Available online at www.dergipark.org.tr/en

INTERNATIONAL ADVANCED RESEARCHES and ENGINEERING JOURNAL

Journal homepage: www.dergipark.org.tr/en/pub/iarej

e-ISSN: 2618-575X

International Open Access

Volume 08
Issue 01

April, 2024

Research Article

A Modified MFCC-based deep learning method for emotion classification from speech

Fatih Şengül [a] ***and Sıtkı Akkaya*** [b,]

[a] *Department of Defense Technologies, Institute Of Graduate Studies, Sivas University of Science and Technology, Sivas 58100, Turkey*
[b] *Department of Electrical and Electronics Engineering, Faculty of Engineering and Natural Sciences, Sivas University of Science and Technology, Sivas 58100, Turkey*

ARTICLE INFO

Article history:
Received 9 October 2023
Accepted 14 March 2024
Published 20 April 2024

Keywords:
Deep learning
Emotion recognition system
Mel frequency cepstral coefficients (MFCCs)
Signal processing

ABSTRACT

Speech, which is one of the most effective methods of communication, varies according to the emotions experienced by people and includes not only vocabulary but also information about emotions. With developing technologies, human-machine interaction is also improving. Emotional information to be extracted from voice signals is valuable for this interaction. For these reasons, studies on emotion recognition systems are increasing. In this study, sentiment analysis is performed using the Toronto Emotional Speech Set (TESS) created by the University of Toronto. The voice data in the dataset is first preprocessed and then a new CNN-based deep learning method is compared. The voice files in the TESS dataset first yielded feature maps using the Mel Frequency Cepstral Coefficient (MFCC) method, and then classification was performed with this method based on the proposed neural network model. Separate models have been created with CNN and LSTM (Long Short-Term Memory) models for the classification process. The experiments show that the MFCC-applied CNN model achieves a better result with an accuracy of 99.5% than the existing methods for the classification of voice signals. The accuracy value of the CNN model shows that the proposed CNN model can be used for emotion classification from human voice data.

1. Introduction

One of the most effective ways of communicating information is the speech. Speech includes verbal content as well as gestures, facial expressions, intonation, and emotional content. Emotional state information to be acquired during speech is also a part of communication. During a speech, it is possible to infer a person's state of mind (emotion), gender, attitude, dialect, and so on. Today, with the development of web and mobile applications, human-machine communication has increased. Voice channels are now used in human-machine communication [1]. Increasingly, voice assistance tools are taking a part in answering questions and fulfilling requests instantly and correctly. as they become more common in our daily interactions. For example, Virtual Personal Assistants (VPAs) usage such as Cortona, Apple's Siri, Google Assistant, and Amazon Alexa is increasing [2]. Although these Virtual Personal Assistants understand the necessary commands by inferring words, they are not good enough at understanding people's emotions and reacting accordingly. Thence, emotion recognition is becoming an investigative field for computer science [3]. People express their emotions using their language. However, in addition to this, qualities such as intonation and speech rate during speech can also be processed and analyzed using signal processing and audio processing [4]. In emotion analysis studies, two approaches are generally applied in the data collection phase. The first one is to attach some sensors to the appropriate parts of the body that give clues about the emotional state. In this approach, data is collected through the interaction of the human body. The other approach is that instead of human-body interaction, human outputs such as sounds or movements are captured using a recording device. The collected data is processed and analyzed for mood [3].

Numerous studies in the literature focus on voice-based sentiment analysis. Some studies on the TESS dataset are as

* Corresponding author. Tel.: *+90 346 217 00 00*; Fax: +90 346 219 16 78.
E-mail addresses: tr.fatih.sengul@gmail.com *(Fatih Şengül)*, sakkaya@sivas.edu.tr *(Sıtkı Akkaya)*
ORCID: 0000-0001-5865-7476 *(Fatih Şengül)*, 0000-0002-3257-7838 *(Sıtkı Akkaya)*
DOI: 10.35860/iarej.1373333

follows. One of them, Venkataramanan et al. made use of The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset in their 2019 research. After preprocessing the audio signals, they performed experiments by creating a four-layer 2D-CNN architecture. Their experiments resulted in the highest accuracy of 68% [2]. In their 2022 research, Donuk et al. introduced a model by designing an LSTM architecture. To assess its performance, they utilized a dataset that combined both the RAVDESS and TESS datasets. The model demonstrated an impressive 88.92% accuracy in classifying eight distinct emotions [5]. In another study, Akinpelu et al. used the Erling Emotional Speech Database (EMO-DB) dataset and the TESS dataset together. As a result of the experiments, they obtained 97.2% accuracy with TESS data and 94.3% accuracy with the EMO-DB dataset [6]. Patel et al. in their 2020 study using the TESS dataset obtained a remarkable accuracy rate of 96% using an autoencoder [7]. In 2021, a speech-to-emotion recognition method using the TESS and RAVDESS datasets is proposed by Asiya et al. In their study, they implemented a deep learning-based classification model of emotions produced by speech based on voice data such as mel spectrogram, chromatogram, and Mel Frequency Cepstral Coefficient (MFCC). That model achieved 89% accuracy using RAVDESS and TESS datasets and various data augmentation techniques [8]. In their study conducted in 2022, Gokilavani et al. performed sentiment analysis with CREMA-D, RAVDNESS, and TESS datasets. They achieved 96% accuracy for the Ravdness dataset, 99% accuracy for the TESS dataset, and 84% accuracy for the Crema-D dataset in their models using the CNN model [9].

In this study, signal processing methods have been used first to classify emotion from voice signals. Different models have been created with deep learning methods to classify the preprocessed voice signals and the most successful method is proposed as a decision support system.

