input space containing 39 features per frame as a input for CNN. We also use the output from CNN as an input of LSTM network.

A huge number of tasks dealt with by humans are influenced by emotional factors. Technology, deep learning, to be precise, has already done a great job in speech recognition. Speech being a primary medium to pass information, we humans can also understand the intensity and mood of the speaker by the speech data generated. This can also be applied for computers to understand the emotion based on what and how the speaker speaks. Music, on the other hand, can have a great application of this especially on the genre classification of music, where we can classify music based on similar emotional deliverance.

1.1 PROBLEMS:

The major problem in speech processing still lies in the selection of features. We have taken spectral MFCC coefficients as our feature. The second issue is in the variation of data while testing in open/wild. One of the reasons lies in the datasets available has audio data with a limited number of speakers and not to mention the constriction of limited accent, various gender pitches and the very humanly fact that everyone has their unique way of being happy or angry.

2.LITERATURE:

It mainly consists of 3 modules, including signal pre-processing, feature extraction, and classification. Feature extraction is an important module that provides the acoustic correlates of emotions in human speech for emotion classification. Although several papers have been published in the literature on finding suitable set of features for emotion recognition, there is still no conclusive evidence to show which set of features can provide the best recognition accuracy.

(2) The basic acoustic features extracted directly from the original speech signals, e.g. pitch and intensity related features, are widely used in speech emotion recognition. Some features derived from mathematical transformation of basic acoustic features, e.g. Mel-Frequency Cepstral Coefficients (MFCC) and Linear Prediction-based Cepstral Coefficients (LPCC) are also employed in some studies. However, not all these features that are taken or modified from those used for speech recognition purposes, are of equal importance for emotion recognition. This prompts the need of feature selection in emotion recognition. To work with a well-selected small feature set, the irrelevant information in the original feature set can be removed. The calculation complexity is reduced with a decreased dimensionality too.

(3) Several feature selection methods have been published in the literature, such as the combination of decision tree and random forest ensemble learning, forward selection, and genetic algorithm. However, one must conduct classification, and evaluate its accuracy repeatedly during selection procedure. To address this issue, a systematic method on feature selection for emotion recognition from speech signal is proposed in this paper. It adopts the idea of Canonical Correlation Analysis (CCA) to select a small set of features that are of most relevance to the emotions. For this purpose, the CCA is used to estimate the linear relationship between the various features and the emotional states. The features with relatively large canonical coefficients are selected and used in recognition. Experiments have been conducted using the LDC database and with the use of the Probabilistic Neural Network (PNN) as the classification method. The numerical results reveal that similar accuracies can

be obtained for the emotional states tested with the use of less than 30% of the features considered. It also brings advantage on the lower computational load required.

(4) Speech emotion recognition is a challenging yet important speech technology. It can be applied to broad areas, such as human-computer interaction, call center environment, and enhancement of speech and speaker recognition performance.

3.METHODOLOGY:

Discrete and dimensional categorization involves CNN model first for the discrete part which involves CNN with classical image classification architecture with a sequence of continuous dropout which is described later.

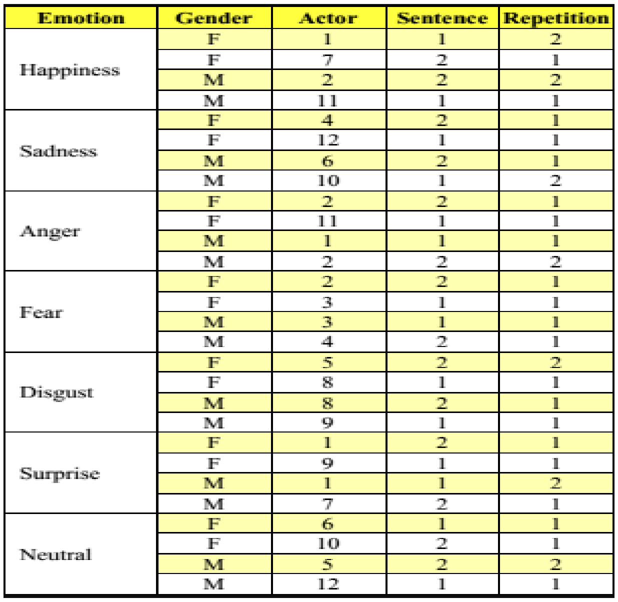
3.1 FEATURES

Spectral features have a sufficiently fine resolution in understanding frequency component of audio. We decided to go with MFCC: Mel-frequency cepstral coefficients. Speech from human is sound generated shaped by vocal tract, tongue etc. we decided to go for MFCC because it resembles the envelope generated in human vocal tract. Mel scale relates pitch/perceived frequency, of an exact tone to its original frequency. Humans are much better at sensing small changes in pitch at low frequencies than they are at high frequencies. MFCC which is on this scale makes our feature similar to closely what humans ear perceive.

