2022 International Conference on Emerging Technologies in Electronics, Computing and Communication (ICETECC) | 978-1-6654-9087-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICETECC56662.2022.10069212

SPEECH EMOTION RECOGNITION USING DEEP LEARNING HYBRID MODELS

Jamsher Bhanbhro
Department of Computer Systems Engineering
Mehran University of Engineering and Technology
Jamshoro, Pakistan
jamsherbhanbhro@gmail.com

Shahnawaz Talpur
Department of Computer Systems Engineering
Mehran University of Engineering and Technology
Jamshoro, Pakistan
shahnawaz.talpur@faculty.muet.edu.pk

Asif Aziz Memon
Department of Computer Science
Dawood University of Engineering and Technology
Karachi, Pakistan
asif.aziz@duet.edu.pk

Abstract— **Speech Emotion Recognition (SER) has been essential to Human-Computer Interaction (HCI) and other complex speech processing systems over the past decade. Due to the emotive differences between different speakers, SER is a complex and challenging process. The features retrieved from speech signals are crucial to SER systems' performance. It is still challenging to develop efficient feature extracting and classification models. This study suggested hybrid deep learning models for accurately extracting crucial features and enhancing predictions with higher probabilities. Initially, the Mel spectrogram's temporal features are trained using a combination of stacked Convolutional Neural Networks (CNN) & Long-term short memory (LSTM). The said model performs well. For enhancing the speech, samples are initially preprocessed using data improvement and dataset balancing techniques. The RAVDNESS dataset is used in this study which contains 1440 samples of audio in North American English accent. The strength of the CNN algorithm is used for obtaining spatial features and sequence encoding conversion, which generates accuracy above 93.9% for the model on mentioned data set when classifying emotions into one of eight categories. The model is generalized using Additive white Gaussian noise (AWGN) and Dropout techniques.**

Keywords—Speech Emotion Recognition, CNN, SER, Stacked CNN

1. INTRODUCTION

Voice signals, the most natural and practical form of human communication, include linguistic data like semantics language type and a wealth of non-linguistic data like facial expression, speech emotion, and so forth. SER [1] has become increasingly important in recent years as artificial intelligence has continued to advance. Researchers are becoming more interested in the study that demonstrates how computers can recognize people's emotions in speech. Speech contains various paralinguistic information, including emotion, which has made speech emotion recognition an appealing research issue in many domains.

There are many factors that make the recognition of emotion from speech signals very challenging. There have been efforts in the last decades, however there aren't many accurate and balanced speech emotion datasets [2]. SER systems face many difficulties. First of all, it takes a lot of time and work to create a high-quality speech emotion database. Second, the database contains a variety of data with diverse speakers, each of whom has a distinct gender, age, language, culture, rhythm, tonality, etc. And finally, while expressing emotions in speech, sentences are commonly utilized rather than specific words. These elements are all crucially significant for SER system..

The readability of the generated text from audio and the accuracy, clarity of the extracted expressed words are the main concerns of traditional speech information processing systems[3]. In addition to the terms and information delivered, the speech signal also conveys the implicit emotional state of the speaker [4]. An excellent SER System that reflects the appropriate speaker's emotions by separating acoustic components is the foundation for a more efficient human-computer interaction. SER systems are useful and have crucial scientific significance in health, machine interactions, and other fields.

Traditionally, hand-crafted and engineered characteristics, such as signal energy, voice pitch, entropy, crossing rate, Mel-frequency cepstral coefficients (MFCC), and chroma-based [5-8], were used to create machine learning (ML) models for speech emotion recognition (SER). Yet how well these models perform depends on the features included. Research is still being done to look into new features and algorithms to predict the dynamics of feature sequences reflecting human emotions, even though it is unknown which characteristics most strongly connect with the different emotions, so that model can easily predict. On the other hand, new deep learning developments and the available processing capacity have enabled the scientific community to develop end-to-end SER systems efficiently.

These algorithms can quickly pick up information from spectrograms or unprocessed waveforms [9, 10], eliminating the need to manually extract a huge number of features, which is a significant advantage. CNN & LSTM models built on spectrograms and raw waveforms have been suggested in recent studies to increase SERs performance [11,12,13,14]. However, building such complex systems requires a large amount of classified training data. Additionally, more labeled training data may make the models more accurate. Furthermore, a lack of labeled training data may cause the models to be overfitted to particular data circumstances and domains, impairing generalization to other new data given for testing.