2. Methodology

The TESS (Toronto emotional speech set) dataset has been used to classify emotion from human voice signals [10]. With the proposed model, human voices in the dataset are classified. In this section, in addition to a review of the dataset used, a detailed analysis of MFCC, the method utilized for the extraction features from voice data, and the deep learning models that are developed for the classification of voice data are presented.

2.1 Dataset

This study focuses on using the TESS dataset to classify emotions from voice signals. The TESS dataset is considered one of the largest datasets of human speech sounds available and has a balanced mix of male and female voice recordings. This balance is advantageous in the training step since other datasets may contain only one gender, which can lead to biased results. The TESS dataset contains a total of 2800 voice files collected from people aged between 26 and 64 [11]. The dataset used in this study includes voice recordings of both male and female speakers expressing seven different emotions in Table 1 which provides an overview of the number of files for each emotion and speaker gender. To showcase the voice waveforms utilized in the study, Figure 1 illustrates some examples. The objective of this study is to classify emotions from voice signals accurately. By utilizing the TESS dataset, which is more balanced than other datasets, the study aims to achieve more reliable results. This study has promising results to foster the advancement of emotion recognition systems applicable in various domains, including affective computing, healthcare, and education.

2.2 MFCC Feature Extraction

In order to solve a problem with machine learning techniques, we need to have the appropriate attributes. However, we may not always have attributes that we can directly use in the problem we are addressing. In such cases, attributes need to be extracted from the data. "Signal Processing" deals with time series [12]. "Image Processing" is the science that deals with visual data such as photos and videos [13]. "Pattern recognition" is the science that aims to extract features from all kinds of signals, which can include both time series and images. When we examine today's studies, it is seen that feature extraction is performed with many different methods and tools. Optimizing the level of features is critical to achieving the highest accuracy in solving the problem [14].

Table 1. Distribution of voice files in the dataset

Gender and Emotion	The Number of Sound Recordings
Angry - Male	200
Disgust - Male	200
Fear - Male	200
Happy - Male	200
Neutral - Male	200
Pleasant Surprise - Male	200
Sad - Male	200
Angry - Female	200
Disgust - Female	200
Fear - Female	200
Happy - Female	200
Neutral - Female	200
Pleasant Surprise - Female	200
Sad - Female	200

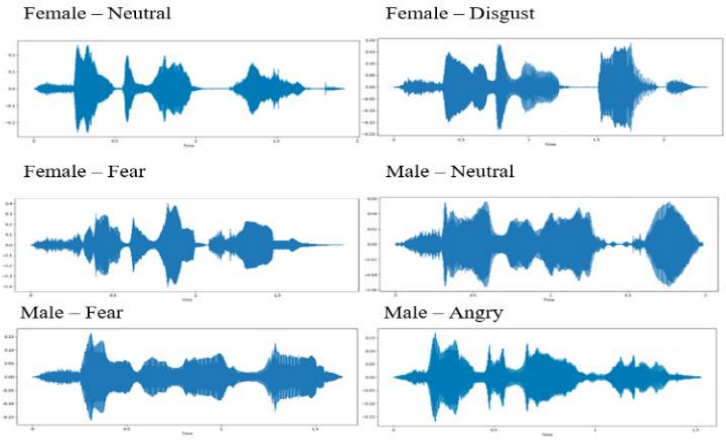

Figure 1. Representation of human voices as a signal

Utilizing too many attributes can lead to increased costs in the machine-learning process. Therefore, it is essential to carefully select the most relevant features for effectiveness. Selecting the right attributes for the purpose is also very important. In general, two categories of features are used in emotion recognition from speech: prosodic features and vocal tract system features. The first category is derived from prosodic data such as Pitch, Energy, and Duration. The second category is related to the voice path, which includes Cepstrum coefficients such as MFCC, Linear Predicted Cepstrum Coefficients (LPCC), Formants, and Discrete Fourier Transform (DFT) harmonics [15].

This study obtains feature maps of human voices using the MFCC method. MFCC is one of the commonly used techniques in the literature for the extraction of features from audio signals. Before classifying human voices in the TESS dataset, preprocessing is crucial to improve the success rate of machine learning methods. For this reason, the MFCC method is used to extract features from voice recordings. In 1980, Davis and Mermelstein were the first to use the MFCCs method [16]. This method is based on human mimics of the way the human ear perceives sound. The human ear sensitivity is linear up to 1 kHz and linear for higher values continues logarithmically. The transition from the real frequency unit Hertz to the frequency unit Mel is provided by (1) [15].

$$mel(f) = 2595 * \log\left(1 + \frac{f}{700}\right) \quad (1)$$

The extraction of features from human vocalizations in the TESS dataset was performed utilizing the Librosa library. This library is developed in the Python programming language. This library is used to perform operations on signal recordings [17]. In this study, the Mel-frequency Cepstrum Coefficient (MFCC) method is used to obtain feature maps of human voices. MFCC is a commonly used method in audio signal processing for feature extraction. Before classifying the human voices in the TESS dataset, it is very important to perform preprocessing to increase the success rate of machine learning methods. For this reason, the MFCC method is used to extract features from the voice recordings. Davis and Mermelstein introduced the MFCC method, which has been found widespread use in speech and voice signal processing, during the 1980s [16]. The MFCC algorithm involves dividing the audio stream into smaller frames using a Hamming window. Spectrums of the frames are then calculated using the Fast Fourier Transform (FFT) and weighted using a Mel scale-based filter bank [18]. Lastly, the Logarithm and Discrete Cosine Transform are applied to calculate the MFCC vector [19]. Figure 2 illustrates the feature processing steps involved in the computation of MFCCs.

