




An efficient algorithm for recognition of emotions from speaker and language independent speech using deep learning

Youddha Beer Singh^{1,2}  · Shivani Goel¹

Received: 11 November 2019 / Revised: 11 September 2020 / Accepted: 22 December 2020 /

Published online: 20 January 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC part of Springer Nature 2021

Abstract

Automatic emotion recognition from speech is a demanding and challenging problem. It is difficult to differentiate between the emotional states of humans. The major problem with this task is to extract the important features from the speech in case of hand-crafted features. The accuracy for emotion recognition can be increased using deep learning approaches which use high level features of speech signals. In this work, an algorithm is proposed using deep learning to extract the high-level features from raw data with high accuracy irrespective of language and speakers (male/females) of speech corpora. For this, the .wav files are converted into the RGB spectrograms (images) and normalized to size (224x224x3) for fine-tuning these for Deep Convolutional Neural Network (DCNN) to recognize emotions. DCNN model is trained in two stages. From stage-1 the optimal learning rate is identified using the Learning Rate (LR) range test and then the model is trained again with optimal learning rate in stage-2. Special stride is used for down-sampling the features with reduced model size. The emotions considered are happiness, sadness, anger, fear, disgust, boredom/surprise and neutral. The proposed algorithm is tested on three popular public speech corpora EMODB (German), EMOVO (Italian), and SAVEE (British English). The accuracy of emotion recognition reported is better as compared to the existing studies for different languages and speakers.

Keywords Human emotion recognition · Spectrogram · DCNN · Learning rate

✉ Youddha Beer Singh
youddhabeer.singh@bennett.edu.in; youddhabeersingh@gmail.com

Shivani Goel
shivani.goel@bennett.edu.in

¹ Computer Science and Engineering Department, School of Engineering and Applied Sciences, Bennett University, Greater Noida 201310, India

² Department of Computer Science and Engineering, Galgotias College of Engineering and Technology, Greater Noida, UP, India

1 Introduction

Emotions are the key components of effective human-computer interaction. Affective computing is a field of science that deals with the understanding of emotions. Sentiment analysis aims to identify emotions that can be used in many systems like recommender systems for improving customer relationships. This paper is focusing on emotion recognition from speech. It's a challenge to generate systems that can communicate with humans via speech [54]. It is important to recognize, respond and analyse the emotional state of humans, car driving system, call centre conversation analysis, robotics, call analysis in case of emergency services such as fire brigade, ambulance, speech to speech translation, and many more. In emotion recognition, one of the main research issues is the extraction of acoustic features from speech signals to get better accuracy. Several features have been mentioned in the literature categorized as Prosodic features, Spectral features, and a combination of both [13].

Acoustic features include intonation, energy, speaking rate, fundamental frequency, duration, intensity, and spectral characteristics. For emotion recognition, there are several machine learning algorithms (MLAs) that have been used to recognize emotions based on the spectral and prosodic features of speech signals. Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), linear discriminant classifiers, nearest neighbourhood classifiers, Support Vector Machine (SVM), and Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNN) are some classifiers that are widely used to recognize emotions based on the acoustic feature of utterances [38, 50]. Recently, researchers have explored approaches for emotion recognition using deep learning with high-level features whereas some researchers have used hand-crafted low-level features to train CNN, RNN, and DNN models to improve the accuracy of emotion recognition.

The accuracy of emotion recognition of these classifiers depends on the selection of spectral and prosodic features of speech and feature extraction techniques used. Noise is also one of the major challenges, that affects the performance of the speech recognition system [37]. The features of speech signal vary across language, culture, speaker and gender. This results in a large number of “hand-crafted” features that can be tackled using deep learning automatically. In this paper, the capability of Deep CNN (DCNN) architecture is investigated in recognizing emotions from actor-based corpora with three different languages: German Emotional speech database (EMODB), British English Surrey audio-visual Expressed Emotion database (SAVEE) and Italian Emotional speech corpus (EMOVO). The proposed work aims to improve the accuracy of emotion recognition for any language and speaker.

The main contributions of this paper are as follows:

- An algorithm for recognition of emotions independent of language and speaker is proposed using DCNN architecture.
- Instead of audio files of actor-based speech corpora, their RGB spectrograms (images) are created and normalized before training. The common size maintained for all spectrograms is 224x224x3 for fine-tuning to DCNN.
- Then optimal features are automatically learned by DCNN architecture with labelled samples. Seven emotions are recognized using 3 databases of different languages with an accuracy better than earlier studies.
- The improvement in accuracy is reported using an optimal learning rate as compared to the random learning rate. The computational cost of the model is reduced using down-sampling by using stride as well as pooling layer in convolution.

The remaining section of the paper is organized as follows: The related work about speech corpora, features of speech, and classifiers used for speech emotion recognition are discussed in brief in Section 2. The details of the proposed algorithm are given in Section 3. In Section 4 experimental details and results are discussed. Finally, conclusions and future scope are summarized in Section 5.

2 Related work

A typical process of speech emotion recognition from speech is divided into two parts i) extraction of relevant and high-level features from the speech signal and ii) selection of classifiers for accurate recognition of emotions from speech signal. In this section, the existing literature about recognizing human emotions from the speech is discussed in terms of speech corpora, features of speech, and classifiers.

2.1 Speech corpora

The performance of human emotion recognition is highly dependent on the quality and type of speech database. The speech corpora can be categorized into three types namely actor-based or simulated, elicited, or induced, or natural. Simulated or actor-based speech corpora are collected from trained and experienced professional artists. Professional artists are asked to speak neutral sentences in many different emotions. Elicited or induced corpora are collected by simulating artificial emotional situation, without informing the speaker [51]. All types of emotions may not be present in the dataset and quality of speech may be low due to overlapping of utterances. Data is collected from real-world data in natural speech corpora. This may include conversation in the call centre, a dialog between doctor and patient, etc. Background noise may result in poor quality speech signals.

In the present work, an actor-based speech corpus is considered because this type of speech corpus is available in most of the languages including all the seven emotions. Many authors have worked on actor-based speech corpora in different languages to recognize the emotions. There are a large number of public and private actor-based speech corpora available. The details of some of the popular actor-based speech corpora are summarized in chronological order in Table 1.

The popular public speech corpora EMODB, SAVEE, and EMOVO are considered to analyse the performance of emotion recognition using the proposed algorithm.

