Speech Emotion Recognition using 1D CNN and LSTM Networks

**Abstract: When it comes to virtual assistants, mental health inspection, and customer service, emotion detection in speech  is a must-have element for interaction between people and machines. Various distinctive characteristics of speech may be used to extract valuable information from audio samples. We want to construct an emotion identification system utilizing the attributes detected in the audio samples.**

**There are various traditional models including SVM, KNN and RF classifiers in machine learning for SER system. Our Proposed model is a combinational work of 1D CNN and LSTM, which includes of a total of four local feature learning blocks(LFLBs) and one long short-term memory(LSTM) layer. The LFLB basically includes of one convolutional layer and one max-pooling layer which are efficient for identifying native correlations and constructing hierarchical correlations. Long term relationships from the provided local features is taken care of LSTM layer.**

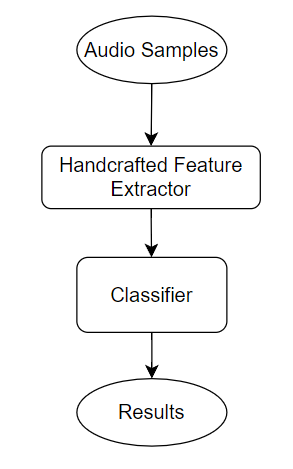
**The amalgamation of convolutional neural network (CNN) and LSTM may exceed the limitations of both networks and are examined using Berlin EmoDB dataset. Outperforming the present conventional models, our proposed method provides a new benchmark for accuracy and efficacy in SER system.**

**Keywords—Speech emotion recognition, convolutional neural network (CNN) and LSTM network, Audio samples, Long short-term memory, Berlin EmoDB**

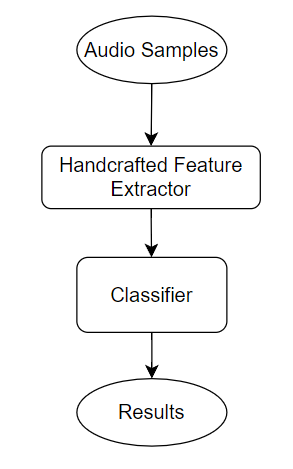
# Introduction

The ability to discern nuanced expressions in spoken language has propelled speech emotion recognition (SER) to unprecedented levels of success in recent years. In contrast to fleeting facial expressions or subtle body language, speech affords an uninterrupted and perpetual glimpse into our interior selves. Due to its quickness and data richness, SER is now heading the charge in emotion detection and identification research. It provides an exhaustive arsenal for comprehending and addressing all human emotions.

The first obstacle that is still being explored in speech analysis arises from an attempt to study the core emotions, which depends on correctly identifying paralinguistic components-auditory elements exclusive of speech content.Speech, we discover, comprises two kinds of information: The material content provided by the words (linguistic) and non-material emotional information encoded in tone and delivery of the auditory communication (paralinguistic). The affects are not simple but they ‘t require an analysis of many auditory parameters—which are usually characterised as continuous (pitch, speech rate) quality (voice quality, laughter), and spectrum (frequency distribution). Despite the fact that academics have looked contributed to a myriad of criteria, the so-called “best-of-the bunch” category has proved almost elusive. All these principles can be considered essential for better decision-making, which brings us to deep learning — a powerful technology allowing automatic discrimination of higher‐level features in emotional utterances and thus mapping speech to emotion. Although manual feature formulation, carefully selected using only sound files could perform well such as varied accuracy, the process is lengthy and labor-intensive is frequently forgotten about lower lying patterns. This feature extraction is further expedited by the structure function of deep learning known as hierarchy technique, which unravels innumerable emotional inflections thereof that are beyond an artificially defined semantic domain of hand-crafted procedures. The development of deep learning posed shock to both the area for audio signal processing. It is surprising that researchers are getting these amazing disclosures in various ventures, and also utilizing darns like Convolutional Neural Networks (CNNs), Deep Belief Networks (DBNs) and Short Term Memory (LSTM). [1] Nevertheless, such deep neural networks might frequently function as an “almost black box,” where groundwork in the dark unexplored of their mind-boggling outputs continues to remain veiled in obscurity. To address this veil two main methodologies have prevailed. The data modelling approach advocated by the statistical team is mostly concerned with understanding what contributed to the result, while the deep learning group prioritizes elucidating whether there are any factors that can be revealed in a deeper analysis, which is focused on creating algorithms that are specialized predictors, known as an “algorithmic model.” Although DNNs are not transparent in the way they learn high-level function, in specific applications their performance is a far cry as to how these were before (see Fig.1).



