1. Souha AYADI[1], 2. Zied LACHIRI[2]

University of Tunis el Manar,Tunisia (1)(2), Signal Image and Information Technology Laboratory , SITI,
National Engineering school of Tunis

doi:10.15199/48.2024.07.12

Audio Emotion Recognition based on Song modality using Conv1D vs Conv2D

***Abstract**. Audio emotion recognition is a very advanced process of detecting emotions from different forms of signals. The form of modality presented in this article is Audio-Song. The goal is to create different neural network architectures capable of recognizing the emotions of a song performer. The database used for this purpose is the RAVDESS database. We compared the performance of Conv1D with Conv2D, where MFCC is used for the feature extractor for both neural network architectures. The accuracies obtained are 83.95 and 82.47% respectively. The better of the two models is Conv1D regarding the accuracy result obtained and the complexity of the model, where it seems that the Conv1D model is less complex than the Conv2D model.*

***Streszczenie.** Rozpoznawanie emocji dźwiękowych to bardzo zaawansowany proces wykrywania emocji na podstawie różnych form sygnałów. Formą modalności przedstawioną w tym artykule jest utwór audio. Celem jest stworzenie różnych architektur sieci neuronowych zdolnych do rozpoznawania emocji wykonawcy utworu. Bazą danych wykorzystywaną w tym celu jest baza danych RAVDESS. Porównaliśmy wydajność Conv1D z Conv2D, gdzie MFCC jest używane do ekstraktora cech dla obu architektur sieci neuronowych. Uzyskane dokładności wynoszą odpowiednio 83,95 i 82,47%. Lepszym z obu modeli jest Conv1D pod względem uzyskanego wyniku dokładności i złożoności modelu, gdzie wydaje się, że model Conv1D jest mniej złożony niż model Conv2D.(**Rozpoznawanie emocji dźwiękowych w oparciu o modalność utworu przy użyciu Conv1D i Conv2D**)*

Keywords: Song emotion recognition, Conv1D, Conv2D.
Słowa kluczowe: Rozpoznawanie emocji w utworze, Conv1D, Conv2D.

Introduction

Automatic emotion recognition is the process of detecting and identifying human emotions. The human ability to express different emotions and recognize them while interacting with other human beings has always been of interest to the field of research. where can always bring new ideas and new perspectives. Especially in the way humans can express themselves, where there are two main ways: facial expressions that appear on the face and audio signals that transmit frequencies contain different information. In this article, we are interested in the audio modality. Where audio can also be rich with different types of forms, such as speech [10], music [9] and acoustics [2]. And also, each form can be studied differently from a particular point of view, depending on the problem to solve. In our case, the form of choice is Audio-Song, which means detecting and classifying emotions when performing a song. This idea is inspired by the RAVDESS database which contains a song type database. Advanced algorithms are created to automatically detect and recognize emotions to enable effective human-machine interaction.

The best known algorithms are the convolutional neural network [14] and the recurrent neural network [16]. Based on these two neural networks, many different architectures are created in order to improve the ability to correctly recognize emotions. Such as DCNN [14] which is basically CNN [13] with higher number of layers. In addition, LSTM [15], which is a special type, belongs to RNN and is characterized by the ability to predict the last information of a long sentence. Moreover, a combination between CNN and LSTM architectures to create a robust model based on the best features of both [3].

In this work, we were interested in creating a deep convolutional architecture based on Conv1D [11] and Conv2D [7] separately, to be able to compare them. However, any neural network algorithm still requires human intervention by labeling [5] the data, which gives the ability to process and learn from huge amounts of data, giving it a distinct advantage for higher performance.

Since feature extraction is a very delicate process that must be performed correctly before passing through the network, MFCC [6] is our method of choice for both architectures as best used for audio processing based on the known variation of bandwidth frequencies of the human ear. This work is divided into two main sections. The first section describes the architecture of each neural network model by highlighting the tools that will be used inside each model. The second section presents the results obtained in different forms, most likely the results of the accuracy and confusion matrix. At the end, a constructive comparison between the presented architectures will be discussed.

Feature extraction

Mel Frequency Cepstral Coefficient (MFCC) [6] is a very well-known method used for audio feature extraction. It is characterized by the same frequency bandwidth as humans and, more specifically, by the way humans distinguish frequencies. Making it the method of choice for audio tasks [8]. This capability is called MEL scaling. The latter works by dividing the frequency band into sub-bands. Then, apply the discrete cosine transform (DCT) to extract the cepstral coefficients.

MFCC is performed via Librosa, which is a simplified method of extracting MFCC for a certain number of frames by providing different arguments, such as the number of MFCCs and the length of the data. So, to transform waveform data into MFCC, the number of MFCC is set to 32 and the maximum data length is 256, with a 16000 number of samples.