2.3 Deep Neural Network (DNN) Models

In these models, the model performance can be influenced by the size of the data being used. Handling large datasets can impact the efficiency and effectiveness of the model. Working with small or flat data can lead to overfitting problems for models. To address this challenge and improve the model's performance through effective learning, data augmentation methods are employed. [20]. This can be done by using noise addition, time shifting, and changing speed methods in voice data. In this study, data augmentation is applied using noise addition and time-shifting methods to overcome overfitting and to better train the model. Preprocessing is of great importance for high classification success with machine learning. To this end, the MFCC method has been used to obtain feature maps of voice signals including human voices in the dataset.

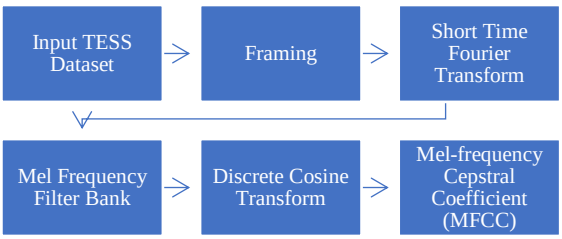

Figure 2. MFCC feature processing steps

Subsequently, these feature maps are fed into the proposed deep models for classification. The classification process has been performed using both CNN and LSTM methods.

The LSTM approach is an extension of the RNN (Recurrent Neural Networks) model and architecture that provides a longer memory span. While RNNs have limited "short-term memory" that utilizes past knowledge in the current neural network, the LSTM method leverages this prior knowledge effectively [21]. In LSTM method artificial neural network structures, the output signal produced in the hidden layer is initialized and is used as one of the values in the next input. The LSTM method is a versatile technique used in several implementations, like speech recognition, anomaly detection in time-series data, handwriting recognition, grammar learning, and music composition [22]. In this study, after obtaining the feature maps from the TESS dataset, the LSTM method is used for classification. The developed LSTM model consists of Dense, Dropout, and Flatten layers and is shown in Table 2. The ReLU layer is used for activation on the obtained feature maps, and the Dropout layer is used to prevent overfitting. Softmax is the preferred choice for classification in the LSTM method. The architectural representation of the LSTM model for emotion classification on speech sounds is illustrated in Figure 3.

The LSTM model is composed of 1 LSTM, 2 Dropout, 1 Flatten, 2 Dense, and 1 Softmax layers. Table 2 shows the parameter, layer, and output shape information of the LSTM model. In deep learning models, the term output shape refers to the ability to efficiently process data sets, which often have variable dimensions. When specifying the dimensions of the output shape, the term 'None' implies that it is flexible depending on the dimensions of the input data.

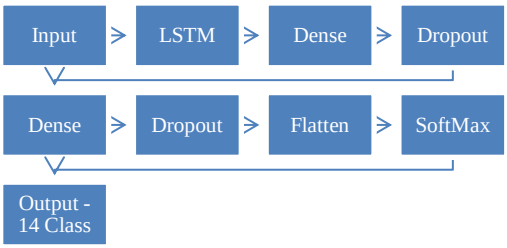

Figure 3. Architectural representation of the proposed LSTM Model

Table 2. Layers and parameters in LSTM model

Layers	Output shape	Activation function	Parameter
LSTM	(None,123)	-	61500
Dense	(None, 64)	Relu	7936
Dropout	(None, 64)	-	0
Dense	(None, 32)	Relu	2080
Dropout	(None, 32)	-	0
Flatten	(None, 32)	-	0
Dense	(None, 14)	-	462

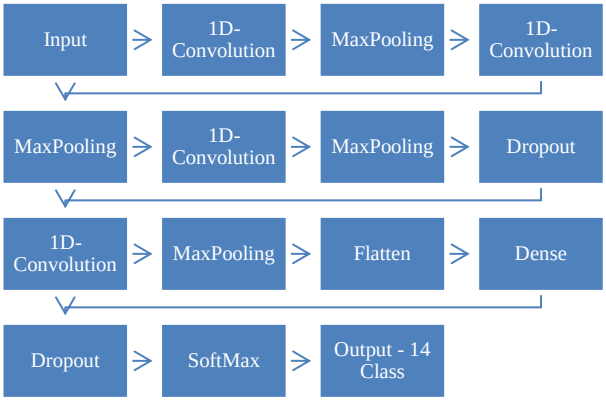

Figure 4. Architectural representation of the proposed CNN Model

One method used for classification is CNN. CNNs have gained widespread adoption in various domains, including image processing, text processing, and audio processing. Their versatile features make them applicable regardless of the number of dimensions [23]. The utilization of 2D-CNN and 3D-CNN has been predominantly applied for image classification purposes, whereas one-dimensional CNNs are commonly employed in various solutions such as signal analysis [24]. In the proposed deep network model for emotion classification from human voices in the TESS dataset, 1D-CNN is used. The proposed model comprises of a 1D-CNN layer, 1D-MaxPooling layer, Dropout layer, Flatten layer, and Dense layer with Softmax activation. The resulting feature maps activate with Relu, and the computational complexity decreases through the utilization of the Pooling layer. During the training process, the Dropout layer has been employed to disable certain nodes in the network to prevent overfitting. The Flatten layer has been responsible for converting the data from a matrix format to a flattened format. The classification process has been accomplished through the utilization of the Softmax layer, which has generated probability values based on the input values. Whichever class is closer to the values obtained, the classification process is completed by placing the data in the relevant class. The CNN model architecture created for emotion classification on speech sounds is given in Figure 4.

When the CNN model is analyzed, it is seen that 4 1D-Convolution, 4 Maxpooling, 1 Dropout, 1 Flatten, 2 Dense, and 1 Softmax layers are implemented. The CNN model's parameters and layer information are presented in detail in Table 3.