2.2 Features

Spectral characteristics like distribution of energy at different parts of the audible frequency range are acoustic features. For a different type of emotion, vocal tract shapes are unique. By using spectral analysis vocal tract shapes can be estimated. To recognize human emotions, spectral features are used [3, 41]. To compute spectral features speech signals are divided into frames (called segments) of length 25-50 ms. Speech signals are assumed to be stable in the specified length. To extract spectral features there are many techniques. Some of the popular techniques are Short-time Coherence Method (SMC), Linear Predictor Coefficient (LPC), One-Sided Autocorrelation Linear Predictor Coefficient (OSALPC) [4] and LP residual [6]. Epoch (Glottal Closure Instance) is very useful in estimating the features vocal tract frequency

Table 1 Details of Actor-based speech corpora

Corpus	Public/Private	Language	Type of file	Total Actors	Number of Utterances	Total samples	Reference
EMODB	Public	German	Audio	10 (5 male+ 5 female)	10	535	[5]
eINTERFACE	Public	English	Audio-Video	42	43	1287	[30]
IITKGP-SESC	Public	Telugu	Audio	10(5 male + 5 female)	15	12,000	[23]
FAU Aibo	Public	German	Audio	51 children (10–13 years)		48,401	[46]
SAVEE	Public	British English	Audio	4 male	15	480	[18]
EMOVO	Public	Italian	Audio	6 (3male + 3 female)	14	588	[8]
RAVDESS	Public	American English	Audio-Video	24(12 male + 12 female)	2	7356	[29]

and pitch. Mel frequency speech power coefficients (MFSPC) is used by Nwe [33]. Normal frequency can be converted to the mel frequency by the Eq. (1).

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

where m is mel frequency and f is normal frequency. Epoch can be extracted using a zero frequency filtered speech signal and LP residual approach [24]. All the linear approaches may not always work correctly for emotion recognition because pitch may not be linear for all human perceptions. Expo Log scale, Mel-frequency scale, and modified Mel-frequency approaches are used to estimate non-linear scales. For recognizing stress emotion, the non-linear scale was reported to be better as compared to linear.

Prosodic features are the pitch of the speech, length of speech, loudness of speech and quality of speech. These features are not useful to drive at the frame level. Therefore, they are extracted at the sentence and utterance level. Pitch is also known as fundamental frequency (F0). The average pitch in case of a female is observed as 210 Hz and in the case of a male, it is observed to be 120 Hz. If the speech signal is quasi-stationary, then there are many different approaches to compute the pitch value. Variation of pitch over the cycle to cycle is known as jitter, and it can be computed by Eq. (2) [14].

$$J = (1/N-1) \sum_{i=1}^{N-1} |T_i - T_{i+1}| \quad (2)$$

where J is jitter, N is the number of cycles and T_i is pitch period. Better recognition of human emotion can be achieved by using autocorrelation-based pitch. The average energy of a speech signal can be computed by Eq. (3) [1].

$$E_n = \sum_{m=n-N+1}^n [x(m)w(n-m)]^2 \quad (3)$$

where E_n is energy value, n is the number of samples, energy distribution, and amplitude in spectrum affect the arousal emotion of a speaker [9]. Therefore, duration and energy features are useful for recognizing human emotions. In general, the male energy level was found to be higher as compared to females in anger emotions. For the same anger emotion, the male speech rate was found to be low as compared to females. Statistical values of pitch include minimum, maximum, range, mean, median, standard deviation, slope minima, slope maxima, kurtosis, skewness, jitter, relative pitch, first-order difference, and so on. Similarly, statistical features of energy and duration include shimmer, duration of voice ratio, speech rate, along with mean, maximum, minimum, standard deviation, and so on. Energy, pitch, and duration contour had been used as dynamics of prosodic features to recognize the seven emotions [20, 40]. It was found that spectral features such as Mel frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC), OSALPC were missing temporal information. For this problem modulation, spectral features were used [42]. While recognizing human emotion from speech, temporal information was used and found to be very useful [53]. A combination of spectral and prosodic features was also considered to recognize human emotions more accurately from speech [25]. For female and male speakers, database critical analysis has been done. Features such as low-MFCC, standard MFCC, and pitch were found to be better in recognizing the emotions from speech [32, 57]. In extracting stable pitch, low-MFCC performed better.

2.3 Classifiers

Many classifiers can be used for the analysis of speech to detect emotions. Earlier, temporal dynamics features of speech signals were captured by using the state transition matrix. At that time, the majority of the researchers used HMM as a classifier to recognize the emotions from speech. The phenomenon was used to recognize the human emotions from speech following the left-to-right sequence, and HMM also adopted a left-to-right structure for human emotion recognition [26, 43]. In an utterance, it is difficult to find emotional cues from the sequential flow. It may be present at the end, middle, or at the beginning in an utterance. To resolve this problem, the ergodic model of HMM was used. In this ergodic model of HMM, it was possible to go from any state to any other state. For human emotion recognition from speech, GMM was a more suitable classifier using global feature extraction as compared to HMM. It is best suited when the data of the dataset is not normally distributed. A boosted GMM was proposed to recognize human emotion from a speech by Tang [48]. Second-order parameters like standard deviation and mean were used to capture hyperplane distribution in GMM. The statistical learning concept introduced a new classifier and regression technique called SVM. SVM gave better results for many pattern recognition applications as compared to other classifiers [44]. There are many other approaches proposed to recognize human emotions from speech. These include the K Nearest Neighbour (K-NN) approach, a supervised learning method that is simplest of all for speech emotion recognition methods [17]. Feature vector contributes more while recognizing emotions from speech.

Artificial Neural Network (ANN) is another efficient classifier to extract the non-linear features in many pattern recognition applications. If the training sample size was low in number, then also ANN gave better results as compared to GMM and HMM. The performance of the ANN classifier totally depends on the number of hidden neurons for each hidden layer and the number of hidden layers. For better human emotion recognition more than one ANNs were also used [10, 45]. RBL was found to be the least used and MLP as the commonly used classifier for emotion recognition from speech. Generalized Feed Forward Neural Networks (GFNN) gave better results as compared to MLP in recognizing human emotions [16]. Other forms of neural networks used for human emotion recognition from the speech were Auto Associated Neural Networks (AANN) and 2D-Neural Network [21]. As compared to the SVM classifier ANN gave better results [36]. The performance of emotion recognition decreased when the number of emotions was increased [56]. Auto-encoder based unsupervised classifier was proposed by Deng [12, 52]. Motamed [31] proposed modified brain emotional learning for emotion recognition from speech. Some researchers extracted the features by using deep neural networks and then used different classifiers. Deep Neural Network (DNN) was introduced for acoustic emotion recognition by Stuhlsatz [47]. The benefit of using deep learning for emotion recognition is the capacity to extract high-level features more accurately. Deep Belief Network (DBN) was used to capture non-linear features [22]. CNN was found to be more efficient in extracting high-level features of images [27]. A similar level of accuracy was achieved by using a neural network on information retrieved from spectrograms in EMODB corpus [39]. The number of classes of emotions recognized was only five. CNN, when combined with the LSTM network was able to learn the representation of speech by itself from raw time representation [49, 55]. The findings of the related work are summarized in Table 2.

