Speech Emotion Recognition Using Deep Learning Models

Team 6

REVATHI BOOPATHI, NIKITHA BRAMADI, JANANI RAVI KUMAR, NEETU RASINGER BABU

DEPARTMENT OF APPLIED DATA SCIENCE, SAN JOSE STATE UNIVERSITY

DATA 255: DEEP LEARNING TECHNOLOGIES

DR. MOHAMMAD MASUM

MAY 9, 2023

Project Outline

MOTIVATION

RESEARCH QUESTION

DATA SETS

EXPERIMENTAL SETTING

PROPOSED ARCHITECTURE

EXPERIMENTAL RESULTS

CHALLENGES

FUTURE RESEARCH DISCUSSION

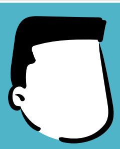
Objective and Motivation

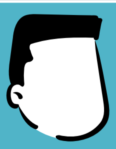
- The speech emotion recognition project aims to develop a deep learning model to recognize emotions from speech signals accurately
- Classify the emotional state of the speaker as one of the predefined emotions, such as **Anger, Disgust, Fear, Calm, Surprise, Happy, Neutral, and Sad**
- The potential applications include human-computer interaction, customer service, and mental health applications, where it could be used to diagnose and treat depression or anxiety conditions.
- The motivation is to develop an automated system to recognize emotions from speech audio.

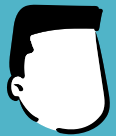
Research Question

- To implement traditional machine learning models and deep learning models to compare accuracy and performance on speech emotion recognition tasks
- To explore features, such as **deep features, statistical, and gender**, to be integrated apart from traditional speech features to improve emotion recognition accuracy.
- To observe the outcomes of imbalanced data, augmented data, and **SER Super Set data**

Datasets

RAVDESS		
Gender		
Actors	12	12
File Count	720	720
Sentences	2	

CREMA-D		
Gender		
Actors	48	43
File Count	3,930	3512
Sentences	12	

SAVEE	
Gender	
Actors	4
File Count	480
Sentences	15

TESS	
Gender	
Actors	2
File Count	2,800
Words	200

SER Super Set Data RAVDESS + CREMA + SAVEE + TESS	File Count	Actors	Female Actors	Male Actors	Sentences/Words	Emotions
	12,162	121	57	64	229	8

Data Processing and Analysis

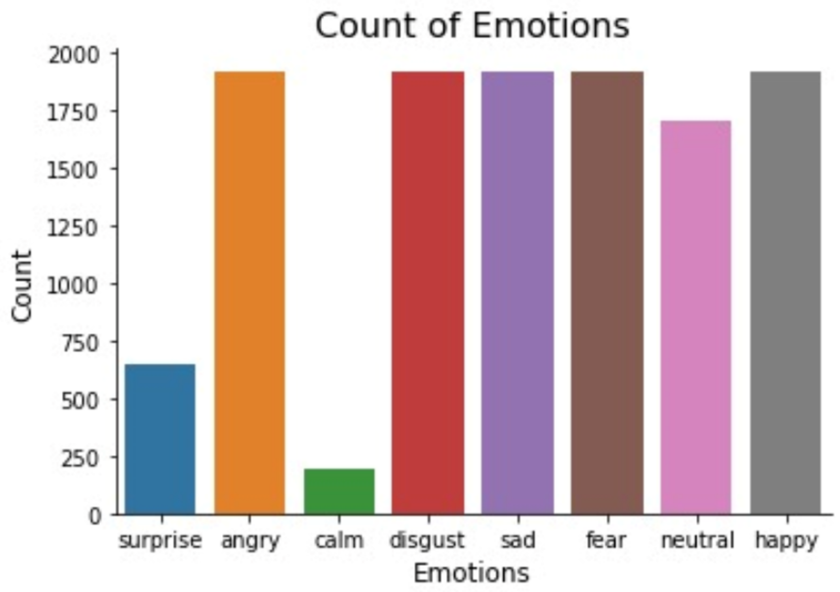
- Working on Raw Audio Data
- Defining Feature Extraction