3. Experimental Result

The Intel DevCloud cloud system has been utilized for preprocessing the voice signals, training the DNN model, and testing the models in this study. To expedite DNN model training and testing, the researchers utilized the Intel oneAPI framework.

Table 3. Layers and parameters in CNN model

Layers	Output shape	Activation function	Parameter
Conv1d	(None,162,256)	Relu	1536
Max_pooling1d	(None,81,256)	-	0
Conv1d	(None,81,256)	Relu	327936
Max_pooling1d	(None,41,256)	-	0
Conv1d	(None,41,128)	Relu	163968
Max_pooling1d	(None,21,128)	-	0
Dropout	(None,21,128)	-	0
Conv1d	(None,21,64)	Relu	41024
Max_pooling1d	(None,11,64)	-	0
Flatten	(None,704)	-	0
Dense	(None,32)	Relu	22560
Dropout	(None,32)	-	0
Dense	(None,14)	-	462

Furthermore, it should be noted that oneAPI is an open, cross-industry, standards-based programming model that supports multiple architectures and vendors. By providing a unified development experience across accelerator architectures, oneAPI aims to enhance application performance, boost productivity, and encourage innovation. The oneAPI platform ensures the improved code to profit by various hardware architectures, including GPUs, multi-core CPUs, or other hardware, all through singular sources [25]. This shortened the experiment times considerably. In the experiments, the TESS dataset is used for the training and testing stages. Different neural network models are created to perform emotion classification from voice signals using different architectures. Accomplishment evaluation of the models is compared on Accuracy, Recall (Sensitivity), Precision,and F-score.

Accuracy

The accuracy of a model is computed by the ratio of the number of correct predictions to the total number of predictions made on the entire dataset(2).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

The symbols TP, TN, FP, and FN state true positive, true negative, false positive, and false negative values, respectively.

Recall

Recall, called sensitivity or true positive rate, is a performance metric that represents the number of true positive predictions made by the model divided by the total number of actual positive instances in the dataset (3).

$$Recall = \frac{TP}{FN} \quad (3)$$

Precision

Precision is a performance metric as given by (4).

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

F Score

The F score is used to show the trade-off between sensitivity and recall. The F score is obtained by equation (5).

$$F\ Score = 2 \times \frac{precision \times recall}{precision+recall} \quad (5)$$

Equations (2), (3), (4), and (5) are utilized to compare the performance of different models. In classification problems, the prediction results can be summarized using a confusion matrix, which is a tabular representation. As shown in Table 4, the confusion matrix can be used to get an idea of the errors made by the classifier. The confusion matrix also gives insight into the error types made by the model.

The study conducted emotion classification by first obtaining feature maps through the MFCC method. The feature maps are then used as input to deep neural network models for classification. 90% and 10% of the dataset are used for training and for testing the models, respectively. The study employed two popular deep learning models, namely CNN and LSTM, for the classification task. The LSTM method, which is widely used in various applications such as sentiment analysis, text generation, and time series, proved to be highly accurate in sentiment analysis from voice. Table 5 provides a comprehensive summary of the LSTM-based model's performance, including precision, recall, and F score. Moreover, the study presented the confusion matrix of the classification test in Figure 5, demonstrating the model's ability to correctly classify samples into their respective emotion classes. Furthermore, the results suggest that the LSTM-based model has successfully achieved accurate emotion classification from voice signals, demonstrating its potential for practical applications in affective computing, healthcare, and education. The precision, recall, and F score provided in Table 5 indicate the model's strong performance in recognizing male and female voices expressing various emotions. Overall, the study's findings provide valuable insights into the effectiveness of deep learning methods, particularly LSTM, for emotion classification from voice signals

When Table 5 is examined for the model created with the LSTM method, it is shown that the most successful classes are Male-Neutral, Male-Sad, Female-Happy, and Female-Sad. The most unsuccessful class is Male - Happy.

Table 4. Confusion Matrix

	Positive Prediction	Negative Prediction
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

.

Table 5. Classification Evaluations Obtained In The LSTM Model

Classes	Recall	Precision	F score
Angry - Male	0.87	0.96	0.91
Disgust - Male	0.74	0.81	0.77
Fear - Male	0.98	0.91	0.95
Happy - Male	0.65	0.58	0.61
Neutral - Male	0.98	0.98	0.98
Pleasant Surprise - Male	0.77	0.72	0.74
Sad - Male	0.97	1.00	0.98
Angry - Female	0.90	0.98	0.94
Disgust - Female	0.95	0.92	0.94
Fear - Female	0.97	0.92	0.94
Happy - Female	0.97	1.00	0.98
Neutral - Female	0.91	1.00	0.95
Pleasant Surprise - Female	0.96	0.91	0.94
Sad - Female	0.97	0.98	0.98

The average accuracy for all classes for the LSTM model is 0.90. Figure 6 illustrates the accuracy and loss curves of the model created using the LSTM method.

When Figure 6 is examined, it becomes evident that the accuracy of the LSTM model is above 90% on the training data, indicating its success during the training phase. However, when examining the test data, the decrease in accuracy suggests that the model is less specialized for the test data, and its generalization ability is limited. Similarly, the loss value in the loss curve increases as it approaches the training data, indicating that the model is inclined to make more errors when applied to the test data.

The effectiveness of the CNN method in performing emotion analysis on voice signals underwent evaluation.