3.2 DATASET

We choose RAVDESS dataset and it contains 7356 files which includes songs and speech. And speech consists of 1440 files. Each audio file was rated on a scale of 10 on intensity, emotional validity, and genuineness. The database contains 12 females, 12 male speakers, in a neutral American accent. The emotion in the speech includes surprise, happy, calm, fearful, sad, disgust and angry. The song section contains happy, calm, sad, angry, and fearful emotions. Every file has two levels of emotional intensity normal and strong.

Surrey Audio-Visual Expressed Emotion (SAVEE) is an acted English dataset. The SAVEE dataset consists of 480 British English utterances. The recordings were done by four male speakers DC, JE, JK, and KL. It consists of 15 sentences for each of the seven different emotions, anger, fear, happiness, disgust, sadness, surprise, and neutral. Each emotion has 60 utterances except neutral with 120 utterances. In this study, only four emotions, including anger, sadness, happy and neutral, were considered. Moreover, 240 utterances were taken for the evaluation of the experiment. The dataset was divided into 70% for training and 30% for testing. To the experiment, a database of emotional speech sampled at a minimum of 48 kHz with a resolution of at least 16 bits per sample was required. Also, the language had to be one of those spoken by our subjects, specifically French or English. Among the many resources currently available, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) Database met these requirements. The speech part of RAVDESS consists in two different English sentences (1: “Kids are talking by the door”, and 2: “Dogs are sitting by the door”) pronounced by 24 different actors (12 male, 12 female). Each sentence is uttered twice (repetition 1 and 2) in seven different emotions (calm, happiness, sadness, anger, fear, disgust, and surprise) and with two different intensities (normal and strong). They are also pronounced twice in an additional “neutral” emotion (obviously in only one intensity).



Emotion	Gender	Actor	Sentence	Repetition
Happiness	F	1	1	2
	F	7	2	1
	M	2	2	2
	M	11	1	1
Sadness	F	4	2	1
	F	12	1	1
	M	6	2	1
	M	10	1	2
Anger	F	2	2	1
	F	11	1	1
	M	1	1	1
	M	2	2	2
Fear	F	2	2	1
	F	3	1	1
	M	3	1	1
	M	4	2	1
Disgust	F	5	2	2
	F	8	1	1
	M	8	2	1
	M	9	1	1
Surprise	F	1	2	1
	F	9	1	1
	M	1	1	2
	M	7	2	1
Neutral	F	6	1	1
	F	10	2	1
	M	5	2	2
	M	12	1	1

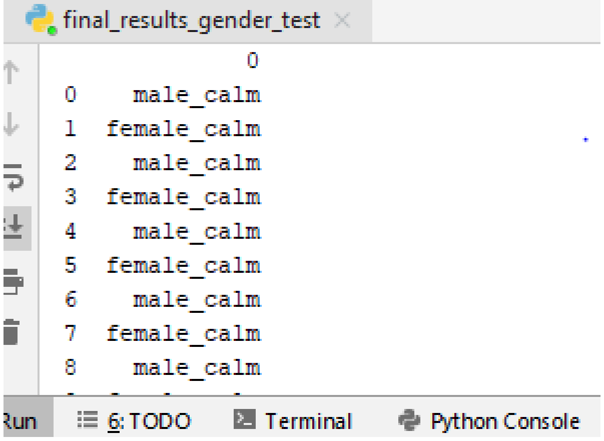
The steps we followed after selecting the datasets are:

Item selection: The “calm” emotion was deemed too close to the neutral one and was therefore not used in the test. Though considered by some theorists as a secondary emotion, disgust was kept because it can be slightly more difficult to detect compared to other emotions and thus might bring to light even moderate impairments. The “strong” intensity was considered too different from spontaneous emotions and was thus not used.

We then looked for a subset of items that was complete (including all actors) and as balanced as possible (one male and one female actor for both sentences). Items considered less accurately acted or with even minor recording issues (wrong level, background noise, abnormal sounds) were rejected.

Data Processing: RAVDESS dataset has the information of each audio's class in its name. we processed the audio and appended to its respective classes. Extracted MFCC floating values and transformed into feature vector and classes.