**Fig.1**. Flowchart of Recognition of Speech Emotion.



**Fig.2.** A flowchart of recognition of speech emotion approach build in this work.

Our concept for speech emotion identification depends on a strong convolutional neural network and long short-term memory (CNN-LSTM) architecture. This network combines four specialized "local feature learning blocks" with different critical layers to extract affective information from speech. Recognizing emotions from spoken words involves accounting for the varying character of sound over time. That's why the LSTM layer plays a key function, capturing long-term dependencies within the audio stream. Unlike prior systems that frequently compress data into a tight corset, our network depends on a modest collection of low-level features, including spectral information, substantially decreasing the amount of data to process. This compact method has an evident benefit - it reduces training time dramatically. And the facts speak for themselves our CNN-LSTM model not only delivers great emotion detection rates but moreover features enhanced generalisation capabilities. This powerful blend of high success rates and vast variety offers promise for appealing applications in healthcare, medical diagnostics, and even managing the complexity of social relationships.

Our work delivers numerous creative contributions: 1) compression of convolutional layer, exponential layer, batch normalization layer and max pooling layer in a LFLB; 2) addition of a layer of LSTM to capture long-term dependencies from sequences of local features, generating CNN LSTM networks following LFLB; 3) experimental substantiation exhibiting, for the first time, the potential of a 1D CNN LSTM system to gain copious emotional aspects straight from raw audio syllables (see Fig.2).

# LITERATURE SURVEY

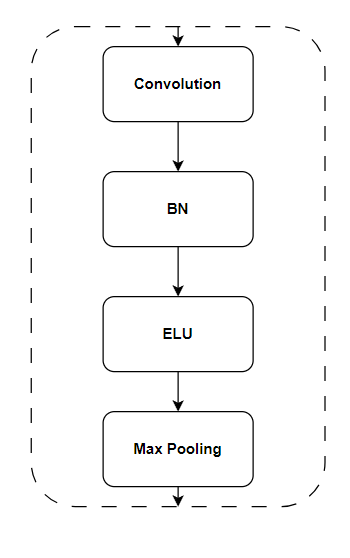
Identifying differentiating features is crucial for speech emotion recognition. Spectrum features, among numerous paralinguistic qualities, find major applicability. Kandali et al. coupled MFCCs with a Gaussian mixture model to differentiate Assamese speech emotions [2]. Milton et al. employed a Support Vector Machine for emotions in the Berlin Emotional Database in 3 stages. Waghmare et al[3]. employed MFCCs to analyze Marathi impassioned speech[10]. Demircan et al. implemented k-NN for Berlin EmoDB [5]. Nalini et al. established a technique employing residual phase and MFCCs with AANN [6]. Chenchah et al. employed HMM and SVM for affective speech [7]. Nalini et al. included information from MFCCs and residual phase for emotion recognition in music employing AANN, SVM, and RBFNN [8]. Despite their utility, most improvised features are low-level.

With a surge in successful Deep Neural Network (DNN) applications, researchers increasingly delve into extracting fundamental affective features. Stuhlsatz et al. innovatively deployed Generalized Discriminant Analysis DNNs layered with limited Boltzmann machines for voice emotion identification, outperforming SVM baselines [14]. Schmidt et al. introduced a deep belief network for music emotion through regression [19], whereas Le et al. produced state-of-the-art results on FAU Aibo using hybrid classifiers [9]. Han et al. increased utterance-level emotion identification using DNNs [11], while Mao et al. merged a semi-CNN architecture with a linear SVM, yielding robust recognition in demanding conditions [16]. Zheng et al. established CNN's superiority over SVM in emotion identification from annotated audio data [4].