From the literature, it is shown that many researchers have used CNN from deep learning approaches for emotion recognition from speech. Using deep learning approaches the accuracy

Table 2 Summary of related work

Ref	Technique used	Dataset(s) used	Features used	Average Accuracy
[26]	QDA, SVM, HMM, LDA	SUSAS	Pitch, MFCC	70%
[43]	KNN	Linguistic Data Consortium	MFCC	66.4%
[48]	BGMM	Random	Pitch MFCC	NS
[44]	SVM	EMODB	MFCC, PPCMCC	82%
[10]	NN		Utterances	80%
[16]	DWT, ANN	Malayalam	MFCC, LPC	68%
[21]	GFFNN	EMODB	LPC	98%
[56]	NN, SVM	eINTERFACE, FAU	Pitch, utterance level	70%,60%
[52]	Fourier parameters	EMODB, ESSDB, CASIA	MFCC	71%
[12]	Universum Auto-encoders	GeWEC		55%
[31]	ANFIS MLP	EMODB	MFCC	72.5%
[47]	DNN, SVM, GerDA	9 datasets	33	52%–80%
[22]	DBN, SVM	IEMOCAP	(pitch, MFCC)	
[39]	NN, SVM	EMO-DB	Pitch, Energy, (MFBs	73%
[49]	SVR, BLSTM-DRNNs	RECOLA	MFCC	73–85.5%
[55]	ID,2D CNN LSTM	EMODB, IEMOCAP	eGeMAPS	68%
[34]	SVM, MLP, k-NN	eINTERFACE05, EMODB, SAVEE, EMOVO	log-mel spectrograms	95%
[2]	CNN	EMODB, Korean dataset	–	60–84%
[7]	ACRNN	IEMOCAP, EMO-DB	MFCC	79%
[19]	CNN	SAVEE, EMO-DB, DES, MES	Mel-spectrogram	
[15]	DNN	eINTERFACE, SAVEE	TEO, MFCC	93%
[35]	CNN, LSTM	IEMOCAP,EMOVO,SAVEE, EMODB, EPST, RAVDESS, TESS	GeMAPS feature set	60–61%
[28]	GoogLeNet	RAVDESS, EMODB, IEMOCAP	spectrograms	Max (69%) among all
[11]	pQPSO(GMM)	EMODB, SAVEE, IEMOCAP	MFCC+LPC+ PMVDR+Pitch	67.10%,72.55%, 67.20%

ACRNN-3D Attention-based CRNN, BGMM-Boosted GMM, eGeMAPS -Extended Geneva minimalistic acoustic feature set, GeWEC-Geneva Whispered Emotion Corpus, GFFNN-Generalized FFNN, LIF- Local Invariant Features, MFB- Mel-frequency filter banks, SVR- Support Vector Regression, TEO-Teager Energy Operator.

is increased but the computational cost of the model is also increased because of large pre-trained architecture. A few researchers have developed approaches for emotion recognition from speech using spectrograms as input [2, 7, 15, 19].

In this work, an algorithm for emotion recognition using DCNN architecture is proposed. For down-sampling, stride as well as pooling layer is used in convolution. It reduced the computational cost of the proposed DCNN based model and increased the recognition accuracy for emotion recognition. It is evaluated on speech corpora EMODB, EMOVO, and SAVEE. The comprehensive explanation of the proposed model is discussed in the successive section.

3 Proposed algorithm

3.1 Pseudocode of the proposed algorithm

In this section, an algorithm is proposed using DCNN architecture for emotion recognition from speech. The pseudo-code to implement the proposed algorithm is described in Table 3.

The .wav files are read from collected speech corpora EMODB, SAVEE, and EMOVO and converted into spectrograms. The detailed process is depicted in Fig. 1.

3.2 Details of DCNN architecture

In this algorithm, the audio files from speech corpora are converted into the spectrograms and normalized to size 224x224x3 in the first step. As spectrograms hold more information that cannot be extracted from the speech signals, it improves the accuracy of emotion recognition. DCNN architecture learns high-level features from RGB spectrograms. DCNN architecture has convolutional layers, a max-pooling layer and fully connected layers. To calculate the probability of each emotion fully connected layers are fed to a softmax classifier. Special stride is used for down sampling the output features rather than the pooling layers. By convolution layer, optimal features (a combination of spectral and prosodic features like pitch, MFCC, and energy) are automatically learned by labelled samples. The details of the DCNN architecture used are shown in Fig. 2.

Table 3 Pseudo code to implement the proposed algorithm

Input: Audio files (Labelled actor based speech corpora)

Output: Recognize the emotion as output (like happiness, sadness, anger, fear, disgust, boredom/surprise, and neutral)

Step 1: Read .wav files from the speech corpora.

Step 2: Get spectrograms of each .wav file by a Short-Time Fourier Transform (STFT) of the wave signal.

Step 3: Convert all spectrograms to size 224x224x3.

Step 4: Divide the data sets into

Train=80% (of all data sets)

Test =10% (of all data sets)

Validation=10% (of all data sets)

Step 5: Train the deep learning model and save as stage-1 (i.e. freeze)

Step 6: Using the Learning Rate (LR) range test, find an optimal learning rate.

Step 7: Unfreeze stage-1 and train the deep learning model with an optimal learning rate and save as stage-2 (i.e. freeze).

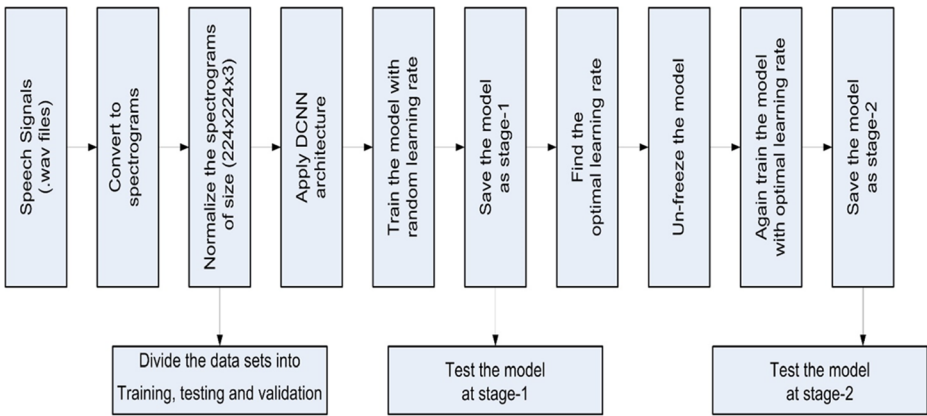


Fig. 1 Steps of the proposed algorithm

In this DCNN architecture, first convolutional layers (Conv-1) have 16 filters of size (7×7) that are applied to the normalized RGB spectrograms of size $(224 \times 224 \times 3)$ with stride (2×2) and same padding. The output feature map obtained from Conv-1 goes through max-pooling (2×2) with stride (1×1) . Similarly, the second convolutional layer (Conv-2) has 32 filters of size (5×5) with stride (2×2) . Conv-3 has the same number of filters as Conv-2 of size (3×3) with stride (2×2) and the same padding. Conv-4 and Conv-5 have 64 filters of size (3×3) with stride (2×2) and Conv-5 has the same padding. In the same way, Conv-6 has 128 filters and Conv-7 has 256 filters of size (3×3) with stride (2×2) and the same padding. In this DCNN architecture activation function, ReLU is used after each convolutional layers to rectify the output features map defined as:

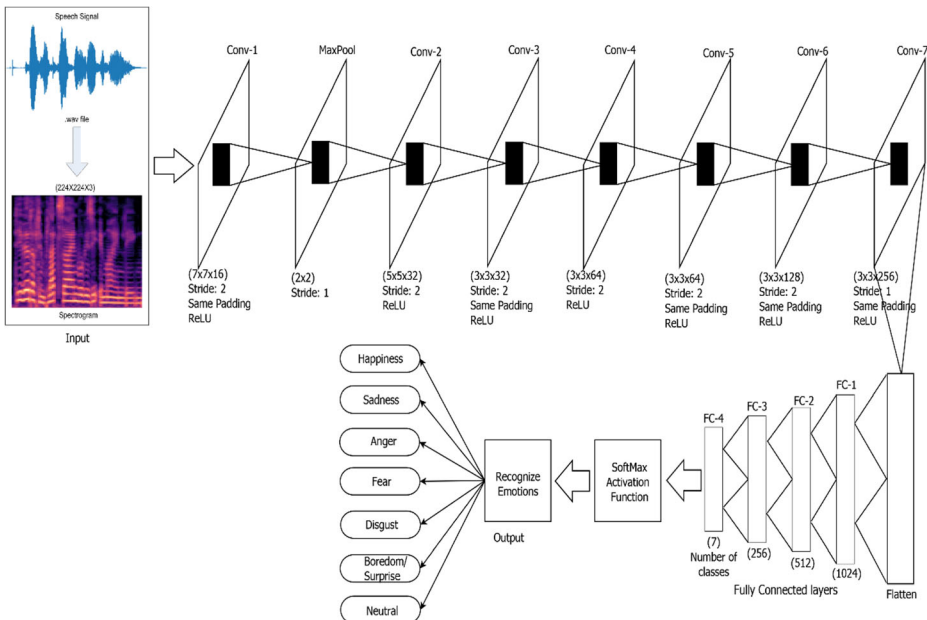


Fig. 2 DCNN architecture details

Table 4 Main Parameters and their value

Parameter Name	Parameter value
Learning rate	Find optimal value during training
Momentum	0.9
Decay	1e-6
Eps	1e-5
solver type	SGD

and $f(z) = 0, \text{ if } z <= 0$
 $f(z) = z, \text{ if } z >= 0$ i.e. $f(z) = \max(0, z)$

To regularize the DCNN model the rectified output layers are followed by batch normalization with momentum 0.9. The last convolutional layer Conv-7 is followed by a flatten layer. The flattening layer is fed as input to the fully connected layers. The first fully connected layer (FC-1) has 1024 neurons and the last fully connected layer (FC-4) has 7 neurons as the number of classes. FC-1 is followed by a 20% dropout ratio. Finally, to calculate the probability of each emotion FC-4 layer is fed to a softmax classifier. The main parameters used for the experiment during training the model are shown in Table 4.

3.3 Converting speech signal into spectrograms

Read the .wav files from collected speech corpora EMODB, SAVEE, and EMOVO and then convert all files into spectrograms. Steps to get the spectrograms from the .wav file given in detail in Table 5.

The parameters used for STFT are frame size (25 ms), overlapFac=0.5 (50% overlap), window=np.hanning (window type like Hamming, Kaiser, etc.). Mathematically it is calculated as in Eq. (4).

$$Xm(w) = \sum_{n=-\infty}^{\infty} x(n)w(n-mR)e^{-jwn} \tag{4}$$

where $Xm(w)$ Discrete Time Fourier Transform (DTFT), $x(n)$ is input signal at time n , $w(n)$ is window function. Some sample spectrograms of each speech corpora of each emotion are shown in Fig. 3. where the vertical axis represents the frequency and the horizontal axis represents the time.

In step 3 all the spectrograms are normalized to size 224x224x3. In step 4 spectrograms are divided into training, testing, and validation set in the ratio of 80%, 10%, and 10% respectively. Then training and validation of the DCNN model is performed using the proposed algorithm and saved (i.e. frozen) as stage-1 in step 5. From the stage-1 model using LR range

Table 5 Steps to get spectrograms

Input: .wav file from speech corpora
Output: spectrograms (image file)
Step-1: Divide the speech signal into 25 milliseconds. (Framing)
Step-2: Apply Hamming windowing
Step-3: Short Time Fourier Transform (STFT) of .wav files
Step-4: scale frequency axis logarithmically
Step-5: plot spectrograms

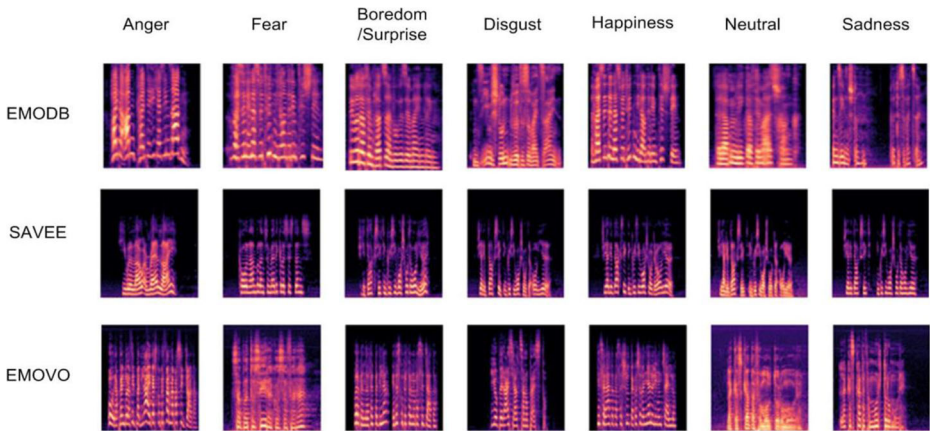


Fig. 3 Spectrograms of each type of emotion for EMODB, SAVEE, and EMOVO corpus

Test, the optimal learning rate is found in step-6. The stage-1 model is un-frozen and trained again with an optimal learning rate and saved as a stage-2 model in step 7. The details of the experiments and results are discussed in section 4.

4 Experimental details and results

In this section, the proposed algorithm is applied for emotion recognition from the speech on EMODB, EMOVO, and SAVEE datasets. All the experiments are performed on standard windows 10 laptop, Intel(R) Core™ i5-7200U CPU@2.70 GHz and 8 GB RAM, ×64-based processor. The performance of the proposed algorithm is compared with classifiers using handcrafted features and recent deep learning approaches. Detailed experimental results are discussed in the coming subsequent sections.

