SPEECH EMOTION RECOGNITION

OUR TEAM

Elif Aybüke
COŞKUN
201811018

Furkan
DURAN
201811027

Abdullah
ÖZDER
202011410

Şima
KAYISI
201811043

İhsan
BARDAKCI
201717007

Assist. Prof. Dr. Ayşe Nurdan SARAN

CONTENT

1 **PROBLEM**
What brought us here?

2 **ANALYSIS**
What we learned and designed as a result of our researches?

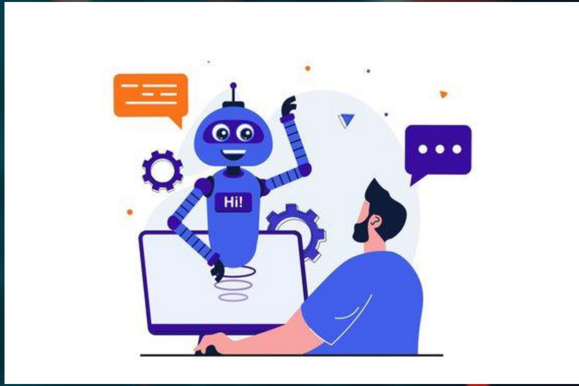
3 **DIFFERENCE**
What is our difference?

4 **TECHNIQUES USED**
What were the techniques and methods used?

5 **SER SUCCESS RATES**
Accuracy rate and method comparison of the system?

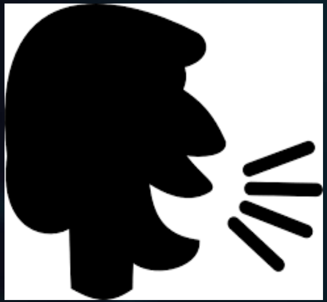
6 **INTERFACES AND CONCLUSIONS**

INTRODUCTION



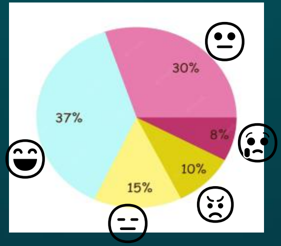
PROBLEM

What brought us here?



Speech Analysis

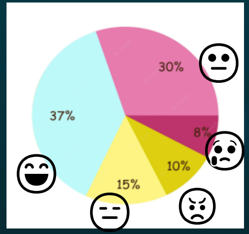
Evaluate of Speech Data



Analysis Result

Text Analysis

Evaluate of Text Data



Analysis Result

ANALYSIS

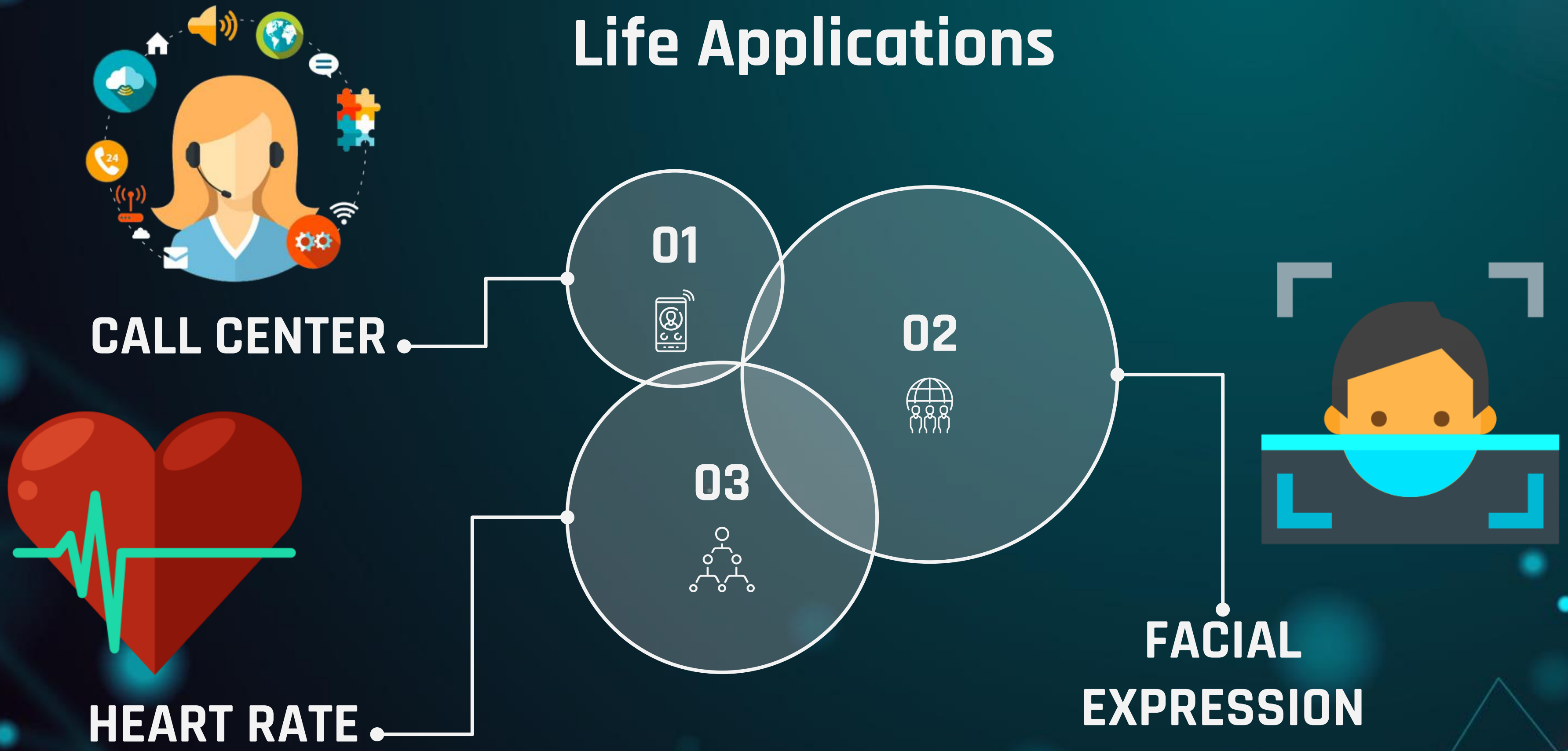
What we learned and designed as a result of our researches?