Model description

The architectures that will be presented in this section are based on a convolutional neural network. where in the latter different models are included and could be applied depending on the task demand. In this case, we chose to apply Conv1D and Conv2D separately. Due to the nature of the database, which is an audio song, it could be processed either by one or two dimensional convolution. by applying both, we could observe the performance of each and compare between each of them.

Conv1D model

The architecture of Con1D is built on the basis of traditional layers, as shown in table I. The presented Con1D model is based on two neural network architectures separated and placed in sequential form. The idea behind building two models successively is to use the output of the first architecture as input to the second architecture to correct the behavior of the first and save time and speed up the training process.

Table 1. The parameters of the sensor

Name	Type	Output Shape
conv1d_input	Input layer	32,256,1
conv1d	Conv1D	32,256,64
activation	Activation	32,256,64
dropout	Dropout	32,256,64
flatten_1	Flatten	524288
dense_1	Dense	6
activation_1	Activation	6
conv1d_1_input	Input layer	32,256,1
conv1d	Conv1D	32,256,64
dropout_1	Dropout	32,256,64
flatten_2	Flatten	524288
dense_2	Dense	128
dropout_2	Dropout	128
dense_3	Dense	6

Table 1 contains the name of each layer, its type and itsshape. The first conv1D architecture starts with the input of the conv1d layer and takes the size (32, 256,1), which indicates the backets configuration, input configuration and number of channels, respectively. The shape of each layer changes based on receiving the output of each previous layer. The second layer is a ReLU activation function layer in introduce the linearity to the system. The third layer is a dropout layer which is used to save the system from collapsing by ignoring certain subsets of nodes. The output will be flattened into vectors before going through classification. The final two-layer, dense and softmax activation function, are used to display the classification results. Moving to the next layer block, the same process will be repeated, removing only the ReLU activation function and replacing the final softmax activation function with a dense layer to strengthen the output classification results.

Conv2D model

The same process is used to create the Conv2D model as described in Conv1D. Only the Conv2D model is more complicated and deeper than the Conv1D model, as shown in Table II, due to the different parameters needed to be used in this architecture and the accompanying layers that must support the internal Conv2D layers. The first Conv2D architecture starts with three Conv2D layers as well as the maxpooling2D layer. Which always assemble in two complementary layers.

A dropout layer is then replaced to facilitate the calculation and reduce the complexity of the system. Then flatten the output to a single vector. the fully connected layer consists of a dense layer, a dropout and another dense layer to improve the first classification before going through the next Conv2D architecture. The second part of the architecture uses less layer than the first part, where we only use a single Conv2D layer with the maxpooling2D layer. For the rest, we kept the same structure for the fully connected layer.

Results and discussions

A confusion matrix [1] is a more perceptible mode of evaluation that provides more information about the model's performance. Because of its organized way of mapping predictions to the original classes. Where the accuracy rate ofeach class can be displayed. Another important method that allows us to visualize the overall performance of a classifier is to plot the ROC curve [4].

Database

The Ryerson Audiovisual Database of Emotional Speech and Songs (RAVDESS) [12] dataset contains 7,356 files. It has two main modalities: Speech and Song. For each modality, there are three categories: Audio only in the form of sound recorder, Video only in mp4 format, Audio-Video combines both sound and visual recorder. To develop our work, we are only interested in using the Audio-Song category belonging to the Song modality. The number of classes detected is six which are: neutral, calm, happy, sad, angry, fear.

Table 2. The parameters of the sensor

Name	Type	Output Shape
conv2d input	Input layer	32,256,1
conv2d	Conv2D	32,254,256
max_pooling2d	Maxpooling2D	15,127256
conv2d_1	Conv2D	13,125,256
max_pooling2d_1	Maxpooling2D	6,62,256
conv2d_2	Conv2D	4,60,256
max_pooling2d_2	Maxpooling2D	2,30,256
dropout	Dropout	2,30,256
flatten_1	Flatten	15360
dense_1	Dense	128
dropout_1	Dropout	128
dense_2	Dense	6
conv2d_3_input	Input layer	32,256,1
conv2d_3	Conv2D	32,256,256
max_pooling2d_3	Maxpooling	16,128,256
dropout_2	Dropout	16,128,256
flatten_2	Flatten	524288
dense_3	Dense	128
dropout_3	Dropout	128
dense_4	Dense	6

Conv1D results

The results obtained after testing the presented Conv1D model are presented in figures 1 and 2. Where the accuracy results obtained for neutral, calm, happy, sad, angry, fear are 89%, 93%, 84%, 68%, 81%, and 83% respectively.