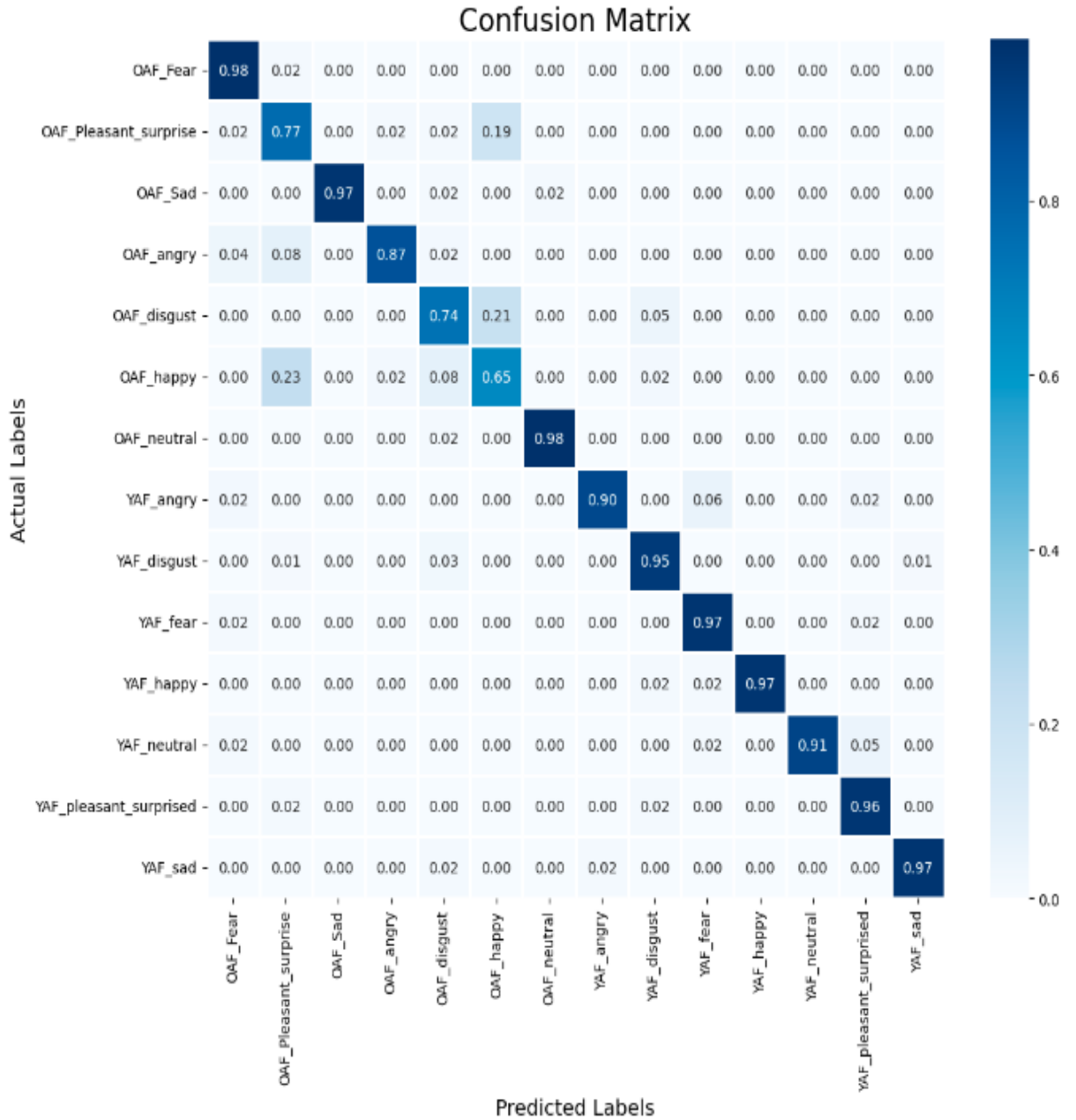

Figure 5. Confusion matrix obtained in the LSTM model.