Dimensional Data Preparation: We used the landmark of SAVEE dataset and tagged value of valence and arousal for RAVDESS dataset resulting in a mapping like this which is used for one hot encoding of two 3-class vector (high/mid/low). The following calculation is rounded to the nearest integer (ceiling) and used just to extrapolate to a slightly larger scale. Original RAVDESS speech items are sampled at 48 kHz. The sequences of operations performed to limit their audio bandwidth to standard telephony bandwidths (full band, super wideband, wideband, narrowband MSIN and narrowband IRS) are illustrated in Fig. 1 and explained in the five following subsections. These sequences of operations are in line with what can be found in processing plans used to select and characterize speech coding standards. They also condition audio signals in much the same way as a speech codec would. In addition to being bandlimited, all test items are normalized at the P.56 level of −26 dBovl (dB relative to overload) and played at the 48 kHz sampling rate.



```
               0
0    male_calm
1  female_calm
2    male_calm
3  female_calm
4    male_calm
5  female_calm
6    male_calm
7  female_calm
8    male_calm
```

Designed CNN architecture: A merged deep CNN which consists of one 1D CNN branch and one 2D CNN branch is constructed in this paper. The 1D CNN is designed to learn deep features from audio clips, and the 2D CNN is built to learn deep features from log-mel spectrograms. CNN has many distinguishing excellences [21]. The properties of spatially local connectivity and shared weights allow CNN to perform the function of the learning filters. Furthermore, the input data of CNN needs relatively little pre-processing when compared with that of other deep architectures.

The task of this paper is to learn deep emotional features from different structural data to recognize speech emotion. The raw audio clip, which contains complete information, has begun to be used in a range of speech-related applications. Log-mel spectrogram, as a type of state-of-the-art spectral features, also has been widely used in recognizing speech emotion. So raw audio clips and log-mel spectrograms are adopted in our experiments to recognize speech emotion.

A CNN architecture can be built by stacking multiple different layers. The building blocks of CNN are some distinct types of layers such as a convolutional layer, pooling layer, ReLU layer, loss layer, LSTM layer, and fully connected layer. The convolution layer and the pooling layer are the core layers of CNN. The convolution layer plays the role of feature extractors and learns the local features which restrict the receptive fields of the hidden layers to be local. When a convolution kernel moves along the input to the convolution, it forms a feature map. Therefore, the number of features maps a convolutional layer has is equal to the number of convolution kernels. The pooling layer, which makes the features robust against noises and distortion, performs the non-linear down-sampling function and reduces the resolution of the features. The aim of building a merged CNN architecture is to learn high-level features from different-dimensional data using different-dimensional CNN branches. The designed architecture consists of one 1D CNN branch and one 2D CNN branch. The 1D CNN branch is used to learn deep features from single-dimensional data and the 2D CNN branch is adopted to learn high-level features from 2D data.

Hyperparameter Optimization: The aim of hyperparameter optimization is to choose a set of hyperparameters for a deep architecture, usually with the goal of optimizing the performance of the architecture on an independent dataset. The most common algorithms for hyperparameter optimization are grid search, random search, gradient-based optimization, and Bayesian optimization. In some experiments, Bayesian optimization has achieved a substantial increase in performance, compared with other methods. Hence, it is adopted in our experiments to select the hyperparameters of the proposed deep architectures.

Transfer Learning Approach: To decrease the training time, a transfer learning approach is adopted to train the merged deep CNN. This approach is always used to train a large target network without overfitting. When the features learned by a base network, namely the designed 1D CNN and 2D CNN in this paper, are transferred to the target network or the merged deep CNN, the common features do not need to be extracted by the target network again. So, the training time of the target network is reduced. The designed architectures are all trained on the same databases. Therefore, the training of merged CNN architecture can be speeded up by transfer learning approach. The merged CNN architecture used in our experiments has two branches, one is a 1D CNN branch, and the other one is a 2D CNN branch. The 1D CNN branch has some same layers as the designed 1D CNN. When the first *n* layers of the trained 1D CNN are copied to the 1D CNN branch of the merged CNN, the learned features of this 1D CNN are transferred to the 1D CNN branch of the merged CNN. So does the 2D CNN branch. After the completion of feature transfer, the merged CNN needs to be fine-tuned. The Bayesian optimization method is also adopted, which has been introduced in the previous section.