Our method isolates itself from the prior indicated initiatives. The constructed 1D CNN and LSTM networks effectively obtain hierarchical native and global characteristics, enhancing speech emotion recognition. In contrast, many data models are confined to obtaining only low-level information for categorization. Previous algorithmic strategies, whether DBN-based or CNN-based, generally rely on learning a given type of emotion-related attribute for recognition.

# PROPOSED METHODOLOGY

Uncovering more discriminating emotion components serves as a primary target for researchers in speech emotion recognition. Speech attributes, designated as handmade or taught, differ in extraction approaches. Handcrafted features are meticulously designed with exhaustive descriptions of function. Conversely, learning features from deep networks like RBM-based DNNs [12, 14] and CNNs [13] exhibit extraordinary predictive potential. Consequently, the trend towards using deep features for predictions continues to acquire momentum.



**Fig.3.** A flowchart of the LFLB layer. BN: batch normalization, ELU: exponential linear

## **Deep feature learning**

The combination of the Local Feature Learning Block (LFLB) and Long Short-Term Memory (LSTM) permits for obtaining the individual as well as aggregate characteristics from unprocessed audio samples. The Convolution layer in LFLB is particularly successful at analyzing grid values and extracting sequential characteristics from surrounding input components. In contrast, LSTM is notably adaptable for managing sequences, where each learnt feature depends on the preceding output components. The collaboration of CNN and LSTM enables the accumulation of detailed features that compose both precise localized information and larger enduring linkages.

Learning of Local Features:

A unique  Local Feature Learning Block (LFLB), functioning as an alternative for CNN, is meticulously constructed for retrieving emotional data. Each LFLB encompasses a layer of convolution, an exponential linear unit (ELU) layer, a batch normalization (BN) layer, and a max-pooling layer, as represented in Fig. 3. The convolution layer and pooling layer, illustrated in Fig. 4, function as the fundamental components of the LFLB. The convolution layer displays relatively local interconnection and associated weights, allowing it to operate as the learning kernel [39–42]. The BN layer normalizes activations at each iteration, enhancing deep network performance and stability. This normalization processes retain mean activation closeness to 0 and activation standard deviation proximity to 1 [37]. The layer of Linear unit in exponential form specifies the output, introducing below zero values that drive the average of activations near to zero, accelerating learning and delivering improved recognition accuracies [38]. The pooling layer enhances feature resistance against distortion and interference, with max-pooling being the most often applied non-linear function. It divides the given data into discrete segments and delivers the utmost value for every segment [67]. Configuration of the  learning block may be task-specific, especially in the variables of the convolution layer along with max-pooling layer. The convolution layer functions as a local feature extractor. Upon input, data undergoes convolution with kernels along the width and height of the input volume, providing a feature map employing dot product calculations. In the case of a 1D convolution layer processing a signal s(n), the output o(n) arises from convolving s(n) with a randomly initialized convolution kernel k(n) of size l in our experiment.

(1)

Next, the features endure input into the BN layer, where activations from the previous layer are normalized at each iteration. The normalisation layer offers a modification that aligns the average value of the problematic characteristics closer to zero and the variation of these features near to a single. Upon feeding the normalized features into the linear unit of exponential layer, the produced features may be represented as follows:

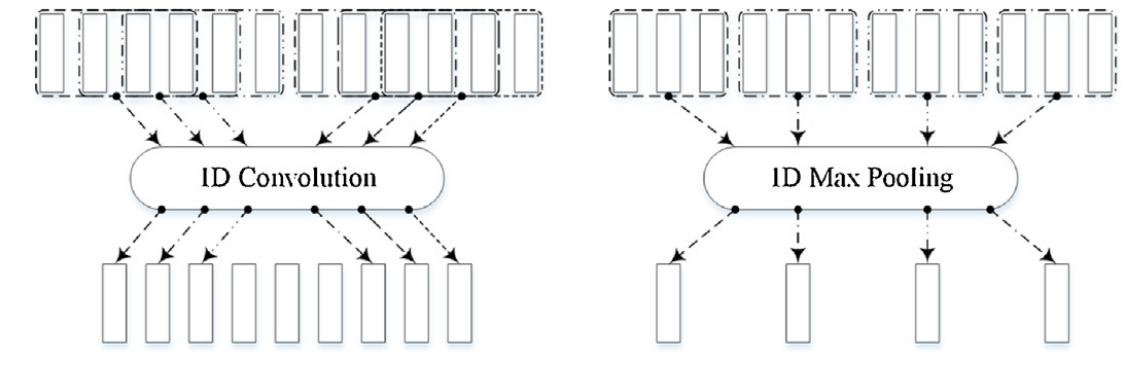
(2)

Here, and represent the *i-th* output feature at the *l-th* layer and the *j-th* input feature at the *(l-1)-th* layer, respectively, with representing the kernel between the *i-th* and *j-th* features.

