2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP) | 978-1-6654-5116-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ATSIP55956.2022.9805924

A combined CNN-LSTM Network for Audio Emotion Recognition using Speech and Song attributs

Souha AYADI[1]
University of Tunis el Manar
National School of Engineers (ENIT),
dept. electrical engineering
Signal Image and Information
Technology Laboratory (SITI)
BP 37,le Belvèdère,1002,Tunis Tunisia
souha.ayadi@enit.utm.tn

Zied Lachiri[2]
University of Tunis el Manar
National School of Engineers (ENIT),
dept. electrical engineering
Signal Image and Information
Technology Laboratory (SITI)
BP 37,le Belvèdère,1002,Tunis Tunisia
zied.lachiri@enit.utm.tn

***Abstract*—Audio emotion recognition is a very active topic. It has many applications in various domains like health care , information technology and defence. This paper presents a hybrid CNN-LSTM Deep Network for audio emotion classification using speech and song description.The model combines the temporal modelling ability of a long shirt term memory (LSTM) with the CNN ability to learn invariant features. We evaluate our experements on RAVDESS Database that contains audio song and audio speech formats. We show the affectiveness of both CNN and LSTM neural networks for feature extraction and memorizing a long sentence in order to provide the stability and a good performance of the model by using MFCC which reinforces the idea of using frames instead of waves to improve the training process. We compare the obtained results for both modalities for the same amount number of emotions which are "happy", "sad", "angree", "neutral", "calm" and "fear".**

***Index Terms*—Song , Speech, MFCC, Conv1D-LSTM.**

I. Introduction

Emotion recognition is a very large field due to the branching out of the topics and the difference between the modality format of the database such as audio, image and video. There are several types of audio such as speech, music, and acoustics. All types of audio differ in how the same phrase is spoken which influences the timing and frequency of the emotion. Several researchers have studied the influence of music on human emotions [7]. Thammasan et al. [22] proposed an approach based on brainxave signals for continuous music emotion classification. Malik et al. [17] proposed a CNN and RNN architecture for two-dimensional emotional space of valence and arousal. Moreover, Liu et al. [14] presented a CNN aarchitecture based on MFCC for multiclass label classification.

While several ideas and several studies revolve around the Speech Emotions Recognition (SER) in order to solve some problems like the redundant features that have close correlations [15] and implementing different architectures [1] [24]. Such as Fayek et al. [1] that used frame based formulation based on minimal speech processing and end-to-end deep learning to evaluate feed forward and recurrent neural network architectures. Also, Li et al. [13] presented an end-to-end multitask learning with self attention by estracting fautures directly from speech and classify emotions along with gender. While, Nwe et al. [19] proposed a text independent method for speech emotion classification. Furthermore, Liw et al. [15] proposed a feature selection method by removing the redundant features and used an extreme learning machine (ELM) to improve the performance of emotion recognition. And, Yao et al. [24] presented three fusion neural networks architecture DNN, RNN and CNN for speech emotion classification.

All the presented methods agree on improving the performance of the model for better accuracy result. In this work, we are interested by convolutional neural networks (CNN) [14] that is considered a good architecture for extracting features by convolving the input data with kernels to get the main labels. And Long Shirt-Term Memory (LSTM) [12] which is an architecture stemming from a Recurrent Neural Network (RNN) which addresses the RNNs gradient disappearance and explosion problems. Our goal is to combine CNN and LSTM architectures and apply them to different audio formats to compare the performance of the same model for these formats.

This article presents an audio emotion classification by presenting a new Conv1D-LSTM architecture for Song and Speech conducted on RAVDESS database. This model uses MFCC for feature extraction by considering the database as frames instead of waves for facilating the training process. The combined model based on Conv1D and LSTM to improve results and use the best criteria of Conv1D and LSTM. This work is organized as follows. In section 2, we present a novel Conv1D and LSTM conbined architecture that is applied on song and speech audio data that uses the audio spectrogram as an input by using MFCC for feature extraction. In section 3, we discuss the results and compare the song results and the speech results. Finally, the conclusion along with the future

978-1-6654-5116-1/22/$31.00 ©2022 IEEE

work.