This study aims to use the RAVDNESS dataset to train two hybrid models, obtain method accuracy and compare the model's performances on slide change of training technique.

The following section of this study summarizes earlier research studies on SER and includes methodology, results, and conclusion.

2. RELATED WORK

Three essential steps of a conventional SER system are, preprocessing, feature extractions, and recognition or classification. SER system's accuracy depends upon the performance of correct feature extractions and correct classification. Hidden Markov-based SER techniques generate functions based on the probability of output using a gaussian distribution which helps to maintain nonlinearity or dynamic features proposed by Tin New [15].

However, because speech signals contain a variety of emotional states, the method necessitates numerous HMM, which increases the training computation, and makes it challenging to recognize emotions; however, overall accuracy becomes down. Herve et al. [16] developed a method for determining posterior probability utilizing large-scale computation to address the inaccuracy of Hidden Markov models by estimating all prior probabilities using processed contingent posterior probabilities, which helps to increase prior probabilities accuracies. LSTM is more beneficial for extensive acoustic modeling. According to the authors [17], they made suggestions by simulating long-term dependent characteristics of audio sequences in different layers of networks. Their created network model has a greater accuracy rate and produces an strong correlation with extracting meaningful information. Others [18] made an approach using layer-based CNN for speech recognition. They have used the reduction technique by adding an attention layer that finds significant weights for time frames to learn quickly. The attention layers help to focus on specific consequences, and so it helps to increase the recognition rate. This makes sure that crucial information doesn't get lost in speech. The technique excels in the basic convolution neural network in terms of object identification accuracy.

To remove the low accuracy issue, we have proposed a technique using deep learning algorithms that works perfectly. We have addressed different techniques that help the model perform much better. Unlike traditional or NLP methods, which first convert speech to text before training a model on text, these approaches use textual data as the input for the models. Compared to these techniques, CNN works much better; we have produced a model by using a neural network. The model is trained with Mel spectrograms of the audio. We have preprocessed audios so that the abnormal condition could easily be removed before giving input to the model that can produce over or underfitting issues. Perfect spectrograms are generated for input. But we have used advanced processing techniques like adding noises so that model can work as a generalized model.

The model's architecture contains four stacked convolutional blocks attached sequentially. Features learned from one block are used as input to the other; hence, till the last block, only essential features are kept, producing perfect classification rates. Each Conv block has the same configuration. There are four layers in each block; the convolution layer applies a filter to capture only specific information description of this layer is mentioned in the methodology. Normalization tries to normalize features, and the activation layer activates the output of the normalization layer using Relu function. Finally, pooling and drop-out layers help to reduce dimensions and only keep features that can be sent forward, as the last block is LSTM, which helps to memorize and, because of memorizing, current, past, and based on neighbors' features, the final decision can be taken with complete perfection. We have used RAVDNESS [19, 20, 21] dataset and processed 1440 audios for model generation. The main advantage of this created model is that classification can be done in simple speeches and on audio songs. Accuracy is better as compared to other models if audio signals have some noise that represents model as best in generalization aspect.

3. METHODOLOGY

3.1 Dataset

Emotional Database

RAVDESS is a database of audio and video with data on emotions. 24 professional audio utterances. The accent used in this database is North American. The fact that this dataset includes songs and basic speeches is an advantage. Songs include feelings of sadness, neutrality, fear, happiness, anger, surprise, and disgust, as do voices that display these emotions. There were two emotional intensity levels (strong and normal) for each expression and also for neutral one. Ten times each actor was given ratings for emotional validity, intensity, and sincerity for collecting 7356 recordings(1440 only audio files) and 247 untrained study volunteers from North America supplied ratings.

The data set is divided into train, validation, and test sets in the order listed: (80,10,10). The Standard Scaler is used to scale dataset.

