Time-Frequency Image-based Speech Emotion Recognition using Artificial Neural Network

Neha Dewangan[1*,] Kavita Thakur[1], Sunandan Mandal[1], Bikesh Kumar Singh[2]

[1]Pt. Ravishankar Shukla University, Raipur, 492010, India

[2]National Institute of Technology, Raipur, 492010, India

Abstract

Automatic Speech Emotion Recognition (ASER) is a state-of-the-art application in artificial intelligence. Speech recognition intelligence is employed in various applications such as digital assistance, security, and other human-machine interactive products. In the present work, three open-source acoustic datasets, namely SAVEE, RAVDESS, and EmoDB, have been utilized[1-3]. From these datasets, six emotions namely anger, disgust, fear, happy, neutral, and sad, are selected for automatic speech emotion recognition. Various types of algorithms are already reported for extracting emotional content from acoustic signals. This work proposes a time-frequency (t-f) image-based multiclass speech emotion classification model for the six emotions mentioned above. The proposed model extracts 472 grayscale image features from the t-f images of speech signals. The t-f image is a visual representation of the time component and frequency component at that time in the two-dimensional space, and differing colors show its amplitude. An artificial neural network-based multiclass machine learning approach is used to classify selected emotions. The experimental results show that the above-mentioned emotions' average classification accuracy (CA) of 88.6%, 85.5%, and 93.56% is achieved using SAVEE, RAVDESS, and EmoDB datasets, respectively. Also, an average CA of 83.44% has been achieved for the combination of all three datasets. The maximum reported average classification accuracy (CA) using spectrogram for SAVEE, RAVDESS, and EmoDB dataset is 87.8%, 79.5 %, and 83.4%, respectively[4-6]. The proposed t-f image-based classification model shows improvement in average CA by 0.91%, 7.54%, and 12.18 % for SAVEE, RAVDESS, and EmoDB datasets, respectively. This study can be helpful in human-computer interface applications to detect emotions precisely from acoustic signals

Keywords: Time-Frequency Image; Neural Network; Automatic Speech Emotion Recognition; Acoustic Signal; Grayscale Image Feature

1. Introduction

Emotions are an individual's feelings about a situation. It is the body's physical and emotional response to a person's thoughts and feelings. People express emotions in identical situations differently. People's expressions on their faces and their voices usually reflect their feelings. For essential. Different emotions can be expressed as happiness, sadness, anger, anxiety, cheerfulness, excitement, lonely, helplessness, annoyance, etc. Additionally, it can be categorized as positive and negative emotions. The primary emotions are classified into six categories: happy, sad,

Corresponding authors: (E-mail: dewanganneha92@gmail.com)

anger, fear, disgust, and surprise[7]. Human uses many types of gestures, including speech, as a means of communication. Speech is the most common, easiest, and natural form of communication.

Communication can be done in other ways as well, but it lacks emotions such as text messages, without proper emotions, can induce misunderstanding. So, emojis were introduced, which replicated the emotions. Using emojis in text messages conveys our emotions. When people speak, their emotions are reflected in their voices, facilitating better communication. One of the unique physiological processes is speech generation. Therefore, the inherent emotional state in the speech can also help to detect the mental and physical health of human[8].

Automatic Speech Emotion Recognition (ASER) is the process of recognizing emotions in speech. ASER uses speech analysis and machine learning to create an automated system that can detect the emotions of human beings from their voice for various purposes. Acoustic feature extraction plays a prominent role in the ASER system to analyze a speaker's voice and determine the speaker's emotional state.

A variety of methods are used to classify emotions from the acoustic signal, such as Prosodic features, Mel-frequency cepstral coefficients, Pitch Frequencies, vocalization duration, and spectrogram. ASER-based systems can be applied to make supportive tools for some areas such as healthcare, digital assistant-based customer service, marketing and other human-machine interactive services.

The main challenge in ASER is extracting hidden features embedded in the acoustic signal. For this, various methods are applied to extract features. Prosodic features, voice quality features, spectral features, and Teager energy features are the different types of speech features8. Time domain, as well as frequency domain feature extraction techniques for acoustic signals, were reported by many researchers. Some researcher uses spectrogram of the acoustic signal. The present method extracts emotional features from time-frequency images (spectrogram). The t-f image is a visual representation of the time component and frequency component at that time in the two-dimensional space, and differing colors show its amplitude. Some recent studies using spectrogram-based emotion recognition with various datasets and classifiers are reported briefly below.

