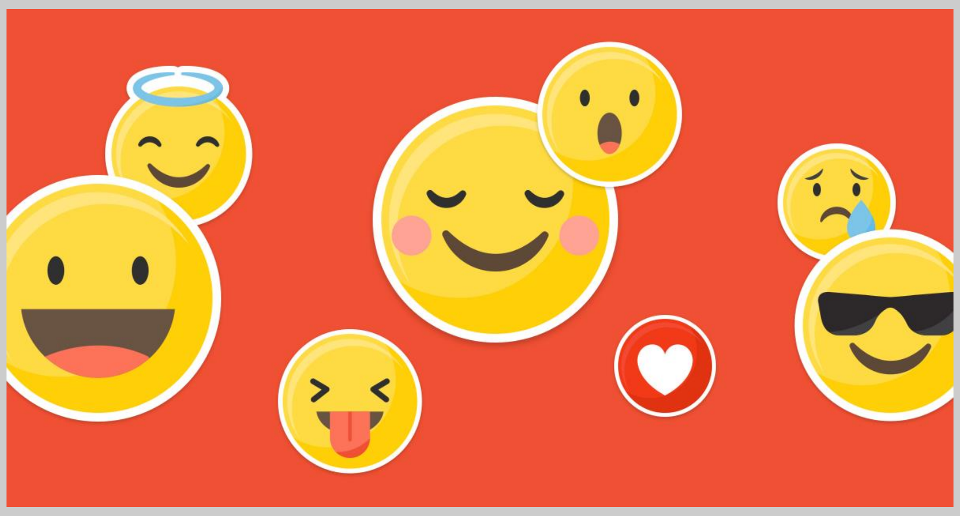
Speech Emotion Recognition

Furkan Duran – Elif Aybüke Coşkun
– Abdullah Özder – İhsan Bardakcı – Şima Kayısı
Dr. Ayşe Nurdan Saran

Çankaya University, Department of Computer Engineering

Abstract

This study deals with the development of Speech Emotion Recognition System in the field of human-computer interaction. The Speech Emotion Recognition system runs an audio analysis model from the audio file that the user uploads. At the same time, a text analysis model works, which converts the sound in this audio file to text and analyzes it. The SER system successfully presents the results of these two models to the user. At the same time, the SER system analyzes the text entered by the user on the "Text" page with the text analysis model and produces a result. Emotional states include various categories such as happy, sad, excited, frustration, neutral, and angry.



Introduction

Emotional expressions play a big role in communication between people today. However, understanding and interpreting emotional states correctly remains a challenge, especially in digital environments. In this context, we have developed the Speech Emotion Recognition, a system designed to analyze and identify emotions from uploaded audio files or text input. The system aims to provide users with accurate and real-time emotion analysis, encompassing emotions such as happy, sad, neutral, excited, frustration and angry.