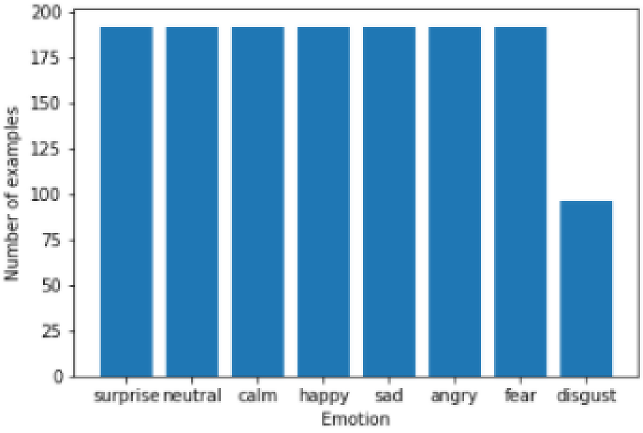


Figure-1 Dataset Visualization

3.2 Pre-Processing

The short sample size of the dataset makes the denser network containing various convolution blocks overfit on training data and underfit on others. As a result, the data augmentation mechanism is included in our design. However, it would be challenging to produce more real samples. The addition of significant noise will create learning issues, and adding small noise components will not make the model more generalized. We need a proper mix of noise. A channel model known as AWGN [21] assumes that the only interference with communication is the linear addition of broadband or white noise with a constant spectral density (measured in watts per hertz bandwidth) and Gaussian amplitude distribution. The model does not consider Fading, frequency selectivity, interference, nonlinearity, and dispersion.

Signals are loaded at a 48 kHz sample rate and shut off between [0.5, 3] seconds. The signal is padded with zeros if it is less than 3 seconds long.

The calculated MEL spectrogram is utilized as an input for the model, for the model spectrogram is split into seven chunks.

3.3 Model

i) CNN – LSTM

The dataset is loaded in this model, then Mel Spectrogram is calculated using Librosa python. MEL spectrum is then divided into seven chunks as mentioned in preprocessing (Mel Spectrogram is found using a hamming window with width 512 and hop lengths 256). Mel Spectrogram, after dividing into chunks was given as input for better learning of 2d CNN [22] with time distributed [23] layers (in a fashion of stack) with four conv blocks used. Figure 2 represents model architecture. The description of four sequential 2dConv blocks is as follows:

In the first block, channel one was used as an input, and channel 16 for output with padding & stride one and kernel size 3. In the first block, max pooling was done with a kernel size of 2 and stride size of 2. And 16 channels/dimensions are used from the last layer to the norm layer.

In the second block channels, 16 were used as input, and channel 32 for output with padding & stride one and kernel size 3 [24]. In the first block, max pooling [25] was done with a kernel size of 4 and stride size of 4. And 32 channels/dimensions were used from the last layer to the norm layer.

The other two blocks also have the same configuration as the second block; only Batchnorm2d varies 64 for the third and128 for the fourth block, and each block has a dropout of the same rate of 0.4 except for the last LSTM [26] block.

After the above configuration, the model uses whatever is found after flattening LSTM is combined with the Linear SoftMax Layer [27,28] (which will help to recognize several inputs that must be the same as hidden nodes in the LSTM layer). And finally, the fully connected layers give the loss function/output and the activation function [29, 30], which will provide the probability of the emotion (classification). The next step in coding was training the model just created and then validating and verifying the model.

Each convolution block has almost the same configuration as mentioned. More critical is the mapping of layers inside convolution; the first layer is the convolution layer applies a filter on the input spectrogram, finds features, and sends it to the batch normalization layer. The BN layer normalizes the output of the conv layer, and then the Activation layer activates outcomes of the BN layer using the Relu activation function. Finally, dimensions are reduced so that only important features can be considered. The Max Pooling layer helps to reduce dimensions. The last layer of the conv block is the dropout layer which dropout neuron and send the most important features obtained to the LSTM block. LSTM block provides the advantage of keeping all the outputs of singular images obtained from CNN block processing. LSTM helps to remember the correct sequences of the classes and their predictions.

Initially model could have performed better on testing data; AWGN noise was added to remove overfitting because results on training samples were accurate, while on testing samples, results were not correctly predicted. The addition of AWGN and the dropout of neurons helps to overcome this issue.