Özseven investigated the effects of texture analysis methods and spectrogram images on speech emotion recognition[9]. In their work, four different texture analysis methods were used to obtain features from the spectrogram images of the speech signals. Also, acoustic features were studied to compare texture analysis methods with acoustic analysis methods. They achieved 82.4%, 60.9%, and 64.6% success rates for texture analysis method and 82.8%, 56.3%, and 74.3% success rates for acoustic analysis methods for EMO-DB(Berlin Database of Emotional Speech), eNTERFACE'05 and SAVEE(Surrey Audio-Visual Expressed Emotion) databases, respectively using SVM classifier. When comparing SER performance based on approach, the acoustic analysis outperforms the texture analysis by 0.4% for EMO-DB and 9.7% for SAVEE and underperforms by 4.6% for eNTERFACE'05. Badshah et al. proposed a model for SER using a spectrogram with the deep convolutional neural network[6]. The EmoDB dataset has been utilized in their work, and the acoustic signals have been converted to spectrogram images. They used three convolutional layers and three fully connected layers to extract suitable features from spectrogram images. The Softmax layer performs the final classification for seven emotions embedded in acoustic signals of the EmoDB dataset. They achieved an overall classification accuracy of 83.4% for all seven emotions.

Hajarolasvadi and Demirel extracted 88-dimensional vectors from acoustic signals of SAVEE, RML, and eNTERFACE’05 databases[10]. They also obtained each signal's spectrogram and then applied k-means clustering to all the extracted features to get keyframes. Then, the corresponding spectrogram of keyframes is encapsulated in a 3-D tensor form, which works as an input of 3-D CNN. The 3-D CNN consists of 2 convolutional layers and a fully connected layer for classifying six emotions, anger, disgust, fear, happy, sad, and surprise, in the dataset mentioned. They achieved 81.05%, 77% & and 72.55% overall classification accuracy using SAVEE, RML(Ryerson Multimedia Laboratory), and eNTERFACE'05 datasets, respectively. MELBP variants of spectrogram images have been used to recognize emotion from the acoustic signal by Mohammed and Hasan[11]. They converted the emotional acoustic signals into 2D spectrogram images, and then four forms of Extended Local Binary Pattern (ELBP) were generated to extract the emotional features from spectrogram images. ELBP provides information about direction and variation in amplitude intensities for the given emotions; as a result, more effective feature vectors were captured. In this paper, a Multi-Block of ELBP (MELBP) using the histogram is proposed to highlight the important features of the spectrogram image. Here, Deep Belief Network (DBN) is used to classify the emotions from extracted features. For the well know SAVEE dataset, they achieved 72.14% accuracy.

Sönmez and Varol developed a lightweight, effective speech emotion recognition method called multi-level local binary pattern and local ternary pattern abbreviated as 1BTPDN[12]. This method first applied a one-dimensional local binary pattern (1D-LBP) and a one-dimensional local ternary pattern (1D-LTP) on the raw speech signal. Then 1D discrete wavelet transform (DWT) with nine levels was utilized to extract the features. Out of 7680 features, 1024 features are selected using neighbourhood component analysis (NCA). Using a third-degree polynomial kernel-based support vector machine as a classifier, they obtained success rates of 89.16%, 76.67%, and 74.31% for EMO-DB, SAVEE, and EMOVO (an Italian emotional speech database) databases, respectively. Wang extracted texture image information from a spectrogram of the speech signals to sense emotions embedded in speech[13]. Two open-source emotional datasets have been used in their work, namely EmoDB and eNTERFACE, along with their self-recorded dataset (KHUSC-EmoDB) for the cross-corpus method. Firstly, the speech signals of the dataset mentioned above are converted into a spectrogram; after that, it is transformed into a normalized grayscale image, and then a cubic curve is used to enhance the image contrast. The features were extracted by Laws' Masks, based on the principle of texture energy measurement and SVM is used as a classifier. Their experimental results show that the correct classification rates range from 65.20% to 77.42%.

Contribution of the paper

1. We implemented and evaluated t-f image-based multiclass emotional state classification model using the BPANN classifier.
2. We evaluate and compare the SER model's performance using various datasets.

The rest of the paper is arranged in the following section: The material and Methods section includes a brief discussion about datasets, time-frequency images, feature extraction, BPANN classifier, and multiclass ASER model. The next section is results and discussions, followed by the conclusion.

