IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features

MEHMET BİLAL ER

Department of Computer Engineering, Faculty of Engineering, Harran University, Şanlıurfa, Turkey

Corresponding author: Mehmet Bilal Er (bilal.er@harran.edu.tr)

ABSTRACT The problem of recognition and classification of emotions in speech is one of the most prominent research topics, that has gained popularity, in human-computer interaction in the last decades. Having recognized the feelings or emotions in human conversations might have a deep impact on understanding a human's physical and psychological situation. This study proposes a novel hybrid architecture based on acoustic and deep features to increase the classification accuracy in the problem of speech emotion recognition. The proposed method consists of feature extraction, feature selection and classification stages. At first, acoustic features such as Root Mean Square energy (RMS), Mel-Frequency Cepstral Coefficients (MFCC) and Zero-crossing Rate are obtained from voice records. Subsequently, spectrogram images of the original sound signals are given as input to the pre-trained deep network architecture, which is VGG16, ResNet18, ResNet50, ResNet101, SqueezeNet and DenseNet201 and deep features are extracted. Thereafter, a hybrid feature vector is created by combining acoustic and deep features. Also, the ReliefF algorithm is used to select more efficient features from the hybrid feature vector. Finally, in order for the completion of the classification task, Support vector machine (SVM) is used. Experiments are made using three popular datasets used in the literature so as to evaluate the effect of various techniques. These datasets are Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Berlin (EMO-DB) and Interactive Emotional Dyadic Motion Capture (IEMOCAP). As a consequence, we reach to 79.41%, 90.21% and 85.37% accuracy rates for RAVDESS, EMO-DB, and IEMOCAP datasets, respectively. The Final results obtained in experiments, clearly, show that the proposed technique might be utilized to accomplish the task of speech emotion recognition efficiently. Moreover, when our technique is compared with those of methods used in the context, it is obvious that our method outperforms others in terms of classification accuracy rates.

INDEX TERMS Speech emotion recognition, Deep learning, Hybrid features, Pre-trained CNN

I. INTRODUCTION

Speaking is the basic means of human interaction which is fast and efficient. During the speech, air flows through the trachea from the lungs to the larynx, and this air flow creates speech signals by vibrating the vocal cords [1]. Research on topics about speech emotion recognition received much more attention from people in recent years. Recognition and measurement of human emotions, automatically, has been one of the up-to-date research areas in the fields ranging from Biomedical engineering and psychophysiology to computer engineering and artificial intelligence [2]. Emotions, to a large extent, along with many meaningful attitudes, are special and powerful mental activities that can be comprehended by simple observations from an outsider. All bodily activities, like speaking, facial expressions, and body movements, constitute the basic building blocks of a human's emotional state [3]. Identity information and emotional states of the speaker are transferred via voice signals across other people [4]. Emotion recognition has been seen to be used more prevalently in human-computer interaction due to the increasing demand of people on smart systems and increasing data processing speed and performance of computers. Autonomous speech emotion recognition systems, fundamentally, simulate human emotions through the use of a computer, and then, features like accentuation, intonation, and pause employ spectrum-based features for matching them with the target emotions. At its core, a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3043201, IEEE Access

speech emotion recognition system is made up of three stages: speech data preprocessing, extraction of emotion features, and emotion classification, respectively [5]. During the course of the speech, it is known that since people might be affected by their physical conditions in their own environment and outer world conditions surrounded them, emotions inherently may exhibit diversity and variation. Therefore, a powerful categorization architecture and speech emotion features involving crucial knowledge are two of the important components of emotion recognition. Speech emotion recognition is a major challenge that attracts researchers because of several applications like voice surveillance, E-learning, clinical studies, detection of lies, entertainment, computer games and, call centers. However, to a large extent, advanced machine learning techniques are compelling tasks, as well. Despite a large number of researches done and advances taken in emotion recognition in recent years, it is still not quite known what the most appropriate method is likely to be. This situation is induced by the subjectivity of emotions. What we mean by the subjectivity is that two different people can recognize the same emotions in distinct ways, and thus, in return, this may cause uncertainty on defining a basic emotion class. Additionally, there is a great deal of ambiguity about determining the most convenient emotional features. Furthermore, there is no predefined feature set that is assigned for the recognition of emotions [6]. Besides that, the presence of background noise in sound recordings, which is caused by real-world sounds, can significantly have a huge impact on the efficiency of a machine learning model [7]. In conventional approaches to recognition of speech emotions, features representing the acoustic content of speech are extracted. Various types of machine learning techniques are employed for comprehending the relation among the extracted features of speech and predetermined emotion tags. In these studies, SVM, hidden Markov models (HMMs) and neural networks are employed. SVMs offer relatively better predictions by putting less effort, while, on the other hand, it is tedious to construct and train neural networks and hidden Markov models. It also requires high computing power and time. Now, deep learning models are used to solve recognition problems such as face recognition, voice recognition for the internet of things, and speech emotion recognition [8]–[10]. One of the main advantages of deep learning techniques is, for example, the automatic selection of important features inherent in audio files with a particular emotion in the task of recognizing speech emotions.

The rest of this paper is organized as follows. In Section 2, the speech voice classification studies in the literature are reviewed and the differences between them are tried to be revealed. In Section 3, the material and the proposed method are introduced. In Section 4, experimental applications related to the dataset and classification of speech sounds used in research are given. In Section 5, the results of our proposed method are discussed.

II. RELATED WORKS

In order to accurately recognize the emotion in speech data, it is very significant to extract features that correctly symbolize the emotional side of the speech signals. Some considerable research is done in the context, including the analysis and synthesis of the speech, will be explained in this section. Generally, the essential feature parameters utilized in the speech emotion recognition system can be separated into two categories in terms of conventional features and deep features. Features extracted from Convolutional Neural Network (CNN) layers are generally used as deep features [11], [12].

In [13], to recognize emotions from speech, a method that is based on MFCC features and Gaussian mixture model classifier is proposed. In [14], Berlin (EMO-DB) dataset is selected for the classification of speech sounds. To classify seven different emotions in the dataset, 3 staged SVM classifier is used. MFCC features are extracted from all the 535 records, which exist in the dataset, and nine separate statistical evaluations are done on the features. 10-fold cross-validation is performed for the training phase and for the testing phase of data. Performance analysis is done using confusion matrix and an accuracy rate of 68% is achieved. In [15], MFCC are searched and these features are classified through Linear Discriminant Analysis (LDA). Also, in the paper, a database that belongs to artificial emotional Marathi speech is used. Data samples are collected from the actors and actresses who play in 5 Marathi films, and emotions that produce Marathi phrases, which can be used in daily intercommunication and can be paraphrased in all emotions, are simulated. Data samples are categorized into 5 basic categories as Happy, Sad, Anger, Afraid and Surprise. In [16], the pre-processing required for recognizing emotions from the speech data is performed. MFCC features are extracted from speech signals. For the classification, K- Nearest Neighbors (K-NN) algorithm is utilized. In [17], four emotional situations are treated to be classifying emotions in speech. For this purpose, features are extracted by utilizing Linear Frequency Cepstral Coefficients (LFCC) and (MFCC) which are the sound features of emotional speech. In addition, those features are classified by utilizing the Hidden Markov Model (HMM) and SVM. In [18], to recognize emotions from speech, features like energy, zero-crossing rate and fundamental frequency are utilized. An average of 56.46% recognition rate is achieved over a dataset that includes seven different emotions. In [19], a method for recognizing Multilingual speech emotion is presented. In this context, first of all, two hundred fifteen acoustic features extracted from the emotional speech. Secondly, feature selection has been made to develop a common set of standard acoustic parameters for multiple languages. Finally, emotions are categorized into essential