The BN(·) function generalizes the information collected from the layer of convolution. Additionally, represents the activation function of ELU network, represented as:

(3)

The additional alpha constant, represented by ˛, is stated to have a value above zero (α > 0), where e signifies Euler’s

**Fig.4.** The right side figure represents the kernel of 1D convolution (size :4, stride: 1) and the left side figure represents the kernel of 1D max pooling (size: 3 , stride: 3)

number. Subsequently, These attributes undergo input into the max-pooling layer. This layer conducts a non-linear down-sampling technique, therefore diminishing the resolution of the features. The outcome of the features obtained by the max-pooling layer may be described as :

(4)

The pooling region is represented as Ωk with index k, the output and input features of the l-th max-pooling layer are represented as and ( with index k and p.)

Learning Global Features:

Long Short-Term Memory (LSTM) is a recurrent neural network (RNN) architecture that we employ to capture long-term contextual connections in an efficient manner [43,44]. LSTM, which was designed with the express objective of uncovering significant correlations within sequences, functions as an excellent complement to the LFLB. By virtue of this integration, the LSTM is able to add contextual dependencies that are gained from the gathered local feature sequences. Functioning with the following four components: an input gate, an output gate, a forget gate, and a cell that preserves the specified type equation.The LSTM may pick, remove, or add information to the block state via a self-recurrent link. The accompanying equations explain the development of an LSTM unit at every timestep t.

The input of an LSTM is and output of the LSTM is ,the connection between the both input and output can be expressed as

(5)

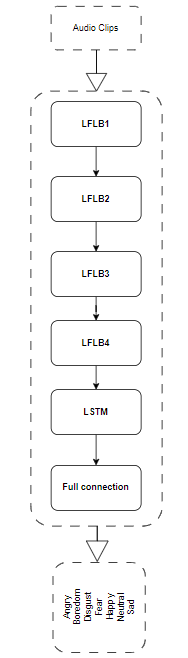
(6)

(7)

(8)

(9)

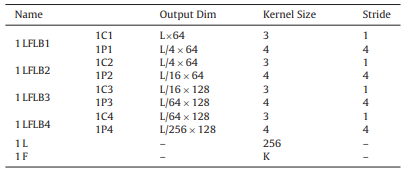
Here, indicates the state of LSTM unit, whereas X, Y, and a correspond to parameter matrices and vectors. Additionally, , , and signify vectors of the gate, whereas denotes a function of sigmoid, and and symbolize hyperbolic tangents. The operator ◦ signifies the product of the Hadamard. In Eqs. (5)–(9), the superscripts l − 1 and l indicate indices for input and output characteristics, respectively. The subscripts in Eqs. (5)–(7) designate the input gate, output gate, forget gate, and cell, respectively, with the subscript g in Eqs. (5)–(7) designating the gate.



**Fig. 5.** Block diagram for the combinational CNN and LSTM networks..

We construct the CNN and LSTM networks by integrating four layers of Local Feature Learning Blocks (LFLBs), one Long Short-Term Memory layer, and one completely connected layer. To distinguish identical building elements or strata, we utilize following coding conventions: 1) the building block network belongs is defined by the preceding number in the designation ; 2) the next integer provides the index of the building block within the networks. The entire design of the both CNN and LSTM networks is depicted in Fig. 5.

Table:  
The network's layer parameters are specified by the output dimension, expressed as L(length) × N( number), where L = duration of the audio clip. The quantity of sensations in the audio clips is proposional to the K (kernel size) of connected layer (1F). Similarly, 1C\* and 1P\* denote the convolutional layer and the max-pooling layer of various 1LFLB blocks, and so forth.



