**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:10.15199/48.2024.06.43

# Speech Emotion Recognition using Dual-Conv2D architecture

***Abstract**. The ability to convey emotions through speech is still of interest to the field of research. Where different neural network architectures have been developed to be able to automatically recognize emotions. In this work, the main objective is to develop an accurate neural network architecture for Speech emotion recognition. The work includes two main parts which essentially concern the use of MFCC as a feature extractor. And present a new technique for creating a CNN architecture based on the use of two separate architectures based on the Conv2D model.*

***Streszczenie.** Możliwość przekazywania emocji za pomocą mowy jest nadal przedmiotem zainteresowania badaczy. Gdzie opracowano różne architektury sieci neuronowych, aby móc automatycznie rozpoznawać emocje. Głównym celem tej pracy jest opracowanie dokładnej architektury sieci neuronowej do rozpoznawania emocji związanych z mową. Praca składa się z dwóch głównych części, które zasadniczo dotyczą wykorzystania MFCC jako ekstraktora cech. Oraz przedstawić nową technikę tworzenia architektury CNN opartą na wykorzystaniu dwóch odrębnych architektur bazujących na modelu Conv2D. (**Rozpoznawanie emocji mowy przy użyciu architektury Dual-Conv2D**)*

**Keywords:** Speech emotion recognition, MFCC, Conv2D.
**Słowa kluczowe:** Rozpoznawanie emocji mowy, MFCC, Conv2D.

## Introduction

Automatic emotion recognition is widely discussed in the research field, where different viewpoints are presented specially when working on speech for feature extraction and emotion classification. Speech is widely observed, especially when it can be affected by human emotions, which can profoundly affect the transformation of information contained in sentences. Artificial intelligence is more specifically interested in the study of the human capacity to produce emotions with the aim of creating a mechanical system capable of detecting and recognizing emotions and also reaching the point where human emotions can be imitated. Different Neural Networks have been used and developed over the last decade to solve different problems. The most famous Neural Networks are the Convolutional Neural Network [2] and the Recurrent Neural Network [12]. The choice between the two indicated by the advantage and ability of each. Where CNN is mainly used for spatial data where the feature can be detected from stable features, such as images. And RNN [10] is mainly used for sequential data, such as audio, because of its ability to recognize the last word based on the previous one. Different dimensions can classify the type of convolutional neural network. The most suitable audio modality is one-dimensional convolution (Conv1D) and two-dimensional convolution (Conv2D). Where several researchers have used these two types of neural networks to solve different problems. For example, Kwon et al.[5] used Conv1D to enhance the speech signal and focus on extracting hidden patterns. Furthermore, Prombut et al. [9] compared Conv1D and Conv2D and proved that the accuracy of Conv2D is better than the accuracy of Conv1D. The idea that will be presented in this article uses an unusual type of neural network which is Conv2d for Speech emotion recognition. Our goal is to frame each wave and pass it through dual-Conv2D architecture.

The work is structured around three main parts. The first part reviews the preparation of data that requires careful processing before passing through the network and highlights the feature extraction method that serves the main idea. The second part gives the reason for choosing a specific neural network architecture, highlights the adaptation method, and focuses on the construction of the architecture.

## Model structure

The development of this work goes through two main parts: first, feature extraction. Second, building a neural network architecture.

## Feature extractor

Mel frequency cepstral coefficient (MFCC) [3] [7] is widely used for speech processing. The reason behind preferring using MFCC is considered closest to the human hearing system and responds better than the linearly spaced frequency bands used in the normal spectrum. Technically, MFCC is used to represent the spectral characteristics of audio to adapt to different machine learning tasks, in our case for speech emotion recognition. The wave signal will be structured in 25 ms frames as standard value. So the frame length for a 16 kHz signal will be 0.025*16,000 = 400 samples. Every 10ms, 160 samples will be counted starting from 0 the first 400 samples. For the next 400 samples, this will start from 160 samples, and so on until the end of the speech file. The main role is therefore that the number of images is divided equally.
For each frame a set of 12 MFCC coefficients is extracted for each 1200 frames.