II. Model Structure

A. Spectrogram extraction

Mel-frequency cepstral coefficient (MFCC) is widely used for audio tasks and mostly known for feature extraction [5] in speech recognition [2], and for music retrieval [18]. The purpose of using MFCC is to imitate the hearing auditory ability. MFCC is obtained by applying on a nonlinear mel-frequency scale, a linear cosine transform of log power spectrum [9]. The detailed calculation steps for feature extraction are as follows:
1) the Pre-emphasis, which means applying a filter to emphasize the higher frequencies [20]. The transfer function for the filter used by Pre-emphases is:

$$H(z) = 1 - bz^{-1} \tag{1}$$

2) Windowing, which is usually a Hamming window that is applied on each frame of the input sequence to reduce the signal to the edges of the frame in order to improve the harmonics [20].
3) DFT spectrum is applied to convert the wendowed frames to magnitude spectrum.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}} \tag{2}$$

$$0 \leq k \leq N-1$$

Where:

- X(k) is the magnitude spectrum that represent the combination of real and imaginary numbers.
- x(n) is the discrete-time signal used to calculate the DFT.
- N is the number of samples that are converted to the frequency domain for each frame.

4)Mel spectrum means to apply a Mel-filter bank, which means a set of band-pass filters, on the Fourier transformed signal. The perceived Mel frequency (fmel) in terms of the physical frequency (f) can expressed as:

$$f_{mel} = 2595 log_{10}(1 + \frac{f}{700}) \tag{3}$$

since we are using the MFCC, filter banks are implemented for the frequency domain. Triangular filters are the filters commonly used in this field, which are Hamming or Hanning filters. To calculate the Mel spectrum, each of the triangular Mel weighting filters is multiplied by the magnitude spectrum, as shown in Eq4.

$$s(m) = \sum_{K=0}^{N-1} |X(k)|^2 H_m(k) \tag{4}$$

$$0 \leq m \leq M-1$$

Where:

- M is the total number of triangular Mel weighting filters.
- Hm(k) is the weight that depends on k and m.
- k is the energy spectrum.
- m is the output band.

5) Discrete cosine transform (DCT): the final step is to apply DCT on the transformed Mel frequency coefficients to produce a set of cepstral coefficients. So the final MFCC expression is calculated as:

$$c(n) = \sum_{m=0}^{M-1} log_{10}(s(m)) cos\left(\frac{\pi n(m-0.5)}{M}\right) \tag{5}$$

$$n = 0, 1, 2...; C-1$$

Where:

- c(n) are the cepstral coefficients.
- C is the number of MFCCs.

The MFCC is used For features extraction by transforming the data to a spectrogram based on time and frequency. In this case, MFCC computed 12 MFCC's over 1400 frames as it shown in figure 1.

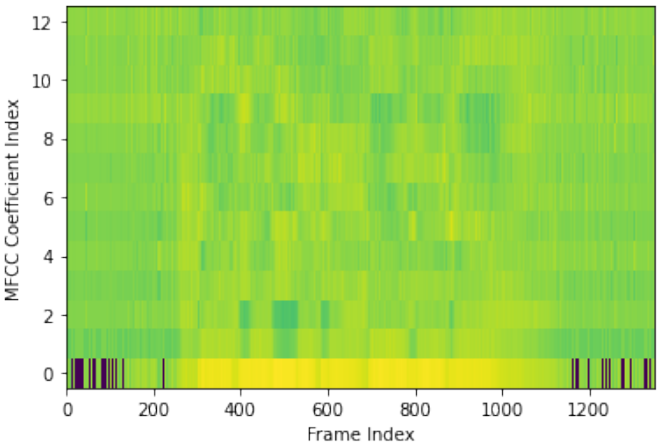

Fig. 1. The representation of the mfcc spectrogram for the "calm" emotion

B. Model Structure for the data

The architecture of the Conv1D-LSTM model based on the classification of layers in a structured form and using few tools to adjust the performance of the model. The architecture of the Conv1D-LSTM model for Song data depicted in fig2 shows that the inputs are fed to the first Conv1D layer with the parameters kernel size (5x256x256) and bias (256), then batch-Normalization by initializing gamma, betta, moving-mean and moving-variance parameters to (256). Next, there is two LSTM layers with the same parameters kernel size (256x1024), recurent kernel (256x1024) and bias (1024). The dropout comes after the LSTM layers, which is a technique used to overcome the problem of overfitting. The dropout rate can be chosen depending on the performance of the model.

