OPTIMIZED ANALYSIS OF EMOTION RECOGNITION THROUGH SPEECH SIGNALS

[1]Sahith,[2]Naresh,[3]Swethan, [4]V. Kakulapati
[1,2,3,4,] Sreenidhi Institute of Science and Technology,
Yamnampet, Ghatkesar, Hyderabad, Telangana -501301
[1]aithasahith0214@gmail.com,[2] ballenaresh456@gmail.com,[3]swethan1502@gmail.com, vldms@yahoo.com

ABSTRACT: An accurate recognition of the user's emotional state is a primary aim of the human interface. The most pressing concern in the field of speech emotion identification is how to efficiently combine the extraction of suitable speech characteristics with a suitable classification engine in a parallel fashion. In this study, the concept of Emotion Recognition through speech signals involves predicting human emotions through speech with a high level of accuracy. This technology improves human-computer interaction, although it is challenging to predict emotions due to their subjective nature and the difficulty of annotating audio. SER relies on various factors such as tone, pitch, expression, and behavior to determine emotions through speech. The process involves training classifiers with samples, and the RAVDESS dataset is used as an example in this work. Due to the wide range of vocal dynamics and pitch changes, emotion identification in spoken language is a difficult problem in computer vision. To overcome this, the Convolutional Neural Network (CNN) method is utilized for speech emotion detection; this employs emotion recognition modules and learners to determine the difference between states of happiness, surprise, anger, neutrality, and sorrow. The system's dataset is built from voice signals; the LIBROSA program is used to retrieve attributes from these samples. The highest precision may be attained through Adam optimization.

***Keywords*:** speech, CNN, LSTM, Adam, feeling, classification, optimization, prediction

1. INTRODUCTION

Humans' ability to talk to each other is both our most advanced and most basic way to talk to each other. As air moves from the lungs to the larynx through the trachea, it makes the vocal cords vibrate, which sends speech signals to the brain. Nowadays, have seen a rise in the investigation into the field of emotion recognition via speech. The biological sciences, psychophysiology, computer science, and artificial intelligence are all at the cutting edge of research into how to automatically identify and rate human emotions. People's feelings and other important attitudes can often be figured out just by watching how they act [1].

Emotional nuance is conveyed well via spoken cues. Recently, scientists have been working on developing AI systems capable of identifying emotions based on a speaker's voice output. As the speaker's pitch and the length of each frame, both affect the emotional content of the voice signal, it's vital to evaluate it at various time scales [2].

The emotional classifier must take into account the culturally and environmentally specific emotions that are encoded in each speech stream [3]. It is culturally and contextually specific how the speech makes the listener feel.

CNN networks can normalize input frequencies and pick up on local details, whereas LSTM networks may be trained to extract and learn from acoustic and textual features. Because of the temporal nature of both spoken and written language, LSTM is well-suited for the extraction and learning of acoustic and textual features. Yet, the expansion in the variance of the hidden state components is not due to an intermediate nonlinear hidden layer. Several recent studies have turned to CNNs and LSTM networks [4, 5] to enhance voice emotion detection.

In terms of representation capacity, DL (deep learning)-based techniques are preferable for SER. There have been several successful demonstrations of autonomous SER using deep learning algorithms like CNN and LSTM [6].

Results reveal that the proposed speech emotion detection system performs better on the RAVDESS dataset with optimized features than the current state-of-the-art methods. The article is broken out as follows: In Sec 2, we provide relevant research on the automatic identification of speech emotions. The proposed system is described in detail in Sec 3. The results of the experiment are discussed in Sec 4. The research concludes in Section 5, followed by future research.

2. RELATIVE WORK:

Spectral Regression [7] is a generalized model that makes use of the connections between Extreme Learning Machines (ELMs) and Subspace Learning (SL). The hope was that this model would compensate for the weaknesses of spectral regression-based GE (graph embedding) and ELM. The effectiveness and viability of the methodologies were evaluated in comparison to standard methods by demonstrating their use across 4 speech-emotional corpora. The researchers, Zhaocheng Huang et al., used a token-based, heterogeneous approach to identify depressed speech. Sharp transitions and auditory regions were determined independently and together in fusions of several embedding techniques. Methods developed for identifying depressive disorders and, presumably, other medical conditions with an impact on voice production were implemented.