Speech is the most important and effective main way of human interaction.

There is a transfer of emotion in every person's speech.

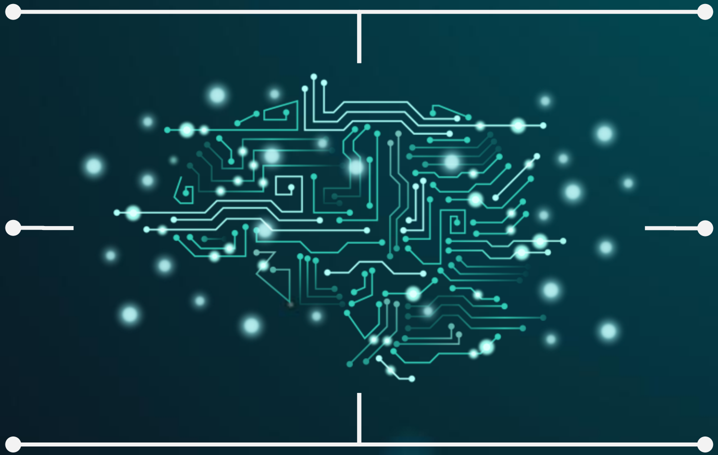
Analyzing speech signals.

Speech Emotion Recognition: Enhancing Daily Life Applications



SYSTEM PURPOSE

Upload File



Machines' understanding of human emotions

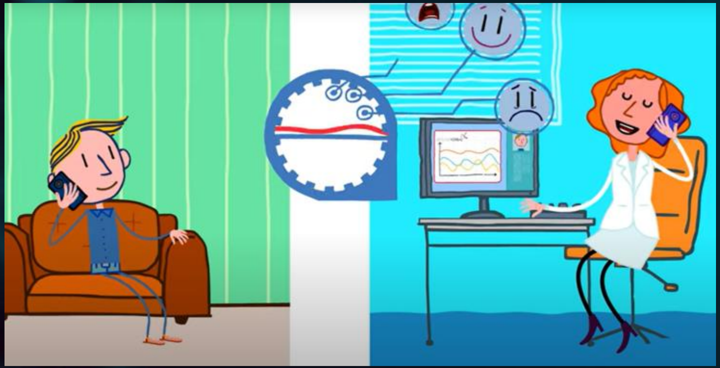
SYSTEM PURPOSE

The aim of this system is to analyze the emotional state of these people as a result of taking the texts that people have spoken or written.

Previous Works About SER

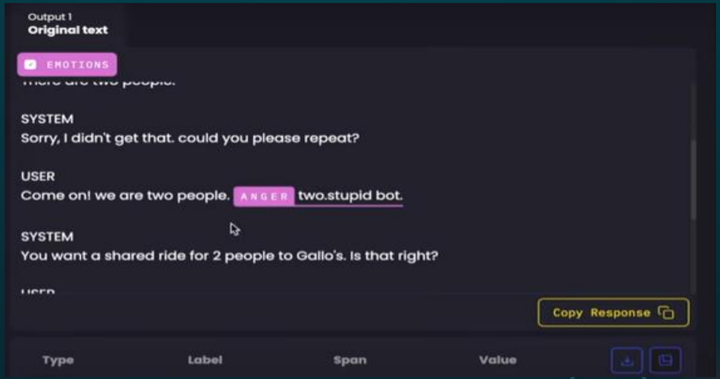
Audio - Based

Traditional SER systems primarily focused on the acoustic properties of speech, such as pitch, volume, and speed. While effective to some degree, these systems often fail to capture the full emotional context, especially in complex or ambiguous situations.



Text - Based

Some projects used text-based emotion recognition. By analyzing words, phrases, and their context, these systems could identify emotions in verbal communication. However, they neglect the tonal aspect of speech, which can sometimes convey more emotion than the words alone.



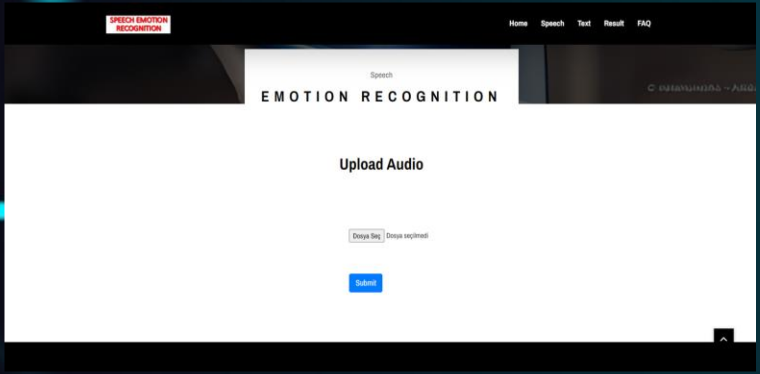
DIFFERENCE

What is our difference?