2. Material and Methods

2.1 Dataset

For this study, we have selected three benchmarked database of different native speakers and different languages, which includes male and female speakers. The datasets used in this paper are discussed briefly below:

2.1.1 SAVEE

Surrey Audio-Visual Expressed Emotion (SAVEE) is a well-known dataset of emotional speech. It contains an audio-visual signal with seven emotions, anger, disgust, fear, happiness, neutral, sadness, and surprise, a total of 480 speech signals in .wav format of four male actors. Each subject's audio-visual signal was recorded for seven emotions, 30 utterances for neutral emotions, and 15 sentences for each remaining emotion. The speech signals were recorded in a visual media lab with 16-bit encoding and a 44.1 kHz sampling rate. All subjects were British English speakers[1].

2.1.2 RAVDESS

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is an open-source dataset that contains speech and song, audio, and video signals of 24 actors (12 male, 12 female). All actors were native North Americans, and their speech signals contained anger, disgust, fear, calm, happiness, neutral, sad, and surprise emotions. The speech signals were recorded in a studio with 16-bit encoding and a 48 kHz sampling rate. RAVDESS dataset contains 60 sentences per actor, i.e., 60x24 = 1440, and all files are in .wav format[2].

2.1.3 EmoDB:

The EmoDB dataset (Berlin Database of Emotional Speech) is created by the Institute of Communication Science, Technical University, Berlin, Germany, which can be openly accessed on their website. This dataset contains speech signals from 10 speakers (5 male, 5 female) on seven emotions: anger, fear, boredom, disgust, happy, sad, and neutral. A total of 535 utterances were recorded with a 48 kHz sampling rate[3].

In the present work, only six common emotions in the above dataset have been selected, and speech signals of emotions, namely anger, disgust, fear, happy, neutral, and sad are used. The number of utterances for each emotion is listed in Table 1. A total of 1870 utterances, i.e., 360, 1056 and 454, are utilized from SAVEE, RAVDESS and EmoDB datasets, respectively.

Table 1 — Emotion and number of utterances selected from the various dataset

S. No.	Dataset → Emotion ↓	SAVEE	RAVDESS	EmoDB
1.	Anger	60	96	127
2.	Disgust	60	192	46
3.	Fear	60	192	69
4.	Happy	60	192	71
5.	Neutral	60	192	79
6.	Sad	60	192	62
	Total =	360	1056	454

2.2 Time-Frequency Image

The time-frequency image is a visual representation of the time component and frequency component at that time in the two-dimensional space, and differing colors show its amplitude. Dark blue shows low amplitude, and bright colors yellow to red show high amplitude (Fig.1). It is also known as a spectrogram, and when used in acoustic signals, it is called voicegram or voiceprints. The speech signal represents a 1-D signal and provides information such as speech rate, amplitude, and space between each sample, giving information about emotions. Similarly, the t-f image represents a 2-D image with color-coded amplitudes, which gives information about emotions embedded in them. The t-f image is usually obtained by applying Fast Fourier Transform (FFT) on the acoustic signal. It begins with the decomposition of the acoustic signals into small time frames. Each frame is converted to the frequency domain from the time domain by applying windowed Short-Time Fourier Transform (STFT), as shown in (eq.1). Here, hamming window with 50% overlapping is used.

$$X(k,t) = \sum_{n=0}^{N-1} x(n)w(n-t)\, e^{\frac{-2\pi jkn}{N}} \quad k=0,1,2,\ldots\ldots N-1 \quad (1)$$

Where X(k,t) is the time-frequency representation of acoustic signal x(n), x(n) is preprocessed Acoustic signal, w(n) represents the Hamming window function, N represents the length of the window function, k represents the corresponding frequency, f (k) = kfs/N , where fs is the sampling frequency

The following equation (eq.2) generates the coefficients of a Hamming window:

$$w(n)=0.54-0.46\cos(2\pi nN), \quad 0\le n\le N \quad (2)$$

The time-frequency image is a very reliable form to extract features for ASER. It holds rich information which can't be extracted in the time domain or frequency domain alone5. Due to this reason, time-frequency image has been used to improve the study in various fields. In many applications, the time-frequency image has been used to classify sound events, speech recognition, speech emotion recognition, and speaker recognition[14-17]. In this paper, MATLAB © R2021a has been used to replicate acoustic signal into a t-f image. (Fig.1) shows the acoustic signal and their t-f image for six emotions.