4.IMPLEMENTATION:

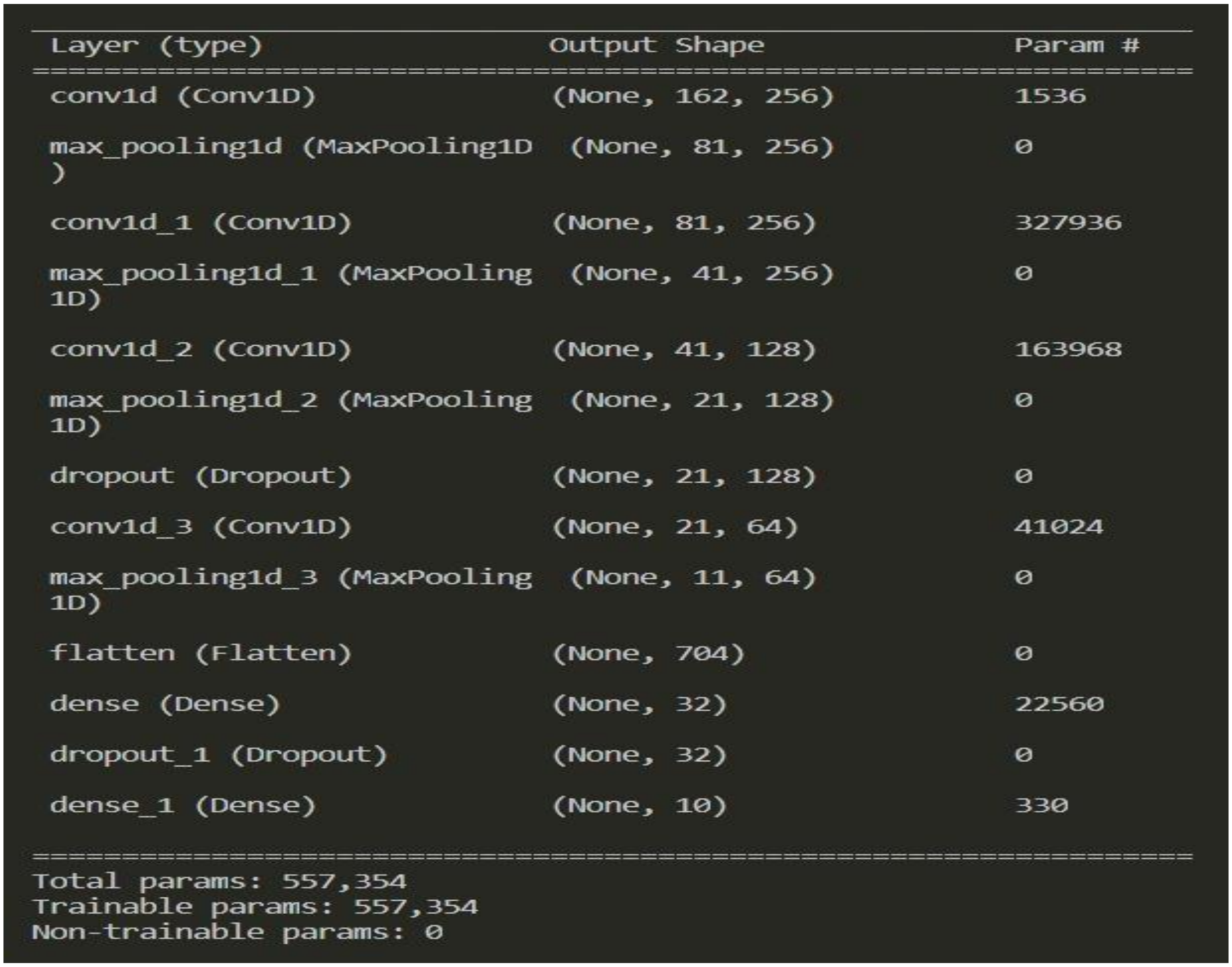
A classification system is an approach to set each speech to a proper emotion class according to the extracted features from speech. There are different classifiers available for emotion recognition. We have used one-dimension CNN with LSTM for classification.

4.1 CONVOLUTIONAL NEURAL NETWORK WITH GLOBAL POOLING

In this section, we focus on learning salient emotional representations from speech using the proposed GCNN that consists of convolutional layers given by one-dimensional temporal convolution filters together with global pooling layers given by global *k*-max pooling. The architecture of the resulting GCNN. Before we proceed to depict the network in detail a brief description of input features to the network is presented. The proposed framework endeavors to use a discriminative Convolution Neural Networks for feature learning schemes utilizing spectrograms produced from speech signals. The stride CNN architecture will have input layers, convolutional layers, a flatten layer, and fully connected layers followed by a SoftMax classifier. The spectrograms will be put into use as they tend to hold rich information. Also, extraction and application of such information when transforming the audio speech signal to text or phonemes are highly unlikely. This capability lets the spectrogram enhance the recognition of emotion. Therefore, the primary idea is to study high-level discriminative features from speech signals, making CNN architecture highly imperative.