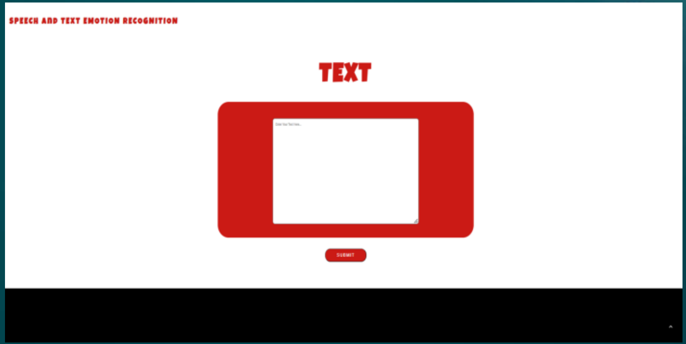
WHAT SETS US APART

Emotion analysis from text and audio files.

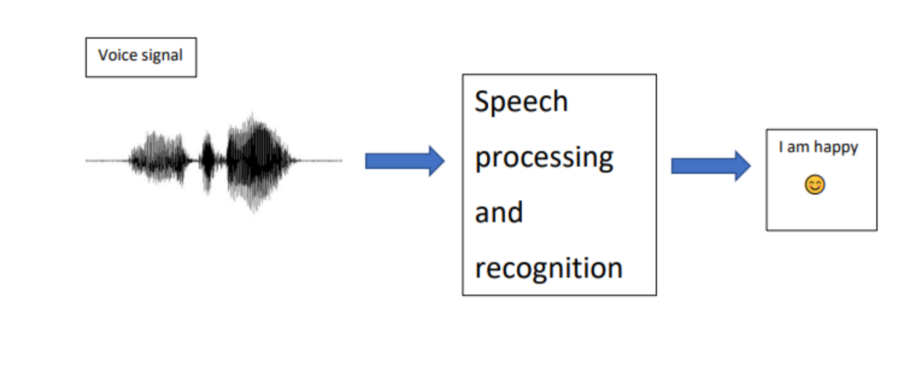
CONTRIBUTIONS



AUDIO



We work from both audio and text files.

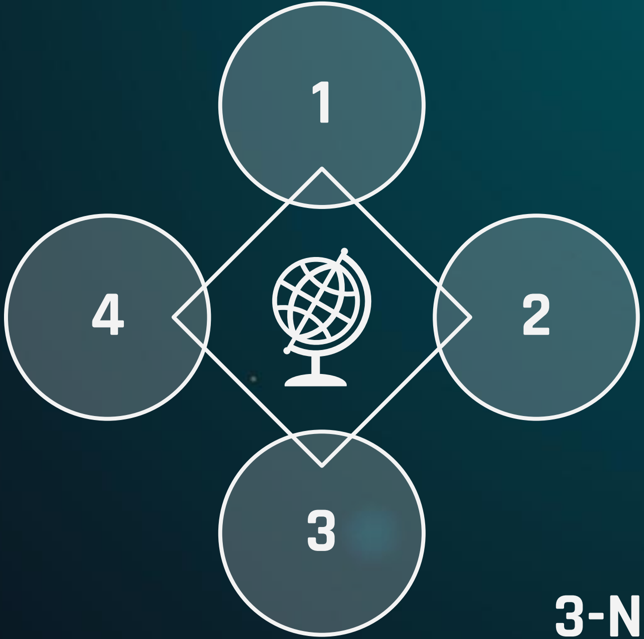


After the audio file is processed, it is converted into text and the mood appears on the screen.