2.3 Feature extraction from grayscale image

Once the t-f image is obtained for every acoustic signal, it is converted into a 400x400 grayscale image. The grayscale image is represented by gray colors with binary values between 0-255. 0 shows the black color, and 255 shows the white color. Using various statistical methods, 472 features are extracted such as First Order Statistics (FOS), Haralick Spatial Gray Level Dependence Matrices (SGLDM), Gray Level Difference Statistics (GLDS), Neighborhood Gray Tone Difference Matrix (NGTDM), Statistical Feature Matrix (SFM), Spectral Texture of Images (STI), Gray Level Run Length Matrix (GLRLM), etc. (see Table 2). Next, these features were given to the back propagation artificial neural network (BPANN) classifier to train and test emotions embedded in acoustic signals[18].

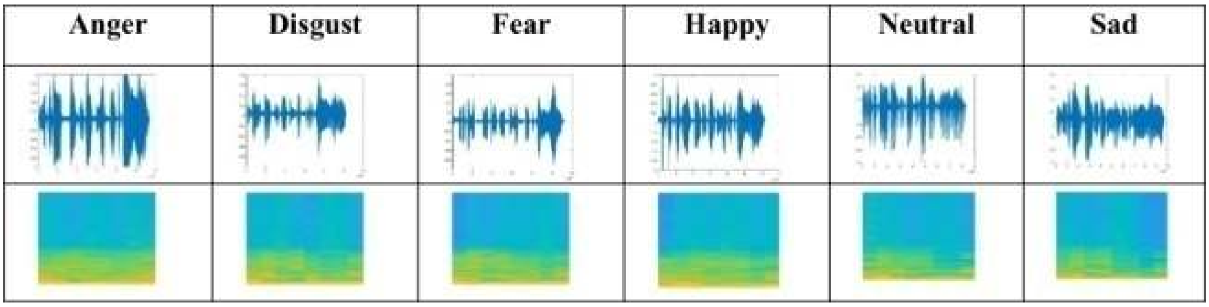


Fig.1— Acoustic signal and t-f image of 6 emotions

Table 2 — List of grayscale image features[18]

Feature Category	Feature Name	No. of Features
Statistical Features	Mean, Variance, median, mode, skewness	5
Haralick textural features	Mean and range values are calculated for features, namely angular second moment, contrast, correlation, a sum of squares, homogeneity, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, information measures of correlation-1, information measures of correlation-2	26
Gray level difference statistics (GLDS)	Homogeneity, contrast, energy, entropy	4
Neighbourhood gray-tone difference matrix (NGTDM)	Coarseness, contrast, busyness, complexity, strength	5
Statistical feature matrix (SFM)	Coarseness, contrast, periodicity, roughness	4

Texture energy measures (TEM)	LL, EE, SS, LE, ES, and LS kernel-based TEM features	6
Fractal dimension texture analysis (FDTA)	FDTA-H1, FDTA-H2, FDTA-H3, FDTA-H4	4
Shape	Area, perimeter, perimeter square per unit area	3
Spectral texture of images (STI)	199 features of spectral energy distribution as a function of radius, 180 features of spectral energy distribution as a function of angle	379
Invariant moments of image (IMI)	MI1-IMI7	7
Statistical measures of texture (SMT)	Average gray level, average contrast, measure of smoothness, third moment, uniformity, entropy	6
Gray-level run length matrix-based properties (GLRLP)	SRE, LRE, GLN, RLN, RP, LGRE, HGRE, SRLGE, SRHGE, LRLGE, LRHGE	11
Texture feature using Segmentation based fractal texture analysis (SFTA) algorithm	SFTA1-SFTA12	12
Total		**472**

2.4 BPANN

The neural network is one of the most broadly used classifiers. In neural networks, there are two algorithms: feedforward and backpropagation. In the present work, backpropagation artificial neural network (BPANN) is utilized to classify emotions. This type of neural network needs supervised learning; it contains one input layer, one output layer, and some hidden layer. In this method, the weight of neurons is changed according to the errors between the actual output value and the expected output value which leads to enhancement in the output. This method repeats the entire dataset in training duration to minimize the error[19].