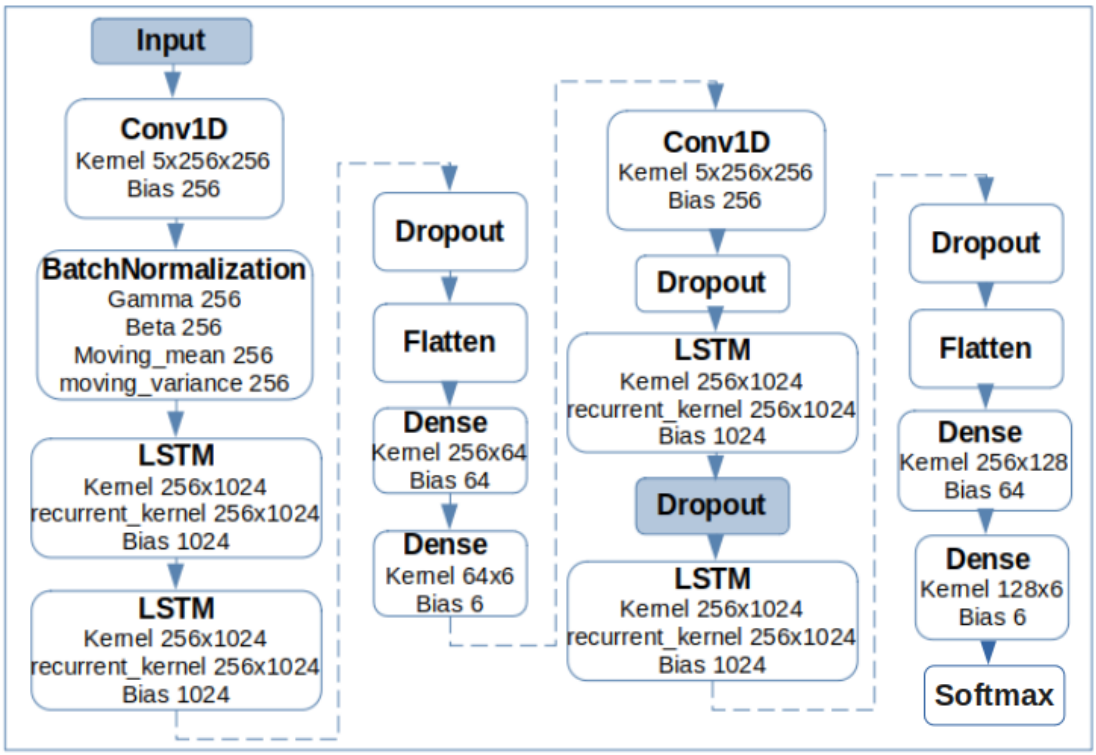

Fig. 2. The architecture of the Conv1D-LSTM model for Song and Speech data. For song, the entire presented layer is used, while for speech the only change is that the dropout between LSTM layers is removed

Then, Flatten is used to flatten the output to one vector. The main purpose of using two dense layers is to help in better classification. The output of the latter is passed to another group of layers which are Conv1D and two LSTMs, between each two layers there is a dropout. The parameters of each layer are the same as the layers before. The last step is to repeat the flatten and classification layers. The same model is used for the Speech data, only without need to the dropout between the LSTM layers as it shown in fig2. The purpose of removing the dropout in this particular layer is that the existence of the dropout confuses the model and renders the precision in a linear form. The architecture ends with a softmax activation function which is used to output the final classification results. The model was compiled using Adam's optimization technique [8] which is known for its computational efficiency and is best used for large data and a large number of parameters. The learning rate for the adam optimizer is 0.01 which is a hyper-parameter that adjusts the speed of the learning phase by setting the step size at each iteration in order to minimize the loss.