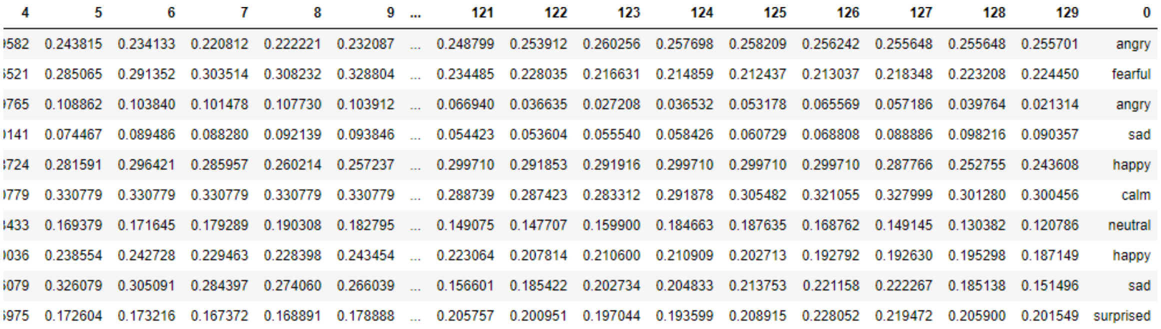
4.1.1 Spectrograms:

Spectrograms represent a signal quality across time at various frequencies present in the specific waveform. Its computation is based on the application of short-term Fourier transform (STFT) to the speech signal, which in turn forms the time-frequency representation. One of the difficult tasks in SER is the dimensioning of the signal using 2D CNN. Therefore, 1D representation of speech signal is modified into a suitable 2D representation for 2D CNN, since we intend to learn high-level features from speech signals using the CNN architecture. Spectrograms are utilized to represent the audios files present in the SAVEE speech emotion database. Sample of the extracted spectrograms of audio files from the SAVEE database depicting the angry, happy, sad, and neutral emotion by applying STFT.



```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 162, 256)          1536

 max_pooling1d (MaxPooling1D  (None, 81, 256)          0
 )

 conv1d_1 (Conv1D)           (None, 81, 256)           327936

 max_pooling1d_1 (MaxPooling  (None, 41, 256)          0
 1D)

 conv1d_2 (Conv1D)           (None, 41, 128)           163968

 max_pooling1d_2 (MaxPooling  (None, 21, 128)          0
 1D)

 dropout (Dropout)           (None, 21, 128)           0

 conv1d_3 (Conv1D)           (None, 21, 64)            41024

 max_pooling1d_3 (MaxPooling  (None, 11, 64)           0
 1D)

 flatten (Flatten)           (None, 704)               0

 dense (Dense)               (None, 32)                22560

 dropout_1 (Dropout)         (None, 32)                0

 dense_1 (Dense)             (None, 10)                330

=================================================================
Total params: 557,354
Trainable params: 557,354
Non-trainable params: 0
```

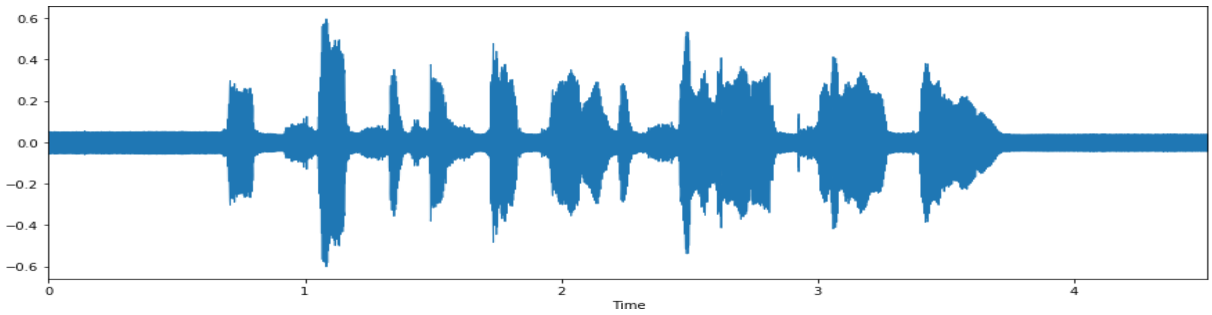
In this work, we first extract acoustic features at frame level with the OpenSmile toolkit which is widely utilized in SER community, including the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and the INTER SPEECH Challenges feature sets. Table I gives the details of these features, where the numbers between brackets indicate the dimensions of the feature. As a result, we get a 238-dimensional feature vector for each frame with the frame length of 60 milliseconds (ms) and frame shift of 10 milliseconds. Besides, for the extra experiments, we also extracted the feature vector with frame length of 30 ms, 90 ms, 120 ms, respectively. For each utterance, a sequence of feature vectors is obtained.