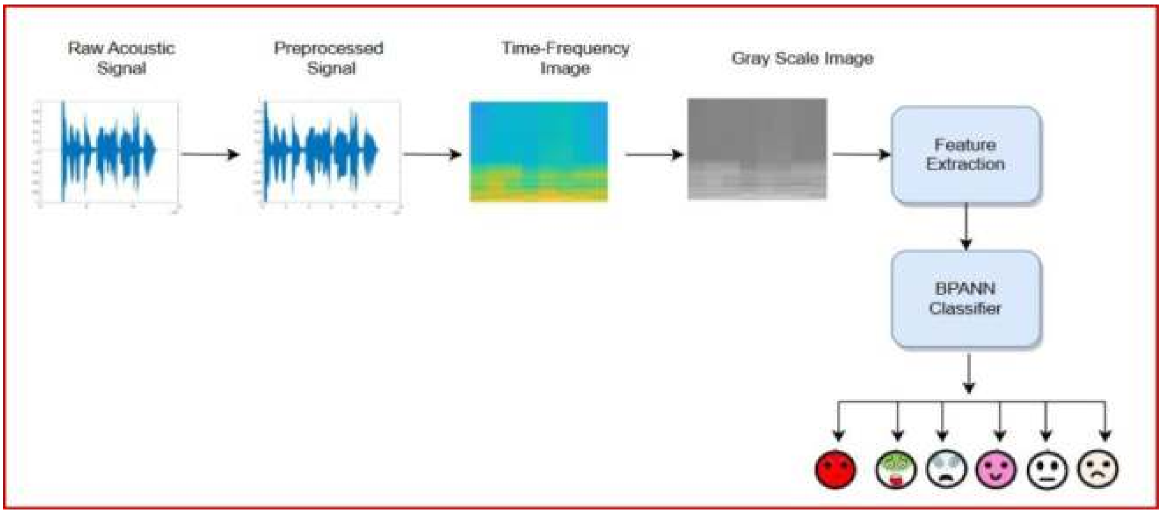


Fig.2 — A t-f image-based Speech Emotion Recognition (SER) model

2.4.1 Experimental analysis of the multiclass SER model

In the present work, three benchmark datasets of the emotional speech signal, namely SAVEE, RAVDESS, and EmoDB, have been utilized to recognize six emotions, i.e., anger, fear, disgust, happy, neutral, and sad. The speech signals of each dataset are firstly preprocessed to remove noise and unwanted signals. It is a necessary part before feature extraction. Here, the preprocessing method abides by two stages.

The first stage is noise removal using a bandpass filter with a 20 Hz to 20 kHz frequency range. In the second stage, the silence part of the speech signal has been removed to decrease the frame length and unwanted signals. After that, speech signals are transformed into 2-D time-frequency images using the FFT method. The t-f image is then converted into a 400x400 grayscale image using MATLAB ©R2021a software. The t-f image-based SER model is shown in (Fig.2). 472 features have been extracted from these grayscale images. These features are used as input for the BPANN classifier. Here, we studied and evaluated the performance of the SER model using each dataset separately and with the mixed dataset. Each dataset is divided using a 5-fold data division protocol. This method divides the whole data into 5 equal parts, 4 used for training purposes, and 1 for testing. It repeats itself process 5 times with different validation data[20].

3. Results and Discussion

In the present work, we proposed the t-f image-based SER model for classifying six common emotions (i.e., anger, disgust, fear, happy, neutral, and sad) from SAVEE, RAVDESS, and EmoDB datasets. The average accuracies of the classifier using grayscale features with 5-fold BPANN are shown in (Fig.3). We got the highest average classification accuracy (CA) of 93.56 % for the EmoDB dataset, followed by the SAVEE dataset with 88.60% and for RAVDESS dataset it is 85.5%. Also, the average CA of 83.44% is obtained for the mixed dataset.