4.1 Data sets

In this work, experiments are conducted on three publicly available labelled, actor-based emotional data sets: EMODB, EMOVO, SAVEE. These are summarized below:

- EMODB speech corpus consists of basic seven emotions, i.e. anger, boredom, disgust, fear, happiness, sadness, and neutral. This corpus consists of a total of 535 samples, and the language of this corpora is German. Utterances are recorded by five males and five female actors. All 10 utterances are used in the present work.
- SAVEE is another popular public emotional speech corpus, which is recorded by four male actors. These data sets consist of seven emotions, i.e. happiness, sadness, anger, fear, disgust, surprise, and neutral. There is a total of 15 utterances and 480 samples.
- EMOVO is a publicly available emotional speech corpus. It covers seven categories of emotions namely, happiness, sadness, anger, fear, disgust, surprise, and neutral. It consists of a total of 588 samples recorded by 6 actors- 3 males and 3 females. The language of this corpora is Italian.

The statistics of these emotional speech corpora are summarised in Table 6.

Table 6 Number of samples in speech corpora used

Database	Happy	Sad	Anger	Scared	Neutral	Disgust	Surprised	Bored	Total
EMODB	71	62	127	69	79	46	–	81	535
SAVEE	60	60	60	60	120	60	60	–	480
EMOVO	84	84	84	84	84	84	84	–	588

Table 7 Confusion matrix for emotions prediction on EMOVB at stage-1

Emotional Class	Anger	Bored / Surprise	Disgust	Happy	Neutral	Sad	Scared
Anger	0.82	0.08	0.00	0.00	0.00	0.10	0.00
Bored/ Surprise	0.10	0.60	0.12	0.00	0.18	0.00	0.00
Disgust	0.10	0.17	0.63	0.00	0.10	0.00	0.00
Happy	0.00	0.08	0.02	0.74	0.16	0.00	0.00
Neutral	0.01	0.00	0.00	0.00	0.75	0.15	0.09
Sad	0.00	0.00	0.00	0.00	0.10	0.79	0.11
Scared	0.09	0.00	0.00	0.00	0.01	0.12	0.78
Overall Accuracy							73.08%

4.2 Experimental results

The training of the proposed model for all considered speech corpora (EMODB, SAVEE & EMOVO) are saved as stage-1. The prediction performance of the proposed model at stage-1 is evaluated on EMOVB, SAVEE, and EMOVO datasets to show the efficiency of the proposed algorithm. Tables 7, 8 and 9) show the prediction performance of the proposed model at stage-1 in terms of confusion matrix on the EMOVB, SAVEE and EMOVO datasets respectively.

Tables 7, 8 and 9) show the overall accuracy of the proposed model at stage-1 for EMOVB, SAVEE, and EMOVO speech corpora as 73.08%, 50%, and 57.15% respectively. To improve the accuracy LR range test is used and an optimal learning rate is found. Learning rate variation concerning the number of iterations and learning rate variation concerning loss are shown in Fig. 4.

The learning rate describes how parameters are updating. Here X-axis shows what happens when the learning rate is increased and Y-axis shows the loss. From Fig. 4a) it can be seen, that once the learning rate pass 10^{-5} , the loss is getting worse because fine-tuning is done. Based on

Table 8 Confusion matrix for emotions prediction on SAVEE at stage-1

Emotional Class	Anger	Bored/ Surprise	Disgust	Happy	Neutral	Sad	Scared
Anger	0.60	0.10	0.10	0.00	0.10	0.10	0.00
Bored/ Surprise	0.10	0.40	0.20	0.00	0.10	0.10	0.10
Disgust	0.00	0.20	0.45	0.01	0.20	0.10	0.04
Happy	0.00	0.18	0.02	0.50	0.20	0.00	0.10
Neutral	0.10	0.10	0.10	0.10	0.49	0.02	0.09
Sad	0.10	0.10	0.10	0.00	0.10	0.54	0.06
Scared	0.20	0.10	0.10	0.00	0.02	0.06	0.52
Overall Accuracy							50.00%

Table 9 Confusion matrix for emotions prediction on EMOVO at stage-1

Emotional Class	Anger	Bored/Surprise	Disgust	Happy	Neutral	Sad	Scared	
Anger	0.67	0.00	0.03	0.00	0.10	0.20	0.00	
Bored/Surprise	0.01	0.49	0.10	0.00	0.30	0.10	0.00	
Disgust	0.10	0.09	0.51	0.00	0.20	0.10	0.00	
Happy	0.00	0.20	0.02	0.58	0.10	0.00	0.10	
Neutral	0.10	0.10	0.00	0.03	0.57	0.10	0.10	
Sad	0.20	0.00	0.00	0.00	0.10	0.61	0.09	
Scared	0.10	0.20	0.00	0.02	0.01	0.10	0.57	
Overall Accuracy							57.15%	

the learning rate finder it was decided to pass an optimal learning rate 10^{-5} (for EMOVB). From Fig. 4b) we found that loss is minimum at 10^{-6} and after that loss increases. So, the optimal learning rate was passed 10^{-6} (for EMOVB). In Fig. 4c) it was found that there was no range where either loss rapidly decreased or after that loss got worse. So here we passed optimal learning rate 10^{-6} at which loss is minimum. After finding the optimal learning rate,

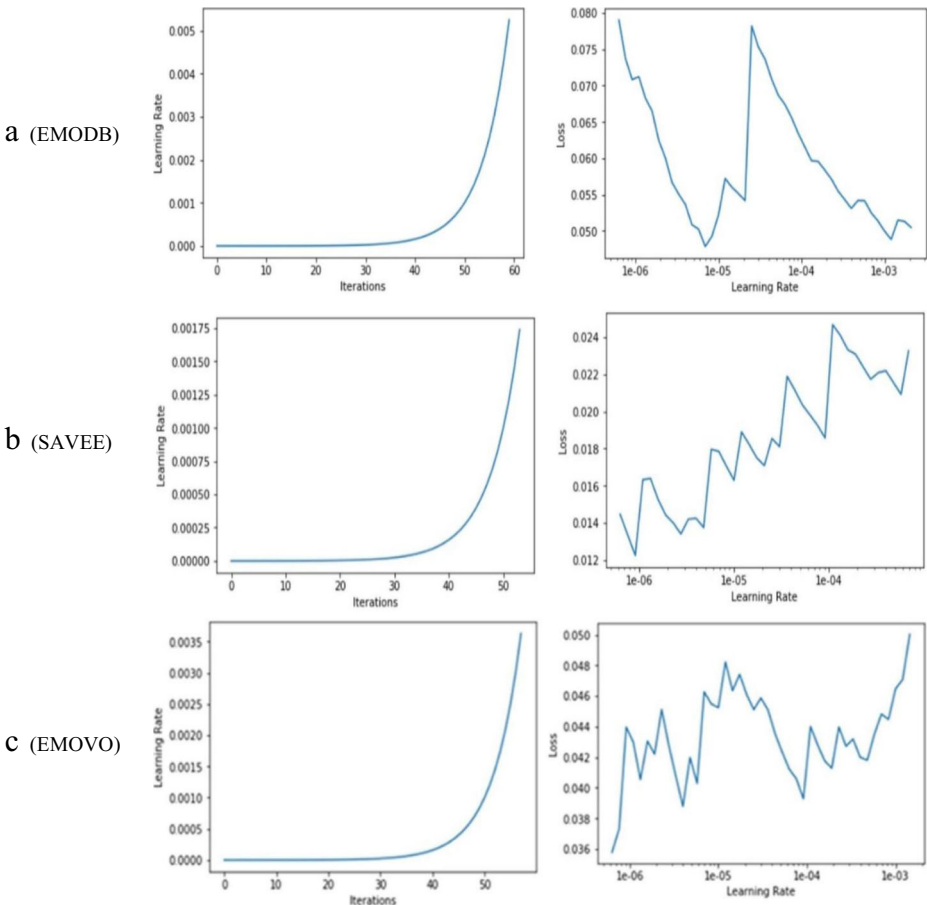


Fig. 4 Learning rate variation for the number of iterations and variation in a loss concerning learning rate