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3043201, IEEE Access

categories by employing logistic model trees. The suggested approach is tested in Japanese, German, Chinese and English emotional speech corpus. The recognition performance is inspected through inter-speaker and inter-group evaluation. In [20], an autonomous system has been introduced to predict the primitives of emotion. Fuzzy logic estimator and system is developed by using acoustic features like energy, speech speed and spectral features that are derived from speech. The approach is tested in two databases. The First database comprises of six hundred and eighty sentences that include three speakers that are in happy, angry, neutral and sad categories. Second database includes more than a thousand expressions recorded from a TV talk show, belongs to speakers that are more than forty-seven, which have authentic emotional expressions. Finally, a general recognition rate of up to 83.5% is achieved by using the K-NN classifier for emotion prediction. In [21], for the determination of seven human emotions (neutral, anger, boredom, disgust, fear, happiness, sadness) by employing speech signals, multilayer perceptron neural network and generalized feed forward neural network are used. The Overall accuracy rate is found as follows: in multilayer perceptron neural network, an accuracy rate of 93% is achieved and in the generalized feed forward neural network, an accuracy rate of 99% is achieved. In [22], they have worked to improve the performance of an emotion recognition system using MFCC and features that are related to energy as constituents of the feature vector. After identifying the frequency range that is influenced by the emotion, normalization is performed using the dynamic time warping-multi-layer perceptron hybrid model. Fast correlation-based filter and variation analysis methods were used in this study to degrade the number of features. Recognition of emotional states is carried out by utilizing the Gaussian mixture model. In [23], a robust emotion recognition approach is presented for the speech signals in noisy environments. Feature size is decreased to 87 from 204 by employing fast correlation-oriented filter feature selection technique. The performance of the suggested technique is measured by using LDA, K-NN, C4.5 decision tree, radial basis function neural networks and SVM classifier. In [24], feature selection techniques are utilized to increase the emotional recognition success and to reduce the amount of work done by using fewer features. A novel statistical feature selection technique is suggested based on emotional changes on acoustic properties. The success of the suggested technique has been checked against other techniques utilized more in the literature. This check depends on feature count and emotion recognition success. With respect to the outcomes gathered, the suggested technique not only provides a major decrease in the number of features, but also increases the classification success. In [25], a combination of spectral features has been extracted from the sound recordings and passed through the feature selection and reduced to the required feature set. It is proven that as compared to single estimators, ensemble learning has a better performance. In recent years, it has been feasible to design and implement deep neural networks with the improvement of computer hardware. An important application of deep neural networks is to extract the speech feature parameters containing deep information about the speech signal and then the classifiers can be trained with the obtained features [26], [27]. In [26], bottleneck features were extracted using the Deep Belief Network (DBN), and then these features were used to train the SVM model to recognize the speech emotions. In [27], a new model is presented. The end-to-end trained model consists of a CNN with 2 layers of Long Short-Term Memory (LSTM), which takes features from the raw signal. The system has fully analyzed the context of the speech signal and outperformed other systems in terms of coefficient of consistency and recognition. In [28], MFCC, chroma-gram, mel-scale spectrogram, Tonnetz representation, and spectral contrast features are extracted from the audio files and employed as inputs for 1D-CNN. In [29], deep emotional features are investigated to recognize speech emotions. CNN and Long short-term memory (LSTM) are configured in order to learn emotional features from speech and log-mel spectrogram. These are one-dimensional CNN-LSTM network and two-dimensional CNN-LSTM network. Experimental outcomes indicate that the devised networks perform well for the task of recognizing speech emotions. Particularly 2DCNN-LSTM network, performs better than Deep Belief Network (DBN) and CNN, which are traditional approaches. In [30], to effectively improve speech emotion recognition performance, a novel speech emotion recognition technique is presented that depends on Deep Neural Network (DNN), decision tree and SVM. It is focused on wherewith to find more prominent speech emotional features and wherewith to create an efficient recognition model. In [31], a hybrid method, consisting of three steps, is proposed for the classification of speech emotions. The spectral features and prosodic features are combined in the feature extraction stage. This hybrid feature vector has been derived from both the speech signal and the glottal waveform. Also, first and second-order derivatives of feature vectors are used to enrich the dimension. In the next step, the feature size is reduced by quantum-based particle swarm optimization to reduce dimensionality. In the last stage, the classification process is made. In [32], in order to increase the classification performance in the emotion recognition problem, a hybrid feature vector consisting of combining prosodic features and spectral features has been proposed. The proposed method was tested on two open access datasets, and five collective learning algorithms were used

to train the data. The results obtained from the proposed method show that hybrid features are effective in speech emotion recognition problem. In [33], a hybrid deep neural network model has been proposed for heterogeneous acoustic features that reduce classification performance in speech emotion recognition problem. The proposed architectural feature extraction module consists of a merge module and a fusion network module. A fusion network is used after the discriminants of heterogeneous features are obtained. SVM is chosen as the classifier, and the results from the experiments show that the proposed architecture improves the classification performance. Besides, bio-inspired computational models can give very effective results in sound processing studies [34]. In [35], unlike traditional feature-based classification methods, a method that works directly on the speech signal is presented. Using the Liquide State Machine, a bio-inspired computational model, it is aimed to automatically recognize speech emotions. The method used was tested on the Emo-DB dataset and a high rate of classification performance was achieved.

In this study, both acoustic and deep features are used to classify speech emotions. A hybrid feature vector is created by both extracting deep features from the original sound signal's spectrogram images and acoustic features from voice records. The purpose of combining acoustic features and deep features is to improve classification performance. The fundamental contributions of this study are as follows:

- A novel technique is proposed for speech emotion recognition problem.
- It is shown that the emotional content of speech can be classified by both acoustic and deep features.
- The classification accuracy rate is increased by virtue of the hybrid feature vector.
- The impact of pre-trained deep networks in speech emotion recognition problem is demonstrated.

III. PROPOSED METHOD

The method presented in this study consists of acoustic features, deep features, pre-trained CNN and SVM combined model. In many studies, acoustic and deep features are used separately [11], [12], [16], [17]. In this study, acoustic and deep features are combined to improve the semantic information of the emotion features in the speech. Acoustic features alone can lose some useful information about speech emotion patterns. Specifically, the patterns of emotion present in speech signals cannot be well mined. Also, the concept of acoustic features in sound signals is based on people's subjective assumptions. Deep features from deep networks are more comprehensive than acoustic features. Both theories and experiments have shown that deep learning can extract a lot of valuable information from auditory signals [36]. In order to take advantage of deep learning, pre-trained deep networks are used as feature extractors. Pre-trained deep networks used in the study; VGG16, ResNet18, ResNet50, ResNet101, SqueezeNet and DenseNet201. The main reason for using these pre-trained CNNs is to show that the depth of CNN has or does not an effect to the performance. As a result of combining acoustic and deep features, a hybrid feature vector is obtained. Also, the ReliefF feature selection algorithm is used to select the most effective features from hybrid feature vector. The last part of the proposed model is the SVM classifier. SVM classifier is trained with the hybrid feature vector, as it is often more effective in higher dimensional domains.

Details of each step of the proposed are explained in the subsections below. The proposed method is given in Figure 1. This model can be summarized as follows:

- Acoustic features are extracted from pre-processed sound signals.
- Spectrogram images of the signals are extracted and data augmentation is applied.
- CNNs pre-trained on the ImageNet dataset are used to extract features from the speech spectrogram.
- Pre-trained CNNs used for feature extraction: VGG16, ResNet18, ResNet50, ResNet101, SqueezeNet and DenseNet201.
- Spectrogram images are given as input to pre-trained networks and deep features are extracted.
- A hybrid feature vector is obtained by combining the obtained deep features and acoustic features.
- ReliefF is applied for feature selection.
- SVM classifier is trained with the hybrid feature vector.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3043201, IEEE Access

