Artificial Intelligence and Social Computing, Vol. 72, 2023, 111–120
https://doi.org/10.54941/ahfe1003279

Emotion Recognition From Speech via the Use of Different Audio Features, Machine Learning and Deep Learning Algorithms

Alperen Sayar[1], Seyit Ertuğrul[1], Tunahan Bozkan[1], Fatma Gümüş[1], and Tuna Çakar[2]

[1]Tam Finans Faktoring A.Ş, Sisli, Istanbul 34360, Turkey
[2]MEF University, Sarıyer, Istanbul 34240, Turkey

ABSTRACT

In this study, different machine learning and neural network methods for emotion analysis from speaking are examined and solutions are sought. Audio consists of a large number of attributes. It is possible to make emotion analysis from sound using these attributes. Root Mean Square Energy (RMSE), Zero Crossing Rate (ZCR), Chroma, Mel Frequency Cepstral Coefficients (MFCC), Spectral Bandwith, Spectral Centroid properties were investigated for speech-free mood prediction. Ravdess, Save, Tess, Crema-D datasets were used. The data sets were voiced in German and English by 121 different people in total. The datasets consist of audio files in wav format containing 7 emotional states as happy, sad, angry, disgusted, scared, surprised and neutral. By using the Librosa library, features were obtained from the audio files in the datasets. The features were used in various machine learning and neural network models and the results were compared. When the classification results are examined, 0.68 for Support Vector Machines, 0.63 for Random Forest Classification, 0.71 for LSTM and 0.74 F-1 for Convolutional Neural Networks.

Keywords: Voice analysis, Speech emotion recognition, Audio features, Classifiers, Machine learning

INTRODUCTION

Communication has been the basis of information exchange since the existence of human beings. Words and emotions follow each other to make communication more accurate, clear, and understandable. Depending on the emotional state of people, there are some physiological changes such as body movements, blood pressure, pulse, and tone of voice. While changes such as heart rate and blood pressure are detected with a special device, changes such as tone of voice and facial expression can be understood without the need for a device. Machines are often used for emotion predictions. (Suha Gokalp et al., 2021). Speech is one of the fastest and most natural communication methods between people. For this reason, researchers have started to use speech signals to make human-machine interaction faster and more efficient.

Speech signals have a complex structure that can contain much information at the same time, such as the speaker's age, mood, gender, physiology, and language. Emotion recognition studies without speech try to obtain semantic information from the sound signal during speech. (Suha Gokalp et al., 2021). This study aims to determine the emotional state of the speaker using speech signals. In academics, Speech Emotion Recognition has become one of the most wondered and investigated research areas (Jain et al., 2020). Along with the studies carried out in recent years, various studies have been carried out on the mood analysis of the speaker using machine learning, and thanks to these studies, great developments have been experienced in this field. However, it is a difficult task to analyse the mood from the sound waves of the speaker, because the sound consists of many parameters and has various features that must be taken into account. For these reasons, choosing the appropriate and correct features for emotion recognition without speech is the critical and perhaps the most important point of this study.

Machine learning basically means that a computer has the ability to automatically perform a task using data and learning methods. The computer uses statistics, various probability algorithms, and neural networks to learn and successfully complete these tasks. In the continuation of the study, parameters of various datasets and algorithms are given to create a machine-learning model.

Various approaches have been successfully applied for speech emotion recognition to date. In this article, various features of sound waves and various machine learning algorithms and neural networks are used for Speech emotion recognition. In order to increase the accuracy and success of the study, 4 different speech databases were combined.

DEVELOPING THE SPEECH EMOTION RECOGNITION SYSTEM DESIGN

Speech emotion recognition generally consists of three parts: feature extraction, feature selection, and classification (Shadi Langari et al., 2020).

DATA PROCESSING

Sample rate, in music and audio technology, indicates how many times per second an audio file or signal is measured. A higher sampling rate means higher sound quality and an audio file with more detail. Sample rate is usually specified as a few thousand or million sampling points per second. Higher sample rate can contain higher frequency values and provides higher sound quality. The sample rate used in this project is 22.05 kHz.

Hop length is a term used in music and audio technology when processing an audio file or signal. Hop length specifies the time interval after an audio file or signal has been measured once. This is used in conjunction with the sample rate, which is used to measure the frequency range of an audio file or signal. Together with the sample rate, it describes the frequency width of an audio file or signal and determines the sound quality. The hop length value used in this project is 512.