Gender Feature
Male: 1 , Female: 0

Acoustic Features
Zero Crossing Rate
Chroma STFT
Mel Frequency Cepstral Coeffient
Root Mean Square
MelSpectrogram

Statistical Features
Mean
Variance
Skewness
Kurtosis
spectral_centroids.mean()
spectral_bandwidth.mean()

- Dropping Imbalanced classes when creating SER Superset Data
  - × Calm
  - × Surprise



Experimental Design - Prerequisites

- **Data Augmentation Techniques Applied**
  - Stretch
  - Pitch Shift
  - Low-Pass Filter
  - Noise

- **K-fold Cross-Validation by creating folds by actors**
  - RAVDESS – 6 folds (4 actors each)
  - SER Superset data – 11 folds (11 actors each)

RAVDESS	Fold 1: Actor 1 - 4	Fold 4: Actor 13 - 16
	Fold 2: Actor 5 - 8	Fold 5: Actor 17 - 20
	Fold 3: Actor 9 - 12	Fold 6: Actor 21 - 24

TT 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	**Fold 6**
TT 2	Fold 1	Fold 2	Fold 3	Fold 4	**Fold 5**	Fold 6
TT 3	Fold 1	Fold 2	Fold 3	**Fold 4**	Fold 5	Fold 6
TT 4	Fold 1	Fold 2	**Fold 3**	Fold 4	Fold 5	Fold 6
TT 5	Fold 1	**Fold 2**	Fold 3	Fold 4	Fold 5	Fold 6
TT 6	**Fold 1**	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6

Avg. Test Accuracy = Avg(TT 1 + TT 2 + TT 3 + TT 4 + TT 5 + TT 6)

Experimental Design Cont.

- Initially RAVDESS is used to create Train, Test and Validation data
- 6 folds were generated with RAVDESS Dataset
- Data Augmentation is performed on any dataset considered
- **Deep Feature Extractor:** Conv1D, Con2D – 32 features extracted
- Statistical – 6 features extracted
- Gender – 1 feature extracted
- **Combined Features Classifier Model:** MLP, LSTM, BiLSTM

Hardware Setting
GPU environment – Kaggle - NVIDIA GPU P 100
Libraries – Librosa, Keras, Sklearn, PyAudio, Scipy
Language - Python

Hyper-parameters:

- **ReduceLROnPlateau** with min-lr = 0.0000001
- Optimizer – **Adam**
- Loss – **Categorical Cross Entropy**
- Activation Function – **ReLU, Softmax**
- Batch Size – **64**

Experimental Design Cont.

Model: "sequential"

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 (MaxPooling1D)	(None, 41, 256)	0
dropout (Dropout)	(None, 41, 256)	0
flatten (Flatten)	(None, 10496)	0
dense (Dense)	(None, 32)	335904
dropout_1 (Dropout)	(None, 32)	0
dense_1 (Dense)	(None, 12)	396
dropout_2 (Dropout)	(None, 12)	0
dense_2 (Dense)	(None, 6)	78