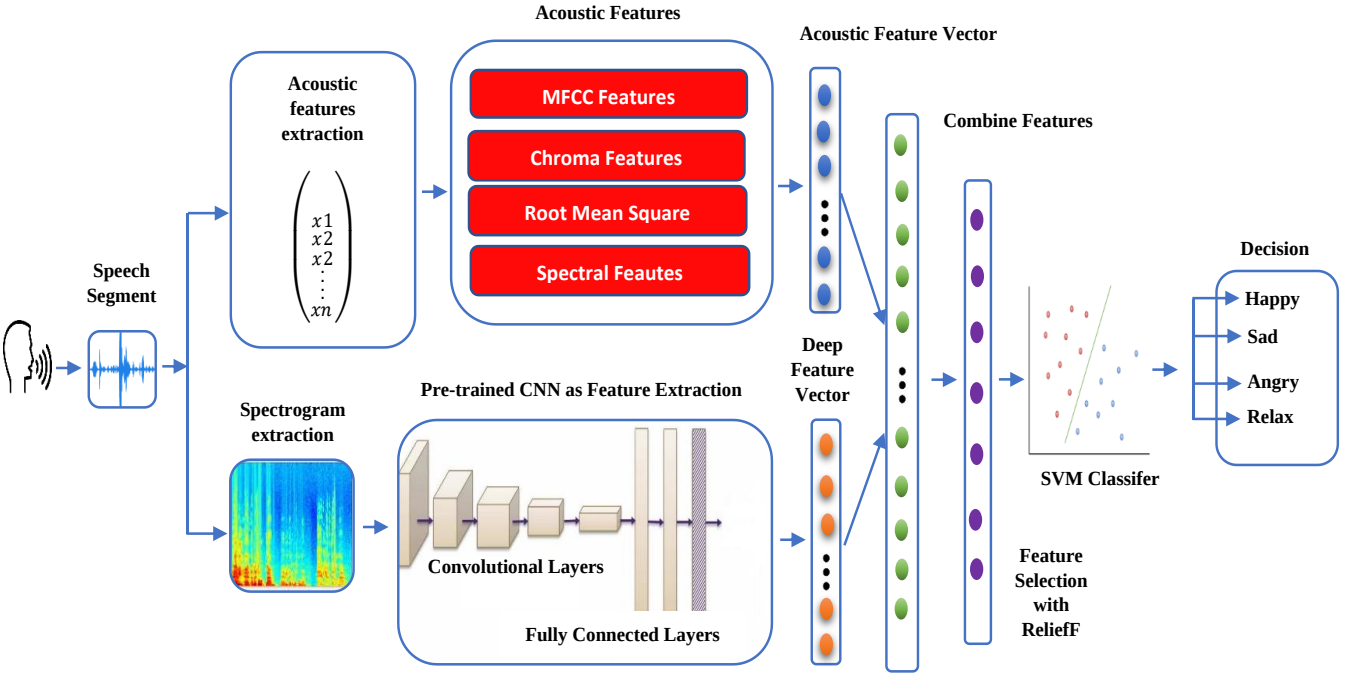


Figure 1. Proposed Method

A. ACOUSTIC FEATURE EXTRACTION

The purpose of the acoustic analysis is to separate the speech signal into its components and to present parametric measurements of these components. Acoustic features are physical properties in terms of frequency, loudness and amplitude. Speech signals are pre-processed before acoustic features are extracted. While recording, samples of speech may contain unwanted information such as noise, depending on environmental factors. In this study, the Butterworth filter is used to remove noise in speech samples. Also, the speech signal is divided into frames of 30ms with an overlap of 15ms. LibROSA toolbox is used to extract acoustic sound features from different groups. LibROSA is a widely used library for music and sound analysis. Extracted acoustic features are Root Mean Square energy (RMS), MFCC, Chroma, Spectrum centroid, Spectral entropy, Skewness, Attack time and Zero crossing rate. The size of the acoustic property vector obtained is 32.

RMS: It is the measure of the loudness of an audio signal. It is found by calculating the square root of the sum of the mean squares of the amplitudes of the sound samples. RMS formula is given in Equation 1.

$$x_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x^2{}_i} = \sqrt{\frac{x^2{}_1 + x^2{}_2 + \ldots + x^2{}_n}{n}} \qquad \text{(Eq.1)}$$

Spectrum centroid: It is usually associated with a measure of the brightness of a sound and is a measure of where the center of mass of the spectrum is. Higher centroid values indicate higher frequency values [37].

Spectral Entropy: The probabilities of the power spectrum components of the signal are taken into account when calculating this value. The normalized power distribution in the frequency domain of the signal is evaluated as the probability distribution [38].

Skewness: Indicates the degree of asymmetry of a distribution around its mean, and it is the average skewness coefficient of the spectral distribution in the Lower frequency bands [39].

MFCC: It is based on human hearing perception and is one of the most used feature extraction methods in the field of sound processing. MFCC features are based on obtaining distinctive values for speakers by imitating the frequency selectivity of the human ear [40]. The steps to be taken to extract MFCC features are given in Figure 2.

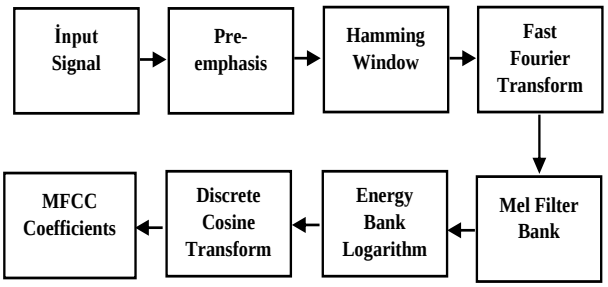


Figure 2. Steps of MFCC [41]

Conversion between Mel scale (M) and frequency scale (Hz) can be done using equations 2 and 3 given below.

$$m = 295 \log_{10}\left(1 + \frac{f}{700}\right) \qquad \text{(Eq.2)}$$

$$f = 700\left(10^{\frac{m}{295}} - 1\right) \qquad \text{(Eq.3)}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3043201, IEEE Access

One way to express this perceptual structure is to use a triangular band-pass mel filter bank. Mel Frequency Kepstrum coefficients are obtained by applying discrete cosine transform after filter bank. MFCC is calculated according to equation 4.

$$MFCC_i = \sum_{k=1}^{20} X_k . Cos\left[i.\left(\frac{k-1}{2}\right).\frac{\pi}{20} \right] \quad i = 1, 2, \ldots M \qquad \text{(Eq.4)}$$

Attack Time: It is the estimation of the time it takes for a signal to reach to its peak. A simple way to define and calculate this feature is to predict the time duration of the range of the phase where the signal's amplitude rises.

Chroma: The notes relate to the energy density around them and provide important information about the harmonic content of the sound. There are 7 notes in western music, and since the two notes are divided into two equal parts, except for the E and F notes, 12 features can be obtained by taking the sounds in between.

Zero-crossing Rate: It is the rate of the transition of a signal through the zero lines, that is, the change of the sign. The X-axis shows how many times the signal has passed, it can be used as an indication of noise as well as frequency.

B. SPECTROGRAM EXTRACTION

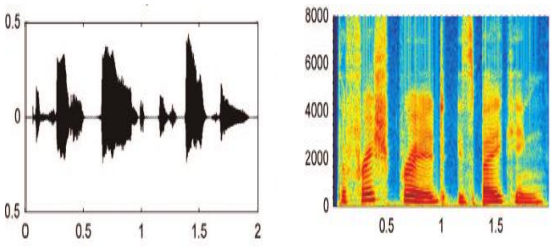
There may be silence at some moments during the conversation and no emotion may arise. This is a factor that makes emotion recognition difficult and the moments when no emotion is felt should be filtered. However, since the duration of the speech sound recordings in the datasets used is not very long, the entire part of the speech sound recordings was taken into account in spectrogram extraction. Since speech signals are not static signals, the signal must be processed in small frames. Speech signals are first divided into 30 ms frames in order to obtain the spectrograms. Also, each frame comes out in such a way that it overlaps a part of the previous frame. The overlap ratio of the frames is chosen as 50% of a frame. Windowing is applied after framing is applied on a signal. The aim is to prevent discontinuity that may occur at the extreme ends of each frame. The widely preferred "Hamming window" technique is employed in this study. Hamming Window minimizes unwanted radiation from the extreme-ends of the regions of the signal [42]. It is also the function that makes the signal convenient for the Fourier transform. The algorithm of the hamming function is as follows: The hamming window is multiplied by the framed sound signal and then the windowed signal is finally obtained. Hamming window formula is given in equation 5.

$$w[n] = 0.54 - 0.46\cos\left(2\pi\frac{n}{N}\right), \qquad \text{(Eq.5)}$$

N : Window length

Fourier transform is performed on the signal after the Hamming window is performed. The Fourier transform transforms the signal from the time domain to the frequency domain. Fast Fourier Transform (FFT) is used in this study. In this step, each frame consisting of N samples is passed from the time domain to frequency domain by performing Fast Fourier transform. A set with N samples is defined as in equation 6. At the last stage, the power spectrograms of the signals which Fourier transform is applied, are extracted. An example speech audio signal and spectrogram image are given in Figure 3.

$$X_n = \sum_{k=0}^{N-1} x_k e^{-2\pi jkn/N}, \quad n = 0, 1, 2, \ldots\ldots, N-1 \qquad \text{(Eq.6)}$$



(a) (b)

Figure 3. Illustration of Speech Sound Signal and Spectrogram (a) Speech Sound Signal, (b) Spectrogram

C. DATA AUGMENTATION

In situations where the original data size is limited, data augment is needed to surmount this issue of data shortage. Data augment is the production of extra training data samples by performing a series of deformations on the data in the training dataset. The essential basis in data enhancement is that the labels of the new data created by deformations applied to the tagged data are not changed. [43]. There are many methods for data augment, such as rotating the image at different angles, horizontally rotating and vertically rotating, adding noise and color manipulation to the image. In this study, 2 different processes were applied to the signal before extracting the spectrogram and these processes are explained below:

- Background Noise: Added random noise in the range of [0.1, 0.5] to the sound samples.
- Time Shifting: The sound is shifted from the starting point and the original length is preserved. Each sample shifted from the starting point to 0.3 seconds. The sound is shifted from the starting point and the original length is preserved. Each sample shifted from the starting point to 0.3 seconds.

The sound samples obtained as a result of increasing data augmentation were added to the original dataset as additional training examples. In Algorithm 1, the pseudo-code of the data augmentation algorithm is given.