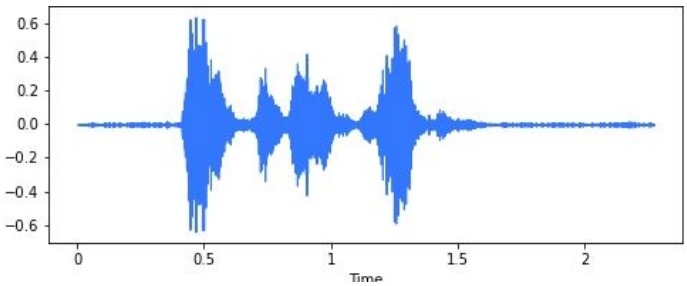

Figure 1: Sound wave in angry.

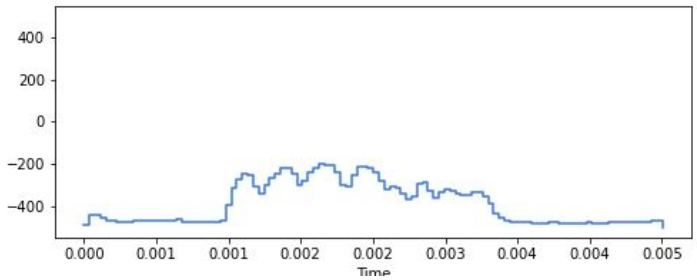

Figure 2: MFCC in neutral.

Frame length, in music and audio technology, refers to the time interval in which an audio file or signal is measured once. Frame length is used along with the sample length to use within the frequency range of the audio file or signal. The frame length value used in this project is 2048.

The Fourier transform is a mathematical operation to find the frequency spectrum of a signal. This process allows temporal patterns of a signal to be expressed over a frequency spectrum. In this way, the amplitudes and phases of the frequency components in the signal are determined and the characteristics of the signal are examined with this information.

MFCC (Mel Frequency Cepstral Coefficients) is a feature vector often used in audio processing applications. MFCCs represent audio based on perception of human auditory systems. In MFCC, the frequency bands are positioned logarithmically(i.e on the Mel Scale) which approximates the human auditory system's response more closely than the linearly spaced frequency bands of FFT or DCT (Goh C,Leon K, Gold B,Morgan N,) The MFCC feature vector is calculated over the frequency spectrum of the audio signals and includes the mel frequencies coefficients (cepstrum) in the audio signals. This cepstrum is logarithmically transformed over the frequency spectrum of the audio signals and then the Fourier transform of this logarithmic transform is taken. The values obtained as a result of these operations are converted into MFCC feature vectors. MFCC feature vectors help identify words and phrases contained in audio signals, and these features make it easier to classify audio signals. The number of mfcc value used in this project is 128.

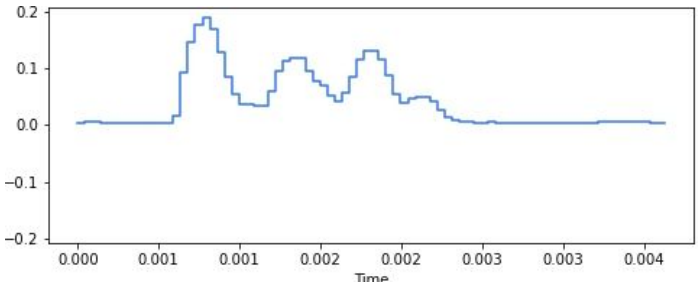

Figure 3: RMSE in happy.

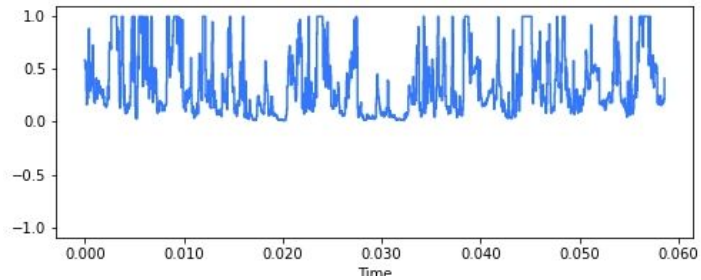

Figure 4: ZRCR in sad.

Pitch refers to the loudness of a melody or tone of voice in music. Pitch is usually expressed with sounds such as do, re, mi, fa, sol, la, si and is based on the working principle of musical instruments. Pitch is determined by the number of periodic oscillations of a sound and is usually measured in Hertz (Hz).