III. Results and Discussion

A. Database

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [16]. There exist 24 recorders for each forme of the data: audio speech, audio song ,video speech and video song. there are 12 actors and 12 actresses who perform in audio and visual recorders. There is six emotions expressions in the song format audio and video "Neutral, calm, happy, sad, angry and fearful", and eight emotions expressions in the audio speech and video speech "Neutral, calm, happy, sad, angry, fearful, surprise and disgust". The modality formats of the database are: Audio-only, Audio-Video, Video-only. Each modality format contain from 1 to 24 recorders, except that one file is missing for the Actor number 18 for the song, to be only 23 files. Since we are interested in emotion recognition

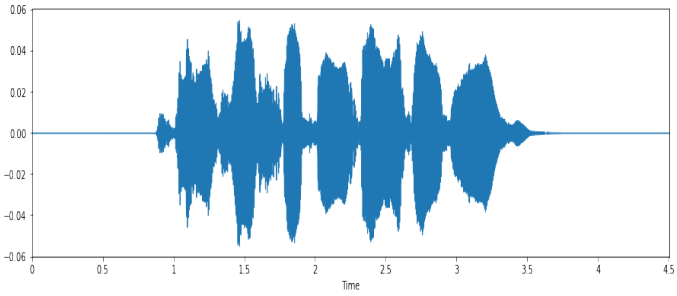

Fig. 3. Example of wave visualisation for "calm" emotion

from audio, we only used the first modality format from the database "Audio-only". The duration for each audio recorder does not exceed 5 seconds as shown in figure 3. The labels are extracted directly from an organized emotion files, each file contains the data belonging to an exact emotion, so that the label is the name of the file. The data is treated as frames with size [32,256]. The distribution of the labels for each emotion

for the song is 182 labels for ”happy, sad, angry and fearful” emotions and 138 labels for ”Neutral and calm”emotions, as it shown in fig4.

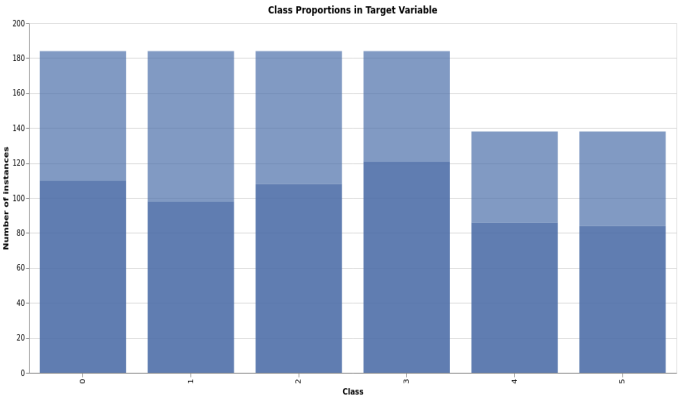

Fig. 4. Class proportion for Song data

The distribution of the labels for each emotion for the speech is 186 labels for ”happy, sad, angry and fearful” emotions and 142 labels for ”Neutral and calm”emotions, as it shown in fig5.

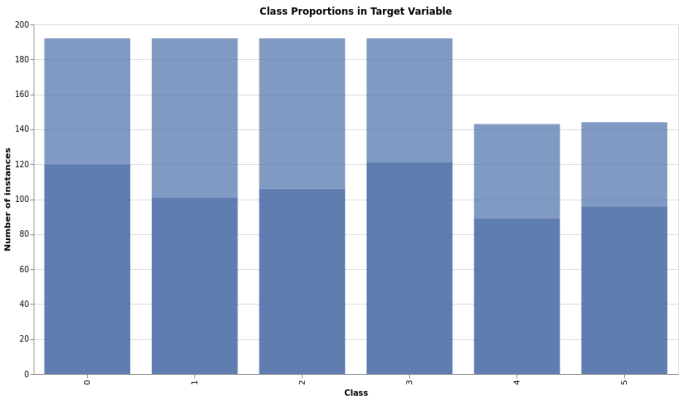

Fig. 5. Class proportion for Speech data

B. Results

The results are presented as a confusion matrix and as accuracy curve, as it shown in figures 6, 7, 8 and 9 for six emotions (”happy”, ”sad”, ”angree”, ”neutral”, ”calm” and ”fear”)for both song and speech. The achieved training accuracy result for song is 90.28% and the achieved validation accuracy result is 73.33%. While the accuracy results for speech are 64.14% and 53.32%, respectively for training and validation. To compare both results, for audio song, the model seems to show better results for the 15 epochs, as the epochs increses the model for the training phase starts to learn faster than the validation phase, but still keeps the performance stability. While for the audio speech, While for audio speech, the model appears to be more stable with equivalence of training and validation performance regarding the same progress steps.