Algorithm 1 Data Augmentation Algorithm

Input: *sp* : Spectrogram
Sr: Sample Rate
bg: Background Noise
tm: Time Shifting,
bgnr: Background Noise Range
tsh : Time shifting rate

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3043201, IEEE Access

Algorithm:

```
Initialize and assign input parameters (sp,sr, bg, tm,bgnr,tsh)
sp = read_folder('foldername/'+Filename');
 for i:= 1 to length(sp) do
   data = Read Spectrogram File(sp.Name);
    bg = add.random.bgnoise(data, bgnr);
    tm = Time.shifting(data,  sr * tsh)
   Write spectrogram( bg_noise.jpg, bg);
   Write spectrogram(time_shifting.jpg, tm);
end for
```

D. DEEP FEATURE EXTRACTION FROM PRE-TRAINED CNN MODELS

Pre-trained CNNs typically have a mathematical structure consisting of three types of layers. These layers are convolution, pooling and fully connected layers [44]. Convolution and pooling layers perform feature extraction, while fully connected layers send the extracted features to the final output for classification. Filters with a certain height and width in the convolution layer are moved from left to right over the input image. Convolution formula is given in equation 7. 'M' refers to the property map given in the equation, and 'w' refers to the convolution core of (x, y) size.

$$M(i,j)=(R*w)(i,j)=\sum_x\sum_y R(i-x,j-y)\,w(x,y) \qquad \text{(Eq. 7)}$$

The pooling layer is another building block of pre-trained CNNs and is applied after the convolution process. Its main goal is to reduce the number of parameters and computations in the network. The most common approach used in pooling is maximum pooling. After the features down-sampled by the convolution and pooling layers are created, these features are linked to the fully connected layer. The output feature maps of the final convolution or pooling layer are typically flattened, that is, converted into a one-dimensional (1D) number sequence or vector. These features are then linked to one or more fully connected layers. Here each input is linked to each output and each neuron has a learnable weight. The last fully connected layer typically has the same number of output nodes as the class number, and classification is done on this layer. Different classifiers can be used in the last layer.

In this study, pre-trained networks on ImageNet dataset such as VGG16, ResNet18, ResNet50, ResNet101, SqueezeNet and DenseNet201 are used as deep feature extractors. The spectrogram images of the speech signals are given as input to these networks. Deep features are the features obtained from the layers of CNN before the classification layer as a result of giving an image to the CNN [45]. The pre-trained VGG16 deep network architecture is developed by Simonyan and Zisserman in the ILSVRC 2014 contest. It is basically a deep network, consisting of thirteen convolutional layers and three fully connected layers. There are fourty-one layers in total with Maxpool, fully connected layer, ReLu layer, Dropout layer and Softmax layer. The image to be presented to the input layer has 224x224x3 pixels. The last layer is the classification layer [46]. The fc7 layer of the VGG16 architecture has 4094 neurons and this layer is chosen as the deep feature extractor layer. ResNet has many variations and some of the commonly used ones are ResNets18, ResNet50 and ResNet101. ResNet won the ILSVRC contest held in 2015 with the lowest error rate of 3.37%. In this study, pre-trained ResNet (ResNet18, ResNet50, ResNet101) models are used. These networks have 18 (72 substrates), 50 (177 substrates) and 101 (347 substrates) layers, respectively. All of these networks have fc1000 (1000 fully connected layers) and this layer is used as the deep feature extractor layer. SqueezeNet is a smaller network designed as a more compact alternative to AlexNet. It has almost 50 times less parameters than AlexNet but works 3 times faster. This architecture was introduced by Lonola *et al.* in 2016 [47]. SqueezeNet generally consists of an independent convolution layer (conv1), eight fire modules (fire2–9) and finally the conv10 layer. The image to be included in the input layer is 227x227x3. SqueezeNet architecture has 1000 neurons in the pool10 layer and this layer is chosen as the deep feature extraction layer. DenseNet is one of the new architectures developed for object recognition and was introduced by Huang et al. in 2017. DenseNet architecture is very similar to ResNet, with some basic differences. DenseNet201 model is used in this study. In DenseNet201 model, there are direct connections to all layers after all previous layers. The default input size for this model is 224x224. The DenseNet201 architecture has 1408 neurons in the conv5_block16 layer and this layer is chosen as the deep feature extraction layer. Spectrogram images size is set to 224x224x3 for the input layer of VGG16, ResNets and DenseNet201 networks, and 27x227x3 for SqueezeNet.

The proposed feature extraction is given in Figure 4. Scatter plots of deep features obtained from ResNet101 are given in Figure 5-7. In order to obtain the scatter plots of the deep features given in Figures 5-7, the spectrogram images of the audio signals are given as input to ResNet101 and the fc1000 layer is used as feature extractor.

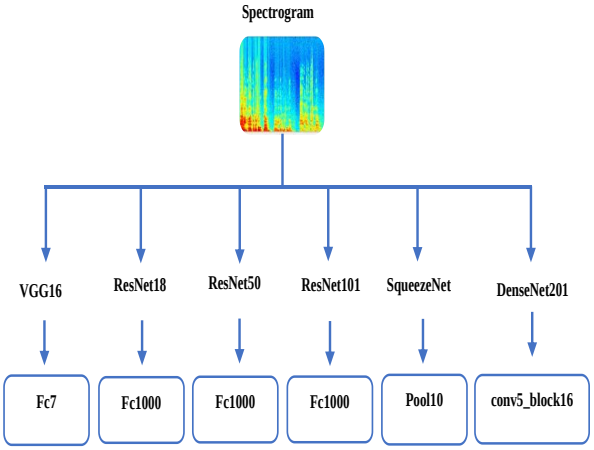


Figure 4. Deep Feature Extraction in Pre-trained Networks

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3043201, IEEE Access

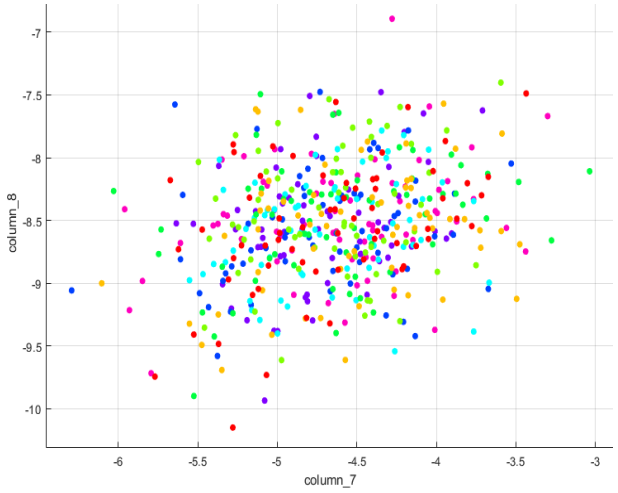


Figure 5. 2D representation of deep features for RAVDESS

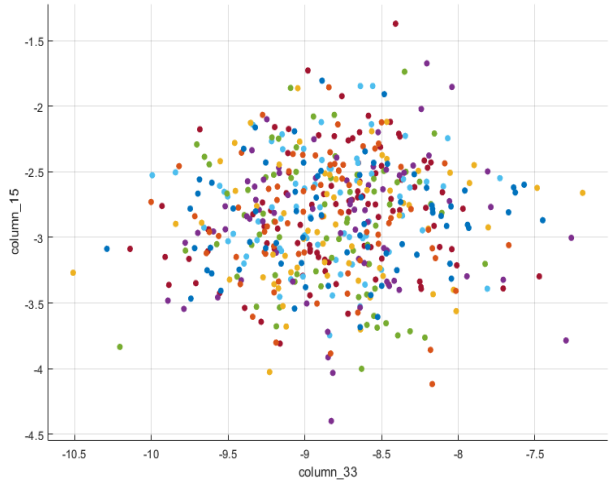


Figure 6. 2D representation of deep features for EMO-DB

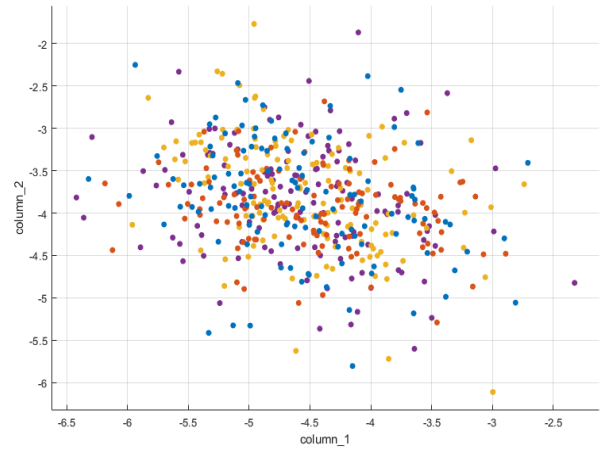


Figure 7. 2D representation of deep features for IEMOCAP

E. FEATURE SELECTION WITH RELIEFF

In this study, ReliefF feature selection algorithm was used to select effective hybrid features. Feature selection is an important area of research in machine learning and data mining. The ReliefF is one of the most important feature selection algorithms that gives successful results in many feature selection applications. ReliefF is an improved version of the Relief statistical model. The Relief method takes a sample from the dataset and performs the feature selection process by creating a model that depends on the proximity of the relevant sample with other samples in its own classes and its distance from different classes [48]. The update formula of the weight coefficient in the ReliefF algorithm is defined as follows:

$$W[K] = W[K_0] - \frac{\sum_{J=1}^{K} diff(A, x_i, H)}{mk} + \sum_{C \neq Class(x_j)} \frac{p(C)}{1 - p(Class(x_i))} \cdot \frac{\sum_{j=1}^{k} diff(A, x_i, M_j(C))}{mk} \quad \text{(Eq. 8)}$$

In algorithm 2, the pseudo-code of the feature selection algorithm with ReliefF is given. The features obtained from VGG16's fc7, ResNet's fc1000, SqueezeNet's pool10 and DenseNet201's conv5_block16 layer were combined with the acoustic features and passed through the ReliefF feature selection algorithm. The number of features obtained before and after feature selection is given in Table 1.