Table 10 Confusion matrix for emotions prediction on EMODB at stage-2

Emotional Class	Anger	Bored / Surprise	Disgust	Happy	Neutral	Sad	Scared
Anger	0.93	0.00	0.00	0.00	0.03	0.04	0.00
Bored/ Surprise	0.00	0.71	0.10	0.00	0.10	0.09	0.00
Disgust	0.00	0.09	0.70	0.01	0.10	0.10	0.00
Happy	0.00	0.00	0.02	0.88	0.10	0.00	0.00
Neutral	0.00	0.00	0.00	0.00	0.91	0.00	0.09
Sad	0.00	0.00	0.00	0.00	0.10	0.90	0.00
Scared	0.00	0.00	0.00	0.00	0.01	0.10	0.89
Overall Accuracy							84.62%

the stage-1 model was unfrozen. And the model was trained again with an optimal learning rate and saved as stage-2. The prediction performance of the proposed model at stage-2 was evaluated on EMODB, SAVEE, and EMOVO datasets to show the efficiency of the proposed algorithm. Tables 10, 11, 12) show the prediction performance of the proposed model at stage-2 in terms of confusion matrix on the EMODB, SAVEE and EMOVO datasets.

Tables 10, 11, 12) show the overall accuracy of the proposed model at stage-2 for EMODB, SAVEE, and EMOVO speech corpora as 84.62%, 75%, and 69.65% respectively. This also shows that better accuracy has been achieved in stage-2 as compared to the stage-1 model. The summary results of the proposed model at stage-1 and stage-2 (Mean + Standard deviation) are shown in Table 13.

The results show that an average of 16.35% better accuracy is achieved at stage-2 as compared to stage-1. For SAVEE corpus the improvement is highest (25%). From the results of the proposed algorithm, it can be concluded that selecting an optimal learning rate is very important as compared to a random learning rate.

4.3 Performance comparison of the proposed algorithm with state-of-the-art

The performance comparison of the proposed algorithm with other state-of-the-art algorithms are given in Table 14, which outperforms the existing results over EMODB, SAVEE, and EMOVO datasets using spectrograms as input.

From Table 14, it is clear that for EMODB corpus, the proposed algorithm stage-2 gives better results (84.62%) accuracy as compared to SVM (71.12%), K-NN (63.74%), and MLP (81.32%) [34], Random Forest (77.18%), Alexnet (81.33%), and Decision Tree (72.82%) [2],

Table 11 Confusion matrix for emotions prediction on SAVEE at stage-2

Emotional Class	Anger	Bored / Surprise	Disgust	Happy	Neutral	Sad	Scared
Anger	0.85	0.15	0.00	0.00	0.00	0.00	0.00
Bored/ Surprise	0.20	0.66	0.14	0.00	0.00	0.00	0.00
Disgust	0.00	0.12	0.68	0.01	0.19	0.00	0.00
Happy	0.00	0.00	0.10	0.75	0.15	0.00	0.00
Neutral	0.00	0.13	0.10	0.00	0.77	0.00	0.00
Sad	0.10	0.00	0.00	0.00	0.00	0.78	0.12
Scared	0.10	0.00	0.00	0.00	0.00	0.14	0.76
Overall Accuracy							75.00%

Table 12 Confusion matrix for emotions prediction on EMOVO at stage-2

Emotional Class	Anger	Bored /Surprise	Disgust	Happy	Neutral	Sad	Scared
Anger	0.79	0.00	0.00	0.00	0.10	0.11	0.00
Bored/Surprise	0.10	0.61	0.19	0.00	0.10	0.00	0.00
Disgust	0.00	0.12	0.62	0.16	0.00	0.10	0.00
Happy	0.00	0.08	0.12	0.70	0.10	0.00	0.00
Neutral	0.10	0.18	0.00	0.00	0.72	0.00	0.00
Sad	0.10	0.00	0.00	0.00	0.16	0.74	0.00
Scared	0.19	0.00	0.00	0.00	0.01	0.11	0.69
Overall Accuracy							69.65%

Table 13 Accuracies at stage-1 and stage-2

Speech Corpus	Average Accuracy (%)		Accuracy (%) gain at stage-2 as compared to stage-1
	Stage-1	Stage-2	
EMODB	73.08± 6.23	84.62± 3.32	11.56
SAVEE	50.00± 4.86	75.00± 2.01	25.00
EMOVO	57.15± 5.23	69.65± 2.54	12.50

3D CRNN (82.82%) [7], CNN(84.50%) [19], CNN-LSTM (69.72%) [35], GoogLeNet (72.55%) [28], and pQPSO (82.82%) [11].

For SAVEE corpus, proposed algorithm stage-2 gives better results (75% accuracy) as compared to SVM (72.39%), K-NN (53.37%) and MLP (71.17%) [34], CNN (69.00%) [19], DNN (59.70%) [15], CNN-LSTM (72.66%) [35] and pQPSO (60.79%) [11].

Table 14 Performance comparison of the proposed algorithm with state-of-the-art methods

	Methods	Input	Average Accuracy (%)		
			EMODB	SAVEE	EMOVO
Özseven T [34]	SVM	Features extracted	71.12	72.39	60.40
	K-NN	with openSMILE	63.74	53.37	39.05
	MLP		81.32	71.17	58.58
Badshah et al. [2]	Random Forest	MFCC	77.18	–	–
	Alexnet		81.33	–	–
	Decision Tree		72.82	–	–
Chen et al. [7]	3D CRNN	Mel-spectrograph	82.82	–	–
Huang et al. [19]	CNN	TEO	84.50	69.00	–
Fayek et al. [15]	DNN	MFCC	–	59.7	–
Parry [35]	CNN-LSTM	GeMAPS feature set	69.72	72.66	53.24
Lee et al. [28]	GoogLeNet	Spectrogram	72.55	–	–
Daneshfar et al. [11]	pQPSO(GMM)	MFCC+LPCC+ PMVDR+Pitch	82.82	60.79	–
Proposed Algorithm	Stage-1	Spectrogram	73.08	50.00	57.15
	Stage-2	Spectrogram	84.62	75.00	69.65

And for EMOVO corpus, proposed algorithm stage-2 gives better results (69.65% accuracy) as compared to SVM (60.40%), k-NN (39.05%), and MLP (58.58%) [34], CNN-LSTM (53.24%) [35].