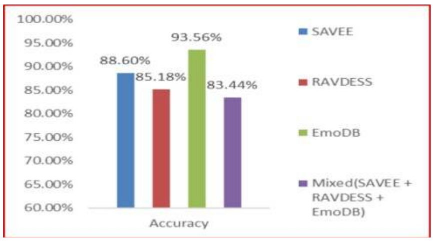


Fig.3 — Classification accuracy of BPANN-based SER model for various datasets

The true positive rate (TPR) and false positive rate (FPR) of six emotions of all three datasets used in the present work are shown in Table 3. The TPR represents the correct positive results during the test among all positive samples. While the FPR gives the amount of incorrectly positive results out of all the negative samples during the test, the FPR tells you how frequently those results occur. TPR should be high, and FPR should be low. Here, the highest TPR of 100% is obtained for the sad emotion of the EmoBD dataset. For the RAVDESS dataset, sad emotion also shows the highest TPR of 87.37%, while for the SAVEE dataset, a TPR of 91.67% is the highest for neutral emotion. The lowest FPR is obtained for the sad emotion of the EmoDB dataset, which is 0.53%. For other datasets, the lowest FPR is 1.39% for the sad emotion of SAVEE and 1.88% for the neutral emotion of RAVDESS. In overall analysis, the EmoDB dataset gives the highest CA and TPR, and lowest FPR from all three datasets utilized in this work.

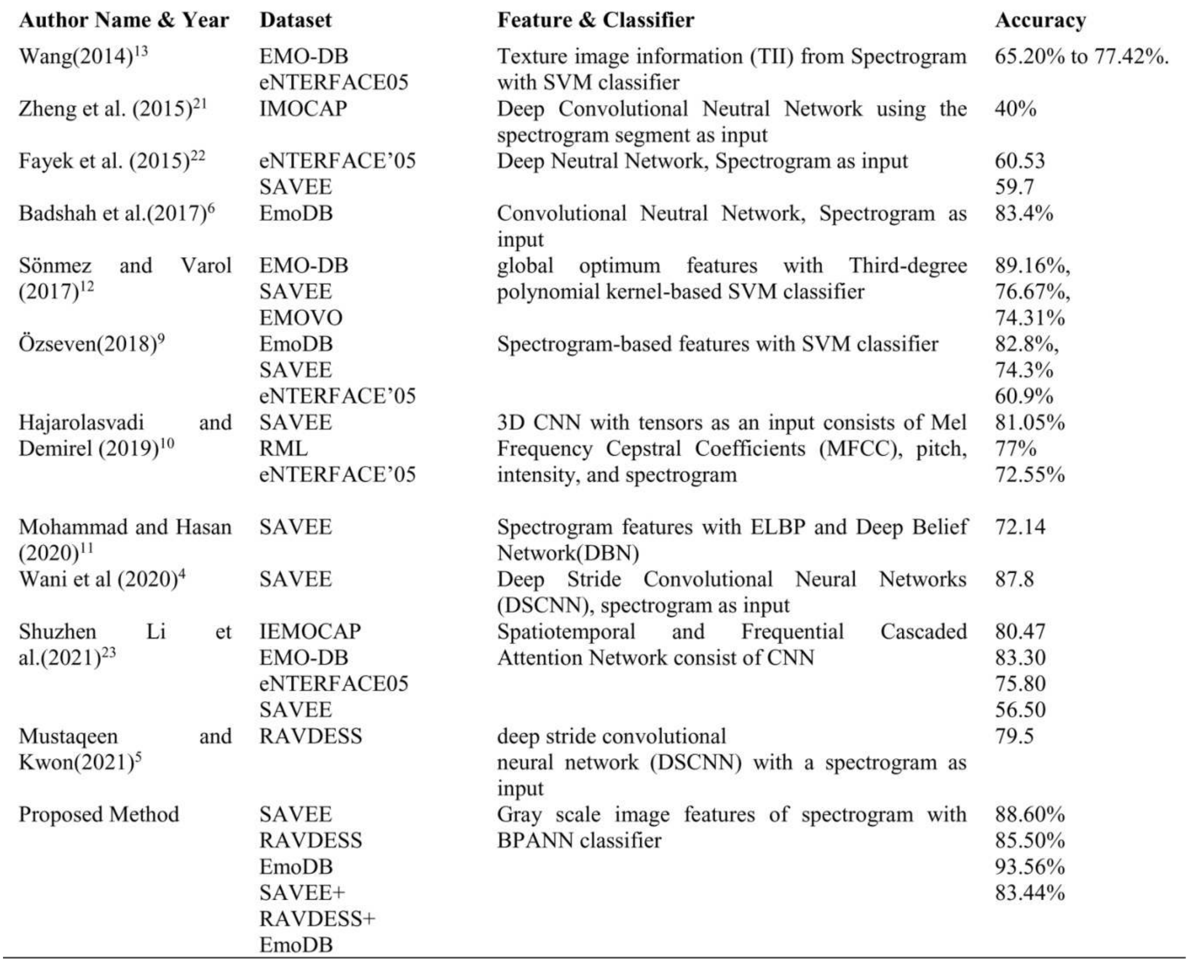
Table 3 — TPR & FPR for six emotions of SAVEE, RAVDESS, EmoDB