TABLE 1. Number of Features Before and After Feature Selection

Features	Number of Features Before Feature Selection	Number of Features After Feature Selection
VGG16 {fc7}+Acoustic features	4128	3248
ResNet18 {fc1000}+Acoustic features	1032	756
ResNet50 {fc1000}+Acoustic features	1032	843
ResNet101 {fc1000}+Acoustic features	1032	642
SqueezeNet {pool10}+Acoustic features	1032	794
DenseNet201{conv5_block16}+Acoustic features	1440	965

Algorithm 2 Feature selection with ReliefF

Input: *f* : Feature vector for train examples
Output: *w*: Predicted features
Algorithm:

1: set all weights w[K]:=0;
2: **for** *i*:= 1 to **do**
3: randomly select an sample
4: *find k-nearest hits sj*
5: ***for each*** *class C class(zi)* ***do***
6: ***for*** *K:=1 to k* ***do***
7: $W[K] = W[K_0] - \frac{\sum_{J=1}^{K} diff(A, x_i, H)}{mk} +$
8: $\sum_{C \neq Class(x_j)} \frac{p(C)}{1 - p(Class(x_i))} \cdot \frac{\sum_{j=1}^{k} diff(A, x_i, M_j(C))}{mk}$
9: **end for**
10 **end for**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3043201, IEEE Access

F. CLASSIFICATION

Classification is the final stage of emotion recognition. The classifier was trained by combining the extracted acoustic and deep features. Softmax is usually employed in the classification layer of the pre-trained model. In this study, we used SVM for classification. The SVM classifier was introduced by Vapnik in 1995 [49]. SVM can solve both linear and nonlinear problems and come up with better results for many practical problems. Our main purpose is to create a separating line or hyperplane between the data of the two classes. The hyperplane separates any size of data linearly. This hyperplane can be two-dimensional, or it can have a dimension of more than two. SVM creates parallel regions by creating two parallel lines. Moreover, it separates the space in one pass to create straight and linear regions. SVM tries to find the widest range, i.e. the largest margin, between the categories, and thus, the two categories are divided with a gap between them where the gap is as wide as possible. This hyperplane is responsible for the partition. It is known that the classification process will be more successful if there is a broader gap between the two classes in SVM.

Input vector {xi, i = 1, ..., n} should belong to one of the classes, yi{−1, 1}. Additionally, a hyperplane can be defined as follows:

$$w_0 x + b_0 = 0 \quad \text{(Eq.9)}$$

Here, w is the weight vector, x is the input vector, and b is a bias. For any given pair of w and b, data can be linearly separated when one of the following cases occurs:

$$w.x_i + b \geq 1 \quad if\ y_i = 1 \quad \text{(Eq.10)}$$

$$w.x_i + b \leq 1 \quad if\ y_i = -1 \quad \text{(Eq.11)}$$

The kernel method is used to solve a nonlinear problem with a linear classifier. The input data is converted into a high-dimensional space with the Φ function, where the kernel function K is defined as follows:

$$k(x, x') = (\Phi(x), \Phi(x')) \quad \text{(Eq.12)}$$

-Linear

$$k(x_i, x_j) = x_i.x_j \quad \text{(Eq.13)}$$

Polynomial

$$k(x_i, x_j) = (x_i.x_j + 1)^d \quad \text{(Eq.14)}$$

Here, d is the degree of the polynomial.

IV.EXPERIMENTAL APPLICATIONS

A. DATASET

In this study, three different voice datasets are utilized, which are widely used by researchers in emotion recognition. These datasets are as follows; RAVDESS, EMO-DB and IEMOCAP. RAVDESS is selected as one of the datasets for our model due to its large availability. This dataset contains audio and visual recordings of 12 male and 12 female actors pronouncing English sentences with eight different emotional expressions. Only speech samples are used for this study. Emotion tags in the dataset are: sad, happy, angry, calm, fearful, surprised, neutral and disgust. Besides, the total number of records in the dataset is 1440 [50]. The second dataset we use in our research is EMO-DB, which is widely used by researchers in the field of speech-based emotion recognition and allows us to make more extensive comparisons with previous studies. The dataset contains 535 audio outputs divided into 7 emotion classes in German. Emotion classes in the dataset are; anger, sadness, fear/anxiety, neutral, happiness, disgust, and boredom [51]. The third dataset we use is the IEMOCAP dataset that is produced from improvised data. This dataset consists of audio, video and facial movement samples collected from five pairs of male and female actors. The audio files of the data series are divided into ten emotion classes: angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted and other. We use the sound samples in the IEMOCAP dataset to measure the performance of the proposed frame on improvised data. In addition, we consider only 4 emotion classes (angry, happy, neutral, and sad) in the IEMOCAP dataset in this study. The number of audio files in four classes is 889 [52].

B. EXPERIMENTAL RESULTS

The implementation of the proposed method is done on a machine that has i7 2.50GHz processor, 12GB memory and NVIDIA 940M GPU hardware properties. The source code required for the application is prepared using MATLAB R2018a. 224x224 size input images are required for VGG16, ResNet and DenseNet201 networks, while 227x227 size input images are required for SqueezeNet. Spectrogram images obtained from the voice recordings in the dataset are automatically resized to this size. Deep features have been extracted from VGG16's fc7, ResNet's fc1000, SqueezeNet's pool10 and DenseNet201's conv5_block16 layer. A hybrid feature vector was created by combining these obtained deep features with acoustic features.

Thirteen different feature vectors are created as:

- Acoustic features
- DenseNet201{conv5_block16}
- DenseNet201{conv5_block16}+Acoustic features
- ResNet18{fc1000
- ResNet18{fc1000}+Acoustic features
- ResNet50{fc1000}
- ResNet50{fc1000}+Acoustic features
- ResNet101{fc1000
- ResNet101{fc1000}+Acoustic features
- SqueezeNet{pool10}
- SqueezeNet{pool10}+Acoustic features
- VGG16{fc7}
- VGG16{fc7}+Acoustic features

Linear kernel SVM is used for classification. Using the 10-fold cross-validation technique, the data were separated for testing and training. The performance of the proposed method is evaluated based on the accuracy rates as shown

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3043201, IEEE Access

in Tables 2-4. The first column of Table 2-4 shows the feature set used, and the second shows the classification success.

Classification results for the RAVDESS dataset are given in Table 2. The best classification result for the RAVDESS dataset was obtained from ResNet101 {fc1000}+Acoustic features as 79.41% with ReliefF feature selection. Also, VGG16{fc7}+Acoustic features with ReliefF reached to an accuracy rate of 74.41%, ResNet18{fc1000}+Acoustic features reached to an accuracy rate of 75.38%, ResNet50 {fc1000}+Acoustic features with ReliefF reached to an accuracy rate of 78.26%, SqueezeNet{pool10}+Acoustic features with ReliefF reached to an accuracy rate of 75.81% and DenseNet201{conv5_block16}+ Acoustic features with ReliefF reached to an accuracy rate of 77.46%. The confusion matrix for the best classification result for the RAVDESS dataset is given in Figure 8. Classification results for the EMO-DB dataset are given in Table 3. The best classification result for the EMO-DB dataset was obtained with ReliefF feature selection reached to an accuracy rate of 90.21% from ResNet101{fc1000}+Acoustic features. In addition, VGG16{fc7}+Acoustic features with ReliefF reached to an accuracy rate of 90.12%, ResNet18{fc1000}+Acoustic features reached to an accuracy rate of 85.47%, ResNet50 {fc1000}+Acoustic features with ReliefF reached to an accuracy rate of 88.67%, SqueezeNet {pool10}+Acoustic features with ReliefF reached to an accuracy rate of 88.63% and DenseNet201{conv5_block16}+Acoustic features reached to an accuracy rate of 87.29%. The confusion matrix for the best classification result for the EMO_DB dataset is given in Figure 9. The classification results for the IEMOCAP dataset are given in Table 4. The best classification result for the IEMOCAP dataset was obtained with ReliefF feature selection as 85.37% from VGG16{fc7}+Acoustic features. In addition, ResNet18{fc1000}+Acoustic features with ReliefF reached to an accuracy rate of 83.47%, ResNet50{fc1000}+Acoustic features with ReliefF reached to an accuracy rate of 82.74%, ResNet101{fc1000}+ Acoustic features with ReliefF reached to an accuracy rate of 82.36%, SqueezeNet{pool10}+Acoustic features reached to an accuracy rate of 83.87% and DenseNet201{conv5_block16}+Acoustic features with ReliefF reached to an accuracy rate of 82.37%. The confusion matrix for the best classification result for the IEMOCAP _DB dataset is given in Figure 10.

