MEKATRONIKA
e-ISSN: 2637-0883
VOL. 4, ISSUE 1, 88 – 95
DOI: https://doi.org/10.15282/mekatronika.v4i1.8626

ORIGINAL ARTICLE

Development of Audio-Visual Speech Recognition using Deep-Learning Technique

How Chun Kit[1], Ismail Mohd Khairuddin[1,*], Mohd Azraai Mohd Razman[1], Anwar P.P. Abdul Majeed[1] and Wan Hasbullah Mohd Isa[1]

[1]Faculty of Manufacturing and Mechatronic Engineering Technology, Universiti Malaysia Pahang, 26600 Pahang, Malaysia.

ABSTRACT – Deep learning is a technique with artificial intelligent (AI) that simulate humans' learning behavior. Audio-visual speech recognition is important for the listener understand the emotions behind the spoken words truly. In this thesis, two different deep learning models, Convolutional Neural Network (CNN) and Deep Neural Network (DNN), were developed to recognize the speech's emotion from the dataset. Pytorch framework with torchaudio library was used. Both models were given the same training, validation, testing, and augmented datasets. The training will be stopped when the training loop reaches ten epochs, or the validation loss function does not improve for five epochs. At the end, the highest accuracy and lowest loss function of CNN model in the training dataset are 76.50% and 0.006029 respectively, meanwhile the DNN model achieved 75.42% and 0.086643 respectively. Both models were evaluated using confusion matrix. In conclusion, CNN model has higher performance than DNN model, but needs to improvise as the accuracy of testing dataset is low and the loss function is high.

ARTICLE HISTORY
Received: 6th May 2022
Revised: 3rd June 2022
Accepted: 27th June 2022

KEYWORDS
Audio-Visual
Speech Recognition
Deep-Learning
Emotion
Spectrogram

INTRODUCTION

Audio-visual speech recognition (AVSR) is a method that helps machines to recognise a word or sentence through both audio or visual inputs. Characteristic derived from both audio speech and visual video are used in order to predict the next phrase spoken by a person. AVSR is better than Automatic Speech Recognition (ASR) because it can not only recognise the speech, but it also can be a lot more accurate in getting the words correct by visual-aiding spectrogram. Given in a noisy situation, ASR cannot recognise the sentences accurately due to the frequency of the sound keeps overlapping by another source of audio. This is similar to human trying to communicate with each other when it was raining. The other person cannot hear clearly due to the heavy rain making noise and overlap the frequency of the speaker.

Multimodality human–computer interaction can be beneficial from audio-visual speech recognition. For a long time, developed machines that are capable of creating or comprehending fragments of natural languages has been extremely challenging. This is due to humans and machines have different medium and sense things differently. Although ASR technology can receive speeches under quiet condition, it still lack of ability to predict correct words through noise. Addition input or data is required for the system to determine the words more accurately. AVSR can increase the speech prediction accuracy by analysing the spectrogram of audio signal. Based on the different features of spectrogram, the system can identify the words spoken by the speaker.

In this research, two deep learning model was proposed, which is Convolutional Neural Network (CNN) and Deep Neural Network (DNN), to recognise the emotion through a speech. Resampling, mixing, cutting and padding is applied our dataset, RAVDESS Emotional speech audio before splitting into training, validatation and testing datasets. Data augment also have been applied to the training dataset for increasing the learning ability of our models.

This paper is organized as follows: Section 2 briefly presents the previous studies of AVSR system. Section 3 explains the data gathering, data processing, development of deep learning model and training script. Section 4 will discuss about the performance of models. Section 5 concludes the paper.

RELATED WORK

This part provides a brief overview of previous studies about AVSR system using deep learning technique. [1] have conducted a study on audio-visual feature fusion via deep neural networks for automatic speech recognition. This study aims to generate effective bimodal features from audio and video stream inputs. The model used for this study is Deep Neural Networks embedded inside Hidden Markov Model (DNN-HMM). The feature extraction used in this are including Mel Frequency Cepstral Coefficients (MFCC), Deep Bottleneck Features (DBNF), concatenation of MFCC and DBNF sets, Basic bimodal deep autoencoder (DNN-I), Modified bimodal deep autoencoder (DNN-II), Tuned bimodal deep autoencoder (DNN-III), and Discriminatively-tuned bimodal deep autoencoder (DNN-IV).

Another DNN model related study was conducted by [2]. The study focus on uses both Data Augmentation (DA) and Ensemble Method (EM) approaches to improve the accuracy of prediction inside a system. Medium Language Model (MLM) and Large Language Model (LLM) was built as the deep learning model for the study. Through average, argmax, and LLR fusion, the best DNN model using original features, perturbed features, and VTLP features is integrated.

*CORRESPONDING AUTHOR | Ismail Mohd Khairuddin | ✉ ismailkhai@ump.edu.my

The studies related to CNN model were read and learnt. For instance, [3] have proposed a bimodal convolutional neural network (CNN) framework based on visual feature creation. This CNN framework is focused in creating an AVSR system that can achieve both audio and visual modal equality during the testing of the system.

In [4], a fine-tuned FaceNet is used as a teacher network to bring comparable pairs closer together in a learnt embedding space, while DenseNet201 is used as a student network and is pre-trained on Imagenet to imitate the fine-tuned FaceNet's outputs. [5] have also presented an emotion recognition system based on emotional Big Data and a deep learning technique. The model used for this project is convolutional neural network (CNN) and consecutive extreme learning machines (ELMs). Big Data is used for training and testing the proposed model. In Big Data, there are only fixed sentences and brief phrases, however in the eNTERFACE database, there are six varied sentences.