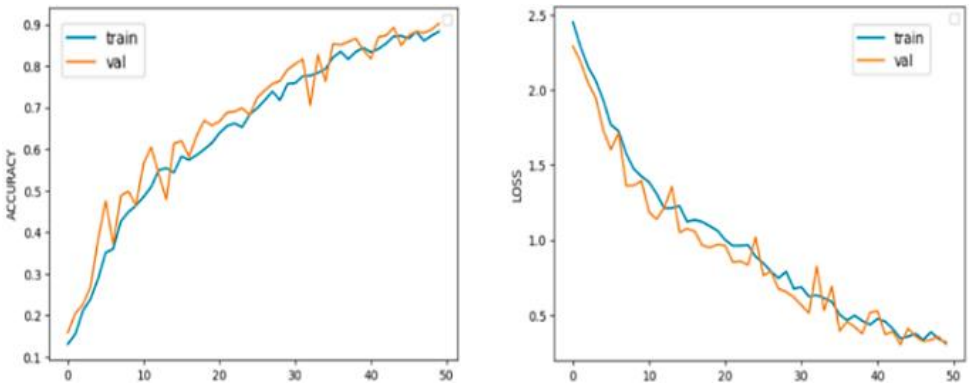

Figure 6. Accuracy and loss curves for the LSTM model

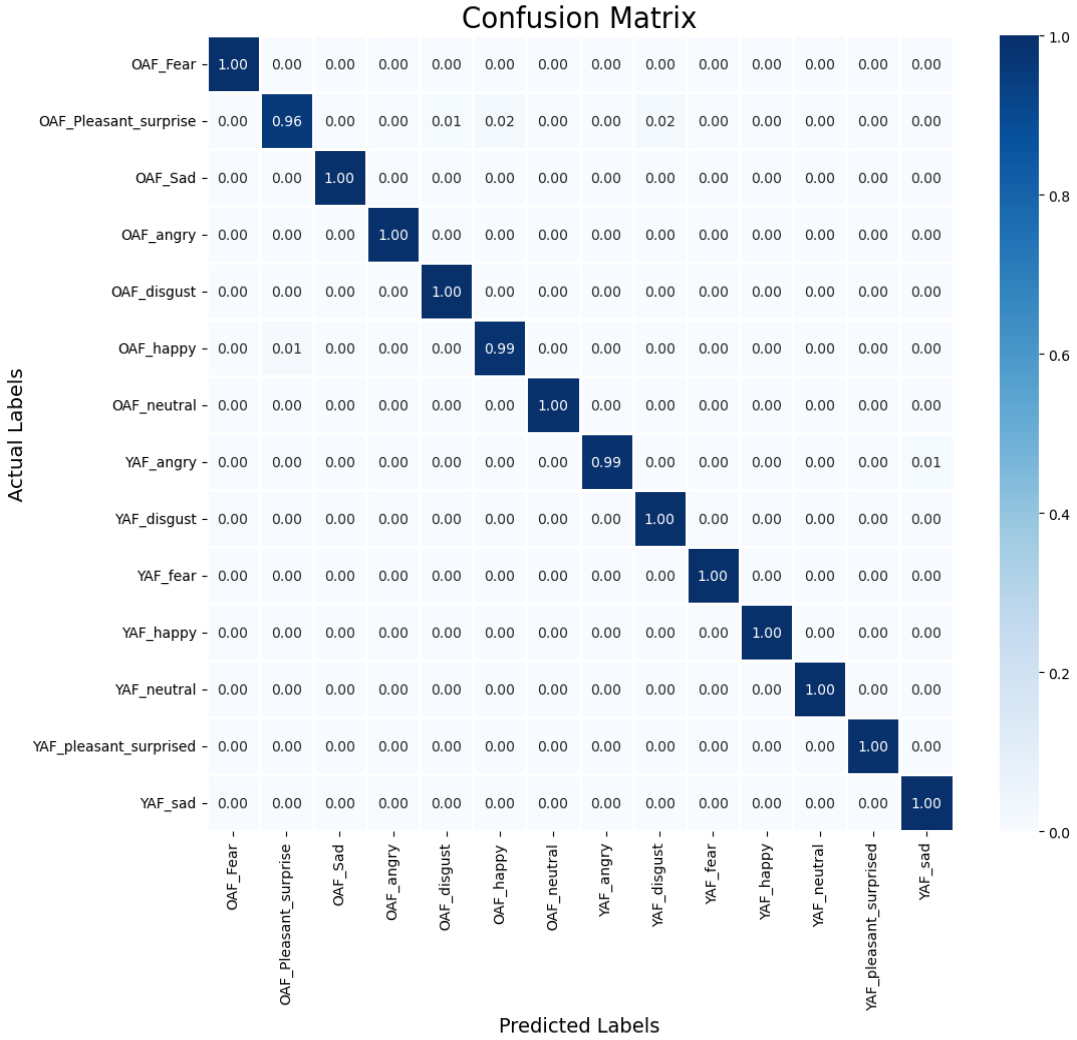

Figure 7. Confusion matrix obtained in the CNN model

Following the creation of a model using the CNN method, testing is conducted, and the resulting confusion matrix is presented in Figure 7.

Moreover, the accuracy, sensitivity, and F Score evaluations of the CNN method, which achieved high accuracy rates, are given in Table 6.

When the test results of the CNN model are analyzed in Table 6 and the confusion matrix is analyzed in Figure 7, it is seen that 10 out of 14 classes achieved full success. The remaining 4 classes achieved results close to full success.

In the deep neural network model created using the CNN method, the average accuracy value in all classes is 0.995. Figure 8 illustrates the accuracy and loss curves of the CNN method, which exhibited superior classification performance compared to the LSTM method.

In the training phase, as illustrated in Figure 8, the CNN model exhibits an accuracy of over 99% on the training data. This indicates superior performance when compared to the LSTM model.

Table 6. Classification Evaluations Obtained In The CNN Model

Classes	Recall	Precision	F score
Angry - Male	0.99	1.00	1.00
Disgust - Male	0.99	0.96	0.97
Fear - Male	1.00	1.00	1.00
Happy - Male	0.96	0.99	0.98
Neutral - Male	1.00	1.00	1.00
Pleasant Surprise - Male	0.96	0.93	0.95
Sad - Male	0.99	1.00	1.00
Angry - Female	0.99	1.00	1.00
Disgust - Female	0.99	0.99	0.99
Fear - Female	1.00	1.00	1.00
Happy - Female	1.00	1.00	1.00
Neutral - Female	1.00	1.00	1.00
Pleasant Surprise - Female	0.99	1.00	1.00
Sad - Female	1.00	1.00	1.00

Furthermore, the consistent decrease in the loss value during training suggests that the CNN model effectively mitigates overfitting. Lastly, the accuracy value remains relatively stable when applied to the test data, indicating the model's strong generalization ability. These results underscore the CNN model's superior efficacy compared to the LSTM model.

To assess the effectiveness of the our approach, we compare the performance with similar studies conducted on the TESS dataset in the existing literature. A comparative analysis of studies providing deep learning models for speech emotion recognition is presented in Table 7.

It is evident that our model outperforms most of the previous studies on the same datasets. In deep network architectures, the preprocessing and the amount of data used during training are very important for accuracy. In this study, feature extraction took place using the MFCC method. In addition, data augmentation operations are applied to diversify the dataset. This helps to reduce overfitting and increase the generalization ability of the model. In addition to these factors, the architecture of our CNN model is also effective in its high success. The model encloses three convolutional layers, two max-pooling layers, two fully connected layers, and a dropout layer. The local features are extracted from the voice data by convolutional layers, while the size of the feature maps is decreased by the max pooling layers. The other layers, fully connected layers, enable the classification of speech into predefined categories by learning complex features from the extracted data. For these reasons, our results are more successful than previous studies despite the larger number of classes. The CNN model performance is remarkable as it achieves error-free classification results in 10 out of 14 emotion and gender classes. In particular, in male classification, the model showed error-free classification in various emotional states and achieved F1 scores of 1.00 in categories such as neutral, fear, angry, and sad.