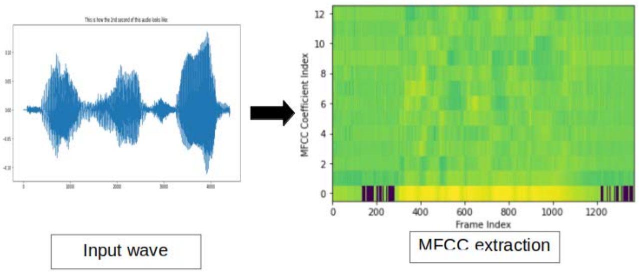

Fig.1. Input wave goes through MFCC for feature extraction

## Neural Network architecture

The architecture of the neural network is based on a convolutional neural network, where we have used a specific type of convolution which is Conv2D. The number of layers and their positions are chosen after evaluating the performance together and observing their performance. To begin, the input will be sent to the first Conv2D convolutional layer. The kernel size of the latter is (3,3,1) and the bias value is 256. The second layer is a max pooling layer which is a suitable layer to always use after the conv2D layer.Two other Conv2D layers contained a kernel with size (3,3,256) and bias count of 256 as well as a Max pooling layer. A dropout was then placed to reduce the complexity of the neural network and make the calculation easier. Then the output will be flattened and transformed into a single vector to go through the final stage of classification. The classification layer is a softmax layer composed of two dense layers separated by another dropout layer to facilitate the transaction of information between the layers responsible for decision making. Sometimes complicating the neural network is not a solution for better performance. So we create another Conv2D architecture that takes the output of the previous neural network and trains it in a way that removes all ambiguities and strengthens the system for better classification. The second architecture consists of a Conv2D layer with a kernel of size (3,3,1) and bias number 256, as well as a maxpooling layer. Ended by a dropout layer to reduce the complexity of calculations and reduce errors as much as possible. The output will be flattened to pass through two density layers separated by a dropout layer, as for the first architecture.

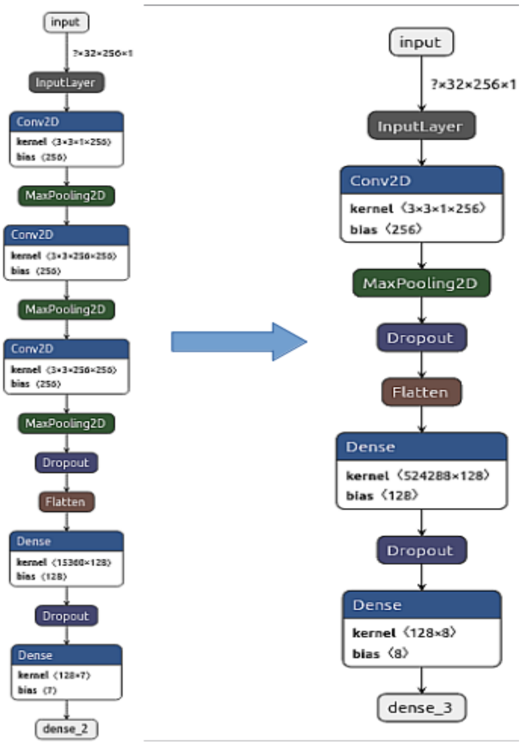

Fig.2. Dual-Conv2D architecture

## Results and discussion

### Preprocessing of the Database

The Ryerson Audio-Visual Database of Emotional Speech and Songs (RAVDESS) is a database created for the purpose of emotion recognition studies. Different modalities belong to this database which are audio and visual. Where contains 24 professional actors (12 female, 12 male) performing a specific phrase ”kids are playing at the door” in two different ways, in song and in speech. The number of emotions belonging to speech is eight: calm, happy, sad, angry, fearful, surprised and disgusted expressions. And the number of emotions belonging to Song is six: calm, happy, sad, angry, and fearful emotions. This database contains three modality formats: audio only as waveform, audio-video as mp4, and video only as video without sound. Actor number 18’s recorder is missing the number of files. The modality form used in this work is audio only, more specifically Speech form. Preparing the database consists of separating the wave recorders belonging to each emotion and grouping them into different folders labeled by the name of the emotion. So that the training process is supervised. The number of classes used in the Audio-Speech part of the database is seven: neutral, happy, sad, angry, fearful, surprised and disgusted. We concatenate calm emotion with neutral emotion because of the similarity between them.