Dataset	Emotion	TPR (%)	FPR (%)
SAVEE	Anger	90.00	3.00
	Disgust	85.00	4.00
	Fear	88.33	1.67
	Happy	88.33	2.00
	Neutral	91.67	1.67
	Sad	88.33	1.39
RAVDESS	Anger	87.13	3.72
	Disgust	86.11	3.02
	Fear	84.24	3.13
	Happy	82.65	3.12
	Neutral	85.26	1.88
	Sad	87.37	3.00
EmoDB	Anger	95.20	2.18
	Disgust	93.33	1.50
	Fear	85.71	1.87
	Happy	88.57	1.60
	Neutral	94.67	0.81
	Sad	100.00	0.53

Table 4 shows the performance comparison of different works done on spectrogram features for speech emotion recognition. In this table, many researchers worked on different datasets of emotional signals viz. EmoDB, eNTERFACE, IMOCAP, SAVEE, EMOVO and RAVDESS. Here, only three datasets are participating in the comparison i.e., SAVEE, RAVDESS and

EmoDB. Using Deep Stride Convolutional Neural Network (DSCNN) Wani et al.'s work show the highest average classification accuracy (CA) for the SAVEE dataset with 87.8%4. The highest average classification accuracy (CA) of 79.5% for the RAVDESS dataset using Deep Stride Convolutional Neural Network (DSCNN) is reported by Mustaqeen and Kwon5. The highest average classification accuracy (CA) for the EmoDB dataset6 is 83.4%, reported by Badshah et al. using a convolutional neural network. The proposed t-f image-based model with Back Propagation Artificial Neural Network(BPANN) shows better results with average classification accuracy (CA) of 88.6%, 85.5%, and 93.56% for SAVEE, RAVDESS, and EmoDB datasets, respectively. This shows improvement in maximum reported average CA by 0.91%, 7.54%, and 12.18 % for SAVEE, RAVDESS, and EmoDB datasets, respectively.

Table 4 — Performance comparison with other research works



Author Name & Year	Dataset	Feature & Classifier	Accuracy
Wang(2014)[13]	EMO-DB eNTERFACE05	Texture image information (TII) from Spectrogram with SVM classifier	65.20% to 77.42%.
Zheng et al. (2015)[21]	IMOCAP	Deep Convolutional Neutral Network using the spectrogram segment as input	40%
Fayek et al. (2015)[22]	eNTERFACE'05 SAVEE	Deep Neutral Network, Spectrogram as input	60.53 59.7
Badshah et al.(2017)[6]	EmoDB	Convolutional Neutral Network, Spectrogram as input	83.4%
Sönmez and Varol (2017)[12]	EMO-DB SAVEE EMOVO	global optimum features with Third-degree polynomial kernel-based SVM classifier	89.16%, 76.67%, 74.31%
Özseven(2018)[9]	EmoDB SAVEE eNTERFACE'05	Spectrogram-based features with SVM classifier	82.8%, 74.3% 60.9%
Hajarolasvadi and Demirel (2019)[10]	SAVEE RML eNTERFACE'05	3D CNN with tensors as an input consists of Mel Frequency Cepstral Coefficients (MFCC), pitch, intensity, and spectrogram	81.05% 77% 72.55%
Mohammad and Hasan (2020)[11]	SAVEE	Spectrogram features with ELBP and Deep Belief Network(DBN)	72.14
Wani et al (2020)[4]	SAVEE	Deep Stride Convolutional Neural Networks (DSCNN), spectrogram as input	87.8
Shuzhen Li et al.(2021)[23]	IEMOCAP EMO-DB eNTERFACE05 SAVEE	Spatiotemporal and Frequential Cascaded Attention Network consist of CNN	80.47 83.30 75.80 56.50
Mustaqeen and Kwon(2021)[5]	RAVDESS	deep stride convolutional neural network (DSCNN) with a spectrogram as input	79.5
Proposed Method	SAVEE RAVDESS EmoDB SAVEE+ RAVDESS+ EmoDB	Gray scale image features of spectrogram with BPANN classifier	88.60% 85.50% 93.56% 83.44%

4. Conclusions

This study's proposed model is based on time-frequency images of acoustic signals to recognize emotions. This model shows improvement by 0.91%, 7.54%, and 12.18 % for SAVEE, RAVDESS, and EmoDB datasets, respectively, as compared to the maximum reported accuracies for the same datasets. The time-frequency image-based study is the most prominent in the area of ASER. Much research has already been done in the ASER application area for human-computer interfaces. Still, it needs more improvement for much more perfection and accuracy and a less complex structure so it can be used widely at a low cost.