To recognize voices from different corpora, the Transfer Linear Subspace Learning (TLSL) framework [8] might be used. Strong representations of attributes over corpora are what TLSL hopes to extract and place in the trained estimated subspace. Current transfer learning methods, which only look for the most transferable parts of a trait, benefit from this development. TLSL is much superior to the other transfer learning methods based on the transformation of characteristics. However, a major limitation of TLSL is that it prioritizes seeking the transportable components of traits while ignoring the less useful portions. Emotional voice recognition in several databases is made possible with the help of TLSL.

Researchers in [9] used LSTM to analyze the feelings communicated by long stretches of text and found that LSTM performed better than conventional RNN in making these determinations. CNN and LSTM [10] have been used to extract high-level properties from raw audio data for use in speech emotion recognition. The greater processing power and storage capacities of LSTM make it a potential solution to the gradient explosion or disappearance problem faced by standard RNNs. As speech is a nonlinear time-series transform signal and text information is tightly related to temporal context, the LSTM network is well-suited for retrieving and classifying acoustic and textual characteristics that model in context and aid in comprehending the value of features. Nevertheless, there is no nonlinear hidden layer in the center, which would provide additional variability for the hidden state components [11].

Their automated evaluation method [12] is used narrative speech data from a Cantonese-speaking PWA diagnosed with aphasia. The linguistic abnormalities in the aphasic speech were identified by analyzing the textual features of the speech data. The Siamese network's learned text features were shown to have a substantial correlation with the AQ scores. The performance of automated speech recognition (ASR) on disordered speech and other languages has to be improved, and a substantial collection of such data should be accumulated for the automatic categorization of aphasia subtypes before this approach may be used more widely.

3. METHODOLOGY

The system is given a set of training data that includes labels for an expression, and it is also given the option to get weight training for the network. A sound is read as participation. Strength normalization is then performed on the audio. To prevent the training performance from being negatively impacted by the presenting order of the examples, normalized audio is employed for the Convolutional Network's instruction. It is via this training method that the sets of weights are produced that provide optimal outcomes. The dataset provides the system with pitch and energy during testing, and then, using the trained network's final weights, it provides the associated emotion. There are five possible expressions, and each of them has a corresponding numerical result. Machine learning classifiers, including CNN and LSTM, were explored, as were a few others. CNN performed the best in our tests, with an 82% success rate. Using prediction has enhanced the click-through rate, but more importantly, it automatically learns from the data without any human domain expertise.

Table 1: The comparative analysis of existing systems

S. No	Methodology and tools	Accuracy	Results
1	CNN Classifier	Approximately 82%	It is good when clustering
2	LSTM Classifier	62%	It performs average

Adam: Adam, an enhanced gradient descent approach that incorporates adjustable learning rate and momentum, is another option. Similar to AdaDelta and RMSProp, Adam also preserves an exponential decay average of previously squared gradients, while also keeping an approximate mean of past gradients.

4. IMPLEMENTATION RESULTS

4.1. Data collection: The RAVDESS dataset has a total of 1440 files, with 24 professional actors (12 men and 12 women) and 60 trials per actor. The performers use a neutral North American dialect to deliver two paired lines while evoking seven distinct feelings via their voices: serenity, happiness, sadness, anger, fear, surprise, and disgust. There are three different expressions available for each emotion: normal, strong, and neutral.

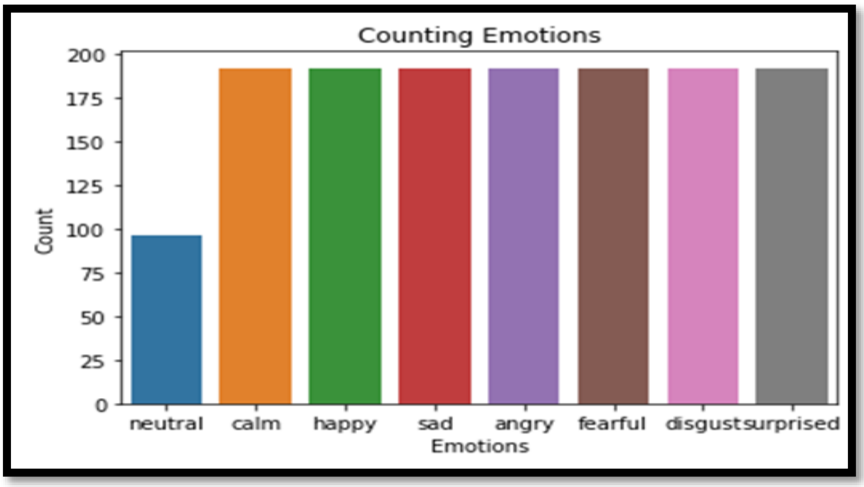


Fig-1 The count of different emotions in the data set.

