Journal of Computer Science
https://dergipark.org.tr/en/pub/bbd

ISSN,e-ISSN: 2548-1304
Volume:IDAP-2022, pp:137-143, 2022
https://doi.org/ 10.53070/bbd.1174033
Research Paper

# Emotion Recognition System from Speech using Convolutional Neural Networks

Metehan Aydin[*1], Bulent Tugrul[1], Yilmaz Ar[1]

[1]Department of Computer Engineering, Ankara University, Ankara, Turkey

(eng.metehan.aydin@gmail.com, btugrul@eng.ankara.edu.tr, ar@ankara.edu.tr)

Received:Sep.08,2022 Accepted:Sep.16,2022 Published:Oct.10,2022

***Abstract***— Emotions can affect human behaviors directly. This situation makes people want to learn the emotion states of the other people they are in touch. The emotion state information can be used in lots of areas in order to improve efficiency. It is a challenging task and requires a wide working pipeline starting from data acquisition to classification. Today, many researchers work in order to recognize emotions using different techniques including text analyzing, body movement analyzing, facial expressions and voice. In this work, we proposed an approach for this problem. Our approach uses human voice and makes classification using a convolutional neural network. The paper explains how our recognizer pipeline is created and how it works in detail.

***Keywords :*** *emotion recognition, voice recognition, speech recognition, convolutional neural networks*

## 1. Introduction

Emotions have an important role in human life. They affect how humans think. This situation results that emotions affect human decisions and furthermore human behaviors (Andrade and Ariely, 2009). People who can recognize their emotions can be happier and more successful in their lives. On the other hand, people who can know the emotion state of the people they are in contact can set strong relations. This is also same for companies. The companies which understand the emotion state of their customers can achieve their goals more successfully. These situations make emotion detection a research field in computer science.

There are many applications of emotion recognition systems in real life ranging from education to production. Some applications of emotion recognition systems are given below.

- Education: Emotion states of students are important for learning process. In order to learn the emotion state of students in a lecture, their emotions can be recognized (Tonguç and Ozkara, 2020).
- Automotive: Emotion state of a driver can affect how he drives his vehicle on traffic. In order to learn his emotion state, emotion recognition is used (Zepf et al., 2020).
- Aerospace: Emotion state of a pilot and cabin crew affect a flight. In order to learn and improve their emotion states, emotion recognition is used (Roza and Postolache, 2019).
- Security: Emotional states of customers in banks or ATMs are important for security. In order to capture the emotion states of customers, emotion recognition is used (Saste and Jagdale, 2017).
- Production: Employee's emotions are a vital for productivity. If they don't feel good, their productivity can be decreased. In order to detect and improve their emotions, emotion detection can be used (Subhashini and Niveditha, 2015).

People express their emotions using their languages. But in addition to this, their bodies can give data about their emotions. These data can be obtained in their hearth rate, blood pressure, body language, faces and voice. Then, the data can be processed using image processing, signal processing, lexical analyzing, body movements analyzing and voice processing (Kołakowska et al., 2013).

In data acquisition, there can be two approach. The first one is that some sensors are connected to appropriate part of the body that gives clues about the emotion state (Kanjo et al., 2019). In this approach, data are collected with the interaction of human body. In order to detect an emotion, identical sensors must be connected to same parts of the body. The other approach is that, instead of interaction of human body, the outputs of a human like voice or movements are captured using a recorder (Sebe et al., 2006) Then, when emotion state is wanted to be

detected, a device with same capability of the data acquisition device is used for input. The input is processed and the emotion is detected.

In this work, our purpose is recognizing emotion state of a person. As we mentioned above, emotion state of people can be recognized with different ways. Our work focuses on emotion recognition from a human voice. The paper is structures as follows. In Section 2, we define out problem. Four related works are reviewed and given in Section 3. The dataset which is selected for this work is described in Section 4. In Section 5, the proposed approach is explained in details. Section 6 shows the results of the experiments and there is a conclusion in Section 7.