A proper fusion is achieved by connecting four Local Feature Learning Blocks (1LFLB1, 1LFLB2, 1LFLB3, and 1LFLB4) with one Lstm layer and one fully connected layer. This network is explicitly designed to generate deep information from crude audio vignettes and therefore one dimension convolution along with fitting aggregation kernels in LFLB are obligatory. Each of the LFLB’s convolution kernels keeps a constant depth of three, stride parameter level one and is assigned with same padding. In the case of 1LFLB1 and 1LFLB2, there are 64 convolution kernels whereas in the third and fourth LFLBs (1LFLB3 and 1LFLOORU), the number increases to bext a hundred twenty eight. The aggregation operations in each LFLB are done by a max-pooling that follows a length of 4. Complete parameters of this design are in the given table1 The first layer is a softmax classifier essentially as the highest point used for emotion recognition provided by obtained data.

As this is one-dimensional network, and assuming the audio sample being represented by a vector of size 1XD, as it passes through LFLBs local features are detected. In post-reshaping, the LFLB4 features are put into the single LSTM layer (1L). This follows as a result of holding explicitly unaltered in an assortment that has multiple dimensions which is a summation of non-explicit grouping with autoregressive distinctive homes. To this end, the LSTM layer recognizes conditional relations between assimilated local features (see Fig. 6). As a result, the patterns outputted by LSTM layer translate to both local information and long term contextual reflections.

Thereafter, the inherited characteristics sustain cropping through connection to several 1F tiers producing a one-one relationship between 1L layer.

(10)

## **Data pre-processing**

The performance of the carefully constructed CNN LSTM networks was investigated using the openly available emotional speech dataset, Berlin EmoDB. The selected database comprises only of performed emotive speech datasets, wherein the participating actors communicated pre-determined words filled with particular emotions.

The Berlin Emotional Database (EmoDB) from 2005 encompasses seven discrete emotions, each accompanied by a nearly similar quantity of utterances, giving a comprehensive assessment of classification accuracy. This resource comprises classified audio samples and accompanying analytic results. Ten experienced performers offered these depicted emotional remarks, including emotions of wrath, tedium, disdain, fear, happiness, neutrality, and sorrow. The dataset comprises 535 phrases collected from everyday speech, providing interpretability across all relevant emotions [20].

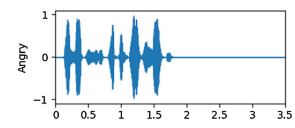
According to identifying emotions, the whole audio samples contained in the specifies database are applied for training. Audio fragment sampling rate is set at 16 kHz. The duration of the presented fragments is constant and takes 8 seconds for unprocessed sound fragments. For the audio clips that are above 8sec, they to cut appropriately and those shorter than this size of time are padded for wrought up a length eight picture. The figure below shows a 128.00 bit vector, which is concerned with the audio sample in order to express it at the sampling frequency of 16 kHz. Consequently, the vectors with 128000 bits formed a sense to the one-dimensional CNN LSTM network implemented in our experimental platform.

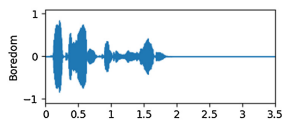
## **Experimental results**

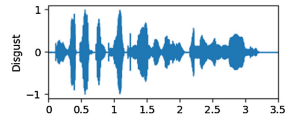
EmoDB experiments can be speaker-dependent and include independent data of this database. With the subsequent processing before which the embedded costs of effective instrumental content are driven out and save a 1D CNN LSTM as its primary tool to retain such affective contents from expressing audio recordings. Despite the fact that deep networks are usually termed \(‘black box method,\) they tend to function as predictive devices and not an expositive aim of describing algorithms determining these structures. Hsowever, we clarify that such an element is pertinent to our trained CNN LSTM network-based configuration, more precisely predictive ones rather than interpretative which are typically used for experimentation.