RMSE (Root Mean Square Energy) is a feature often used in audio processing applications and helps to measure the power levels of audio signals. The RMS value is calculated as the square root of the mean of the square of the temporal samples of the audio signals. This process is used to measure the power levels of the audio signals, and thanks to these values, it becomes easier to detect the words and expressions contained in the audio signals. We use the mean value of RMSE in this work.

ZRCR (Zero Crossing Rate) is a feature often used in audio processing applications and helps to measure the number of zero crossings of temporal samples of audio signals. This feature makes it more effective to measure the characteristics of the frequency spectrum of audio signals, and thanks to these features, it becomes easier to detect the words and expressions contained in the audio signals. We use the mean value of ZRCR in this work.

MODELING

Unsupervised learning aims either to discover similar sample groups in the data or to determine the distribution in the data space by identifying hidden patterns in the data. It uses unlabelled data to identify patterns. Clustering and Association are types of unsupervised learning.

Supervised learning, unlike unsupervised learning, works with labelled data. Its main purpose is to predict the correct label for unlabelled data. In this learning type, both input and output variables are presented.

A supervised learning algorithm can be shown simply as:

$$Y = f(x)$$

X: input value

Y: predicted output

Classification and regression are subcategories of supervised learning.

Classification is used for categorical outputs. Regression is used for continuous outputs.

In a regression task, the purpose of the model is to predict and understand the significant relationship between dependent and independent variables. It is a predictive statistical process and uses a continuous function to evaluate how outputs change for given inputs.

In machine learning, classification is the process of categorizing items based on a pre-categorized training dataset. The classifiers used in the reviewed articles and planned to be used in the project are defined as follows.

SVM is a machine learning algorithm used to solve classification problems with two or more classes. The purpose of the Support Vector machine is to find a hyperplane in an N-dimensional (N-feature number) problem space. There are many ways to find this out. However, the goal is to find a plane with the maximum distance between the data points of both classes. (Gandhi, 2021) Due to the high training cost, it is not preferred for data sets with high data size, it is generally used for medium or small data sets. Sequential minimal optimization has been developed to eliminate this problem. SMO is a training algorithm developed for DVM. It was developed as a solution to the high computation and memory usage problem of DVM. It is one of the most used methods for DVM. It performs well for linear DVM. In summary, SMO is based on solving the variables to be optimized in pairs in subspaces and solving the pairs formed with different combinations. (Akpinar, 2021)

Decision trees are one of the widely used machine algorithms in regression and assumption problems. Basically, it aims to reach according to the answers given to yes or no questions. The decision tree classifier starts with the root node and has a decision node and a leaf node. Decision nodes are used for decision making or classification and have multiple nodes and leaf nodes are outputs of decision nodes (Shaik Zuber et al.).

RNN (Recurrent Neural Network) is a type of artificial neural network that can be used in audio processing. RNN is designed to process the data sequence with its recursive connections between hidden layer activations in neighboring time steps (Schuster and Paliwal, 1997). In voice processing applications, RNN is often used for operations such as text-to-speech conversion, voice recognition, and voice customization. In our study, we used 3-layer LSTM, which is a type of RNN, together with ModelCheckpoint and ReduceLROnPlateu methods.

The CNN (Convolutional Neural Network) algorithm is specially designed for tasks such as image recognition and classification. Today, it is frequently used in applications such as sound processing. This algorithm takes audio

signals as input and tries to guess which words or phrases are present in the audio. In our study, 6-layer CNN was used together with ModelCheckpoint and ReduceLROnPlateu methods.

RESULTS

The obtained results from the model development stage indicate promising findings. First of all, different feature extraction methods were applied including Root Mean Square Energy (RMSE), Zero Crossing Rate (ZCR), Chroma, Mel Frequency Cepstral Coefficients (MFCC), Spectral Bandwith, Spectral Centroid properties for understanding speech-free mood prediction. There were different datasets (Ravdess, Save, Tess, Crema-D) combined for modelling and the whole dataset contained voiced in German and English by 121 different people in total. Moreover, the datasets consist of audio files in wav format containing 7 emotional states as happy, sad, angry, disgusted, scared, surprised and neutral.

These emotional states has also been used as labels within this combined dataset. The mentioned features were extracted from the audio files and classification models were developed to predict the correct labels using the extracted features. The classification results as F1 score have been 0.68 for Support Vector Machines (shown in Table 1), 0.63 for Random Forest Classification (shown in Table 2), 0.71 for LSTM (shown in Table 3) and 0.74 for Convolutional Neural Networks (shown in Table 4).