## 2. Problem Definition

Although emotion recognition is so important and have a wide area of usage, its applications generally not portable, they need some equipment in order to realize their function. The main reason of this situation is that in order to train their models, they use complicated sensors or complicated equipment. The equipment is generally expensive and this situation somehow blocks the usage of emotion detection systems or costs more than its needs. In addition, these, the applications which are not portable enough result that the applications can't be used effectively. In order to use them, user must be set up the system. Setting up the system is takes long time. This situation also restricts the usage area of emotion detection system.

The objective of our work is eliminating the drawbacks of the situations which is mentioned above. In order to eliminate them, we proposed to use a speech recognition system which is trained by a deep neural network. Speech recognition systems are widely used and a many researchers are contributing to development of it. In order to recognize a speech, there must be a device which has a microphone and a microprocessor. Today, most of people have smart phones or computers and all of these devices provides a microphone and a microprocessor. An emotion recognition system which can be executed with these devices can be accessible easily and will be portable enough.

In recent years, convolutional neural networks are a widely discussed and developing topic in computer science. They offer high accuracy rates and they have a potential to handle so many types of problems. We think that these advantages and strengths of convolutional neural networks can solve our problem effectively. The solution will be discussed in the next sections of the paper.

## 3. Related Works

There is many research about emotion recognition. The topic has an increasing popularity in computer science. In this section, we select four related works which are different from each other and explained them in an understandable manner.

### 3.1. Emotion Recognition by Speech Signals

In this work (Kwon et al, 2003), emotion recognition is implemented using voice data. The work is implemented using different features, different datasets and different classifiers. The results of the work are compared them and tried to select the one which has best performance.

In the work, features selection is handled in the following way. Log energy, formant, mel-band energies, and mel-frequency cepstral coefficients (MFCCs) are selected as the base features. They selected these features from works (Scherer, 1996; Tato et al., 2002) which are done before. Then, velocity/acceleration of pitch and MFCCs features were added in order to form feature streams. Totally, there are 15 features. In order to classify these features, different classifiers are used these classifiers are support vector machine (SVM), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and HMM classifiers.

The work is done in two different datasets and the results are given with respect to these databases. These datasets are text-independent SUSAS database, and the speaker-independent AIBO database. In SUSAS database, two different experiments are done. In the first experiment, there are two emotions. These are stressed and neutral. In this experiment, best results are taken using HMM based classification with %96.3 accuracy. In the second experiment, four classes which represents neutral, angry, Lombard and loud emotions are used. The best results of this experiment is taken by Gaussian SVM classifier with %70.1 accuracy. In AIBO database, the experiment is done using five classes. The classes are representing five different emotions, angry, bored, happy, sad, neutral. The best result is taken by Gaussian SVM classifier with the accuracy of %42.3.

### 3.2. Speech Emotion Classification and Recognition with different methods for Turkish Language

In this work (Bakır and Yuzkat, 2018), emotions are recognized using voice data. The language used in voices are Turkish. The purpose of the project is creating an automatic emotion recognition system from speech or voice.

The work uses Mel Frequency Cepstral Coefficients (MFCC) and Mel Frequency Discrete Wavelet Coefficients (MFDWC) feature extraction methods. In addition to these, Support Vector Machine (SVM) is used

in order to spectral features of the data. Then, these features are trained using different methods. These methods are Gauss Mixture Model (GMM), Artificial Neural Network (ANN), Dynamic Time Warping (DTW), Hidden Markov Model (HMM) and hybrid model (GMM with combined SVM).

The work is done using a database which is created for this work. The database has five emotions. These emotions are anger, fear, sadness, happiness and neutral. There are approximately 3000 samples in the database. These samples are taken from 25 females and 25 males. The language of these samples are Turkish. Two experiments are done. In each experiment, there are two categories; female and male. In the first experiment, features are extracted using SVM. The best result for male category is taken by HMM classifier with accuracy of %75.71 and the best result for female category is taken by HMM classifier with the accuracy of %77.63. In the second experiment, MFCC feature extraction method is used with five different feature vector. The best result for male category is taken by HMM classifier with the accuracy of %77.25 and the best result for female category is taken by HMM classifier with the accuracy of %75.68.