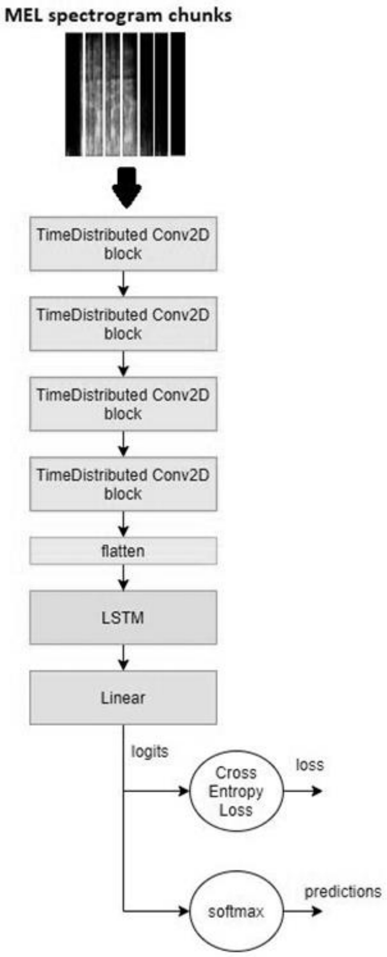
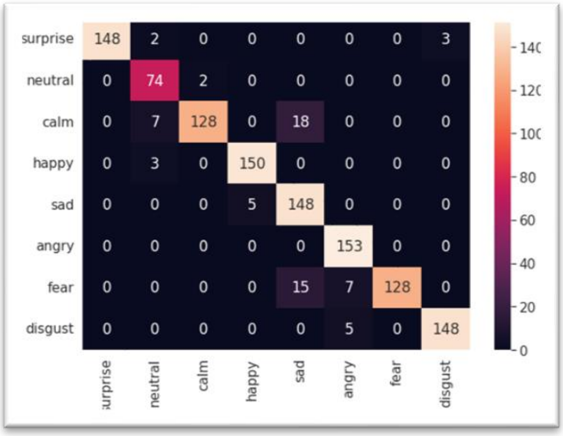

MEL spectrogram chunks
TimeDistributed Conv2D block
TimeDistributed Conv2D block
TimeDistributed Conv2D block
TimeDistributed Conv2D block
flatten
LSTM
Linear
logits
Cross Entropy Loss
loss
softmax
predictions

Figure 2: Flow Diagram of First Model

4. Results

The following are the model results, and it produces the best results. There are 1147 audio clips for training (80% of the dataset). As the dataset is balanced, it contains the same clips for all the categories. Results produced by the model are represented in the following confusion matrix.



	urprise	neutral	calm	happy	sad	angry	fear	disgust
surprise	148	2	0	0	0	0	0	3
neutral	0	74	2	0	0	0	0	0
calm	0	7	128	0	18	0	0	0
happy	0	3	0	150	0	0	0	0
sad	0	0	0	5	148	0	0	0
angry	0	0	0	0	0	153	0	0
fear	0	0	0	0	15	7	128	0
disgust	0	0	0	0	0	5	0	148

Figure 3: Confusion Matrix of Model

As mentioned, loss before AWGN is more and after addition of AWGN reduces loss. The following figures (figure-4&5) show the loss results before and after adding AWGN. (Images are directly taken from software- that's why there are grids)

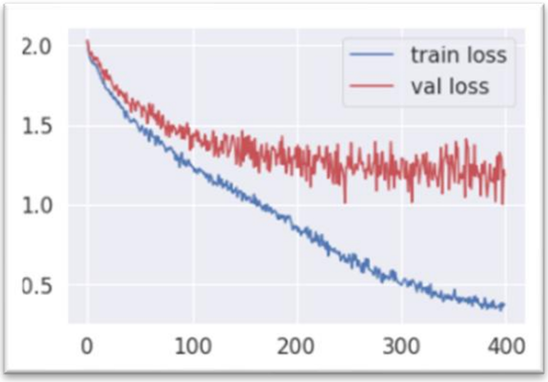


Figure 4: Loss Before AWGN

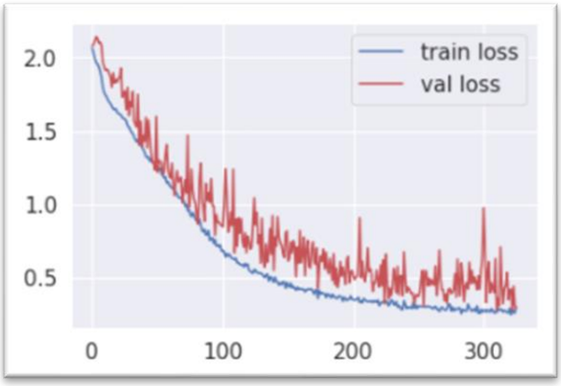


Figure 5: Loss After adding AWGN

The table below provides a thorough description of the confusion matrix. Both the right and wrong predictions for each class are mentioned. Except in the case of fear, all classes are correctly predicted.