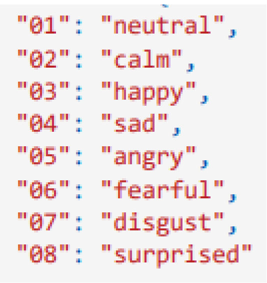
4.2 Data pre-processing: In Python, the librosa library is commonly used for audio signal processing tasks such as feature extraction from sound files. Librosa allows for the extraction of various features including MFCC (mel-frequency cepstral coefficients), chroma features, and mel spectrograms. Before extracting features, it is common to pre-process the audio data by removing noise using the noise reduction functions provided by librosa.

MFCCs are commonly used audio features that capture information about the spectral shape of the audio signal. They are computed by first applying a Fourier transform to the audio signal to obtain the power spectrum. In order to approximate the sampling rate of the audio signal, the power spectrum is converted to the mel-scale. Finally, the logarithm of the mel-spectrogram is transformed using a cosine transform to obtain the MFCCs.

Chroma features capture information about the pitch class of the audio signal. They are computed by first dividing the audio signal into short frames, typically 20-30 milliseconds long. In order to assign each frame's power spectrum to one of 12 pitch classes (matching the notes in the Western music system), a series of triangle filters are used. The chroma features are then obtained by summing the power spectrum values within each pitch class.

Mel spectrograms are similar to traditional spectrograms, but they use a mel-scale to represent the frequency axis instead of a linear scale. Mel spectrograms are commonly used as input features for machine learning models in audio classification and other audio-related tasks.
Overall, the use of librosa and these various feature extraction techniques enables the analysis of audio data and the development of various audio-related applications.

Table 2: Assigned numbers to different emotions



```
"01": "neutral",
"02": "calm",
"03": "happy",
"04": "sad",
"05": "angry",
"06": "fearful",
"07": "disgust",
"08": "surprised"
```