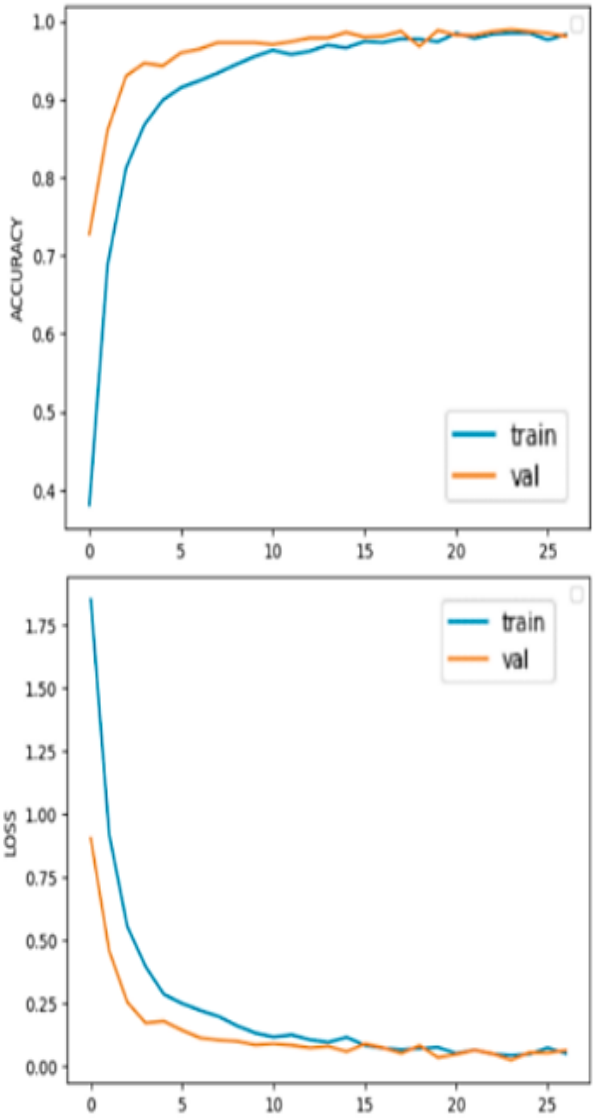

Figure 8. Accuracy and loss curves for the CNN model

Similarly, in the female classification, the model demonstrated high accuracy with error-free classification in emotions such as neutral, angry, happy, fear, pleasant surprise, and sad, recording an F1 score of 1.00. This accuracy value demonstrates the robustness and adaptability of the model in discriminating complex emotional nuances and gender categories across a wide range of classes.

Table 7. Comparative study on TESS dataset

Reference	Method	No. of Class	Accuracy (%)
Venkataramanan et al. [2]	CNN	14	68
Donuk et al. [5]	LSTM	8	88.92
Akinpelu et al. [6]	MLP&SVM	6	97.20
Patel et al. [7]	CNN	7	96
Asiya et al. [8]	CNN	8	89
Gokilayani et al. [9]	CNN	7	99
Proposed method	**A new CNN model**	***14***	***99.5***

4. Conclusions

Sentiment analysis has gained immense popularity in recent years with companies, organizations, and governments using it to understand people's reactions to various topics. Sentiment analysis can be carried out for different types of data, comprising textual and auditory content. In this study, sentiment classification is conducted on voice signals using the TESS dataset. This dataset consists of 14 classes, including 7 emotions and gender classification (male-female). Short-Time Fourier Transform and MFCC methods are used for feature extraction. Diversity in the dataset is ensured by data augmentation. This helps to lessen overlearning and enhance the model generalization ability. After data augmentation, deep neural network models are trained for classification. While 90% of the dataset was used for training, 10% was used for testing. The performances of two deep learning models, namely CNN and LSTM, are compared. The LSTM model achieved an accuracy of 90% with 14 classifications, while the CNN model outperformed with a 99.5% accuracy rate, achieving complete success in 10 out of 14 classes, with near-complete success in the remaining 4 classes. The proposed CNN model aims to classify gender and emotion from human voices in the TESS dataset. The model's architecture embodies three convolutional layers, two max-pooling layers, two fully connected layers, and a dropout layer. The convolutional layers extract increasingly complex features from the data, while the max pooling layers help to abates the dimension of the data and improve the model's generalization ability. The fully connected layers learn a complex representation of the data and map it to the output classes. The dropout layer assures a prevention of overfitting by randomly dropping out some of the neurons during training. Comparing the success rate of the proposed model with other similar works in the literature yields that the proposed model achieves a higher accuracy rate, which is one of the highest rates reported in recent studies using the TESS dataset. Therefore, the results of proposed method demonstrate the effectiveness of deep learning methods, especially the CNN model, in achieving high accuracy rates in sentiment analysis using voice data.

Declaration

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The authors also declared that this article is original, was prepared in accordance with international publication and research ethics, and ethical committee permission or any special permission is not required.

Author Contributions

Fatih Şengül conceptualized and conducted the study, authored the manuscript, and actively contributed to the entire research process. Sıtkı Akkaya provided valuable guidance and supervisory support throughout the study, offering insightful feedback and contributing to the refinement of the research methodology. Both authors collaboratively wrote and reviewed the manuscript. The authors are listed in alphabetical order.