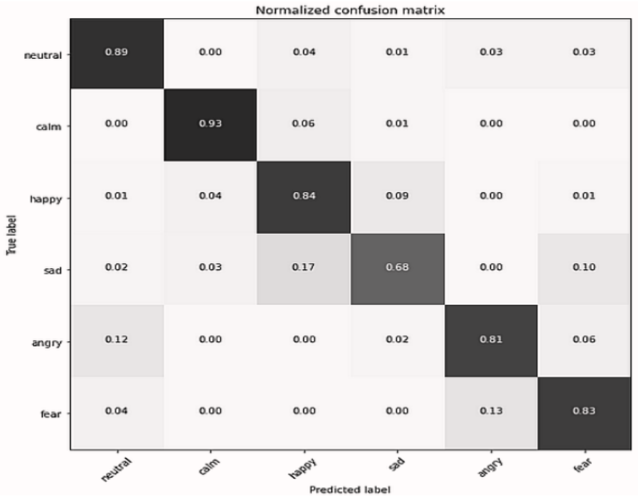

Fig.1. Confusion Matrix of Conv1D model

In this presented case, the ROC curve shows the true positive rate of training accuracy, validation accuracy, training loss, and validation loss, as shown in Figure 2. Where the results are 93.57%, 83.95%, 29.91% and 47.92% respectively. The shape of the curve shown in figure 2 shows the performance stability of the neural network model where the training accuracy and validation accuracy increase together at the same time.

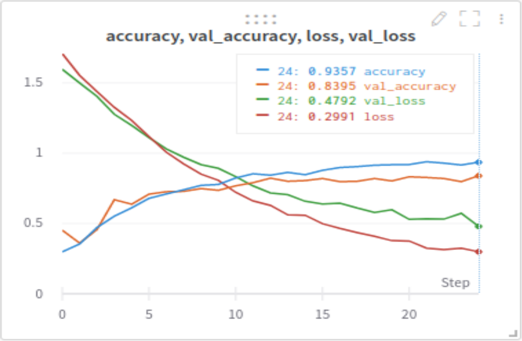

Fig.2. Visualization of the accuracy results of Conv1D Model

Conv2D results

The achieved results of the presented Conv2D model are shown in figures 3 and 4. Where the obtained accuracy results for neutral, calm, happy, sad, angry, fear are 97%, 92%, 62%, 83%, 73%, and 85% respectively.

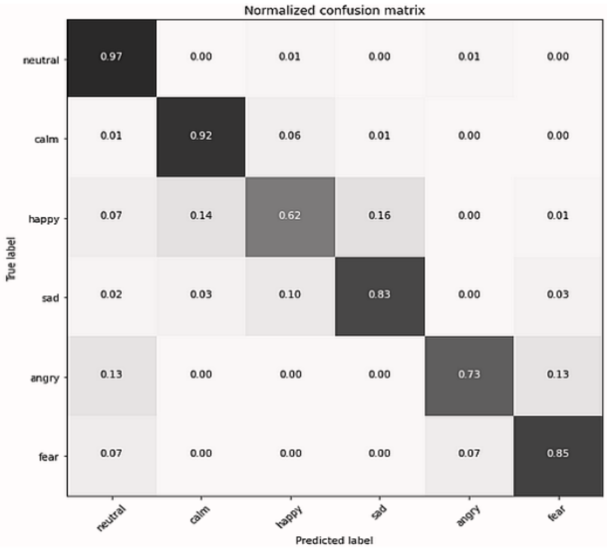

Fig.3. Confusion Matrix of Conv2D Model

The visualization of the accuracy results of the Conv2D model is shown in figure 4. Where the evolution of precision over 25 epochs shows stability in performance. This also shows that there is no such problem as overfitting or underfitting. Where the training accuracy , the validation accuracy, the training loss and the validation loss are 92.42%, 82.47%, 49.04% and 28.57% respectively.

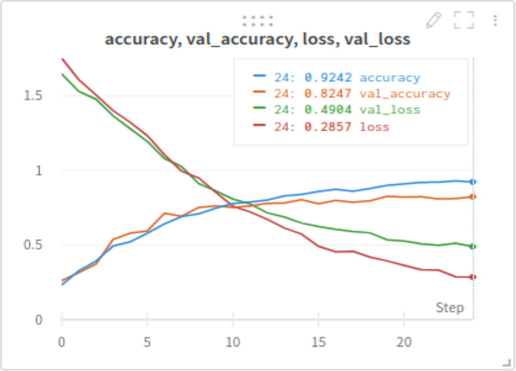

Fig. 4. Visualization of the accuracy results of Conv2D Model

Comparing the results of Conv1D and Conv2D, both models manage to achieve good results in terms of accuracy and performance stability. Where the accuracy result for Conv1D is 83.95% and the accuracy result for Conv2D is 82.47%. The difference is that the first model achieved a higher accuracy than the second model by 1.48%. Which is considered a very close achievement between the two. The complexity of the model has a very important role in these results, comparing the construction of the two models, the second model is deeper regarding the number of layers and the intervention of the parameters used. It can be said that a simple model like Conv1D can obtain a better result than a deeper model like Conv2D.