References

[1] Haq S, Jackson PJ, Edge J, InProc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'08), 2008.

[2] Livingstone SR, Russo FA, PloS one. 2018 May 16;13(5):e0196391, DOI :10.1371/journal.pone.0196391

[3] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B, InInterspeech 2005 Sep 4 (Vol. 5, pp. 1517-1520).

[4] Wani TM, Gunawan TS, Qadri SA, Mansor H, Kartiwi M, Ismail N, In2020 6th International Conference on Wireless and Telematics (ICWT) 2020 Sep 3 (pp. 1-6). IEEE. DOI: 10.1109/ICWT50448.2020.9243622

[5] Kwon S, Sensors. 2019 Dec 28;20(1):183.DOI: 10.3390/s20010183

[6] Badshah AM, Ahmad J, Rahim N, Baik SW, In 2017 international conference on platform technology and service (PlatCon) 2017 Feb 13 (pp. 1-5). IEEE. DOI: 10.1109/PlatCon.2017.7883728

[7] Ekman P, Friesen WV, Ellsworth P, Elsevier; 2013 Oct 22.

[8] Akçay MB, Oğuz K, Speech Communication. 2020 Jan 1;116:56-76.DOI: 10.1016/j.specom.2019.12.001

[9] Özseven T, Applied Acoustics. 2018 Dec 15;142:70-7. DOI: 10.1016/j.apacoust.2018.08.003

[10] Hajarolasvadi N, Demirel H, Entropy. 2019 May 8;21(5):479. DOI: 10.3390/e21050479

[11] Mohammed SN, Hassan AK, Int J Intell Eng Syst. 2020;13(5):257-66.

[12] Sönmez YÜ, Varol A, IEEE Access. 2020 Oct 16;8:190784-96. DOI: 10.1109/ACCESS.2020.3031763

[13] Wang KC, Sensors. 2014 Sep 9;14(9):16692-714. DOI: 10.3390/s140916692

[14] Mao Q, Dong M, Huang Z, Zhan Y, IEEE transactions on multimedia. 2014 Sep 29;16(8):2203-13. DOI: 10.1109/TMM.2014.2360798

[15] Yu D, Seltzer ML, Li J, Huang JT, Seide F, arXiv preprint arXiv:1301.3605. 2013 Jan 16. DOI: 10.48550/arXiv.1301.3605

[16] Dennis J, Tran HD, Li H, IEEE signal processing letters. 2010 Dec 20;18(2):130-3. DOI: 10.1109/LSP.2010.2100380

[17] Lee H, Pham P, Largman Y, Ng A, Advances in neural information processing systems. 2009;22.

[18] Singh BK, Verma K, Thoke AS, Procedia Computer Science. 2015 Jan 1;46:1601-9. DOI: 10.1016/j.procs.2015.02.091

[19] Goh AT, Artificial intelligence in engineering. 1995 Jan 1;9(3):143-51. DOI: 10.1016/0954-1810(94)00011-S

[20] Browne MW, Journal of mathematical psychology. 2000 Mar 1;44(1):108-32. DOI: 10.1006/jmps.1999.1279

[21] Zheng WQ, Yu JS, Zou YX, In2015 international conference on affective computing and intelligent interaction (ACII) 2015 Sep 21 (pp. 827-831). IEEE. DOI: 10.1109/ACII.2015.7344669

[22] Fayek HM, Lech M, Cavedon L, In2015 9th international conference on signal processing and communication systems (ICSPCS) 2015 Dec 14 (pp. 1-5). IEEE. DOI: 10.1109/ICSPCS.2015.7391796

[23] Li S, Xing X, Fan W, Cai B, Fordson P, Xu X, Neurocomputing. 2021 Aug 11;448:238-48. DOI: 10.1016/j.neucom.2021.02.094