Total params: 665,850
Trainable params: 665,850
Non-trainable params: 0

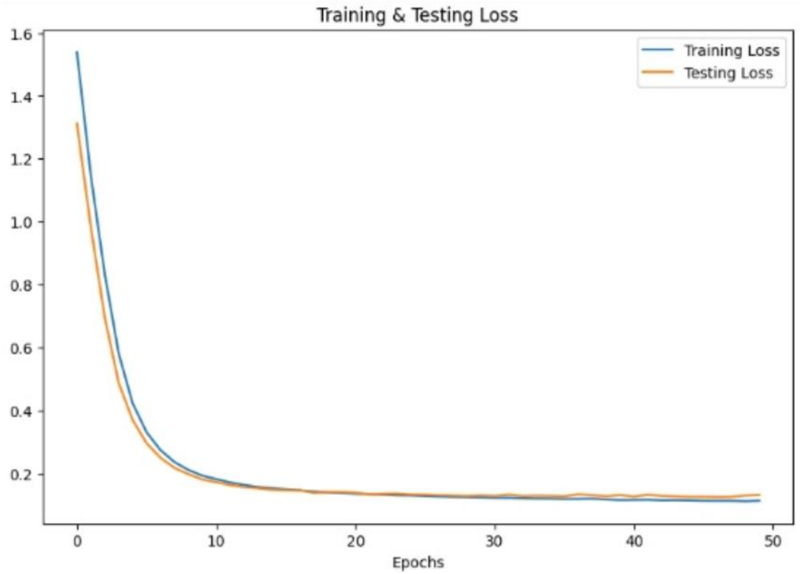
Conv1D Deep Feature extractor

MLP Dense Layer Classifier

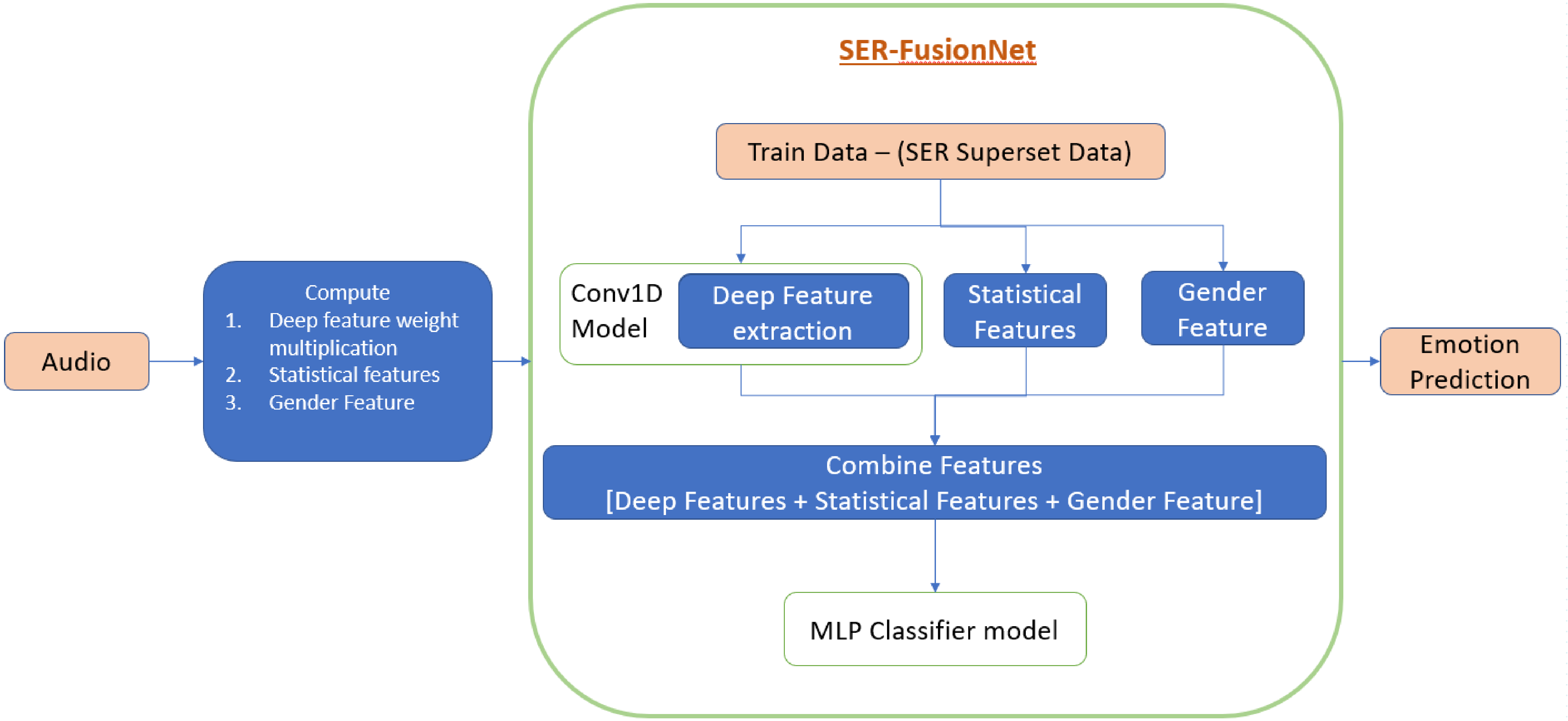
Model: "sequential_4"