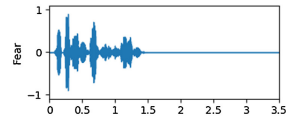
Various procedures are implemented to reduce the risk or quantity of overfitting in our experiments. Overfitting is a challenge since it could contribute to worse predictions when given with novel sample data. In circumstances of overfitting, the model prefers to recall the training data instead of enhancing its prediction skills. The occurrence of overfitting may be related to several variables, such as excessive complexity in a deep network or overtraining. An surplus of model degrees of freedom during network training also contributes to overfitting [66]. To combat this difficulty, we utilize batch normalization [18, early halting [15], and model selection . These techniques jointly attempt to repair and reduce overfitting concerns.

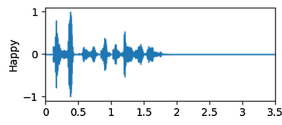
The database data was randomly separated into the training set and the testing set, covering 80% and 20% data respectively. Consistent results from the trials illustrate the reliability of the created 1D CNN LSTM networks in recognizing speech emotion. Our major aim is to obtain good generalization performance and accuracy in voice emotion recognition. Consequently, only the best-performing models are preserved in the trials. The accuracy of validation works as a critical indication for determining the capabilities for generalization of the training model. The recording of the model happens when the accuracy of validation peaks during the training of the constructed networks, assuring that the trained model not only matches the data in experiments but also displays amazing prediction ability in identifying speech emotion.

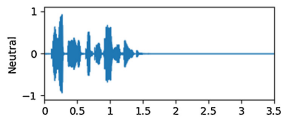
The given normal correctnesses in this inquire about don't reflect the furthest achievable correctnesses, given the differing qualities of models prepared within the aiming systems. To maintain a strategic distance from overfitting, as it were the best-performing models with predominant expectation abilities are considered. The normal exactnesses and approval exactnesses given in this think about are computed from these chosen models. Table 2 presents the test discoveries done on Berlin EmoDB, showing the execution of the finest prescient and well-fitted models. Within the trials, recording happens when the approval precision ceases extending amid preparing, guaranteeing the show achieves more prominent expectation execution (see Fig. 7). The chart illustrates that, amid the period of farthest approval exactness, the preparing exactness does not continuously surge. On the off chance that a circumstance happens where the approval precision brings down but the preparing precision reliably increments, meaning overfitting, early ending is done to end the training process.

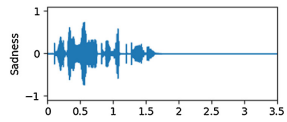












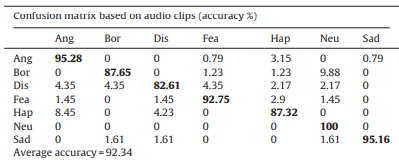
**Fig. 6.** Sample Waveforms of various emotions present in EmoDB data set.

Implementing early halting is a deliberate strategy to reduce overtraining and enhance the model's prediction capabilities. Once a selected monitor, such as training accuracy or validation accuracy, ceases to indicate progress, the model's training is discontinued. In our testing, we meticulously assess validation accuracy for increased prediction performance, setting the patience parameter to eight epochs. This assures that the training terminates when the validation accuracy no longer exhibits improvement, signifying the attainment of superior prediction performance by the network.

Table 2 illustrate the outstanding detection accuracies obtained in recognizing emotions within the Berlin EmoDB dataset. During our analysis, we discovered situations where agreeable statements were challenging to discern, especially without language context. This accords with our empirical observations. However, the intended networks effectively recognized the ebullient emotion in Berlin EmoDB. The differences in cultural, environmental, and educational variables among speakers undoubtedly contribute to the different chance of detecting or misrecognizing the same speech emotion across distinct cultures.

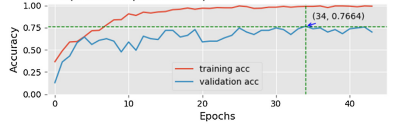
The data given in this table illustrates the designed network's capacity to effectively learn emotional properties from audio samples for speech emotion recognition. A comparative analysis using several known feature representations and methodology indicates that the constructed 1D CNN LSTM network performs well in terms of average accuracy. Notably, Table 8 reveals that the network has the greatest average accuracy when applied to the audio samples of Berlin EmoDB.

Table 2: The table demonstrates the speaker-dependent experiments in the form of confusion matrix on Berlin EmoDB (first column represents the different emotions present in the dataset).