ADVANTAGES

1-EASE OF USE

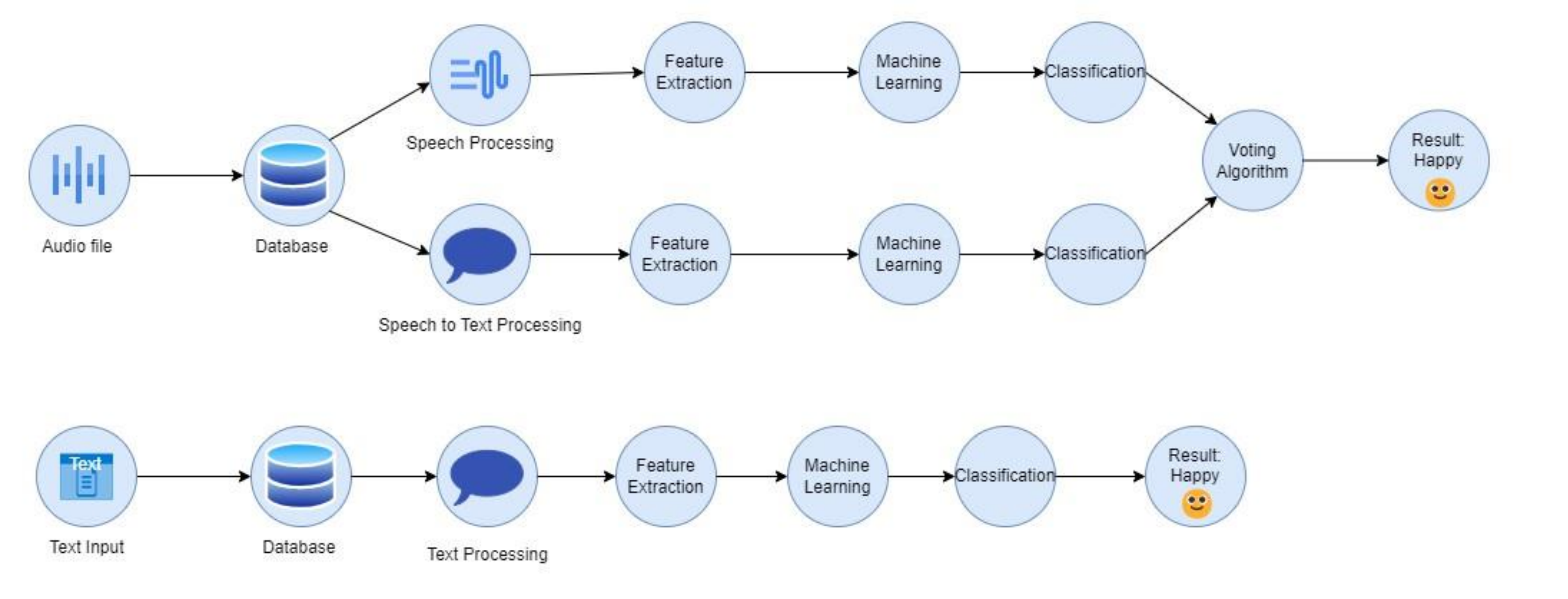
2- FEWER ERRORS



3-NO REQUIREMENT OTHER THAN PC

4- SAVING ON TIME

FLOWCHART

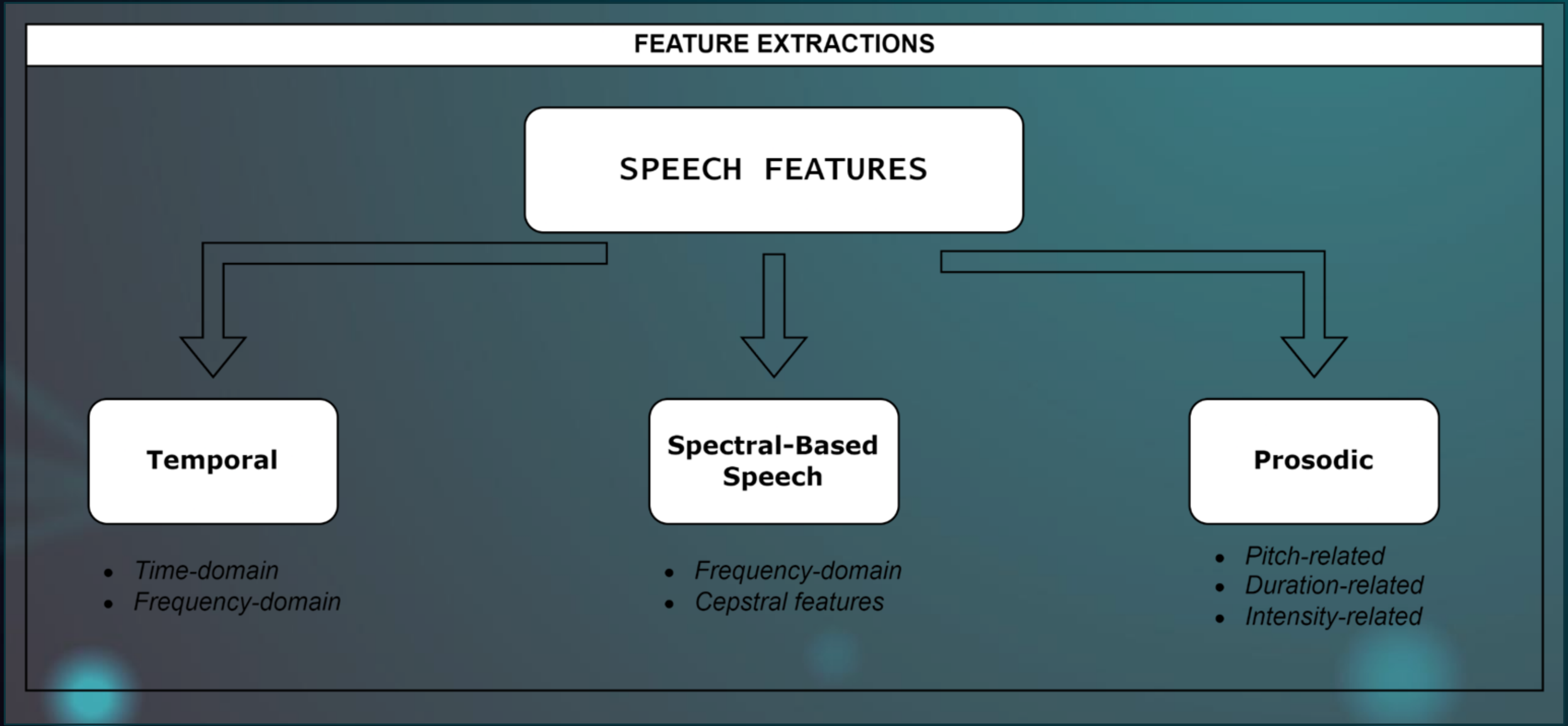

Speech Processing
Feature Extraction
Machine Learning
Classification
Audio file
Database
Voting Algorithm
Result: Happy
Speech to Text Processing
Feature Extraction
Machine Learning
Classification
Text
Text Input
Database
Text Processing
Feature Extraction
Machine Learning
Classification
Result: Happy

FEATURE EXTRACTIONS

The speech features can be divided into 3 main categories:

- Temporal Features
- Spectral-Based Speech Features
- Prosodic Features

FEATURE EXTRACTIONS



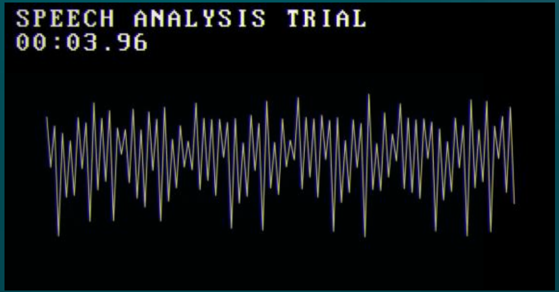
FEATURE EXTRACTIONS

Mean	$\mu=\frac{1}{N}\sum s(t)$
Variance	$\alpha^2=\frac{1}{N}\sum(s(t)- \mu)^2$
Kurtosis	$K=(m_4/(m_2^{2}))$
Skew	$S=(m_3/(m_3^{3/2}))$
Latency to Amp. Ratio	$t_{max/}\ S_{max}$
Absolute Amp.	$\lvert s_{max}\rvert$
Abs. Latency to Amp ratio	$\lvert t_{max/}\ S_{max}\rvert$

Temporal Features:

- Time - Domain Features:
  - Amplitude: The instantaneous magnitude of a sound signal.
  - Zero-crossing rate: The rate at which the sound signal crosses the zero level.
  - Energy: The energy content of the sound signal.
  - Duration: The total duration of the sound.
- Frequency - Domain Features:
  - Mel-frequency cepstral coefficients (MFCCs): Temporal features used to represent the spectral content.
  - Spectral centroid: The spectral center frequency of the sound signal.
  - Spectral flux: The rate of change of spectral content over time.

FEATURE EXTRACTIONS

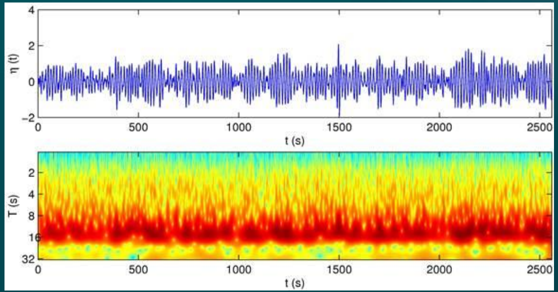


Spectral Features:

- Pitch - Related Features:
    - Spectral centroid: The spectral center frequency of the sound signal.
    - Spectral rolloff: The value of the frequency at which a certain percentage of the energy in the sound signal has passed.
    - Spectral harmonicity: A feature representing the harmonic structure of the signal.

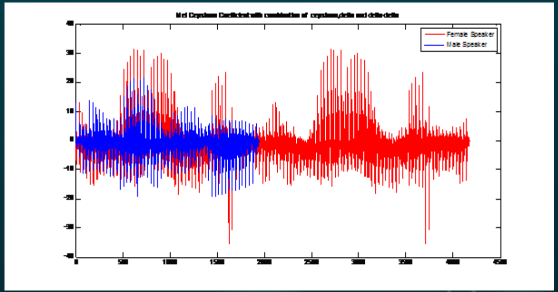
- Cepstral Features:
    - Mel-frequency cepstral coefficients (MFCCs): Temporal features representing the spectral content of the sound signal.
    - Linear predictive coding (LPC) cepstral coefficients: Features used for signal prediction in the sound signal.

FEATURE EXTRACTIONS



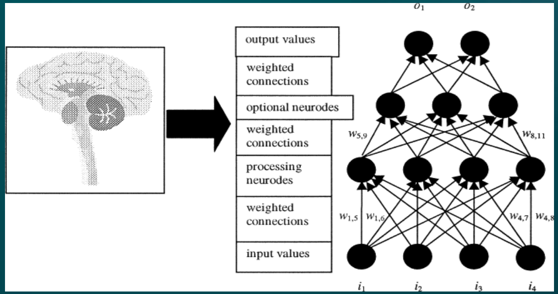
Prosodic Features:

- Frequency - Domain Features:
  - Pitch range: The frequency range of the sound signal.
- Duration - Related Features:
  - Phoneme duration: Total duration of phonemes in the sound signal.
  - Pause duration: Duration of pauses in the sound signal.
  - Speech rate: Rate of speech in the sound signal.
- Intensity - Related Features:
  - Phoneme duration: Total duration of phonemes in the sound signal.
  - Pause duration: Duration of pauses in the sound signal.
  - Speech rate: Rate of speech in the sound signal.

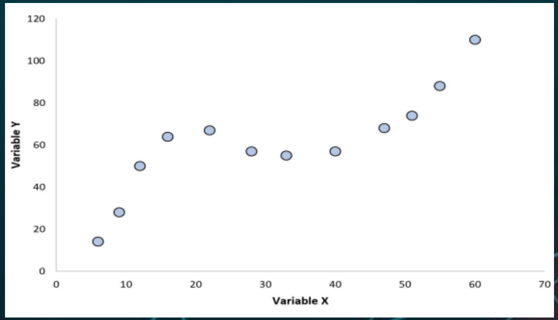


MACHINE LEARNING

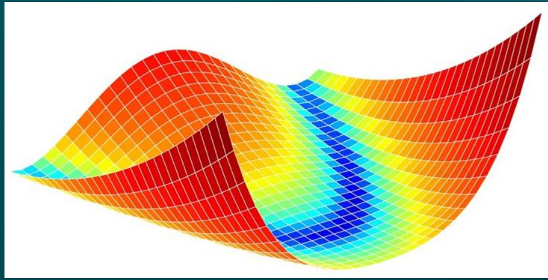
Technical Approach:



- Model Architecture:
  - In this code block, an Artificial Neural Network (ANN) model is utilized. ANN is a widely used model in machine learning.
  - The model consists of sequential layers. Each layer takes the outputs of the previous layer as input and generates new features through fully connected operations.
  - Activation functions shape the outputs of neurons in each layer, enabling the model to learn non-linear relationships.