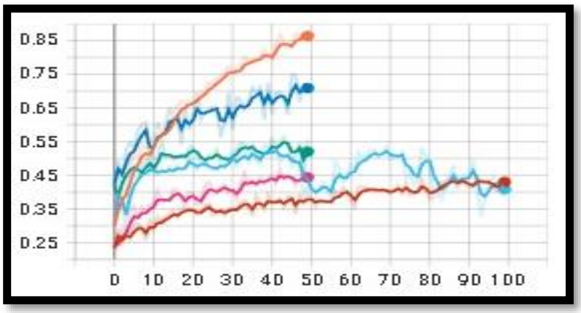


Fig. 2: The comparison of proposed model accuracy

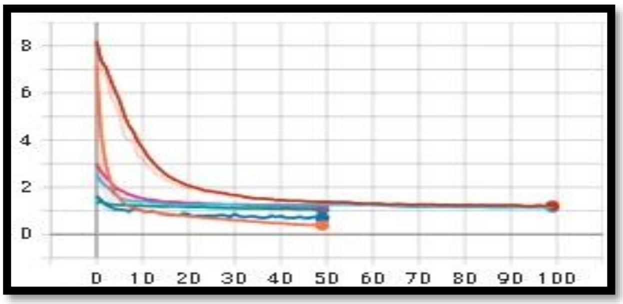
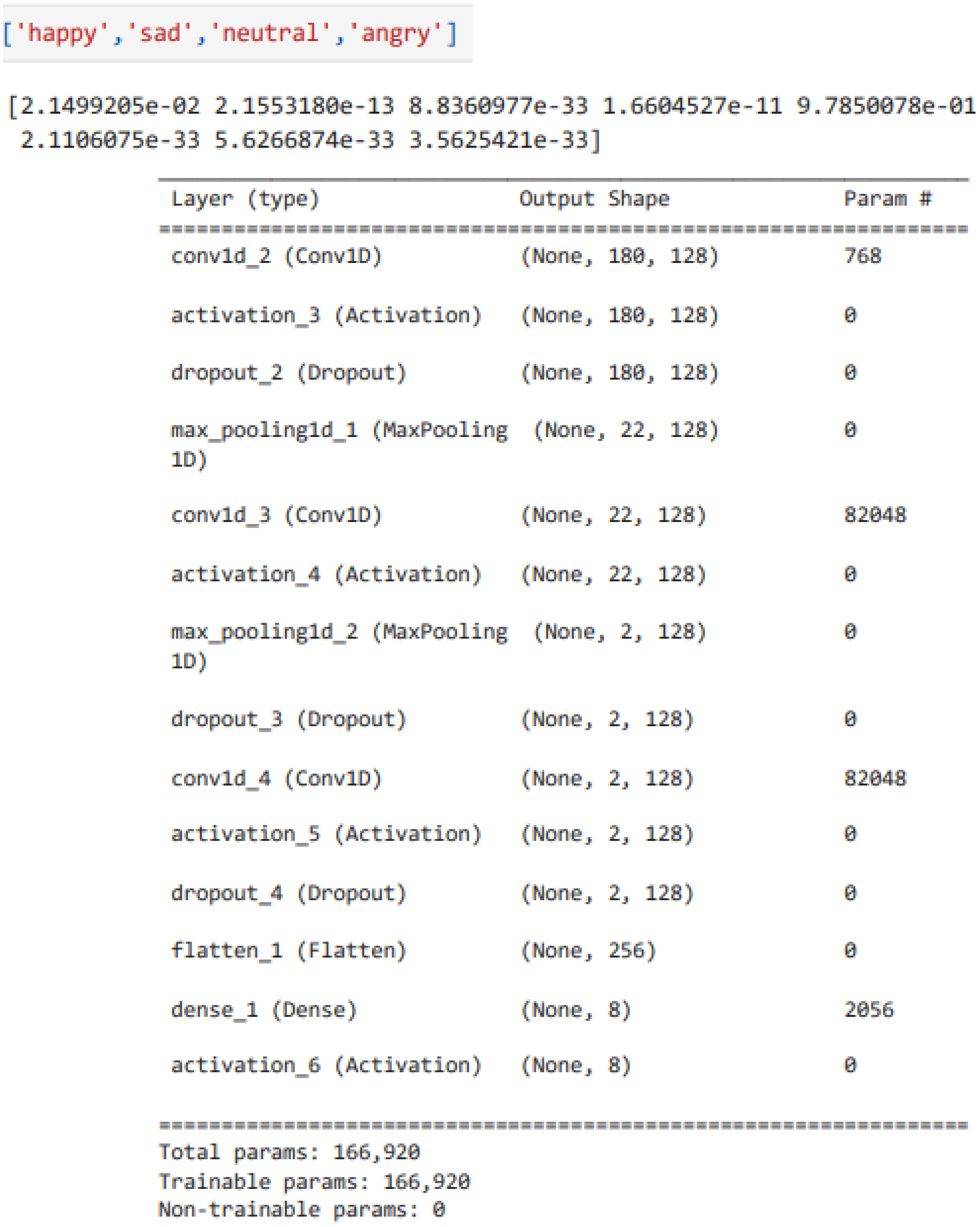


Fig. 3: The comparison of the proposed model loss function



```
['happy','sad','neutral','angry']
```

```
[2.1499205e-02 2.1553180e-13 8.8360977e-33 1.6604527e-11 9.7850078e-01
 2.1106075e-33 5.6266874e-33 3.5625421e-33]
```

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d_2 (Conv1D)           (None, 180, 128)          768

 activation_3 (Activation)   (None, 180, 128)          0

 dropout_2 (Dropout)         (None, 180, 128)          0

 max_pooling1d_1 (MaxPooling  (None, 22, 128)          0
 1D)

 conv1d_3 (Conv1D)           (None, 22, 128)           82048

 activation_4 (Activation)   (None, 22, 128)           0

 max_pooling1d_2 (MaxPooling  (None, 2, 128)           0
 1D)