Conclusion

A very particular type of modality presented in this work which is the Song audio type. Where the extracted emotions come from the RAVDESS database. The feature extractor used is MFCC, which is the most well-known feature extractor for audio tasks because it has the same variation of the bandwidth frequencies as human hearing. The presented neural network architectures are based on a convolutional neural network, where we have built two different models conv1D and conv2D. And worked on using different parameters to adapt each model to the database and ensure good performance. the accuracy results obtained for the two models are 83.95% and 82.47%, respectively. By comparing the presented models, we can conclude that a simpler architecture of the Conv1D model can achieve a better result than a complicated and deeper architecture of the Conv2D model.

Authors: *Souha AYADI, Signal Image and Information Technology(SITI) Laboratory, Department of Electrical Engineering, National Engineering School of Tunis, Campus Universitaire Farhat Hached el Manar BP 37, Le Belvedere 1002 TUNIS, E-mail: souha.ayadi@enit.utm.tn; Zied LACHIRI, Signal Image and Information Technology(SITI) Laboratory, Department of Electrical Engineering,National Engineering School of Tunis, Campus Universitaire Farhat Hached el Manar BP 37, Le Belvedere 1002 TUNIS, E-mail: zied.lachiri@enit.utm.tn .*

REFERENCES

[1] Wejdan Ibrahim AlSurayyi, Norah Saleh Alghamdi,and Ajith Abraham. Deep learning with word embedding modeling for a sentiment analysis of online reviews. International Journal of Computer Information Systems and Industrial Management Applications, 11:227–241, 2019.

[2] Bagus Tris Atmaja, Akira Sasou, and Masato Akagi. Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. Speech Communication, 140:11–28, 2022.

[3] Souha Ayadi and Zied Lachiri. A combined cnn-lstm network for audio emotion recognition using speech and song attributs. In 2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), pages 1–6. IEEE, 2022.

[4] Subhajit Chatterjee and Yung-Cheol Byun. Eeg-based emotion classification using stacking ensemble approach. Sensors, 22(21):8550, 2022.

[5] Stuart Cunningham, Harrison Ridley, Jonathan Weinel, and Richard Picking. Supervised machine learning for audio emotion recognition: Enhancing film sound design using audio features, regression models and artificial neural networks. Personal and Ubiquitous Computing, 25:637–650, 2021.

[6] Harshit Dolka, Arul Xavier VM, and Sujitha Juliet. Speech emotion recognition using ann on mfcc features. In 2021 3rd international conference on signal processing and communication (ICPSC), pages 431–435. IEEE, 2021.

[7] Pooja Gambhir, Amita Dev, Poonam Bansal, and Deepak Kumar Sharma. End-to-end multi-modal low-resourced speech keywords recognition using sequential conv2d nets. ACM

Transactions on Asian and Low-Resource Language Information Processing, 2023.

[8] 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.

[9] Donghong Han, Yanru Kong, Jiayi Han, and Guoren Wang. A survey of music emotion recognition. Frontiers of Computer Science, 16(6):166335, 2022.

[10] C Hema and Fausto Pedro Garcia Marquez. Emotional speech recognition using cnn and deep learning techniques. Applied Acoustics,211:109492, 2023.

[11] S Jothimani and K Premalatha. Mff-saug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network. Chaos, Solitons & Fractals, 162:112512, 2022.

[12] 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):e0196391, 2018.

[13] Zixuan Peng, Yu Lu, Shengfeng Pan, and Yunfeng Liu. Efficient speech emotion recognition using multi-scale cnn and attention. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3020–3024. IEEE, 2021.

[14] R Raja Subramanian, Yalla Sireesha, Yalla Satya Praveen Kumar Reddy, Tavva Bindamrutha, Mekala Harika, and R Raja Sudharsan. Audio emotion recognition by deep neural networks and machine learning algorithms. In 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), pages 1–6. IEEE, 2021.

[15] Jianyou Wang, Michael Xue, Ryan Culhane, Enmao Diao, Jie Ding, and Vahid Tarokh. Speech emotion recognition with dual-sequence lstm architecture. In ICASSP 2020-2020 IEEE International Conference on
Acoustics, Speech and Signal Processing (ICASSP), pages 6474–6478.IEEE, 2020.

[16] Satya Prakash Yadav, Subiya Zaidi, Annu Mishra, and Vibhash Yadav. Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (rnn). Archives of Computa- tional Methods in Engineering, 29(3):1753–1770, 2022.