DISCUSSION

This modelling study examines intelligent voice emotion recognition systems as opposed to conventional methods that are widely used interview techniques in human resources. One of the major needs in this domain, has been to provide an objective and automatic process to reduce the time and human resource spent on this domain. Our current proposal fulfils this requirement since it reduces the analysis and reporting of the whole session within less than a minute.

Table 1. SVM classification report.

Emotion	Precision	Recall	F1-Score	Support
Angry	0.75	0.81	0.78	1923
Disgust	0.59	0.62	0.61	1923
Fear	0.70	0.56	0.62	1923
Happy	0.65	0.59	0.62	1923
Neutral	0.65	0.69	0.67	1895
Sad	0.66	0.73	0.69	1923
Surprise	0.86	0.88	0.87	652
Accuracy			0.68	12162
Macro Avg	0.70	0.70	0.70	12162
Weighted Avg	0.68	0.68	0.68	12162

Table 2. Random forest classification report.

Emotion	Precision	Recall	F1-Score	Support
Angry	0.66	0.81	0.73	1923
Disgust	0.55	0.53	0.54	1923
Fear	0.73	0.45	0.55	1923
Happy	0.60	0.52	0.56	1923
Neutral	0.59	0.67	0.63	1895
Sad	0.61	0.72	0.66	1923
Surprise	0.86	0.84	0.85	652
Accuracy			0.63	12162
Macro Avg	0.66	0.65	0.65	12162
Weighted Avg	0.64	0.63	0.62	12162

Table 3. LSTM classification report.

Emotion	Precision	Recall	F1-Score	Support
Angry	0.74	0.78	0.76	1923
Disgust	0.69	0.63	0.66	1923
Fear	0.67	0.66	0.66	1923
Happy	0.62	0.64	0.63	1923
Neutral	0.73	0.74	0.74	1895
Sad	0.73	0.74	0.74	1923
Surprise	0.84	0.84	0.84	652
Accuracy			0.71	12162
Macro Avg	0.72	0.72	0.72	12162
Weighted Avg	0.70	0.71	0.70	12162

Table 4. CNN classification report.

Emotion	Precision	Recall	F1-Score	Support
Angry	0.81	0.80	0.81	1923
Disgust	0.64	0.75	0.69	1923
Fear	0.74	0.65	0.70	1923
Happy	0.68	0.73	0.70	1923
Neutral	0.77	0.73	0.75	1895
Sad	0.76	0.72	0.74	1923
Surprise	0.92	0.92	0.92	652
Accuracy			0.74	12162
Macro Avg	0.76	0.76	0.76	12162
Weighted Avg	0.74	0.74	0.74	12162

On the other hand, speech emotion recognition frameworks typically consist of three major components: categorization, feature selection, and feature extraction. Examining in depth the algorithms and characteristics utilized in this project's execution. These components directly correspond to the ones in the relevant academic literature. However, new methods might be applied to provide more fruitful grounds regarding the modelling.

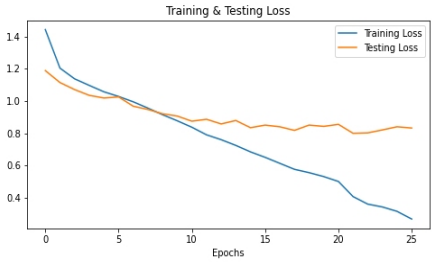

Figure 5: CNN accuracy.

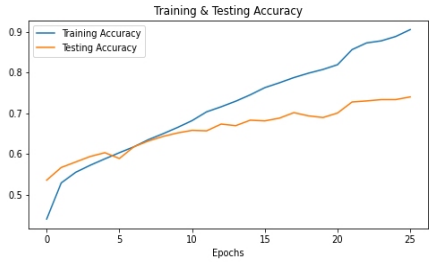

Figure 6: CNN loss.

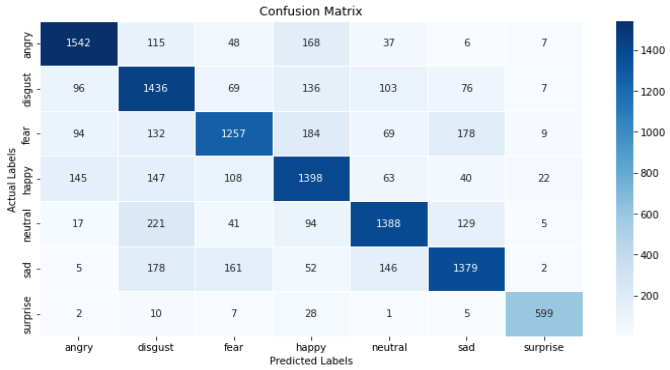

Figure 7: CNN confusion matrix.