.

## **Discussion**

As the condition of our organization, we produced 1D CNN LSTM systems with four last sheets of example squares in addition to a solitary LSTM layer permitting us to exploit both local and global data. As prolonged examination of a phenomenon related with speech synthesis is time-varying, this calls for the temporal characteristics of acoustic transmission to be reflected in it. Using the itake that CNN and LSTM offered simultaneously, we trained our network to obtain EEPO detection. Last but not least, even if the trials effectively revealed a considerable quantity of deep emotional data from the input dataset, characterizing and making human causal actions based on such conveyed emotions has an auditory aspect remains vital. Looking at the construct networks that we can see if extract emerging from this observation is that classifiers based on these networks turned out to be wonderful and convenient for emotion detection, a multiple-layer procedure not represented might demonstrate scale, while being better off with data extraction. Therefore, such networks were effective in speech emotion humanization for scores that were expected and swam considerably little due to diverse assessments.



**Fig.7.** The graph shows the validation and training accuracies on Berlin EmoDB

Table 3: The Comparison of other methods with our network on Berlin EmoDB.

|  |  |
| --- | --- |
| Research Work | Accuracy |
| Huang yongming et al[17] | 65 |
| Zhegwei Huang et al[13] | 68 |
| Our work | 76. |

Over fitting:

Based on our findings, regularization reacted as a potentially substantial strategy to restrict overfitting by setting up limitations on the model’s parameters and boosting the cost function. This is done by placing penalties on the parameters and activity of layers during optimization, entirely incorporated into the loss function of our deep networks. The data demonstrate that regularization does not only generate the convergence of our networks, but it also permits over fitting management.

Out model accordingly contains Batch Normalization (BN) layers to minimize overfitting within our CNN LSTM models. BN provides a dynamic interaction between aspects entering each line, turning conventional statistics files into indefinable quantities. This finding increased the generalization effects of our trials, indicating to make it simple for deep networks to train quicker and have less overfitting.

In addition, the early stalling was carefully employed to mitigate overfitting. Our iterative training procedure aiming at improving the relationship between model and training data, and early termination was necessary role which elevated the model’s generalizability. The selection of the monitored variables was as well capable to influence obtained results while arguing forbearance; they showed a complex approach warranting important safeguards in balancing forces. The experiments are able to present a reason for early stopping whereby the networks have gained more general features and end up predicting better.

## **Conclusion**

This paper discusses 1D CNN LSTMs- also known as speech emotion recognition. The focus is to understand the methods of entire learning for finding host local correlations and global contextual information from out-the-box audio recordings and log mel spectrograms. Image-based local feature learning is possible only using a Local Feature Learning Block (LFLB), which is created from the convolutional layer, batch normalization (BN) layer, exponential linear unit layer and max pooling. With the help of any supporting contextual connections through these local features are modeled and inserted into an LSTM layer so that now they can be perceived by the network. Therefore, the embedded properties attained from the procured CNN LSTM networks combine both local features and long-term contemporaneous relations.

The network's abilities were assessed on a benchmark database, revealing that the constructed CNN LSTM networks excel at learning differentiating characteristics and accumulating high-level abstractions of emotional input. In contrast to prior feature representations and methodologies, the 1D CNN LSTM network demonstrates a superior average accuracy.

On the other hand, though in this study deep networks indicate an advantage for speech emotion classification, many approaches should be considered. In particular, the full description of how emotions are perceived by the already-present networks is missing - how exactly these mentsioned networks perceive feelings remains a black box. While several researchers, as reported in Section E, have contributed a number of the advances in unveiling the black box phenomenon associated with deep learning being a learning approach that this paper is primarily concerned with reports about their efforts earn much of emphasis on deep networks used for picture processing.

The mode of gathering information by the use of speech is different from using photos; this calls for higher recommendations on how to penetrate and understand what’s in the “black box” referred to as deep networks meant only for processing voice. The issue always presented is the increased accuracy would be possible through vocal emotion identification. Particularly, as the network structure becomes more advanced and intelligent, looking into new topologies or learning strategies that can take advantage of broader properties or train superior prediction models is necessary. Moreover, formulating ways to be able to unite several deep characteristics learned by other networks is interesting perspectives as well.