Layer (type)	Output Shape	Param #
dense_12 (Dense)	(None, 12)	504
dense_13 (Dense)	(None, 10)	130
dense_14 (Dense)	(None, 6)	66

Total params: 700
Trainable params: 700
Non-trainable params: 0



Architecture of the Proposed Method: **SER-FusionNet**



Experimental Result on SER-FusionNet

	Predicted Labels	Actual Labels
0	angry	neutral
1	disgust	neutral
2	happy	neutral
3	neutral	neutral
4	happy	disgust
5	disgust	disgust
6	happy	disgust
7	disgust	disgust
8	disgust	disgust
9	disgust	disgust

Deep Feature Extraction Model:
Conv1D, Conv2D

Classifier Model:
MLP, LSTM, BiLSTM

Other Evaluation Metrics:
Precision, Recall, F1 Score

Data Set	Train Data	Test Data	CV	Feature Fusion Accuracy	
				Val	Test
RAVDESS	5-24 Actors	1-4 Actors	No	89.00 %	50.1%
	5 * (4 Actor fold)	1 * (4 Actor fold)	YES – 6 folds	62 ± 2%	46 ± 1%
SER Superset	5-121 Actors	1-4 Actors	No	67.09%	49.22%
	10 * (11 Actor fold)	1 * (11 Actor fold)	YES – 11 folds	67 ± 3 %	38 ± 2 %
	CREMA, RAVDESS	SAVEE, TESS	No	50.79%	26.05%

Literature Review

Author	Related Works	Dataset	Feature Extraction Method	Classification Method	Accuracy
Kawade et al., (2022)	Speech Emotion Recognition Using 1D CNNLSTM Network on Indo-Aryan Database	RAVDESS	Pitch, Energy, ZCR, MFCC	1D CNN, LSTM	87% (Val)
Ullah et al., (2022)	Speech Emotion Recognition Using Deep Neural Networks	RAVDESS, Crema-D, Tess and SAVEE	MFCC, Energy and Related Features, ZCR	1-D CNN	92.62%
Rumagit et al., (2021)	Model Comparison in Speech Emotion Recognition for Indonesian Language	Data collected manually.	Mel-spectrogram, Chroma, and MFCC	SVM, MLP, and Logistic Regression	76.22%
Zhang et al., (2023)	A Deep Learning Method Using Gender-Specific Features for Emotion Recognition	RAVDESS and CASIA	MFCC mean, Fundamental frequency F0, and Spectral Contrast Ratio	CNN , BiLSTM	82.59%

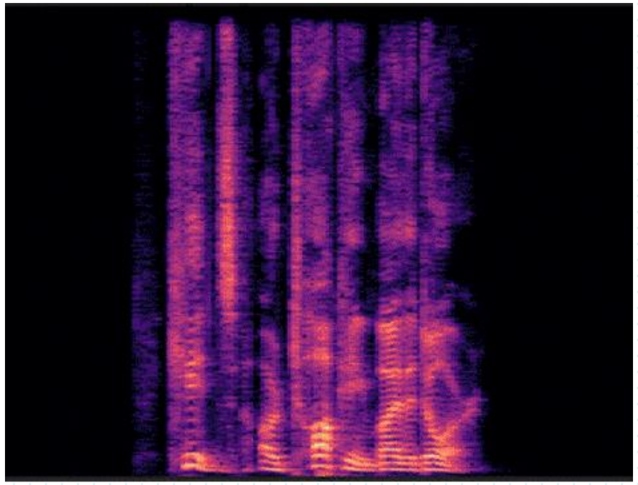
Challenges and Future Research Discussion

Challenges:

- **Data Collation:** Missing single dataset with wide variety of emotions, consistent formatting, and data quantity
- **Data Processing:** Interpret and identify processing techniques of High Dimensional raw audio data

- Spectrogram Image-based Emotion Classification Experiment

Transfer Learning Model	Train, Test Dataset	Accuracy
VGG16	RAVDEES	54%



- Explore including features from images generated into SER-FusionNet
- To increase the precision of emotion identification, use additional modalities along with speech, such as facial expressions or physiological signs.