Within the scope of this study, four common audio processing methods (including rmse, zrcr, chroma, and mfcc) were determined for distinguishing characteristics in speech emotion analysis. For better modelling outputs, new extracted features are necessary to provide results with higher scores. So far, the 6-layered CNN model has provided the highest output among the developed models with a success rate of 74%.

Lastly, as mentioned within the manuscript, four distinct public data sets were utilized for the research. As a consequence of analyzing the data sets, it has been shown that the success rate of mood analysis may differ depending on the spoken language. Thus, it seems that there should be other languages integrated into this model. We are planning to develop a national database

for this that could also be used in the research domains such as understanding the effects of neurophysiological signals.

CONCLUSION

In this study, intelligent systems for speech emotion recognition are examined and a very fundamental model has been developed. The main contribution of this study has been the developed models with different models on the combined datasets. The provided conclusion as a result of this study is as follows: Basically, speech emotion recognition architectures consist of three main parts including classification, selection of features, and extraction of features. As a result, it was understood that rmse, zrcr, chroma, and mfcc are distinctive features in speech emotion analysis. 4 different data sets were used in the project. As a result of the analysed data sets, it has been understood that the success rate in mood analysis may vary according to the spoken language. Thus, regarding the major limitation of this study, new spoken languages should be added to the combination of these datasets to provide a more realistic model for the use of human resources during the interviews, meanwhile one of the major challenges will be with respect to the increasing the performance metrics to reach a more acceptable solution.

REFERENCES

Betül Akpınar, Adaptif Sıralı Minimal Optimizasyon ile Destek Vektör Makinesi, (20, Kasım, 2021).

Bhavan, A., Chauhan, P., & Shah, R. R. (2019). Bagged support vector machines for emotion recognition from speech. Knowledge-Based Systems, 184, 104886.

Bidirectional recurrent neural networks M. Schuster, K. K. Paliwal.

Convolutional Neural Network (CNN) Based Speech-Emotion Recognition. Alif Bin Abdul Qayyum, Asiful Arefeen*, Celia Shahnaz.

Detection and analysis of emotion recognition from speech signals using Decision Tree and comparing with Support Vector Machine Shaik Zuber and K. Vidhya.

GÖKALP, S., & AYDIN, İ. (2021). Farklı Derin Sinir Ağı Modellerinin Duygu Tanımadaki Performanslarının Karşılaştırılması.

Goh C, Leon K (2009) Robust computer voice recognition using improved MFCC algorithm. In: Proceedings of the 2009 international conference on new trends in information and service science, IEEE, pp. 835–840. 22.

Gold B, Morgan N, Ellis D (2011) Speech and audio signal processing: processing and perception of speech and music. Wiley, New Jersey.

Huang, K. Y., Wu, C. H., & Su, M. H. (2019). Attention-based convolutional neural network and long short-term memory for short-term detection of mood disorders based on elicited speech responses. Pattern Recognition, 88, 668–678.

Konuşmadan Duygu Tanıma Üzerine Detaylı bir İnceleme: Özellikler ve Sınıflandırma Metotları Emel Çolakoğlu Serhat Hızlısoy Recep Sinan Arslan.

Langari, S., Marvi, H., & Zahedi, M. (2020). Efficient speech emotion recognition using modified feature extraction. Informatics in Medicine Unlocked, 20, 100424.

Nagesh Singh Chauhan, Naive Bayes, 22, Kasım, 2021).

Pan, Y., Shen, P., & Shen, L. (2012). Speech emotion recognition using support vector machine. International Journal of Smart Home, 6(2), 101–108.

Sethu, Vidhyasaharan and Epps, Julienand Ambikairajah, Eliathamby, "Speech BasedEmotionRecognition," pp. 197–228, September 2015.

The Application of Capsule Neural Network Based CNN for Speech Emotion Recognition Xin-Cheng Wen Kun-Hong Liu Wei-Ming Zhang Kai Jiang.
Wang, K., Su, G., Liu, L., & Wang, S. (2020). Wavelet packet analysis for speaker-independent emotion recognition. Neurocomputing, 398, 257–264.
Yao, Z., Wang, Z., Liu, W., Liu, Y., & Pan, J. (2020). Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLDRNN. Speech Communication, 120, 11–19.