**REFERENCES**

[1] A. Graves, N. Jaitly, A.R. Mohamed, Hybrid speech recognition with Deep Bidirectional LSTM, Autom. Speech Recognit. Underst. (2013) 273–278.

[2] A.B. Kandali, A. Routray, T.K. Basu, Emotion recognition from Assamese speeches using MFCC features and GMM classifier, in: TENCON 2008 - 2008 IEEE Region 10 Conference IEEE, 2008, pp. 1–5.

[3] ] A. Milton, S.S. Roy, S.T. Selvi, SVM scheme for speech emotion recognition using MFCC feature, Int. J. Comput. Appl. 69 (9) (2013) 34–39.

[4] W.Q. Zheng, J.S. Yu, Y.X. Zou, An experimental study of speech emotion recognition based on deep convolutional neural networks, in: International Conference on Affective Computing and Intelligent Interaction IEEE, 2015, pp. 827–831..

[5] S. Demircan, H. Kahramanl, Feature extraction from speech data for emotion recognition, J. Adv. Comput. Netw. 2 (1) (2014) 28–30

[6] N.J. Nalini, S. Palanivel, M. Balasubramanian, Speech emotion recognition using residual phase and MFCC features, Int. J. Eng. Technol. 5 (6) (2013) 4515–4527

[7] ] F. Chenchah, Z. Lachiri, Acoustic emotion recognition using linear and nonlinear cepstral coefficients, Int. J. Adv. Comput. Sci. Appl. 6 (11) (2015).

[8] N.J. Nalini, S. Palanivel, Music emotion recognition: the combined evidence of MFCC and residual phase, Egypt. Inf. J. 17 (1) (2015) 1–10.

[9] ] D. Le, E.M. Provost, Emotion recognition from spontaneous speech using hidden markov models with deep belief networks, in: Automatic Speech Recognition and Understanding (ASRU) IEEE, 2013, pp. 216–221.

[10] V.B. Waghmare, R.R. Deshmukh, P.P. Shrishrimal, G.B. Janvale, Emotion recognition system from artificial marathi speech using MFCC and LDA techniques, in: International Conference on Advances in Communication, Network, and Computing, 2014

[11] ] L. Chen, X. Mao, H. Yan, Text-independent phoneme segmentation combining EGG and speech data, IEEE/ACM Trans. Audio Speech Lang. Process. 24 (6) (2016) 1029–1037.

[12] Y. Kim, H. Lee, E.M. Provost, Deep learning for robust feature generation in audiovisual emotion recognition, Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on 32 (2013) 3687–3691.

[13] Z. Huang, M. Dong, Q. Mao, Y. Zhan, Speech Emotion Recognition Using CNN, ACM Multimedia, 2014, pp. 801–804.

[14] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, B. Schuller, Deep neural networks for acoustic emotion recognition: raising the benchmarks, in: IEEE International Conference on Acoustics, Speech, and Signal Processing IEEE, 2011, pp. 5688–5691

[15] J. Loughrey, P. Cunningham, Using Early Stopping to Reduce Overfltting in Wrapper-Based Feature Weighting, Trinity College Dublin Department of Computer Science, 2005, TCD-CS-2005-41, pp12.

[16] Q. Mao, M. Dong, Z. Huang, Y. Zhan, Learning salient features for speech emotion recognition using convolutional neural networks, IEEE Trans. Multimedia 16 (8) (2014) 2203–2213

[17] Y. Huang, A. Wu, G. Zhang, Y. Li, Extraction of adaptive wavelet packet filter-bank-based acoustic feature for speech emotion recognition, IET Signal Process. 9 (4) (2015) 341–348

[18] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, 2015, pp. 448–456.

[19] ] E.M. Schmidt, J.J. Scott, Y.E. Kim, Feature learning in dynamic environments: modeling the acoustic structure of musical emotion, in: International Symposium/Conference on Music Information Retrieval, 2012, pp. 325–330.

[20] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, A database of German emotional speech, INTERSPEECH 2005 - Eurospeech, European Conference on Speech Communication and Technology (2005) 1517–1520