MACHINE LEARNING

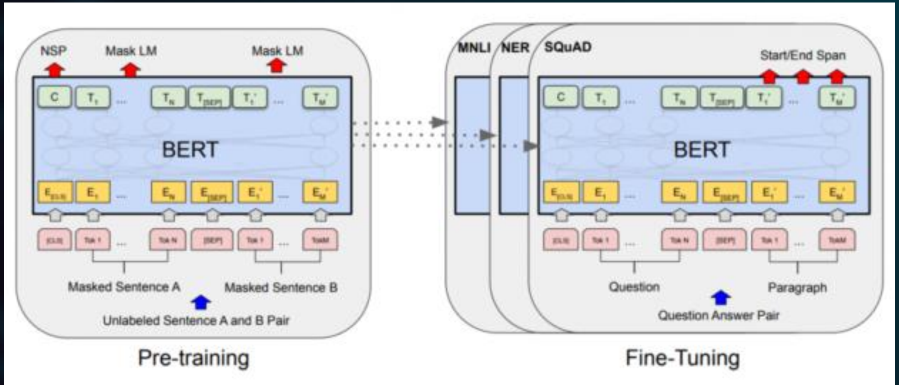


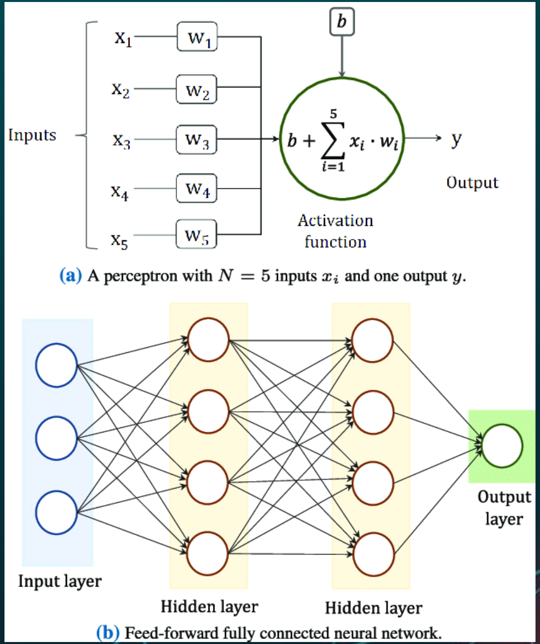
Technical Approach:

- Optimization Algorithm:
  - In this code block, the Adam optimization algorithm is employed to adjust the model's parameters.
  - Adam is a gradient-based optimization algorithm commonly preferred in deep learning models.
  - The optimization algorithm manages parameter updates during the model's training process, aiming to enhance its performance.
- Loss Function and Metrics:
  - Categorical cross-entropy loss is employed for training the model in this code block.
  - Categorical cross-entropy is a commonly used loss function for classification problems. It measures the compatibility between the model's predictions and the true labels.
  - Additionally, the model's performance is evaluated using the accuracy metric. Accuracy represents the percentage of correct predictions made by the model.

Text Emotion Recognition

- Bidirectional Encoder Representations from Transformers
- Artificial Neural Network





(a) A perceptron with $N = 5$ inputs x_i and one output y.

(b) Feed-forward fully connected neural network.

Text Emotion Recognition

- Text Data Preprocessing
- Text Data Converted to Numerical Embedding Representations
- Adding Dense and Dropout Layers on BERT with ANN
- Final Output Layer of Model, Defining Sigmoid Activation Function to Determine Probabilities of Classes

- This method provides effective classification of text data by using BERT model and Artificial Neural Networks (ANN).

Technique and Success Rates Used in Previous Studies

Papers	Dataset	Emotions	Technique	Accuracy (%)
Clustering-Based Speech Emotion Recognition (2020)	IEMOCAP EMO-DB RAVDEES	Angry, Happy, Sad, Fear, Surprise, Neutral	CNN + LSTM	72.25 85.57 77.02
Speech Emotion Recognition with Deep Learning (2020)	RML Dataset	Angry, Disgust, Fear, Happy, Sad, Surprise	Basic AE with SVM Stacked AE with SVM	72.83 74.07
Speech emotion recognition with deep convolutional neural networks (2020)	IEMOCAP EMO-DB RAVDEES	Angry, Disgust, Fear, Happy, Sad, Surprise	CNN LSTM	64.30 71.61 86.1
Multimodal Speech Emotion Recognition and Ambiguity	IEMOCAP (Audio Only)	Angry, Happy, Sad, Fear, Surprise, Neutral	RF XGB SVM MNB MLP	56.0 56.6 33.7 31.3 41.0
Multimodal Speech Emotion Recognition (2019)	IEMOCAP	Angry, Happy, Sad, Fear, Surprise, Neutral	RF XGB SVM MNB LR MLP LSTM ARE (4-class) E1 (4-class) E1	56.0 55.6 33.7 31.3 33.4 41.0 43.6 56.3 56.2 56.6

When the projects done in the past are examined, it has been observed that the feelings of Angry, Happy, Sad, Neutral and Surprised have been studied more intensely.