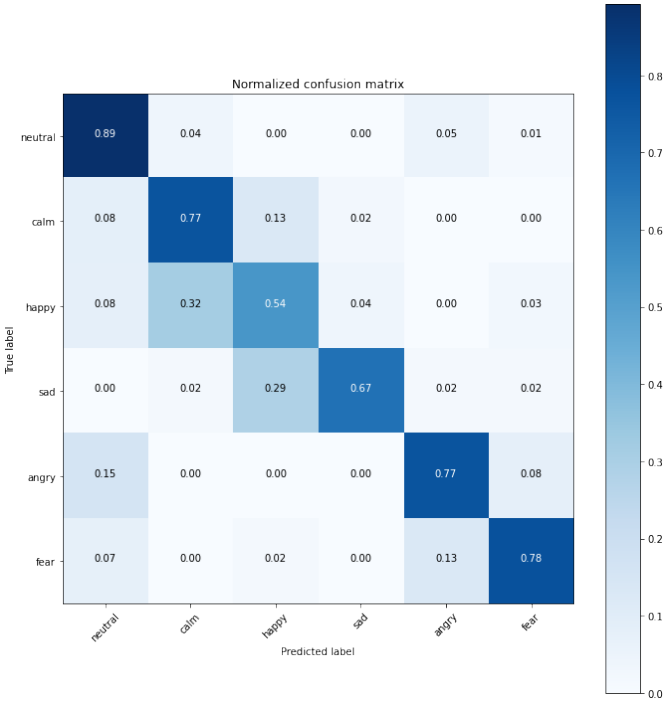

Fig. 6. Confusion Matrix for Song data

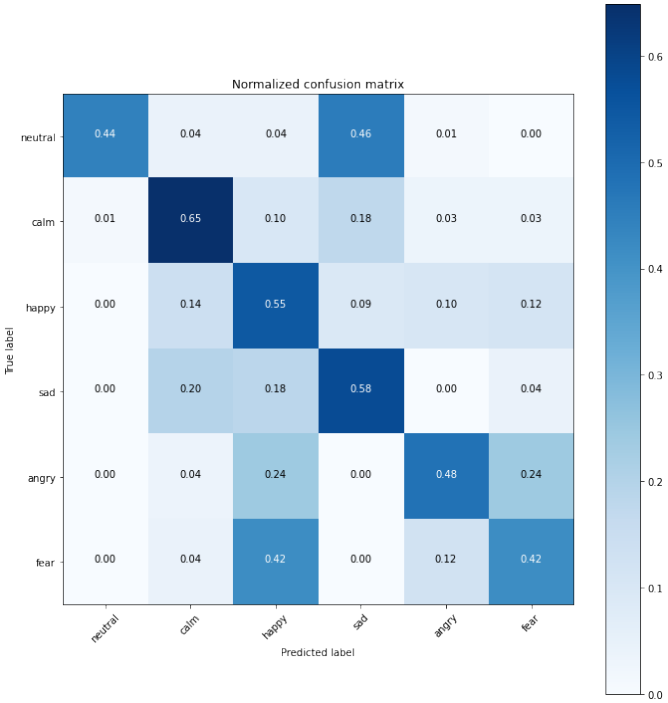

Fig. 7. Confusion Matrix for Speech data

C. Discussion

This work presents a Conv1D-LSTM architecture applied on two different modalities speech and song to compare the results and highlight the parameters used in this model. Our CNN is mostly encouraged by its ability of feature detection