TABLE 2. Classification results on RAVDESS Dataset

Feature Vector	Accuracy %
Acoustic features	66.42
DenseNet201{conv5_block16}	75.43
DenseNet201{conv5_block16}+Acoustic features	76.97
DenseNet201{conv5_block16}+Acoustic features with ReliefF	77.46
ResNet18 {fc1000}	70.25
ResNet18 {fc1000}+Acoustic features	75.38
ResNet18 {fc1000}+Acoustic features with ReliefF	74.56
ResNet50 {fc1000}	73.24
ResNet50 {fc1000}+Acoustic features	76.86
ResNet50 {fc1000}+Acoustic features with ReliefF	78.26
ResNet101 {fc1000}	73.77
ResNet101 {fc1000}+Acoustic features	77.56
ResNet101 {fc1000}+Acoustic features with ReliefF	79.41
SqueezeNet {pool10}	74.23
SqueezeNet {pool10}+Acoustic features	73.46
SqueezeNet {pool10} + Acoustic features with ReliefF	75.81
VGG16 {fc7}	71.15
VGG16 {fc7}+Acoustic features	73.36
VGG16 {fc7}+Acoustic features with ReliefF	74.41

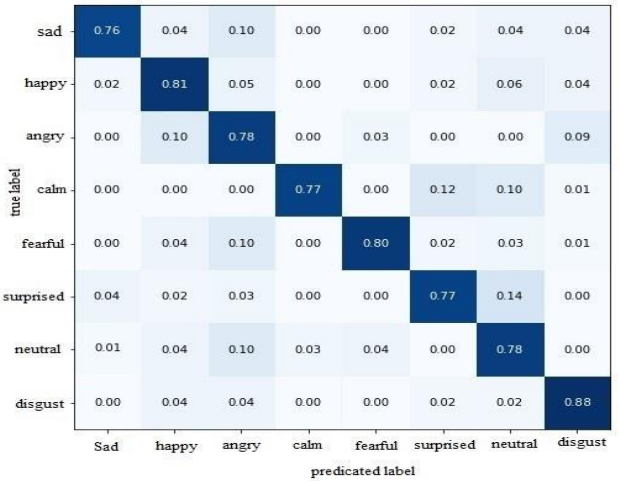


Figure 8. Confusion Matrix for the RAVDESS Dataset

TABLE 3. Classification results on EMO-DB Dataset

Features	Accuracy %
Acoustic features	76.34
DenseNet201{conv5_block16}	85.42
DenseNet201{conv5_block16}+Acoustic features	87.29
DenseNet201{conv5_block16}+Acoustic features with ReliefF	86.48
ResNet18 {fc1000}	85.06
ResNet18 {fc1000}+Acoustic features	85.47
ResNet18 {fc1000}+Acoustic features with ReliefF	85.42
ResNet50 {fc1000}	83.28
ResNet50 {fc1000}+Acoustic features	87.43
ResNet50 {fc1000}+Acoustic features with ReliefF	88.67
ResNet101 {fc1000}	85.58
ResNet101 {fc1000}+Acoustic features	88.98
ResNet101 {fc1000}+Acoustic features with ReliefF	90.21
SqueezeNet {pool10}	86.24
SqueezeNet {pool10}+Acoustic features	87.46
SqueezeNet {pool10}+Acoustic features with ReliefF	88.63
VGG16 {fc7}	85.27
VGG16 {fc7+Acoustic features	89.21
VGG16 {fc7+Acoustic features with ReliefF	90.12

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3043201, IEEE Access

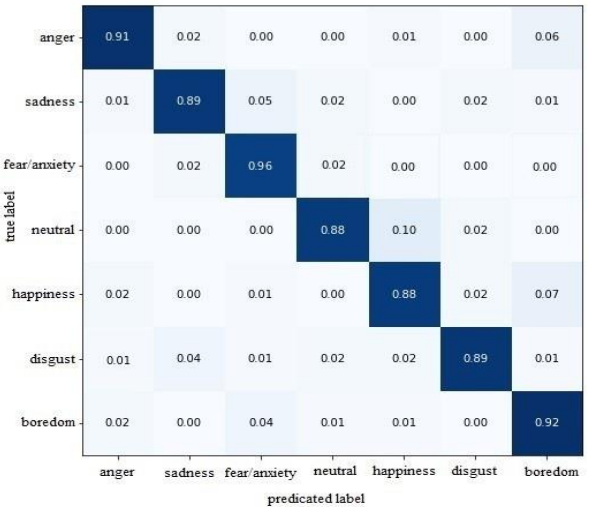


Figure 9. Confusion Matrix for the EMO-DB Dataset

TABLE 4. Classification results on IEMOCAP Dataset

Features	Accuracy %
Acoustic features	71.36
DenseNet201{conv5_block16}	77.51
DenseNet201{conv5_block16}+Acoustic features	80.46
DenseNet201{conv5_block16}+Acoustic features with ReliefF	82.37
ResNet18 {fc1000}	75.50
ResNet18 {fc1000}+Acoustic features	82.46
ResNet18 {fc1000}+Acoustic features with ReliefF	83.47
ResNet50 {fc1000}	78.65
ResNet50 {fc1000}+Acoustic features	81.56
ResNet50 {fc1000}+Acoustic features with ReliefF	82.74
ResNet101 {fc1000}	77.72
ResNet101 {fc1000}+Acoustic features	81.45
ResNet101 {fc1000}+Acoustic features with ReliefF	82.36
SqueezeNet {pool10}	81.45
SqueezeNet {pool10}+Acoustic features	83.87
SqueezeNet {pool10}+Acoustic features with ReliefF	82.68
VGG16 {fc7}	81.26
VGG16 {fc7}+Acoustic features	84.52
VGG16 {fc7}+Acoustic features with ReliefF	85.37

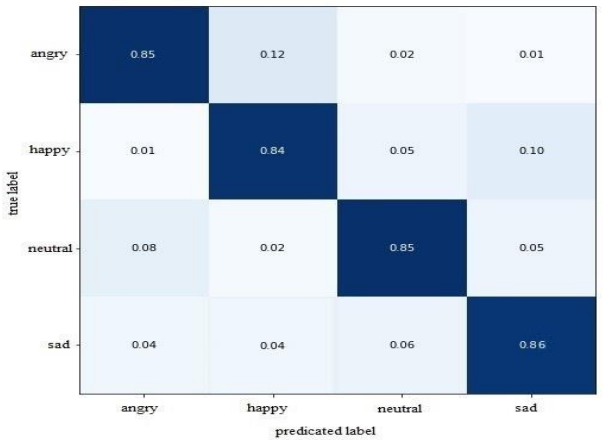


Figure 10. Confusion Matrix for the IEMOCAP Dataset

Training computational complexity of Pre-trained CNNs are shown in Table 5.

Table 5. Computational complexity of CNNs.

Models	Training Time(s) for RAVDESS Dataset	Training Time(s) for EMO-DB Dataset	Training Time(s) for IEMOCAP Dataset
DenseNet201	854	364	647
ResNet18	178	96	148
ResNet50	257	101	189
ResNet101	299	134	207
SqueezeNet	237	82	126
VGG16	372	178	246

Besides, experiments were carried out by the transfer learning method using only spectrogram images. The remaining parameters of the original models, except for the fully connected layers, are preserved and used as initial values. The last layer is set to be the same size as the number of classes in the new data. Hyper parameters are selected as follows:

- minibatch size is 64
- maximum epoch number is 32
- learning rate is 1e-4

The results obtained with the transfer learning method are given in Table 6. According to the data in Table 6, the best classification result for the RAVDESS dataset obtained from ResNet101 is reached to an accuracy rate of 72.34%, The best classification result for the EMO-DB dataset obtained from ResNet50 is reached to an accuracy rate of 85.49% and the best classification result for the IEMOCAP dataset is obtained from ResNet101, which reached to an accuracy rate of 82.76%.

TABLE 6. Results of Transfer Learning Method

Models	Accuracy % for the RAVDESS	Accuracy % for the EMO-DB	Accuracy % for the IEMOCAP
DenseNet201	70.86	83.49	81.76
ResNet18	70.86	84.18	80.43
ResNet50	70.46	85.49	82.15
ResNet101	72.34	83.67	82.76
SqueezeNet	71.76	84.73	80.75
VGG16	71.43	85.14	81.45

ANOVA was used for statistical analysis of the experimental results and the results are given in table 7. According to the ANOVA statistical test, "P" <0.05 is required for a significant relationship between the results obtained by different methods. As it can be seen from the results in Table 7, the “P” value is generally close to zero as a result of comparing different methods, and this shows us that there is a significant relationship between the results.

Table 7. ANOVA Test Results

Compared Results	P Value for RAVDESS Dataset	P Value for EMO-DB Dataset	P Value for IEMOCAP Dataset

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3043201, IEEE Access

VGG16+Acoustic features & ResNet18+Acoustic features	≈ 0	≈ 0	≈ 0
ResNet18+Acoustic features & ResNet50+Acoustic features	≈ 0	≈ 0	≈ 0
ResNet50+Acoustic features & ResNet101+Acoustic features	0.002	0.002	0.002
ResNet101+Acoustic features & SqueezeNet +Acoustic features	0.141	0.141	0.141
SqueezeNet+Acoustic features & DenseNet201+Acoustic features	0.123	0.118	0.128

In this study, the findings we obtain to better evaluate the performance of the proposed method are compared with the results obtained from other methods used in the literature. Table 8 summarizes outstanding studies on the emotional classification of speech sounds.