The Techniques We Worked On For Text and Speech and Our Success Rates

METHOD	DATASET	EMOTIONS	TECHNIQUE	ACCURACY
Emotion Recognition with Speech	RAVDEES	Angry – Disgust – Fear – Happy – Neutral – Sad	SVM + ANN	% 99
Emotion Recognition with Speech	IEMOCAP	Angry – Excited – Frustration – Happy – Neutral – Sad	SVM + ANN	% 42
Emotion Recognition with Speech	IEMOCAP + RAVDEES	Angry – Excited – Frustration – Happy – Neutral – Sad	SVM + ANN	% 54
Emotion Recognition with Speech	IEMOCAP	Angry – Excited – Frustration – Happy – Neutral – Sad	ANN	% 33
Emotion Recognition with Speech	IEMOCAP	Angry – Disgust – Fear – Happy – Neutral – Sad	CNN + RESNET	% 30
Emotion Recognition with Speech	IEMOCAP	Angry – Disgust – Fear – Happy – Neutral – Sad	CNN + RESNET + ALEXNET	% 24
Emotion Recognition with Speech	IEMOCAP	Angry – Disgust – Fear – Happy – Neutral – Sad	DNN + CRNN	% 32
Emotion Recognition with Text	IEMOCAP	Angry – Excited – Frustration – Happy – Neutral – Sad – Fear - Disgust	BERT + ANN	% 26
Emotion Recognition with Text	IEMOCAP	Angry – Excited – Frustration – Happy – Neutral – Sad	BERT + ANN	% 82

The Techniques That We Used For Text and Speech and Our Success Rates

METHOD	DATASET	EMOTIONS	TECHNIQUE	ACCURACY
Emotion Recognition with Speech	RAVDEES	Angry – Disgust – Fear – Happy – Neutral – Sad	SVM + ANN	% 99
Emotion Recognition with Speech	IEMOCAP	Angry – Excited – Frustration – Happy – Neutral – Sad	SVM + ANN	% 42
Emotion Recognition with Speech	IEMOCAP + RAVDEES	Angry – Excited – Frustration – Happy – Neutral – Sad	SVM + ANN	% 54
Emotion Recognition with Speech	IEMOCAP	Angry – Excited – Frustration – Happy – Neutral – Sad	ANN	% 33
Emotion Recognition with Speech	IEMOCAP	Angry – Disgust – Fear – Happy – Neutral – Sad	CNN + RESNET	% 30
Emotion Recognition with Speech	IEMOCAP	Angry – Disgust – Fear – Happy – Neutral – Sad	CNN + RESNET + ALEXNET	% 24
Emotion Recognition with Speech	IEMOCAP	Angry – Disgust – Fear – Happy – Neutral – Sad	DNN + CRNN	% 32
Emotion Recognition with Text	IEMOCAP	Angry – Excited – Frustration – Happy – Neutral – Sad – Fear - Disgust	BERT + ANN	% 26
Emotion Recognition with Text	IEMOCAP	Angry – Excited – Frustration – Happy – Neutral – Sad	BERT + ANN	% 82

How Are the Success Rates of the SER System Compared to Other Articles?

Papers	Dataset	Emotions	Technique	Accuracy (%)
Clustering-Based Speech Emotion Recognition (2020)	IEMOCAP EMO-DB RAVDEES	Angry, Happy, Sad, Fear, Surprise, Neutral	CNN + LSTM	72.25 85.57 77.02
Speech Emotion Recognition with Deep Learning (2020)	RML Dataset	Angry, Disgust, Fear, Happy, Sad, Surprise	Basic AE with SVM Stacked AE with SVM	72.83 74.07
Speech emotion recognition with deep convolutional neural networks (2020)	IEMOCAP EMO-DB RAVDEES	Angry, Disgust, Fear, Happy, Sad, Surprise	CNN LSTM	64.30 71.61 86.1
Multimodal Speech Emotion Recognition and Ambiguity	IEMOCAP (Audio Only)	Angry, Happy, Sad, Fear, Surprise, Neutral	RF XGB SVM MNB MLP	56.0 56.6 33.7 31.3 41.0
Multimodal Speech Emotion Recognition (2019)	IEMOCAP	Angry, Happy, Sad, Fear, Surprise, Neutral	RF XGB SVM MNB LR MLP LSTM ARE (4-class) E1 (4-class) E1	56.0 55.6 33.7 31.3 33.4 41.0 43.6 56.3 56.2 56.6

METHOD	DATASET	EMOTIONS	TECHNIQUE	ACCURACY
Emotion Recognition with Speech	IEMOCAP + RAVDEES	Angry – Excited – Frustration – Happy – Neutral – Sad	SVM + ANN	% 54
Emotion Recognition with Text	IEMOCAP	Angry – Excited – Frustration – Happy – Neutral – Sad	BERT + ANN	% 82