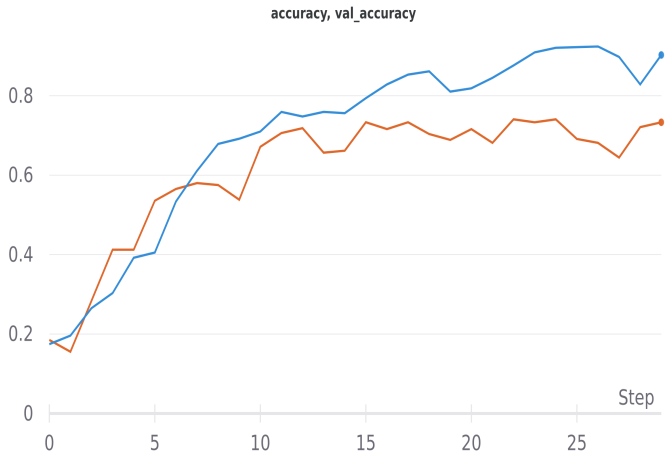

Fig. 8. Accuracy curves for Song data: The blue curve indicates the training accuracy and the orange curve indicates the test accuracy

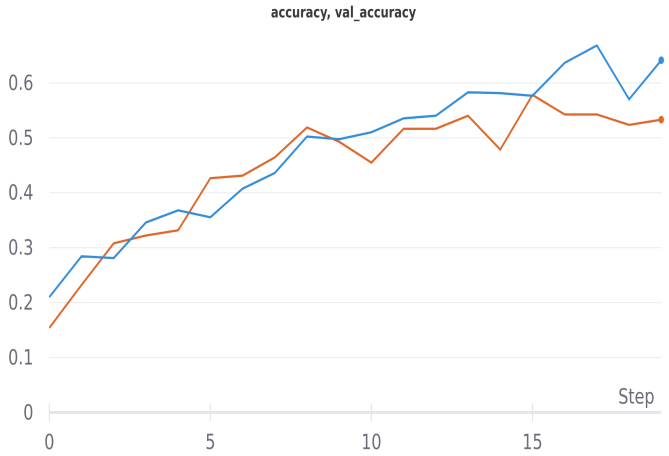

Fig. 9. Accuracy curves for Speech data: The blue curve indicates the training accuracy and the orange curve indicates the test accuracy

which mostly dedicated for visual tasks like image classification, recognition and tracking to achieve a better accuracy [21], and achieved a very good results by using conv1d for audio tasks like removing the noise by using a small filter size [10]. While the LSTM is mostly encouraged by its ability to memorize information from a long sentence [12].

Unlike different architectures was been applied like

TABLE I
SPEECH EMOTION RECOGNITION RESULTS REPORTED IN SEVERAL WORKS

	Speech accuracy results
Garg et al. [3](random forest)	40.27%
Han et al. [4](DNN-ELM)	54.3%
He et al. [6](GAN and VAE)	52.74%
Proposed Conv1D-LSTM model	53.32%

There are some architectures other than CNN and LSTM for speech emotion recognition like Garg et al. [3] who achieved 40.27% for random forest using MFCC, Han et al. [4] who achieved 54.3% using DNN and ELM model, and He et al. [6] who presented an unpaired semi-supervised data augmentation method based on GAN and VAE and achieved 52.74% by assigning labels to data. The proposed model achieved for speech emotion recognition 53.32% which is higher than the random forest method [3] by 13.05% and the method of GAN and VAE [6] by 0.58% . For song emotion recognition, the tasks are different depends on the music sounds and the influence of the music on the listener. Li [11] studied the music sound classification by considering the emotion detection problem as a multilabel classification problem by using SVM model, and the accuracy result is 50% for six main classes instead of thirteen. Wu et al. [23] achieved 49% in terms of F1 score by using HMER model. In this study, our premary goal is to compare between two audio format song and speech by suing the same model. We enhance the emotion classification performance by combining Convolutional neural networks and Long-Short Term memory (Conv1D-LSTM) to exploit the main characteristics of both models. The MFCC ability for feature extraction reinforces the idea of using frames instead of waves, which helps to improve the training process by reducing the time period to achieve a good result. Our model seems to achieve better results for song and a lower result for speech.