Table 8. Performance comparison with Other Approaches

Approach	Features Used	Classifier	Training–Testing Data Splitting	Dataset	Accuracy %
Segokar and Sircar [53]	Continuous wavelet Transform, Prosodic features	SVM	Cross-validation	RAVDESS	60.1
Zeng et al. [54]	Spectrogram	CNN	Cross-validation	RAVDESS	65.97
Bhavan et al. [25]	MFCCs, spectral centroids and MFCC derivatives	Bagged ensemble of SVMs	90% - 10%	RAVDESS	75.69
Proposed model	ResNet101{fc1000}+Acoustic features (Before data augment)	SVM	Cross-validation	RAVDESS	77.26
Proposed model	ResNet101{fc1000}+Acoustic features (After data augment)	SVM	Cross-validation	RAVDESS	79.41
Wang et al. [55]	Fourier parameters, MFCC	SVM (Gaussian kernel)	Not mentioned	EMO-DB	73.3
Kotti et al. [56]	Cepstrum-based features	Linear SVM	Not mentioned	EMO-DB	87.70
Wu et al. [57]	Modulation spectral features (MSFs)	Linear discriminant analysis (LDA)	Cross-validation	EMO-DB	85.84
Proposed model	ResNet101{fc1000}+Acoustic features (Before data augment)	SVM	Cross-validation	EMO-DB	87.68
Proposed model	ResNet101{fc1000}+Acoustic features (After data augment)	SVM	Cross-validation	EMO-DB	90.21
Lee and Tashev [58]	Spectrogram, Frame level features	Recurrent neural network	Cross-validation	IEMOCAP	62.85
Zhao et al. [29]	Mel spectrograms	LSTM	Cross-validation	IEMOCAP	89.16
Proposed model	VGG16{fc7}+Acoustic features (Before data augment)	SVM	Cross-validation	IEMOCAP	82.64
Proposed model	VGG16{fc7}+Acoustic features (After data augment)	SVM	Cross-validation	IEMOCAP	85.37

In [53], they classified speech samples with the SVM classifier using the RAVDESS dataset. The authors employed Continuous Wavelet Transform (CWT) for feature selection. The best classification result obtained through use of 5 fold cross-validation technique and Quadratic SVM, with an accuracy rate of 60.1% In [54], spectrograms produced from a deep neural network (DNN) and speech sounds are used in emotion recognition. Spectrograms are given as input to Gated Residual Networks (GResNets) and an accuracy rate of 65.97% has been achieved. In [25], an SVM Ensemble having a Gauss kernel is suggested. First of all, MFCCs are extracted with spectral centroids to represent emotional speech. Then, an accuracy rate of 75.69% has been accomplished using the wrapper-based feature selection method to obtain the best feature set. In [55], they investigated features based on Fourier parameters (FP) for emotion recognition models. The authors used only 6 of the 7 emotion classes in the EMO-DB dataset, by removing the "disgust" class. FP and MFCC features extracted from the dataset and an average accuracy rate of 73.3% is reached by using the SVM classifier. In [56], binary cascade classification is presented. Many features that are based on energy, pitch, jitter and TEO autocorrelation have been extracted and also the difference in speech patterns due to gender has been investigated. The best emotion recognition accuracy rate achieved by SVM with linear kernel equals to 87.7%. In [57], They only applied traditional machine learning techniques to classify samples in the EMO-DB dataset. The authors proposed a new type of sound features called

modulation spectral features (MSFs). Using (LDA) classifier, they achieved an accuracy rate of 85.8%. In [58], presents a speech emotion recognition system using the recurrent neural network (RNN) model. A powerful learning method with a bidirectional long short-term memory (BLSTM) model has been adopted to extract features of emotional states. In [29], 2D CNNLSTM is used to learn emotion related features from the Mel spectrogram.

V. CONCLUSION

In this paper, a new method based on deep and acoustic features is proposed to recognize human emotions. The proposed method is evaluated on RADVES, EMO-DB and IEMOCAP datasets, which are very popular in the literature. At first, acoustic features and spectrograms are extracted from speech signals. Thereafter, data augmentation process is employed to create additional training data samples for spectrogram images. Subsequently, six pre-trained popular CNNs are used to extract deep features from spectrogram images. These networks are VGG16, ResNet18, ResNet50, ResNet101, SqueezeNet and DenseNet201. Acoustic features and deep features are combined to create a hybrid feature vector. In this way, it is targeted at improving the classification success. Also, ReliefF is used for feature selection. Finally, linear SVM is used for classification. The best performance gained in "ResNet101{fc1000}+Acoustic Features with ReliefF" for the RADVES dataset with an accuracy rate of 79.41, the best performance gained in "ResNet101 {fc1000}+Acoustic features with ReliefF" for the EMO-DB dataset with an accuracy rate of 90.21%, and the best performance gained in "VGG16{fc7}+Acoustic features with ReliefF" for IEMOCAP dataset with an accuracy rate of 85.37%. As shown in Table 2-4, according to the results obtained from the experiments, our proposed method has superior accuracy rates compared with those of previous studies done in the literature. As a consequence, in future studies, it is recommended to develop new techniques for the determination of optimum feature sets in order to improve the accuracy rates of classification.

REFERENCES

[1] P. Schlegel, S. Kniesburges, S. Dürr, A. Schützenberger, and M. Döllinger, "Machine learning based identification of relevant parameters for functional voice disorders derived from endoscopic high-speed recordings," *Sci. Rep.*, vol. 10, no. 1, p. 10517, Jun. 2020, doi: 10.1038/s41598-020-66405-y.

[2] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo, "Recognizing Emotions Induced by Affective Sounds through Heart Rate Variability," *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 385–394, 2015, doi: 10.1109/TAFFC.2015.2432810.

[3] H. Gunes and M. Piccardi, "Bi-Modal Emotion Recognition from Expressive Face and Body Gestures," *J. Netw. Comput. Appl.*, vol. 30, no. 4, pp. 1334–1345, 2007, doi: 10.1016/j.jnca.2006.09.007.

[4] D. Połap, "Model of identity verification support system based on voice and image samples," *J. Univers. Comput. Sci.*, vol. 24, pp. 460–474, Jan. 2018.

[5] G. Lu, L. Yuan, W. Yang, J. Yan, and H. Li, "Speech emotion recognition based on long short-term memory and convolutional neural networks," *Nanjing Youdian Daxue Xuebao (Ziran Kexue Ban)/Journal Nanjing Univ. Posts Telecommun. (Natural Sci.*, vol. 38, pp. 63–69, Oct. 2018, doi: 10.14132/j.cnki.1673-5439.2018.05.009.

[6] V. Garg, H. Kumar, and R. Sinha, "Speech based Emotion Recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers," *2013 National Conference on Communications (NCC)*. IEEE, 2013, doi: 10.1109/ncc.2013.6487987.

[7] K. Han, D. Yu, and I. Tashev, *Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine*. 2014.

[8] S. Mittal, S. Agarwal, and M. J. Nigam, "Real Time Multiple Face Recognition: A Deep Learning Approach," in *Proceedings of the 2018 International Conference on Digital Medicine and Image Processing*, 2018, pp. 70–76, doi: 10.1145/3299852.3299853.

[9] H.-S. Bae, H.-J. Lee, and S.-G. Lee, "Voice recognition based on adaptive MFCC and deep learning," *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*. IEEE, 2016, doi: 10.1109/iciea.2016.7603830.

[10] K. R. Malik, M. Ahmad, S. Khalid, H. Ahmad, F. Al-Turjman, and S. Jabbar, "Image and command hybrid model for vehicle control using Internet of Vehicles," *Trans. Emerg. Telecommun. Technol.*, vol. 31, no. 5, 2019, doi: 10.1002/ett.3774.

[11] A. M. Badshah *et al.*, "Deep features-based speech emotion recognition for smart affective services," *Multimed. Tools Appl.*, vol. 78, no. 5, pp. 5571–5589, 2019, doi: 10.1007/s11042-017-5292-7.

[12] T. Anvarjon, Mustaqeem, and S. Kwon, "Deep-Net: A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features," *Sensors (Basel).*, vol. 20, no. 18, p. 5212, Sep. 2020, doi: 10.3390/s20185212.

[13] A. B. Kandali, A. Routray, and T. K. Basu, "Emotion recognition from Assamese speeches using MFCC features and GMM classifier," *TENCON 2008 - 2008 IEEE Region 10 Conference*. IEEE, 2008, doi: 10.1109/tencon.2008.4766487.

[14] A. Milton, S. Sharmy Roy, and S. Tamil Selvi, "SVM Scheme for Speech Emotion Recognition using MFCC Feature," *Int. J. Comput. Appl.*, vol. 69, no. 9, pp. 34–39, 2013, doi: 10.5120/11872-7667.