### 3.3. Speech Emotion Recognition using Convolutional and Recurrent Neural Networks

This work (Lim and Lee, 2016) recognize emotions from voice data. The purpose of the work is proposing a systematic approach in order to recognize emotions based on concatenated CNNs and RNNs without using any traditional hand-crafted features. In the results, the accuracies of these networks are classified.

In this work, the voice data is converted to two dimensional representation using Short Time Fourier Transform (STFT) with a frame size of 256 and 50% overlap. Then, these data is trained using CNNs, time distributed CNNs and Long Short Term Memory (LSTM) architectures.

The CNN architecture has two convolution and max-pooling layers for learning representation and last layer flattens and connects fully with 1024 nodes. The LSTM network has three sequential stacked layers and has input and output size of 1x128. The time distributed CNNs uses two additional sequential LSTM layers with 1024 nodes in addition to the CNN network used in this work. The Berlin database (Burkhardt et al., 2005) is used in this work. There are seven classes each represents the emotion; happy, boredom, sadness, disgust, fear, anger, neutral. For each of the network, one experiment is done. In the first experiment, The CNN network is used. The f-1 score of this experiment is 86.06 (±2.39). In the second experiment, the LSTM network is used and the f-1 score of this experiment 78.31 (±4.59). The last experiment is done using the time distributed LSTM with the f-1 score of 86.65 (±1.73).

### 3.4. Emotion Recognition in Speech Signal: Experimental Study, Development, and Application

In this work (Petrushin, 2000), emotion recognition is done from voice data. The purpose of the work is that developing a computer agent in order to recognize emotions. The agent is planned to use in a call center. The agent can be used for two purposes. First, they proposed that it can be better to inform call center operator about the emotion state of the customer. Second, e-mails taken from customers are desired to be sorted with their urgency state.

They select create 43 different features using statistical methods for each utterance. Then, they use RELIEF algorithm in order to select the features. Then, these features are trained using K-nearest neighbors, neural network and ensembles of neural network classifiers.

A dataset is created for the work. The data are collected from thirty people. Each people tells given sentences with five times happiness, anger, sadness, and fear and neutral emotions. Three experiments are done for the work. In the first experiment, k-nearest neighbor's method is used and the average accuracy is %55. In the second experiment, a neural network with two layers' network architecture and 8, 10, 14 elements input vectors is created. The average accuracy is %65. In the last experiment, an ensemble neural network is used in which ensemble size from 7 to 15. The average accuracy of the network is %70.

## 4. Dataset

The dataset of the proposed approach is created from two different speech datasets. These datasets are Ryerson Audio-Visual Database of Emotional Speech and Song dataset (RAVDESS) (Livingstone and Russo, 2018) and Crowd Sourced Emotional Multimodal Actors (CREMA-D) (Cao et al., 2014) dataset.

**Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) Dataset:** RAVDESS dataset contains 1440 samples. The samples are created from 24 (12 males and 12 females) different actors. Dataset contains 8 different emotions. These emotions are being neutral, calm, happy, sad, angry, fearful, disgust, surprised.

**Crowd Sourced Emotional Multimodal Actors (CREMA-D) Dataset:** CREMA-D dataset contains 7442 samples. The samples are created from 91 (48 males 43 females) different actors. Dataset contains 6 different emotions which are being neutral, happy, sad, angry, fearful, disgust.

The proposed approach merges these two datasets into one dataset. The merged dataset is going to have 6 emotions and these emotions contain being neutral, happy, sad, angry, fearful and disgust. Calm and surprise emotions from RAVDESS are discarded from the merged dataset. There are 384 total calm and surprised emotion samples in RAVDESS. Thus, the merged dataset will contain 8498 samples which are created from 115 (60 males 55 females) different actors.

## 5. Proposed Approach

This section explains how our proposed approach works. As it mentioned above, RAVDESS and CREMA-D datasets are selected as dataset. The work has two main steps. These steps are feature extraction and training.