## Results and discussions

The results are presented in three forms: ROC curve [6], confusion matrix [13] and accuracy [1] [11]. Where The ROC curve [4] shows the performance of the neural network model for the number of existing classes based on two parameters: the true positive rate and the false positive rate.

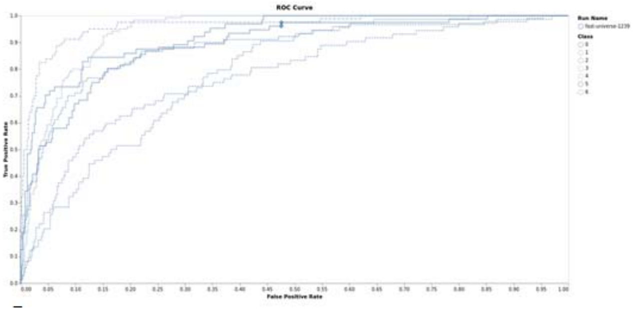

Fig.3. ROC curve of seven classes belonging to Audio-Speech in RAVDESS database

And the confusion matrix [8] shows the prediction results which indicate the actual rate of a detected class in matrix form. Where each row of the matrix represents the occurrence in an actual class while each column represents the occurrences in a predicted class.

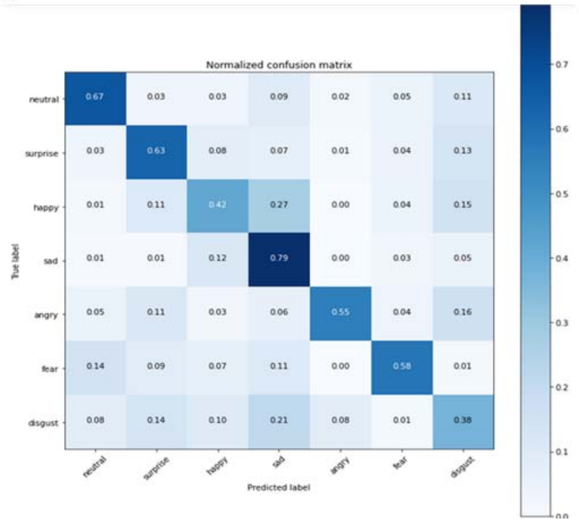

Fig.4. Confusion Matrix

In addition to the confusion matrix, figure 4 shows the performance of conv2D over a total number of 25 epochs, where the accuracy obtained is 59.2%. It seems that the system has a little difficulty correctly detecting the emotions of happiness and disgust, based on figure 3, and shows a lower accuracy rate, which affects the total accuracy rate.

We compare the results with our previous work [4], where we combined Conv1D and LSTM to obtain 53.32% for speech emotion recognition. The accuracy of the results of this work is better by 5.88%. Furthermore, comparing our obtained results with other studies, in which Pham et al. [10] obtained 69.4% accuracy by applying convolutional neural network on RAVDESS database. Also for the same database, Asiya et al. [3] worked on improving the accuracy using data augmentation and achieved 68%.