5 Conclusion and future scope

In this work, we have evaluated our proposed algorithm on three different language labelled speech corpora are considered EMOVB, EMOVO, and SAVEE. All audio files (.wav) from all speech corpora are converted into the spectrograms of size 224x224x3. Training, testing, and validation ratio are considered 80%, 10%, and 10% respectively. Then the deep learning model is considered to recognize the emotions from speech corpus. Deep learning models are developed at two stages: stage-1 and stage-2. At stage-1 the model is developed with a random learning rate and at stage-2 it is developed using an optimal learning rate. The performance of the proposed algorithm is compared with other hand-crafted features using classifiers and between stage-1 and stage-2 also. The recognition rates are found to be 84.62%, 75%, and 69.65% for EMOVB, SAVEE, and EMOVO datasets respectively which are better than existing studies. The proposed algorithm gave better results with any type of language and actor because it uses an optimal learning rate as compared to a random learning rate. Cross-lingual emotion recognition and cross-corpus emotion recognition may be another direction for future research.

References

1. Anagnostopoulos C, Iliou T, Giannoukos I (2012) Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intell Rev* 43:155–177. <https://doi.org/10.1007/s10462-012-9368-5>
2. Badshah A, Rahim N, Ullah N, Ahmad J, Muhammad K, Lee M et al (2019) Deep features-based speech emotion recognition for smart affective services. *Multimed Tools Appl* 78:5571–5589. <https://doi.org/10.1007/s11042-017-5292-7>
3. Bitouk D, Verma R, Nenkova A (2010) Class-level spectral features for emotion recognition. *Speech Commun* 52:613–625. <https://doi.org/10.1016/j.specom.2010.02.010>
4. Bou-Ghazale S, Hansen J (2000) A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Trans Speech Audio Process.* 8:429–442. <https://doi.org/10.1109/89.848224>.
5. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B (2005) A database of German emotional speech, in: *Inninth European Conference On Speech Communication And Technology*
6. Chauhan A, Koolagudi SG, Kafley S, Rao KS (2010, April) Emotion recognition using LP residual. In: *2010 IEEE Students Technology Symposium (TechSym)*. IEEE, pp 255–261. <https://doi.org/10.1109/TECHSYM.2010.5469162>
7. Chen M, He X, Yang J, Zhang H (2018) 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Process Lett* 25(10):1440–1444. <https://doi.org/10.1109/LSP.2018.2860246>
8. Costantini G, Iaderola I, Paoloni A, Todisco M (2014) Emovo corpus: an italian emotional speech database, in. *In international Conference on Language Resources And Evaluation, European Language Resources Association (ELRA)*, 3501–3504.
9. Cowie R, Cornelius RR (2003) Describing the emotional states that are expressed in speech. *Speech Commun* 40(1-2):5–32. [https://doi.org/10.1016/s0167-6393\(02\)00071-7](https://doi.org/10.1016/s0167-6393(02)00071-7)
10. Dai K, Fell HJ, MacAuslan J (2008) Recognizing emotion in speech using neural networks. *Telehealth and Assistive Technologies* 31:38–43

11. Daneshfar F, Kabudian SJ (2020) Speech emotion recognition using discriminative dimension reduction by employing a modified quantum behaved particle swarm optimization algorithm. *Multimed Tools Appl* 79(1):1261–1289. <https://doi.org/10.1007/s11042-019-08222-8>
12. Deng J, Xu X, Zhang Z, Frühholz S, Schuller B (2017) Universum autoencoder-based domain adaptation for speech emotion recognition. *IEEE Signal Process Lett* 24(4):500–504. <https://doi.org/10.1109/lsp.2017.2672753>
13. El Ayadi M, Kamel M, Karray F (2011) Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* 44:572–587. <https://doi.org/10.1016/j.patcog.2010.09.020>
14. Farrús M, Hernando J (2009) Using Jitter and Shimmer in speaker verification. *IET Signal Process* 3:247. <https://doi.org/10.1049/iet-spr.2008.0147>
15. Fayek H M, Lech M and Cavedon L (2015) Towards real-time speech emotion recognition using deep neural networks. In 2015 9th international conference on signal processing and communication systems (ICSPCS) 1–5. IEEE. <https://doi.org/10.1109/ICSPCS.2015.7391796>
16. SA Firoz, SA Raji, AP Babu (2009) Automatic Emotion Recognition from Speech Using Artificial Neural Networks with Gender-Dependent Databases. *International Conference on Advances in Computing, Control, and Telecommunication Technologies, Trivandrum, Kerala* 162–164. <https://doi.org/10.1109/ACT.2009.49>
17. Han K, Yu D, Tashev I (2014) Speech emotion recognition using deep neural network and extreme learning machine. In Fifteenth annual conference of the international speech communication association 223–227.
18. Haq S, Jackson PJ (2011) Multimodal emotion recognition. In: *Machine audition: principles, algorithms and systems*. IGI Global, pp 398–423. <https://doi.org/10.4018/978-1-61520-919-4.ch017>
19. Huang Z, Dong M, Mao Q and Zhan Y (2014) Speech emotion recognition using CNN. In *Proceedings of the 22nd ACM international conference on Multimedia* 801–804. <https://doi.org/10.1145/2647868.2654984>
20. Jawarkar N (2007) Emotion recognition using prosody features and a fuzzy min-max neural classifier. *IETE Technical Rev* 24:369–373
21. Khanchandani KB (2009) MA Hussain, Emotion recognition using multilayer perceptron and generalized feed forward neural network. *CSIR* 68:367–371 <http://hdl.handle.net/123456789/3787>
22. Kim Y, Lee H, Provost EM (2013) Deep learning for robust feature generation in audio visual emotion recognition. *IEEE Int Conference Acoustics, Speech Signal Process, Vancouver*, pp 3687–3691. <https://doi.org/10.1109/ICASSP.2013.6638346>
23. Koolagudi S, Maity S, Kumar V, Chakrabarti S, Rao K (2009) IITKGP-SESC: speech database for emotion analysis. In *international Conference On Contemporary Computing*. 485–492.
24. Koolagudi SG, Reddy R, Rao KS (2010, July) Emotion recognition from speech signal using epoch parameters. In: *2010 international conference on signal processing and communications (SPCOM)*. IEEE, pp 1–5. <https://doi.org/10.1109/SPCOM.2010.5560541>
25. Koolagudi S, Murthy Y, Bhaskar S (2018) Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition. *Int J Speech Technol* 21:167–183. <https://doi.org/10.1007/s10772-018-9495-8>
26. Kwon OW, Chan K, Hao J, Lee TW (2003) Emotion recognition by speech signals. In *Eighth European Conference on Speech Communication and Technology*
27. Le Cun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature*
28. Lee MC, Chiang SY, Yeh SC, Wen TF (2020) Study on emotion recognition and companion Chatbot using deep neural network. *Multimed Tools Appl* 79:19629–19657. <https://doi.org/10.1007/s11042-020-08841-6>
29. Livingstone S, Russo F (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE* 13:e0196391. <https://doi.org/10.1371/journal.pone.0196391>
30. Martin O, Kotsia I, Macq B, Pitas I (2006, April) The eNTERFACE'05 audio-visual emotion database. In: *22nd International Conference on Data Engineering Workshops (ICDEW'06) Atlanta, GA, USA*, pp 8–8. <https://doi.org/10.1109/ICDEW.2006.145>
31. Motamed S, Setayeshi S, Rabiee A (2017) Speech emotion recognition based on a modified brain emotional learning model. *Biol Inspired Cognitive Architect* 19:32–38. <https://doi.org/10.1016/j.bica.2016.12.002>
32. Neiberg D, Elenius K, Laskowski K (2006) Emotion recognition in spontaneous speech using GMMs. In *Ninth international conference on spoken language processing*
33. Nwe TL, Wei FS, De Silva LC (2001) Speech based emotion classification. *Proceedings of IEEE Region 10th International Conference on Electrical and Electronic Technology. TENCON, Singapore*, pp 297–301. <https://doi.org/10.1109/TENCON.2001.949600>
34. Özseven T (2019) A novel feature selection method for speech emotion recognition. *Applied Acoustics* 146: 320–326. <https://doi.org/10.1016/j.apacoust.2018.11.028>
35. Parry J, Palaz D, Clarke G, Lecomte P, Mead, R., Berger M, Hofer G (2019) Analysis of Deep Learning Architectures for Cross-corpus Speech Emotion Recognition. *Proc Interspeech*:1656–1660. <https://doi.org/10.21437/Interspeech.2019-2753>