4	5	6	7	8	9	...	121	122	123	124	125	126	127	128	129	0
582	0.243815	0.234133	0.220812	0.222221	0.232087	...	0.248799	0.253912	0.260256	0.257698	0.258209	0.256242	0.255648	0.255648	0.255701	angry
521	0.285065	0.291352	0.303514	0.308232	0.328804	...	0.234485	0.228035	0.216631	0.214859	0.212437	0.213037	0.218348	0.223208	0.224450	fearful
765	0.108862	0.103840	0.101478	0.107730	0.103912	...	0.066940	0.036635	0.027208	0.036532	0.053178	0.065569	0.057186	0.039764	0.021314	angry
141	0.074467	0.089486	0.088280	0.092139	0.093846	...	0.054423	0.053604	0.055540	0.058426	0.060729	0.068808	0.088886	0.098216	0.090357	sad
724	0.281591	0.296421	0.285957	0.260214	0.257237	...	0.299710	0.291853	0.291916	0.299710	0.299710	0.299710	0.287766	0.252755	0.243608	happy
779	0.330779	0.330779	0.330779	0.330779	0.330779	...	0.288739	0.287423	0.283312	0.291878	0.305482	0.321055	0.327999	0.301280	0.300456	calm
433	0.169379	0.171645	0.179289	0.190308	0.182795	...	0.149075	0.147707	0.159900	0.184663	0.187635	0.168762	0.149145	0.130382	0.120786	neutral
036	0.238554	0.242728	0.229463	0.228398	0.243454	...	0.223064	0.207814	0.210600	0.210909	0.202713	0.192792	0.192630	0.195298	0.187149	happy
079	0.326079	0.305091	0.284397	0.274060	0.266039	...	0.156601	0.185422	0.202734	0.204833	0.213753	0.221158	0.222267	0.185138	0.151496	sad
975	0.172604	0.173216	0.167372	0.168891	0.178888	...	0.205757	0.200951	0.197044	0.193599	0.208915	0.228052	0.219472	0.205900	0.201549	surprised

4.3 PREPROCESSING:

Before using MFCC we make some preprocessing on the data set. All the speech files are with. wav extension, first we compute amplitude values of each file with a sample rate of 16000 sample per second. Then we take a weighted average according to the length of speech files and make them equal by adding zeros to the smaller file to make them equal to the average length file and crop all the larger file for the same purpose. After this process all files became of equal size.



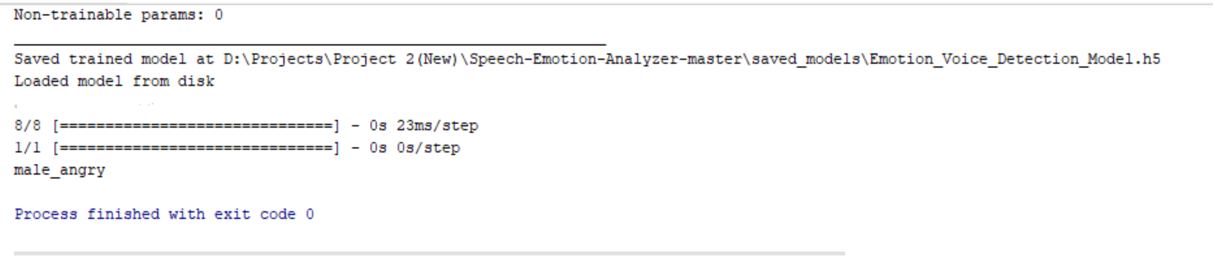
5. EXPERIMENT AND RESULTS:

At first, we divided the whole data set with 80% and 20% data. 80% data were used for training purpose and 20% data were used for validation purpose. After that we have computed the MFCC features with velocity and acceleration for each files of training dataset and test data set also. We provided those extracted features as initial input for convolution neural network. We use CNN with three convolution layers having 32, 16, 8 filter respectively. We have set 500 epochs for our network. We have used “adadelta” function as optimizer and “ReLU” as activation function. In LSTM network we have provided two hidden layer with 50 nodes in first layer and

20 nodes in second layer. We have used “SoftMax” as activation function for the final output nodes. We also used categorical cross entropy for computation of loss. After 100 epochs training accuracy reached at 96% and test accuracy reached at 77%.The merged CNN is evaluated on two different public emotional speech databases. The Berlin emotional database (Berlin EmoDB) and interactive emotional dyadic motion capture (IEMOCAP) database are selected to perform the experiments. Pre-determined sentences with the required emotions in the two selected databases are expressed by invited actors. All the experimental results are obtained on a graphics processing unit equipped with 4 GB of dedicated graphics memory.