 dropout_3 (Dropout)         (None, 2, 128)            0

 conv1d_4 (Conv1D)           (None, 2, 128)            82048

 activation_5 (Activation)   (None, 2, 128)            0

 dropout_4 (Dropout)         (None, 2, 128)            0

 flatten_1 (Flatten)         (None, 256)               0

 dense_1 (Dense)             (None, 8)                 2056

 activation_6 (Activation)   (None, 8)                 0

=================================================================
Total params: 166,920
Trainable params: 166,920
Non-trainable params: 0
```

Fig 4: The CNN-LSTM with ADAM and RMS prop optimization epochs

LSTM accuracy: 51.85%
RMSProp optimizer accuracy:61.8%
Adam optimization seemed to be the best fit for our CNN model from both experiments on optimization functions.
Accuracy: 81.76%
Several learning methodologies, including ADAM (Adaptive Moment Estimation) and RMSPROP (Root Mean Square Propagation) optimization algorithms, are used to evaluate the efficiency of the proposed approach. The suggested model enhances the accuracy of the SER for both the ADAM and RMSPROP algorithms. The RAVDESS dataset offers SER accuracy of 61.8% for the RMSPROP optimization approach and 81.76% for the ADAM algorithm.

5. CONCLUSION

Emotion recognition using speech involves identifying human emotions through speech patterns. To achieve accurate results, it is important to have a high-quality database with clear, noise-free recordings of actors' voices. Various methods for emotion recognition using speech have been developed, including feature extraction from speech samples and the use of CNNs and LSTM to classify emotions. While many different audio features can be used to recognize emotions, feature extraction using MFCCs has proven particularly effective in identifying emotions through speech. In an optimized analysis of emotion recognition through speech signals, the CNN classifier gave more accurate results than LSTM. This result recommends a technique for identifying emotions from an audio clip. The user trains the system by adding recordings of the sound and an emotional label to a database, and the system then through two rounds of training and assessment. The suggested technique performs an outstanding job, with a high accuracy rate, in emotion detection compared to earlier efforts.

6. FUTURE ENHANCEMENT:

In the future, emotional state recognition using multimodal analysis. Emotion recognition accuracy may be estimated with the use of voice signals and other classification algorithms based on RNN algorithms and probabilistic neural network techniques. The use of optimization strategies to enhance model efficiency.

6. REFERENCES:

[1]. M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo, ‘‘Recognizing emotions induced by affective sounds through heart rate variability,’’ IEEE Trans. Affect. Comput., vol. 6, no. 4, pp. 385–394, Oct. 2015, DOI: 10.1109/TAFFC.2015.2432810.

[2]. Y. Sun, G. Wen Ensemble softmax regression model for speech emotion recognition, Multimedia Tools Appl., 76 (6) (2017), pp. 8305-8328.

[3]. M. Ghai, S. Lal, S. Duggal, S. Manik, Emotion recognition on speech signals using machine learning, In Proceedings of International Conference on Big Data Analytics and Computational Intelligence (ICBDAC) (2017), pp. 34-39.

[4]. R. Ge, C. H. Wang, X. Xu et al., “Action recognition with hierarchical convolutional neural networks features and bi-directional long short-term memory model,” *Control Theory & Applications*, vol. 34, no. 6, pp. 790–796, 2017.

[5]. J. Zhao, X. Mao, and L. Chen, “Speech emotion recognition using deep 1D & 2D CNN LSTM networks,” *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[6]. Amin, K.R.; Jones, E.; Babar, M.I.; Jan, T.; Zafar, M.H.; Alhussain, T. Speech emotion recognition using deep learning techniques: A review. IEEE Access 2019, 7, 117327–117345.

[7] Xu, Xinzhou & Deng, Jun & Coutinho, Eduardo & Wu, Chen & Zhao, Li & Schuller, Björn. (2018). Connecting Subspace Learning and Extreme Learning Machine in Speech Emotion Recognition. IEEE Transactions on Multimedia. PP. 1-1. 10.1109/TMM.2018.2865834.

[8]. Song, Peng. (2017). Transfer Linear Subspace Learning for Cross-Corpus Speech Emotion Recognition. IEEE Transactions on Affective Computing. PP. 1-1. 10.1109/TAFFC.2017.2705696.

[9]. D. Li and J. Qian, “Text sentiment analysis based on long short-term memory,” in *Proceedings of the 2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI)*, pp. 471–475, Wuhan, China, October 2016.

[10]. J. Zhao, X. Mao, and L. Chen, “Learning deep features to recognize speech emotion using merged deep CNN,” *IET Signal Processing*, vol. 12, no. 6, pp. 713–721, 2018.

[11]. T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, “Convolutional, long short-term memory, fully connected deep neural networks,” in *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4580–4584, Brisbane, Australia, April 2015.

[12]. Y, Lee T, Kong APH. Automatic Assessment of Speech Impairment in Cantonese-speaking People with Aphasia. IEEE J Sel Top Signal Process. 2020 Feb;14(2):331-345. doi: 10.1109/JSTSP.2019.2956371. Epub 2019 Nov 28. PMID: 32499841; PMCID: PMC7271834.