36. Partila P, Voznak M (2013) Speech emotions recognition using 2-d neural classifier. In *Nostradamus 2013: Prediction, modeling and analysis of complex systems* (pp. 221–231). Springer, Heidelberg. https://doi.org/10.1007/978-3-319-00542-3_23
37. Pervaiz A, Hussain F, Israr H, Tahir MA, Raja FR, Baloch NK, ... Zikria YB (2020) Incorporating Noise Robustness in Speech Command Recognition by Noise Augmentation of Training Data. *Sensors* 20(8):2326
38. Polzehl T, Sundaram S, Ketabdar H, Wagner M, Metz F (2009) Emotion classification in children's speech using fusion of acoustic and linguistic features. in: *Tenth Annual Conference of The International Speech Communication Association*
39. Prasomphan S (2015) Improvement of speech emotion recognition with neural network classifier by using speech spectrogram. *International Conference on Systems, Signals and Image Processing (IWSSIP)*, London, pp 73–76. <https://doi.org/10.1109/IWSSIP.2015.7314180>
40. Rao K S, Reddy R, Maity S, Koolagudi SG (2010) Characterization of emotions using the dynamics of prosodic features. In *Speech Prosody Fifth International Conference*
41. Rao K, Koolagudi S, Vempada R (2013) Emotion recognition from speech using global and local prosodic features. *Int J Speech Technol.* 16:143–160. <https://doi.org/10.1007/s10772-012-9172-2>.
42. Razak AA, Komiya R, Izani M, Abidin Z (2005) Comparison between fuzzy and NN method for speech emotion recognition. *Third International Conference on Information Technology and Applications (ICITA'05)*, Sydney, pp 297–302. <https://doi.org/10.1109/ICITA.2005.101>
43. Sato N, Obuchi Y (2007) Emotion Recognition using Mel-Frequency Cepstral Coefficients. *J Nat Language Process* 14:83–96. https://doi.org/10.5715/jnlp.14.4_83
44. Shen P, Changjun Z, Chen X (2011, August) Automatic speech emotion recognition using support vector machine. In: *Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology*, vol 2. IEEE, pp 621–625. <https://doi.org/10.1109/EMEIT.2011.6023178>
45. Singh YB, Goel S (2018) Survey on Human Emotion Recognition: Speech Database, Features and Classification. *International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Greater Noida (UP), India, pp 298–301. <https://doi.org/10.1109/ICACCCN.2018.8748379>
46. Steidl S, Batliner A, Seppi D, Schuller B (2010) On the impact of children's emotional speech on acoustic and language models. *EURASIP J Audio, Speech Music Process* 2010(1):783954. <https://doi.org/10.1186/1687-4722-2010-783954>
47. Stuhlsatz A, Meyer C, Eyben F, Zielke T, Meier G, Schuller B (2011) Deep neural networks for acoustic emotion recognition: Raising the benchmarks. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague 5688–5691. <https://doi.org/10.1109/ICASSP.2011.5947651>.
48. Tang H, Chu SM, Hasegawa M, Johnson HTS (2009) Emotion recognition from speech VIA boosted Gaussian mixture models. *IEEE International Conference on Multimedia and Expo*, New York, pp 294–297. <https://doi.org/10.1109/ICME.2009.5202493>
49. Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, Schuller B, Zafeiriou S (2016) Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, pp 5200–5204. <https://doi.org/10.1109/ICASSP.2016.7472669>
50. Ververidis D, Kotropoulos C (2006) Emotional speech recognition: Resources, features, and methods. *Speech Commun* 48:1162–1181. <https://doi.org/10.1016/j.specom.2006.04.003>
51. Wahlster W (2013) *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer Science & Business Media
52. Wang K, An N, Li BN, Zhang Y, Li L (2015) Speech Emotion Recognition Using Fourier Parameters. *IEEE Trans Affect Comput* 6:69–75. <https://doi.org/10.1109/TAFFC.2015.2392101>
53. Wu S, Falk TH, Chan W-Y (2011) Automatic speech emotion recognition using modulation spectral features. *Speech Commun* 53(5):768–785. <https://doi.org/10.1016/j.specom.2010.08.013>
54. Yu D, Deng L (2016) *Automatic Speech Recognition*. London Limited, Springer
55. Zhao J, Mao X, Chen L (2019) Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed Signal Process Control* 47:312–323. <https://doi.org/10.1016/j.bspc.2018.08.035>
56. Zheng W, Xin M, Wang X, Wang B (2014) A novel speech emotion recognition method via incomplete sparse least square regression. *IEEE Signal Process Lett* 21(5):569–572. <https://doi.org/10.1109/lsp.2014.2308954>
57. Zhou J, Wang G, Yang Y, Chen P (2006) Speech Emotion Recognition Based on Rough Set and SVM. *5th IEEE International Conference on Cognitive Informatics*, Beijing, pp 53–61. <https://doi.org/10.1109/COGINF.2006.365676>