TABLE II
SONG EMOTION RECOGNITION RESULTS REPORTED IN SEVERAL WORKS

	Song accuracy results	F1 score
Li [11](SVM)	50%	
Wu et al. [23]		49%
Proposed Conv1D-LSTM model	73.33%	

IV. CONCLUSION

The goal is accompliched by combining both Conv1D and LSTM and create a new architecture conducted on RAVDESS database. The modality format of the database allow as to conduct the model on song and speech because of difference of delivering the same sentence with two different ways. The database treated as frames instead of waves format to facilate the training process and enhance the accuracy results. The structure of the model started by applying MFCC for feature extraction and feeding the input data to the new architecture. The main used tools for maintaining the stability of the model are dropout , learning rate and BatchNormalization. The achieved accuracy results are 73.33% for emotions extracted from audio song and 53.32% for emotions extracted from audio speech. The difference between the nature of the frequencies acts on the performance of the model which explain the difference between the achieved accuracy values. For feature work, we will work on combining Conv2D and LSTM.

REFERENCES

[1] Haytham M. Fayek, Margaret Lech, and Lawrence Cavedon. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68, 2017. Advances in Cognitive Engineering Using Neural Networks.

[2] Todor Ganchev, Nikos Fakotakis, and George Kokkinakis. Comparative evaluation of various mfcc implementations on the speaker verification task. In *Proceedings of the SPECOM*, volume 1, pages 191–194, 2005.

[3] Utkarsh Garg, Sachin Agarwal, Shubham Gupta, Ravi Dutt, and Dinesh Singh. Prediction of emotions from the audio speech signals using mfcc, mel and chroma. In *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 87–91. IEEE, 2020.

[4] Kun Han, Dong Yu, and Ivan Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech 2014*, 2014.

[5] Md Rashidul Hasan, Mustafa Jamil, MGRMS Rahman, et al. Speaker identification using mel frequency cepstral coefficients. *variations*, 1(4):565–568, 2004.

[6] Gewen He, Xiaofeng Liu, Fangfang Fan, and Jane You. Image2audio: Facilitating semi-supervised audio emotion recognition with facial expression image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 912–913, 2020.

[7] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proc. ismir*, volume 86, pages 937–952, 2010.

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[9] Shashidhar G Koolagudi, Deepika Rastogi, and K Sreenivasa Rao. Identification of language using mel-frequency cepstral coefficients (mfcc). *Procedia Engineering*, 38:3391–3398, 2012.

[10] Soonil Kwon et al. A cnn-assisted enhanced audio signal processing for speech emotion recognition. *Sensors*, 20(1):183, 2020.

[11] Tao Li and Mitsunori Ogihara. Detecting emotion in music. 2003.

[12] Xiangang Li and Xihong Wu. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. *CoRR*, abs/1410.4281, 2014.

[13] Yuanchao Li, Tianyu Zhao, and Tatsuya Kawahara. Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In *Interspeech*, pages 2803–2807, 2019.

[14] Xin Liu, Qingcai Chen, Xiangping Wu, Yan Liu, and Yang Liu. Cnn based music emotion classification. *arXiv preprint arXiv:1704.05665*, 2017.

[15] Zhen-Tao Liu, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, and Guan-Zheng Tan. Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing*, 273:271–280, 2018.

[16] Steven R. Livingstone and Frank A. Russo. 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):1–35, 05 2018.

[17] Miroslav Malik, Sharath Adavanne, Konstantinos Drossos, Tuomas Virtanen, Dasa Ticha, and Roman Jarina. Stacked convolutional and recurrent neural networks for music emotion recognition. *CoRR*, abs/1706.02292, 2017.

[18] Meinard Müller. *Information retrieval for music and motion*, volume 2. Springer, 2007.

[19] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. Speech emotion recognition using hidden markov models. *Speech Communication*, 41(4):603–623, 2003.

[20] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] Nattapong Thammasan, Koichi Moriyama, Ken-ichi Fukui, and Masayuki Numao. Continuous music-emotion recognition based on electroencephalogram. *IEICE TRANSACTIONS on Information and Systems*, 99(4):1234–1241, 2016.

[23] Bin Wu, Erheng Zhong, Andrew Horner, and Qiang Yang. Music emotion recognition by multi-label multi-layer multi-instance multi-view learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 117–126, 2014.

[24] Zengwei Yao, Zihao Wang, Weihuang Liu, Yaqian Liu, and Jiahui Pan. Speech emotion recognition using fusion of three multi-task learning-based classifiers: Hsf-dnn, ms-cnn and lld-rnn. *Speech Communication*, 120:11–19, 2020.