[15] D. V. Waghmare, R. Deshmukh, P. Shrishrimal, and G. Janvale, *Emotion Recognition System from Artificial Marathi Speech using MFCC and LDA Techniques*. 2014.

[16] S. Demircan and H. Kahramanlı, "Feature Extraction from Speech Data for Emotion Recognition," *J. Adv. Comput. Networks*, vol. 2, no. 1, pp. 28–30, 2014, doi: 10.7763/jacn.2014.v2.76.

[17] F. Chenchah and Z. Lachiri, "Acoustic Emotion Recognition Using Linear and Nonlinear Cepstral Coefficients," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 11, 2015, doi: 10.14569/ijacsa.2015.061119.

[18] Y. Huang, G. Zhang, X. Li, and F. Da, "Small sample size speech emotion recognition based on global features and weak metric learning," *Shengxue Xuebao/Acta Acust.*, vol. 37, pp. 330–338, May 2012.

[19] X. Li and M. Akagi, "Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model," *Speech Commun.*, vol. 110, pp. 1–12, 2019, doi: 10.1016/j.specom.2019.04.004.

[20] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Commun.*, vol. 49, no. 10–11, pp. 787–800, 2007, doi: 10.1016/j.specom.2007.01.010.

[21] K. Khanchandani and M. Hussain, "Emotion recognition using multilayer perceptron and generalized feed forward neural network," *J. Sci. Ind. Res.*, vol. 68, pp. 367–371, Apr. 2009.

[22] D. Gharavian, M. Sheikhan, and F. Ashoftedel, "Emotion recognition improvement using normalized formant

supplementary features by hybrid of DTW-MLP-GMM model," *Neural Comput. Appl.*, vol. 22, no. 6, pp. 1181–1191, 2012, doi: 10.1007/s00521-012-0884-7.

[23] X. Zhao, S. Zhang, and B. Lei, "Robust emotion recognition in noisy speech via sparse representation," *Neural Comput. Appl.*, vol. 24, no. 7–8, pp. 1539–1553, 2013, doi: 10.1007/s00521-013-1377-z.

[24] T. Özseven, "A novel feature selection method for speech emotion recognition," *Appl. Acoust.*, vol. 146, pp. 320–326, 2019, doi: 10.1016/j.apacoust.2018.11.028.

[25] A. Bhavan, P. Chauhan, Hitkul, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowledge-Based Syst.*, vol. 184, p. 104886, 2019, doi: 10.1016/j.knosys.2019.104886.

[26] S. Li and L. Xu, "Research on Emotion Recognition Algorithm Based on Spectrogram Feature Extraction of Bottleneck Feature," *Comput. Technol. Dev.*, pp. 82–86, 2017.

[27] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, doi: 10.1109/icassp.2018.8462677.

[28] D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomed. Signal Process. Control*, vol. 59, p. 101894, 2020, doi: 10.1016/j.bspc.2020.101894.

[29] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, 2019, doi: 10.1016/j.bspc.2018.08.035.

[30] L. Sun, B. Zou, S. Fu, J. Chen, and F. Wang, "Speech emotion recognition based on DNN-decision tree SVM model," *Speech Commun.*, vol. 115, pp. 29–37, 2019, doi: 10.1016/j.specom.2019.10.004.

[31] F. Daneshfar, S. J. Kabudian, and A. Neekabadi, "Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier," *Appl. Acoust.*, vol. 166, p. 107360, 2020, doi: https://doi.org/10.1016/j.apacoust.2020.107360.

[32] K. Zvarevashe and O. Olugbara, "Ensemble Learning of Hybrid Acoustic Features for Speech Emotion Recognition," *Algorithms*, vol. 13, no. 3, p. 70, 2020, doi: 10.3390/a13030070.

[33] W. Jiang, Z. Wang, J. S. Jin, X. Han, and C. Li, "Speech Emotion Recognition with Heterogeneous Feature Unification of Deep Neural Network," *Sensors (Basel).*, vol. 19, no. 12, p. 2730, Jun. 2019, doi: 10.3390/s19122730.

[34] D. Połap, M. Woźniak, R. Damaševičius, and R. Maskeliūnas, "Bio-inspired voice evaluation mechanism," *Appl. Soft Comput.*, vol. 80, pp. 342–357, 2019, doi: https://doi.org/10.1016/j.asoc.2019.04.006.

[35] R. Lotfidereshgi and P. Gournay, *Biologically inspired speech emotion recognition*. 2017.

[36] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *J. Biomed. Inform.*, vol. 83, pp. 103–111, 2018, doi: https://doi.org/10.1016/j.jbi.2018.05.007.

[37] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, 2002, doi: 10.1109/tsa.2002.800560.

[38] A. Toh, R. Togneri, and S. Nordholm, "Spectral entropy as speech features for speech recognition," *Proc. PEECS*, Jan. 2005.

[39] A. Lidy, Thomas; Rauber, "Computing statistical spectrum descriptors for audio music similarity and retrieval," in *MIREX 2006 - Music Information Retrieval Evaluation*, 2006.

[40] C. K. On, P. M. Pandiyan, S. Yaacob, and A. Saudi, "Mel-frequency cepstral coefficient analysis in speech recognition," *2006 International Conference on Computing & Informatics*. IEEE, 2006, doi: 10.1109/icoci.2006.5276486.

[41] C. Paseddula and S. V Gangashetty, "Late fusion framework for Acoustic Scene Classification using LPCC, SCMC, and log-Mel band energies with Deep Neural Networks," *Appl. Acoust.*, vol. 172, p. 107568, 2021, doi: https://doi.org/10.1016/j.apacoust.2020.107568.

[42] S. Gupta, J. Jaafar, W. F. Wan Ahmad, and A. Bansal, "Feature Extraction Using Mfcc," *Signal Image Process. An Int. J.*, vol. 4, pp. 101–108, Aug. 2013, doi: 10.5121/sipij.2013.4408.

[43] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, 2017, doi: 10.1109/LSP.2017.2657381.

[44] D. C. Cireundefinedan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Flexible, High Performance Convolutional Neural Networks for Image Classification," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, 2011, pp. 1237–1242.

[45] R. Paul *et al.*, "Explaining Deep Features Using Radiologist-Defined Semantic Features and Traditional Quantitative Features," *Tomogr. (Ann Arbor, Mich.)*, vol. 5, no. 1, pp. 192–200, Mar. 2019, doi: 10.18383/j.tom.2018.00034.

[46] K. S. and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2014, [Online]. Available: https://arxiv.org/abs/1409.1556.

[47] F. Landola, S. Han, M. Moskewicz, K. Ashraf, W. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50 × fewer parameters and < 0.5 MB model size," in *ICLR 2017*, 2016, doi: 1602.07360.

[48] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," *Inf. Sci. (Ny).*, vol. 282, pp. 111–135, 2014, doi: 10.1016/j.ins.2014.05.042.

[49] "Support Vector Machines, 1992; Boser, Guyon, Vapnik," in *SpringerReference*, Springer-Verlag.

[50] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, p. e0196391, May 2018, [Online]. Available: https://doi.org/10.1371/journal.pone.0196391.

[51] F. Burkhardt, A. Paeschke, and W. F. S. B. W. M. Rolfes, "A database ofgerman emotional speech," in *Ninth European Conference on SpeechCommunication and Technology*, 2005.

[52] C. Busso *et al.*, "IEMOCAP: interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008, doi: 10.1007/s10579-008-9076-6.

[53] P. Shegokar and P. Sircar, "Continuous wavelet transform based speech emotion recognition," in *2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2016, pp. 1–8, doi: 10.1109/ICSPCS.2016.7843306.

[54] Y. Zeng, H. Mao, D. Peng, and Z. Yi, "Spectrogram based multi-task audio classification," *Multimed. Tools Appl.*, vol. 78, Dec. 2017, doi: 10.1007/s11042-017-5539-3.

[55] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech Emotion Recognition Using Fourier Parameters," *Affect. Comput. IEEE Trans.*, vol. 6, pp. 69–75, Jan. 2015, doi: 10.1109/TAFFC.2015.2392101.

[56] M. Kotti and F. Paternò, "Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema," *Int J Speech Technol*, vol. 15, Jun. 2012, doi: 10.1007/s10772-012-9127-7.

[57] S. Wu, T. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Commun.*, vol. 53, pp. 768–785, May 2011, doi: 10.1016/j.specom.2010.08.013.

[58] J. Lee and I. Tashev, *High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition*. 2015.

MEHMET BILAL ER was born in Sanlıurfa, Turkey, in 1988. He received the B.S degree in computer engineering from Eastern Mediterranean University, Cyprus, in 2010, the M.Sc. degree in computer engineering from The Cankaya University, Turkey, in 2013. The Ph.D degree in computer engineering from the Maltepe University, Turkey, in 2019. He is currently Assistant Professor with the School of Computer Engineering, Harran University, Turkey. His research interests include sound processing and deep learning.