In feature extraction, the features of each samples are extracted in order to use for training the model. The feature of a sample is extracted using five different techniques. These techniques are below.

1. Zero Crossing Rate
2. Chroma
3. Mel Frequency Cepstral Coefficient (MFCC)
4. Root Mean Square Value
5. Mel Spectrogram Frequency

At first, Zero crossing rate features are extracted. These features represent at which time a signal changes from positive to negative or vice versa. Then, chroma features of the samples are extracted. The purpose of chroma features are summarizing the harmonic context of an audio. Chroma gives different pitch classes that the audio has. After extracting chroma features, Mel Frequency Cepstral Coefficient (MFCC) features of the samples are extracted. MFCC features of a sample represents the real cepstral of a windowed short time signal. The signal is derived from Fast Fourier Transform (FFT). Then, root mean square value of the data is extracted. A root mean square value is defined as the square root of the mean square of instantaneous values of the voltage signal from the audio. Then, Mel Spectrum Frequency features of the samples are extracted. Mel Spectrums are the spectrums that visualize sounds on the Mel scale.

In order to recognize emotions from voice; first, features are extracted from the samples. Then, a machine learning method is used. In machine learning, there are three main techniques in order to learn or recognize some patterns. These techniques are Supervised Learning, Unsupervised Learning and Reinforcement Learning. In supervised learning: The system is trained with input which is labeled with appropriate kinds until it can detect the relationship between input data and the appropriate output kind. In unsupervised learning, the system is designed to recognize patterns of a dataset that has no labelled information. The output of the system gives the most likely labels in the data. In reinforcement learning, this type of systems uses a reward-penalty method to teach an artificial intelligence system. Agents are defined in environments. Then, rewards for their right actions, penalties for their wrong actions are defined. The agents try to maximize their rewards; thus, they avoid from penalties. In this way, they learn the right actions. In our work, we try to label a sample in order to learn which emotion it is. The characteristics of our problem is best matched with the supervised learning. Thus, we used supervised learning for classification. In our approach we use convolutional neural networks (CNNs). Convolution neural networks (CNNs) are neural networks which designed for image analysis (O'Shea and Nash, 2015). It is mainly used for image classification. There are many applications of CNNs including image recognition, face recognition, and video analysis (Albawi et al., 2017). They are generally used with two dimensional images but they can be also used for one dimensional and three dimensional images. CNNs have three main components. These components are convolution layer, pooling layer and activation function. Convolution is basically a filter to an input that results an output. In a convolution, a linear operation that involves the multiplication of a set of weights with the input is operated. The convolution filter must be smaller than the input data but all the pixels in the image must be filtered and their features must be extracted from the image. In the pooling layer, a two-dimensional filter over each channel of the output of convolution summarize the output. It reduces the width and height of the convolution with keeping the information as much as possible. An activation function takes the output signal of the previous neuron then, converts it to a form which the next cell needs. They basically have two roles, one of them is converting one form to another and the other one is adding non-linearity to the network.

Our convolution network initially has a convolution layer with 256 size, 5 kernel size and 1 stride. ReLU activation function is followed the layer. After the layer, the network has a pooling layer with pooling size 5 and stride 2. The second and third convolution and pooling layers has the same features with the first one. But the size of third convolution layer is 128. After these three layers, there is dropout regularization with %20 rate. After the dropout, there is a convolution and pooling layer with same features as the first layer but its size is 64. Thus, there are four convolution and pooling layers in the network. The output of the last convolution layer is flattened and

dropout regularization with %30 rate is applied to flattened data. Adam optimizer is selected as optimizer for the network.

## 6. Results

The network is trained with 64 batch size for 200 epochs. This experiment is done for three times. Results of the experiment is shown in Table 1.

**Table 1.** Experiments

| Experiment No | Accuracy |
|---|---|
| 1 | %51.46 |
| 2 | %51.32 |
| 3 | %51.00 |

Thus, the average accuracy of the experiments is calculated as %51.26. Training and testing accuracy rate graphs of the experiments are given below. Training and testing accuracy graphs of the experiments are given in Figure 1.