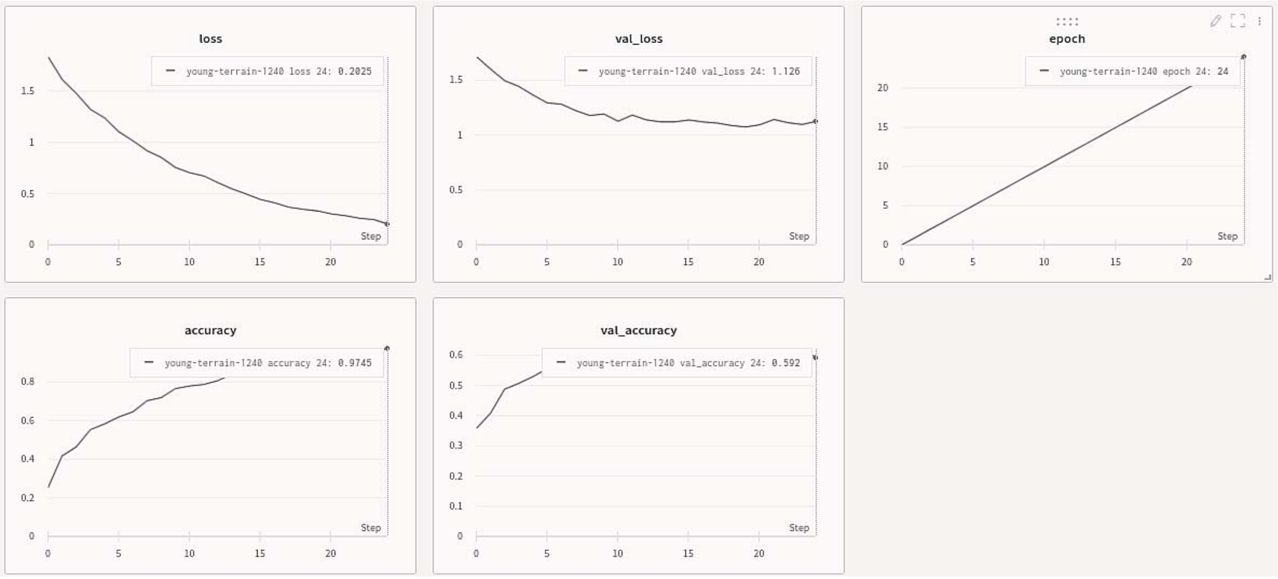

Fig.5. Accuracy curve of seven classes belonging to Audio-Speech in RAVDESS database

## Conclusion

In this paper, we presented a Speech emotion recognition method including collaboration between MFCC for feature extraction and Conv2D architecture to produce reasonable outcomes. This study explored a deep learning paradigm to receive sound from waves and extract features from wave signals using MFCC. The neural network architecture built in this work is based on the use of a Dual-Conv2D. Which means creating two Conv2D successively based on receiving the output from the previous neural network. This presented idea helps reduce the complexity of the architecture, avoids the merging of system shapes, and avoids the overfitting problem. For future work, we carrying to find the best method and focus on improving the accuracy rate.

**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] Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby. Deep learning techniques for speech emotion recognition, from databases to models. Sensors, 21(4):1249, 2021.

[2] Tursunov Anvarjon, Mustaqeem, and Soonil Kwon. Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features. Sensors, 20(18):5212, 2020.

[3] UA Asiya and VK Kiran. Speech emotion recognition-a deep learning approach. In 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), pages 867–871. IEEE, 2021.

[4] 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, 2022.

[5] Manas Jain, Shruthi Narayan, Pratibha Balaji, Abhijit Bhowmick, Rajesh Kumar Muthu, et al. Speech emotion recognition using support vector machine. arXiv preprint arXiv:2002.07590, 2020.

[6] Kittisak Jermsittiparsert, Abdurrahman Abdurrahman, Parinya Siriat takul, Ludmila A Sundeeva, Wahidah Hashim, Robbi Rahim, and Andino Maseleno. Pattern recognition and features selection for speech emotion recognition model using deep learning. International Journal of Speech Technology, 23:799–806, 2020.

[7] Soonil Kwon et al. 1d-cnn: Speech emotion recognition system using a stacked network with dilated cnn features. Computers, Materials & Continua, 67(3), 2021.

[8] Manuel Milling, Alice Baird, Katrin D Bartl-Pokorny, Shuo Liu, Alyssa M Alcorn, Jie Shen, Teresa Tavassoli, Eloise Ainger, Elizabeth Pellicano, Maja Pantic, et al. Evaluating the impact of voice activity detection on speech emotion recognition for autistic children. Frontiers in Computer Science, 4:837269, 2022.

[9] Suprava Patnaik. Speech emotion recognition by using complex mfcc and deep sequential model. Multimedia Tools and Applications, 82(8):11897–11922, 2023.

[10] Minh H Pham, Farzan M Noori, and Jim Torresen. Emotion recognition using speech data with convolutional neural network. In 2021 IEEE 2$^{nd}$ International Conference on Signal, Control and Communication (SCC), pages 182–187. IEEE, 2021.

[11] VM Praseetha and PP Joby. Speech emotion recognition using data augmentation. International Journal of Speech Technology, 25(4):783–792, 2022.

[12] Naris Prombut, Sajjaporn Waijanya, and Nuttachot Promrit. Feature extraction technique based on conv1d and conv2d network for thai speech emotion recognition. In 2021 5th International Conference on Natural Language Processing and Information Retrieval (NLPIR), pages 54–60, 2021.

[13] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. A comprehensive review of speech emotion recognition systems. IEEE access, 9:47795–47814, 2021.