Acknowledgment

We express sincere gratitude to Intel for providing the essential Intel oneAPI toolkit, pivotal in our sentiment analysis research. Access to DevCloud cloud system, generously granted by Intel, facilitated efficient model training and testing, playing a vital role in achieving accurate results. The support and resources from Intel significantly enhanced the quality and effectiveness of our work in sentiment analysis.

Nomenclature

MFCC : Mel Frequency Cepstral Coefficient
LSTM : Long Short-Term Memory
CNN : Convolutional Neural Network
TESS : Toronto emotional speech set

References

1. Liu, K., Wang, D., Wu, D., Liu, Y., and Feng, J., *Speech emotion recognition via multi-level attention network.* IEEE Signal Processing Letters, 2022. **29**: p. 2278-2282.
2. Venkataramanan, K., and Rajamohan, H. R., *Emotion recognition from speech*. arXiv preprint, 2019. p. 1912-10458.
3. Aydin, M., Tuğrul, B., and Yilmaz, A. R., *Emotion Recognition System from Speech using Convolutional Neural Networks.* Computer Science, 2022. p. 137-143.
4. Xu, Y., *English speech recognition and evaluation of pronunciation quality using deep learning*. Mobile Information Systems, 2022. p. 1-12.
5. Donuk, K., and Hanbay, D., *Konuşma Duygu Tanıma için Akustik Özelliklere Dayalı LSTM Tabanlı Bir Yaklaşım*. Computer Science, 2022. **7**(2): p. 54-67.
6. Akinpelu, S., and Viriri, S., *Robust Feature Selection-Based Speech Emotion Classification Using Deep Transfer Learning*. Applied Sciences, 2022. **12**(16): 8265.
7. Patel, N., Patel, S., and Mankad, S. H., *Impact of autoencoder based compact representation on emotion detection from audio*. Journal of Ambient Intelligence and Humanized Computing, 2022. p. 1-19.
8. Asiya, U. A., and Kiran, V. K., *Speech Emotion Recognition-A Deep Learning Approach*, in Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC). 2021. Palladam, India: p. 867-871.
9. Gokilavani, M., Katakam, H., Basheer, S. A., and Srinivas, P. V. V. S., *Ravdness, crema-d, tess based algorithm for emotion recognition using speech*, in 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT). 2022. Tirunelveli, India: p. 1625-1631.

10. Pichora-Fuller, M. K., and Dupuis, K., *Toronto emotional speech set (TESS)*. *Scholars Portal Dataverse*. 2020. *1*.
11. Sun, C., Li, H., and Ma, L., *Speech emotion recognition based on improved masking EMD and convolutional recurrent neural network*. Frontiers in Psychology, 2023. **13**: 1075624.
12. Zhang, C., Mousavi, A. A., Masri, S. F., Gholipour, G., Yan, K., and Li, X., *Vibration feature extraction using signal processing techniques for structural health monitoring: A review*. Mechanical Systems and Signal Processing, 2022. **177**: 109175.
13. Zhang, Y., and Zheng, X., *Development of Image Processing Based on Deep Learning Algorithm*, in 2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC). 2022. Dalian, China: p. 1226-1228.
14. Jastrzebska, A., *Time series classification through visual pattern recognition.* Journal of King Saud University-Computer and Information Sciences, 2022. **34**(2): p. 134-142.
15. Kop, B. Ş., and Bayindir, L., *Bebek Ağlamalarının Makine Öğrenmesi Yöntemleriyle Sınıflandırılması*. Avrupa Bilim ve Teknoloji Dergisi, 2021. **27**: p. 784-791.
16. Davis, S., and Mermelstein, P., *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*. IEEE transactions on acoustics, speech, and signal processing, 1980. **28**(4): p. 357-366.
17. Choudhary, R. R., Meena, G., and Mohbey, K. K., *Speech emotion-based sentiment recognition using deep neural networks*, in Journal of Physics: Conference Series. IOP Publishing. 2022. p. 012003.
18. Yıldırım, M., *MFCC Yöntemi ve Önerilen Derin Model ile Çevresel Seslerin Otomatik Olarak Sınıflandırılması*. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 2022. **34**(1): p. 449-457.
19. Li, F., Liu, M., Zhao, Y., Kong, L., Dong, L., Liu, X., and Hui, M., *Feature extraction and classification of heart sound using 1D convolutional neural networks*. Eurasip Journal on Advances in Signal Processing, 2019. **2019**(1): p. 1-11.
20. Maharana, K., Mondal, S., and Nemade, B., *A review: Data pre-processing and data augmentation techniques*. Global Transitions Proceedings, 2022. **3**(1): p. 91-99.
21. Alpay, Ö., *LSTM mimarisi kullanarak USD/TRY fiyat tahmini*. Avrupa Bilim ve Teknoloji Dergisi, 2020. p. 452-456.
22. Priyadarshini, I., and Puri, V., *Mars weather data analysis using machine learning techniques*. Earth Science Informatics, 2021. **14**: p. 1885-1898.
23. Adem, K. and Kılıçarslan, S., *COVID-19 diagnosis prediction in emergency care patients using convolutional neural network*. Afyon Kocatepe Üniversitesi Fen Ve Mühendislik Bilimleri Dergisi, 2021. **21**(2): p. 300-309.
24. Liu, S., and Chen, M., *Wire Rope Defect Recognition Method Based on MFL Signal Analysis and 1D-CNNs*. Sensors, 2023. **23**(7): p. 3366.
25. Christgau, S., and Steinke, T., *Porting a legacy cuda stencil code to oneapi*, in 2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), 2020. New Orleans, LA, USA: p. 359-367.