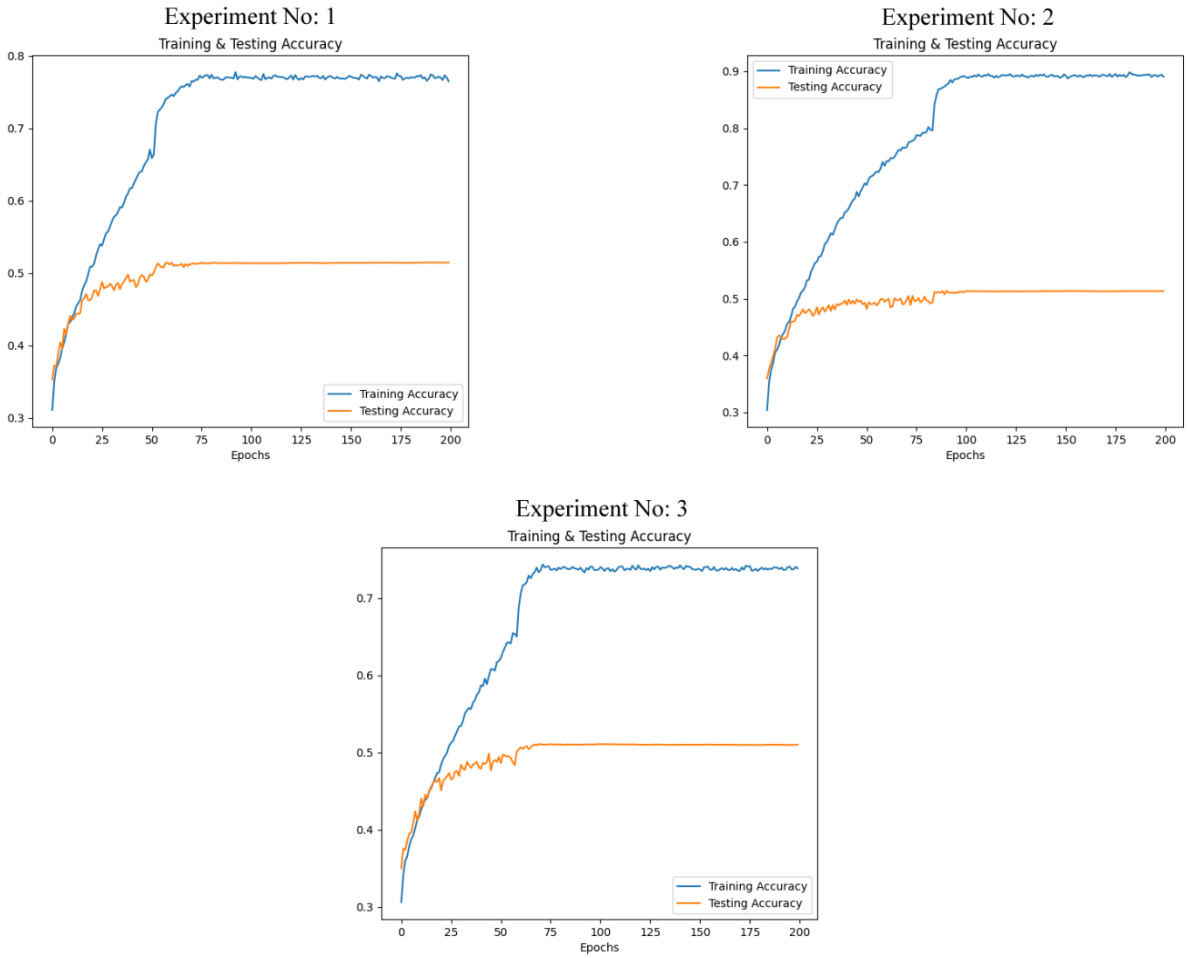


**Figure 1.** Training and testing accuracy graphs

## 7. Conclusion

Emotions are important for human life and human health. Knowing the emotion state of human emotion can be better to make a good relationship with the human. In the recent years, this situation is realized and the emotion state information of a human is started to be used many area ranging from education to aerospace. In order to recognize a human's emotions, there are some methods. These methods can be image processing, gesture analyzing, body pressure and voice recognizing. Image processing, gesture analyzing and body pressures have