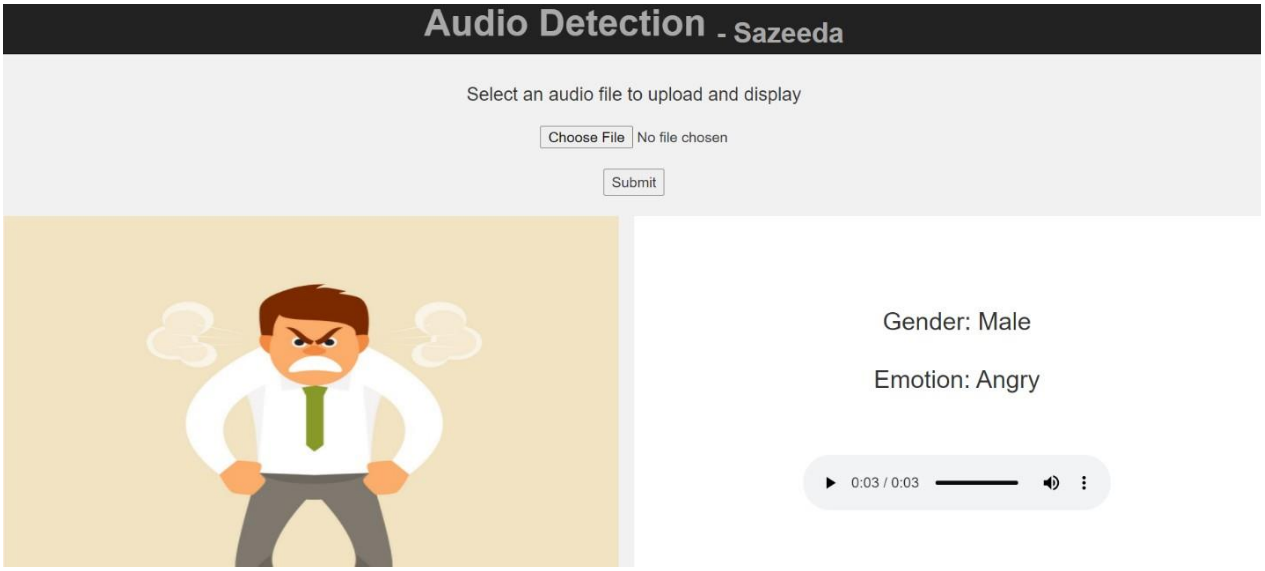
TRAINING PHASE: Test subjects are first asked to listen at least once to every test item, after these items have been processed in the baseline fullband condition. The goal of this phase is to familiarize subjects with the structure and content of test stimuli, and with the way emotions are acted. This phase typically takes less than 5 minutes to complete.At first glance, this training phase could introduce a bias in the study by making signals in the fullband condition easier to recognize for the subjects. But this corresponds to the real-life situation where training is done on acoustic, thus fullband, signals.

TEST PHASE: The test items processed in each of the five bandwidth conditions described in section 3, are presented in a randomized order to the subjects. These can listen to an item as many times as they want before taking a decision regarding the emotion they have recognized. This phase typically takes between 15 and 20 minutes to complete.



```
Non-trainable params: 0
_________________________________________________________________
Saved trained model at D:\Projects\Project 2(New)\Speech-Emotion-Analyzer-master\saved_models\Emotion_Voice_Detection_Model.h5
Loaded model from disk

8/8 [==============================] - 0s 23ms/step
1/1 [==============================] - 0s 0s/step
male_angry

Process finished with exit code 0
```

FINAL WEB APPLICATION: The model which is processed and built is exported into h5 file using FLASK application (Template) where we will be creating a web application. The file is read and the audio file is selected then the service call will take audio file as input and it sends the file for preprocessing, the processed output audio file is sent as input for .h5 file . The output from this web application will be shown in terms of whether the audio is male or female and also it shows the emotion of audio file with a image.