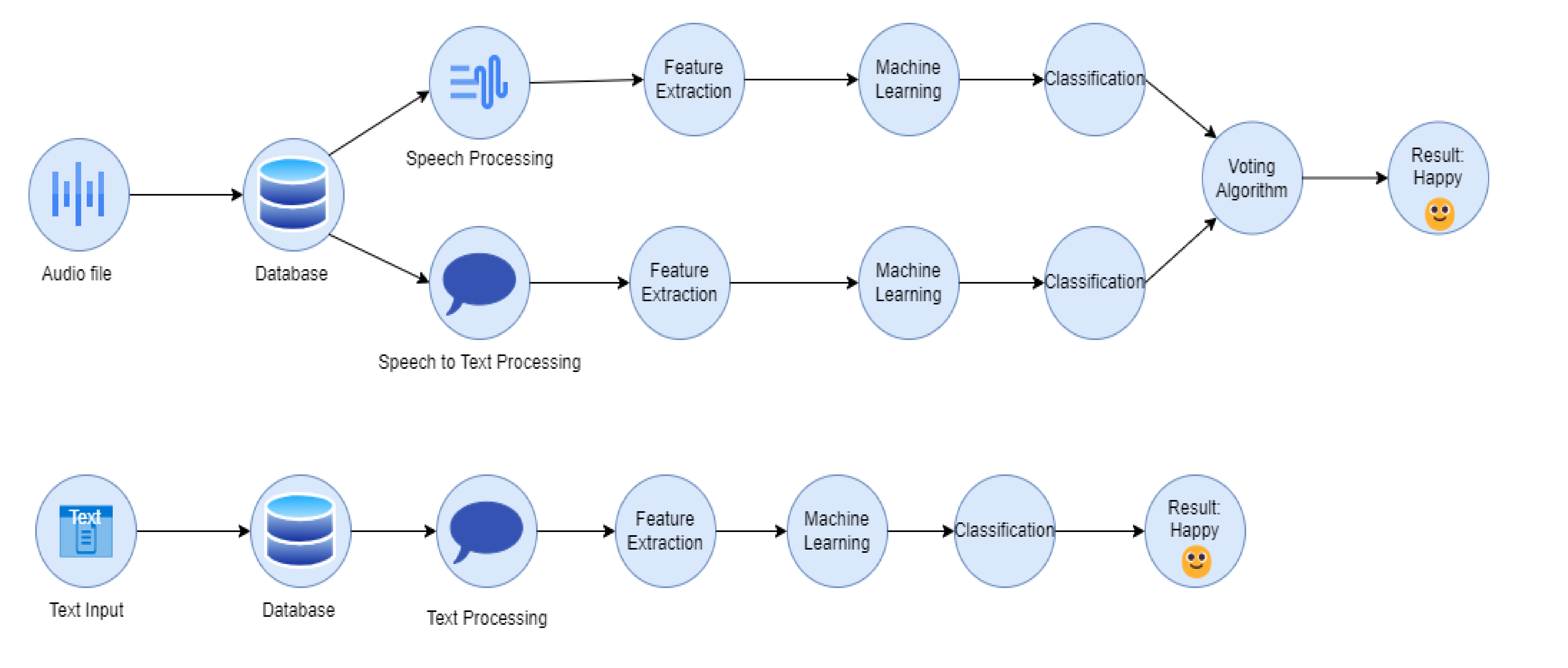


Figure 1 – System Architecture

Solution

We have developed an approach to the SER system, unlike the previous ones, can analyze both text and speech. The user uploads an audio file through a web-based interface, which serves as the primary data source for emotional analysis. The audio file undergoes preprocessing steps.

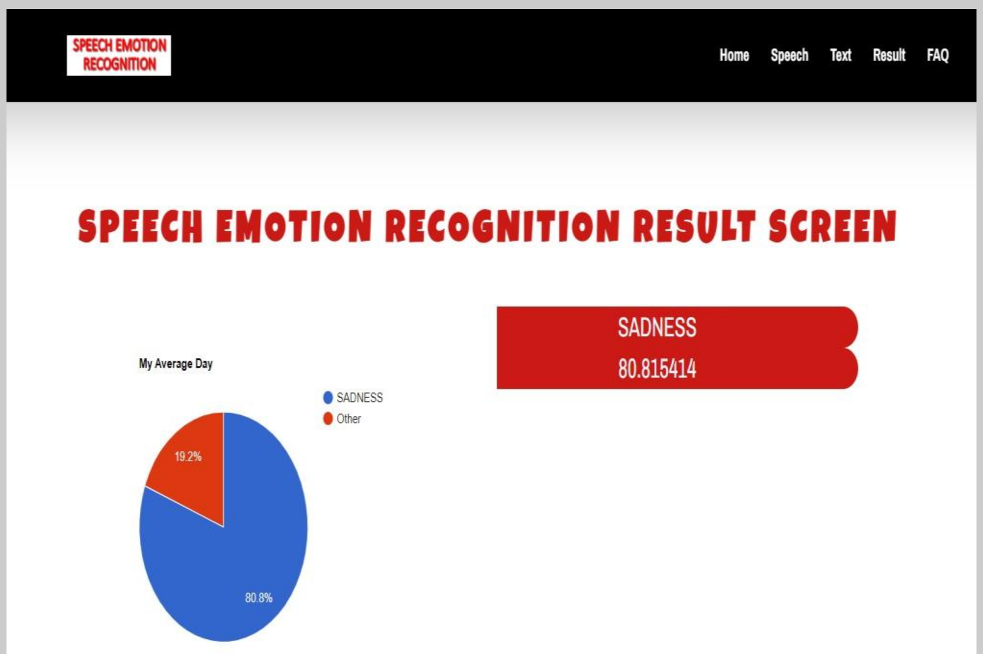
For the Speech Model, the SER system runs an emotion recognition model using machine learning algorithms such as SVM (Support Vector Machine) and ANN (Artificial Neural Network). This model utilizes feature extraction techniques to convert the audio data into feature vectors. ANN and SVM generate the necessary analyses and produce output.

As for the Text Model, the same audio file is processed through a text model by converting it into text. The converted text from the audio file is analyzed and processed using BERT (Bidirectional Encoder Representations from Transformers) and ANN (Artificial Neural Network) techniques, resulting in an output.

The voting algorithm combines the results of speech-based emotion recognition and text-based emotion analysis using a voting algorithm. This algorithm assigns weights to the results of both models and determines the highest-rated emotional label or intensity. For example, among the emotional labels, options like happy, sad, neutral, excited frustration or angry may exist. By developing our project on the web, we aimed to make it easier for everyone to access.

Results & Conclusion

We have obtained results similar to many studies. Of course, the most significant difference and feature of our system, the Speech Emotion Recognition System, is that it combines both text and voice emotion analysis on a system and presents it to the end user over the web. In this way, we have created a system that everyone can access and use. It provides as accurate an analysis as possible.



SPEECH EMOTION RECOGNITION RESULT SCREEN

SADNESS
80.815414

Figure 2 – Finished Product

Acknowledgement

Our teacher, Dr. Nurdan Saran supported us at every stage of the project. In every stage, she guided us to understand the problem and set our goals. Words are not enough to express our gratitude to her.