S.no	Actual	Model Predicted
1	Surprise	96.5% Surprise, 1.75% disgust and neutral
2	Neutral	96.5% Neutral, 3.5% Calm
3	Calm	84.5% Calm, 12% sad, 3.5% Neutral
4	Happy	98.2% Happy, 1.8% Neutral
5	Sad	96.5% Sad, 3.5% Happy.
6	Angry	100%
7	Fear	84.4% Fear, 10% Sad, 3.5% Angry, 1.5% Surprise
8	Disgust	96.5% Disgust, 3.5% Angry

Table-1 Confusion matrix Description

Classification matrices, Accuracy, Recall, and F1-Score of the model obtained from the confusion matrix are mentioned below.

Class	Precision	Recall	F1 Score
surprise	96.7%	100%	0.97
neutral	97.3%	86.04%	0.91
calm	83.6%	98.5%	0.90
happy	98.0%	96.8%	0.97
sad	96.7%	81.7%	0.89
angry	100%	92.7%	0.96
fear	85.3%	100%	0.91
disgust	96.3%	98.01%	0.97
Accuracy Overall	93.4		

Table-2 Performance Matrices

A few main things affect the model's performance; loss before augmentation is much more, and after augmentation, there is much reduction in training losses, so the addition of AWGN gives an advantage. AWGN advantages model generalization. As the model is a hybrid model (CNN+LSTM), the model needed much time for training with epochs 500+, (The average accuracy drops to 60% with epochs under 200. The best emotion classification accuracy of the first is 93.9%.

	Epochs	Training time	Accuracy
Model	200	6 hours	60%
	600	20 hours	80.37%
	1400	38 hours	93.9%

5. Conclusion

It is concluded that the model works perfectly. It's almost impossible to produce such perfect results, with many problems for SER systems due to the varieties of inputs. CNN + LSTM delivers benefits and has higher accuracy rates due to repeated learning and deep feature extractions. The model is more complex, demands significant training, and yields more accurate predictions than the other SERs. The precision of the performance is a result of the balanced dataset addition of augmentation to prevent losses and the careful handling of convolution blocks and layers. It is concluded that using the best hyperparameters, the Mel Spectrogram of the signal used as input by deep learning algorithms can yield excellent results.

[1] R. Anusha, P. Subhashini, D. Jyothi, P. Harshitha, J. Sushma and N. Mukesh, "Speech Emotion Recognition using Machine Learning," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), 2021, pp. 1608-1612, doi: 10.1109/ICOEI51242.2021.9453028.

[2] Y. B. Singh and S. Goel, "Survey on Human Emotion Recognition: Speech Database, Features, and Classification," 2018 International Conference on Advances in Computing, Communication Control and

Networking (ICACCCN), 2018, pp. 298-301, doi: 10.1109/ICACCCN.2018.8748379.

[3] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," in IEEE Transactions on Speech and Audio Processing, vol. 8, no. 4, pp. 429-442, July 2000, doi: 10.1109/89.848224.

[4] S. R. Kadiri and P. Alku, "Excitation Features of Speech for Speaker-Specific Emotion Detection," in IEEE Access, vol. 8, pp. 60382-60391, 2020, doi: 10.1109/ACCESS.2020.2982954.

[5] Y. Zhan and X. Yuan, "Audio post-processing detection and identification based on audio features," 2017 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), 2017, pp. 154-158, doi: 10.1109/ICWAPR.2017.8076681.

[6] Roberts, L. (2020, March 14). Understanding the mel spectrogram. Medium. Retrieved July 20, 2022, from https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

[7] Prabakaran, D., & Sriuppili, S. (2021). Speech Processing: MFCC Based Feature Extraction Techniques- An Investigation. Journal of Physics: Conference Series, 1717.

[8] Shah, Ayush & Kattel, Manasi & Nepal, Araju & Shrestha, D.. (2019). Chroma Feature Extraction.

[9] Mengna Gao, Jing Dong, Dongsheng Zhou, Qiang Zhang, and Deyun Yang. 2019.End-to-end speech emotion recognition is based on a one-dima ensional convolutional neural network. In Proc. ACM ICIAI. 78–82.

[10] [Xi Ma, Zhiyong Wu, Jia Jia, Mingxing Xu, Helen Meng, and Lianhong Cai. 2018. Emotion recognition from variable-length speech segments using deep learning on spectrograms. InProc. INTERSPEECH. 3683–3687

[11] M. Neumann and N. T. Vu, "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 7390-7394, doi: 10.1109/ICASSP.2019.8682541.