Audio Detection - Sazeeda

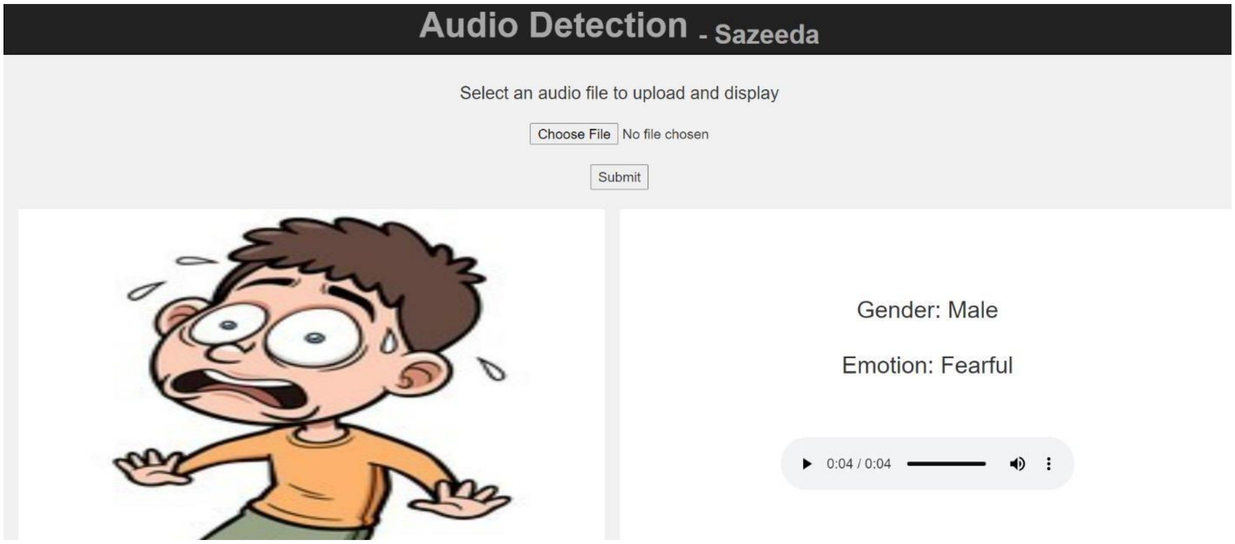
Select an audio file to upload and display

Choose File No file chosen

Submit

Gender: Male

Emotion: Angry



Audio Detection - Sazeeda

Select an audio file to upload and display

Choose File No file chosen

Submit

Gender: Male

Emotion: Fearful

7.APPLICATIONS AND FUTURE WORK:

One of the areas where this work can be applied is in the classical problem of genre classification in music, as the primary task of music is to trigger various emotions in humans as well in some of the other species too, we can classify music based on similar strength and mood of triggering emotions and sectors where human-computer interaction is much required such as customer call center, school and universities, forensics, lie detection and in medicine.

our focus is on the recognition of prototypic expressions of several basic emotions based on displayed emotional utterances in laboratory settings. Subtle, continuous, and context-specific interpretations of affective utterances recorded in naturalistic and real-world settings are clearly more important and more difficult research problems. Feature learning, as an advanced technique to learn a transformation of raw inputs to a representation that can be effectively exploited by a classifier, is well-suited for addressing these challenges. In the future, we plan to extend the proposed method in this paper and evaluate its performance on naturalistic speech data.

Although the merged CNN has obtained high SER accuracy, there are many aspects worthy of further discussion and study:

1. In the process of obtaining a higher predictive model of SER, more new network architectures such as LSTM are also good choices to construct a merged deep network. Perhaps new objective function or optimizer can work and improve the network performance, too.
2. In experiments, we find that the merged CNN with transferred features has memory ability. This ability caused by transfer learning can influence the performance of the target network. It is also a hint of improving the performance of a deep network on a small target database by transferring general features learned from a large base network.

8.REFERENCES:

1. J. Rong, Y-P. P. Chen, M. Chowdhury, and G. Li, "Acoustic features extraction for emotion recognition," IEEE/ACIS international conference on computer and information science, vol. 11, no. 13, pp. 419-424, Jul. 2007.

2. W. Ser, L. Cen, Z. L. Yu, "A hybrid PNN-GMM classification scheme for speech emotion recognition", the I9 International Conference on Pattern Recognition (ICPR), 2008.

3. C. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," IEEE Transactions on Speech and Audio Processing, vol. 13, no. 2, pp293-303, March 2005.

4. Petrushin, V.A., "Emotion recognition in speech signal: experimental study, development, and application", in Proc. of ICSLP 2000, pp. 222-225, 2000.

5. Fernandez, R., Picard, R.W., "Classical and Novel Discriminant Features for Affect Recognition from Speech", in Proc. of INTERSPEECH 2005, pp. 1-4, Lisbon, Portugal, 2005.

6. Luengo, I., Navasm E., Hernaez, I., Sanchez, J., "Automatic Emotion Recognition using Prosodic Parameters", in Proc. of INTERSPEECH 2005, pp. 493-496, Lisbon, Portugal, 2005.

7. Jiang Haihua and Hu Bin, "Recognition of Mandarin Speech and Emotion Based on PCA and SVM [J]", Computer Science, vol. 42, no. 11, pp. 270-273, 2015.