some restrictions and disadvantages. The biggest disadvantage and the restriction is their hardware usage. They use dedicated hardware and this affects the portability and the cost of recognizing emotion. Due to these reasons, in this work, voice recognition method is used in order to recognize emotions. There are some ways in order to recognize emotions from voice. The most appropriate and widely used method is machine learning. But in machine learning, there are three main methods in order to learn and recognize patterns. These methods are supervised learning, unsupervised learning and reinforcement learning. The characteristics of the problem is directed us to use supervised learning with a neural network. Our neural network is a convolutional neural network. A feature vector which contains Zero Crossing Rate, Chroma, Mel Frequency Cepstral Coefficient (MFCC), Root Mean Square Value and Mel Spectrogram Frequencyare is feed as an input of the network and the class label of the sample is taken as output. The dataset of the work is chosen RAVDESS and CREMA-D datasets. In the results, four experiments are done and the model gives %51.26 average accuracy rate.

**References**


Andrade, E. B., & Ariely, D. (2009). The enduring impact of transient emotions on decision making. Organizational behavior and human decision processes, 109(1), 1-8.

Tonguç, G., & Ozkara, B. O. (2020). Automatic recognition of student emotions from facial expressions during a lecture. Computers & Education, 148, 103797.

Zepf, S., Hernandez, J., Schmitt, A., Minker, W., & Picard, R. W. (2020). Driver emotion recognition for intelligent vehicles: A survey. ACM Computing Surveys (CSUR), 53(3), 1-30.

César Cavalcanti Roza, V., & Adrian Postolache, O. (2019). Multimodal approach for emotion recognition based on simulated flight experiments. Sensors, 19(24), 5516.

Saste, S. T. & Jagdale, S. M. (2017). Emotion recognition from speech using MFCC and DWT for security system. International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, pp. 701-704.

Subhashini, R., & Niveditha, P. R. (2015). Analyzing and detecting employee's emotion for amelioration of organizations. Procedia Computer Science, 48, 530-536.

Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., & Wróbel, M. R. (2013, June). Emotion recognition and its application in software engineering. In 2013 6th International Conference on Human System Interactions (HSI) (pp. 532-539). IEEE.

Kanjo, E., Younis, E. M., & Ang, C. S. (2019). Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. Information Fusion, 49, 46-56.

Sebe, N., Cohen, I., Gevers, T., & Huang, T. S. (2006, August). Emotion recognition based on joint visual and audio cues. In 18th international conference on pattern recognition (ICPR'06) (Vol. 1, pp. 1136-1139). IEEE.

Kwon, O. W., Chan, K., Hao, J., & Lee, T. W. (2003). Emotion recognition by speech signals. In Eighth European conference on speech communication and technology.

Scherer, K. R. (1996, October). Adding the affective dimension: a new look in speech analysis and synthesis. In ICSLP.

Tato, R., Santos, R., Kompe, R., & Pardo, J. M. (2002). Emotional space improves emotion recognition. In Seventh International Conference on Spoken Language Processing.

Bakır, C., & Yuzkat, M. (2018). Speech emotion classification and recognition with different methods for Turkish language. Balkan Journal of Electrical and Computer Engineering, 6(2), 122-128.

Lim, W., Jang, D., & Lee, T. (2016, December). Speech emotion recognition using convolutional and recurrent neural networks. In 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA) (pp. 1-4). IEEE.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005, September). A database of German emotional speech. In Interspeech (Vol. 5, pp. 1517-1520).

Petrushin, V. A. (2000). Emotion recognition in speech signal: experimental study, development, and application. In Sixth international conference on spoken language processing.

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5), e0196391.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing, 5(4), 377-390.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In 2017 international conference on engineering and technology (ICET) (pp. 1-6). IEEE.