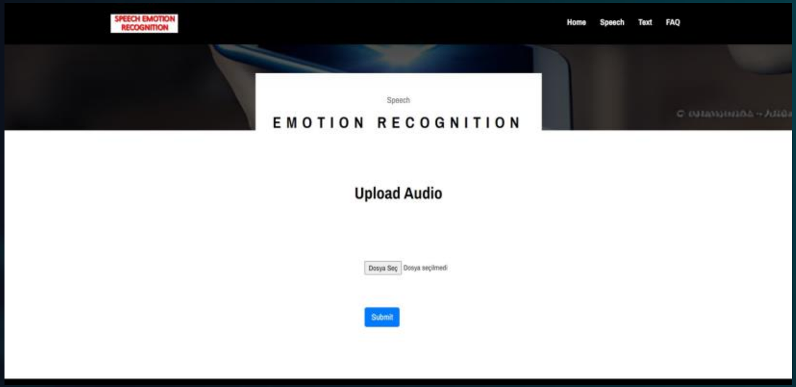
INTERFACES

Interfaces and Flow of our SER System

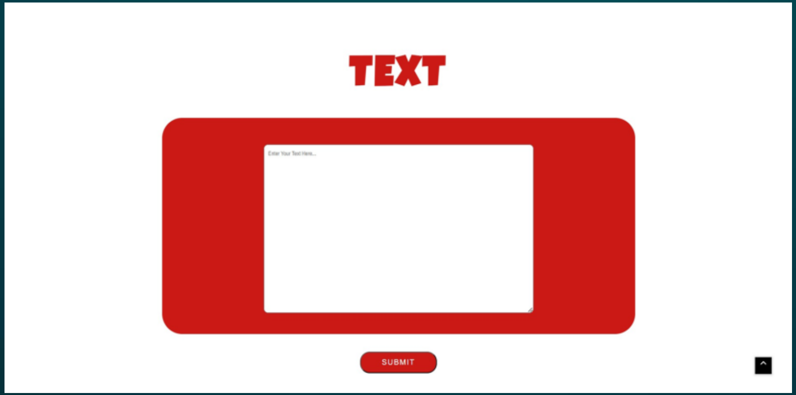
HOME PAGE

INTERFACES

Interfaces and Flow of our SER System



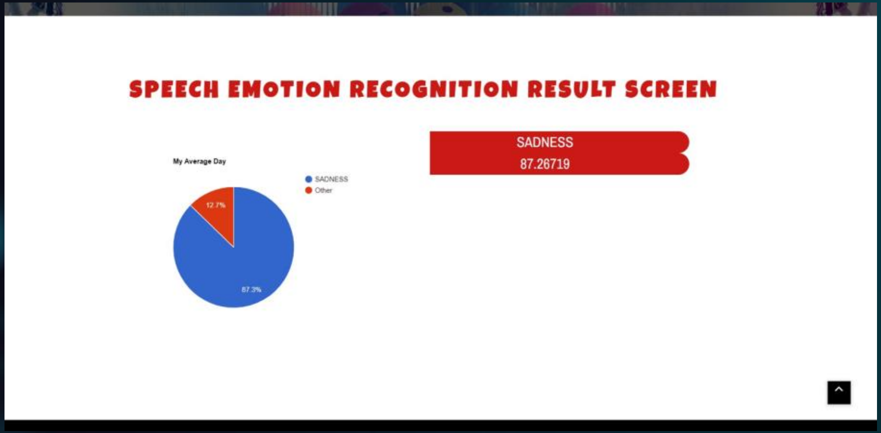
SPEECH PAGE



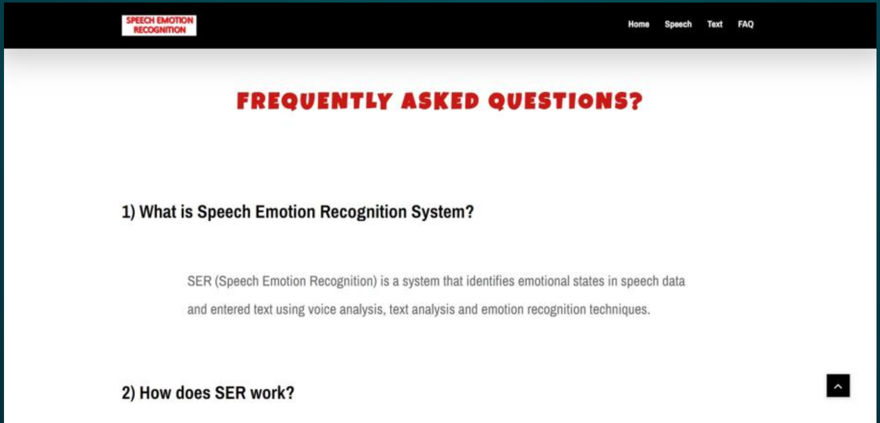
TEXT PAGE

INTERFACES

Interfaces and Flow of our SER System



RESULT PAGE



FAQ PAGE

TECHNOLOGIES

What were the technologies used?

macOS

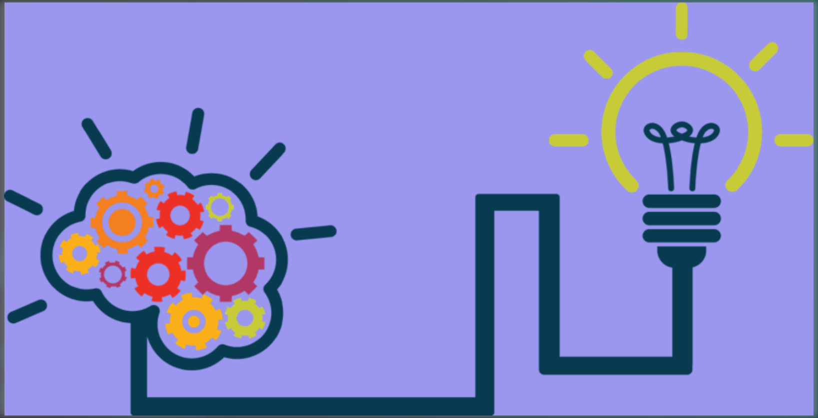
Linux

JS

React

librosa

CONCLUSION



THANK YOU FOR LISTENING!!