[12] D. Eledath, P. Inbarajan, A. Biradar, S. Mahadeva and V. Ramasubramanian, "End-to-end speech recognition from raw speech: Multi time-frequency resolution CNN architecture for efficient representation learning," 2021 29th European Signal Processing Conference (EUSIPCO), 2021, pp. 536-540, doi: 10.23919/EUSIPCO54536.2021.9616171.

[13] Xi Ma, Zhiyong Wu, Jia Jia, Mingxing Xu, Helen Meng, and Lianhong Cai. 2018. Emotion recognition from variable-length speech segments

[14] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention.IEEE ICASSP. 2227–2231.

[15] Nwe T L, Foo S W, De Silva L C. Speech emotion recognition using hidden Markov models[J]. Speech communication, 2003, 41(4): 603-623. (in NETHERLANDS).

[16] Bourlard H, Konig Y, Morgan N, et al. A new training algorithm for hybrid HMM/ ANN speech recognition systems[C]/ / 1996 8th European Signal Processing Conference. Trieste, Italy: IEEE, 1996:1-4. (in Italy).

[17] Kipyatkova I. LSTM-based language models for very large vocabulary continuous Russian speech recognition [M]/ /Speech and Computer. Cham: Springer International Publishing, 2019: 219-226. (in Germany).

[18] Zhang Y Y, Du J, Wang Z R, et al. Attention-based recognition [C]//2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Honolulu, USA: IEEE, 2018: 1771-1775. (in USA).

[19] A. U A and K. V K, "Speech Emotion Recognition-A Deep Learning Approach," 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud) (I-SMAC), 2021, pp. 867-871, doi: 10.1109/I-SMAC52330.2021.9640995.

[20] Livingstone, S. R., & Russo, F. A. (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(5), e0196391.

[21] S. S. Meher and T. Ananthakrishna, "Dynamic spectral subtraction on AWGN speech," 2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN), 2015, pp. 92-97, doi: 10.1109/SPIN.2015.7095302.

[22] A. Mujaddidurrahman, F. Ernawan, A. Wibowo, E. A. Sarwoko, A. Sugiharto and M. D. R. Wahyudi, "Speech Emotion Recognition Using 2D-CNN with Data Augmentation," 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), 2021, pp. 685-689, doi: 10.1109/ICSECS52883.2021.00130.

[23] Eom, Y., & Bang, J. (2021). Speech Emotion Recognition Using 2D-CNN with Mel-Frequency Cepstrum Coefficients. Journal of Information and Communication Convergence Engineering , 19 (3), 148–154. https://doi.org/10.6109/JICCE.2021.19.3.148

[24] Pandey, A. K. (2021, January 24). Convolution, padding, Stride, and pooling in CNN. Medium. Retrieved July 20, 2022, from

[25] Phan, Huy, et al. "Robust Audio Event Recognition with 1-Max Pooling Convolutional Neural Networks." *Interspeech 2016*, ISCA, 2016, pp. 3653–57. *DOI.org (Crossref)*, https://doi.org/10.21437/Interspeech.2016-123.

[26] J. Oruh, S. Viriri and A. Adegun, "Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition," in IEEE Access, vol. 10, pp. 30069-30079, 2022, doi: 10.1109/ACCESS.2022.3159339.

[27] Passricha, Vishal and Aggarwal, Rajesh Kumar. "A Hybrid of Deep CNN and Bidirectional LSTM for Automatic Speech Recognition" *Journal of Intelligent Systems*, vol. 29, no. 1, 2020, pp. 1261-1274.

[28] James, Praveen & Mun, Hou & Vaithilingam, Chockalingam & Tan, Alan & Chiat, Wee. (2018). End to End Speech Recognition using LSTM Networks for Electronic Devices. Journal of Advanced Research in Dynamical and Control Systems. 10. 933-939.

[29] S. -X. Zhang, R. Zhao, C. Liu, J. Li and Y. Gong, "Recurrent support vector machines for speech recognition," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5885-5889, doi: 10.1109/ICASSP.2016.7472806.

[30] Graves, Alex, et al. Speech Recognition with Deep Recurrent Neural Networks. arXiv:1303.5778, arXiv, 22 Mar. 2013. arXiv.org,

[31] Zhang, Yuanyuan, et al. "Attention Based Fully Convolutional Network for Speech Emotion Recognition." 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2018, pp. 1771–75. DOI.org (Crossref), https://doi.org/10.23919/APSIPA.2018.8659587