[6] looked at visual-audio emotion detection utilising the CNN, MTCNN, and SVM models, which were based on multi-task and ensemble learning. The voice feature is recovered using OPENSMILE's Interspeech 2010 Acoustic Feature Set (IS10) acoustic feature, while the face feature is retrieved from the video's initial frame and five following photos. To obtain final emotion result, a blending classifier is trained using SVM and MTCNN.

[7] developed a Mixture of Brain Emotional Learning (MoBEL) model using an audio-visual fusion model with deep learning characteristics. The model involved in this study are 3D – CNN, CRNN and MoBEL. The audio features of the video samples are extracted through converting raw signal into log Mel-spectrogram images, and then applied CRNN model to extract speech emotional features. Following the extraction of both visual and auditory data, the features are merged and supplied into the MoBEL network model, which trains spatial-temporal information from multiple modalities simultaneously.

DATASETS

Data gathering

The data used for this study is from Kaggle website. The dataset is called RAVDESS Emotional speech audio, featuring 24 professional actors, which contains 12 males and 12 females actors vocalizing two lexically-matched statements in a neutral North American accent. Each actor provides 60 utterances, and a total of 1440 utterances is obtained from the dataset. The speech emotions include calm, happy, sad, angry, fearful, surprise, and disgust expressions, and each expression is produced at normal and strong emotional intensity, with a little additional neutral expression.

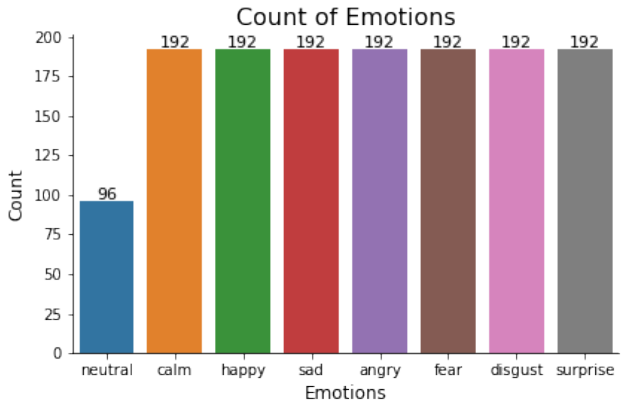

Figure 1. RAVDESS Emotional speech audio dataset emotions' count

Data processing

The dataset is processed through three steps: resampling, mixing and cutting or padding. The data is resampled to 22050 Hz if the orignal sample rate is not equal to 22050Hz. The audio will also mixed to become mono if it previously is stero. And lastly, cutting or padding process will be apply based on the number of signal less than or more than 22050. The comparison between the waveforms and spectrograms of an audio are as below:

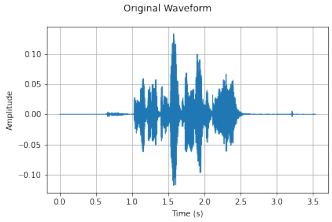

Figure 2. Original waveform

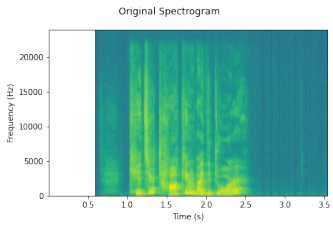

Figure 3. Original spectrogram

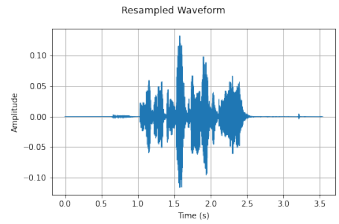

Figure 4. Resampled waveform

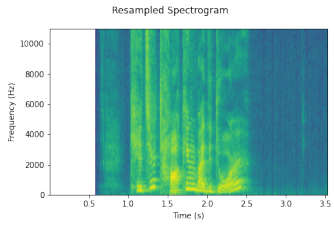

Figure 5. Resampled spectrogram

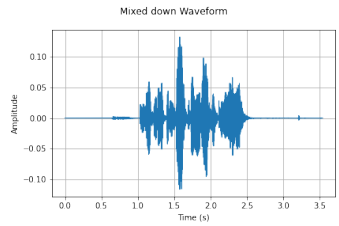

Figure 6. Mixed waveform

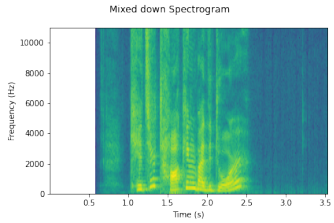

Figure 7. Mixed spectrogram

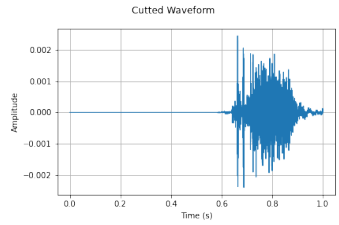

Figure 8. Cutted waveform

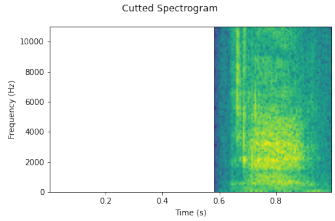

Figure 9. Cutted spectrogram

The processed dataset is divided into training, validation and testing dataset with 80:10:10 ratio. Here is the graphs demonstrate the datasets' classification for each emotion:

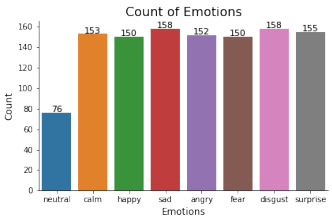

Figure 10. Training dataset's emotions count

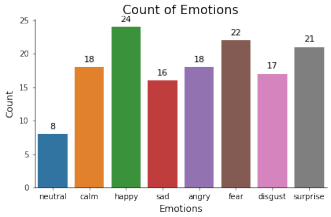

Figure 11. Validation dataset's emotions count

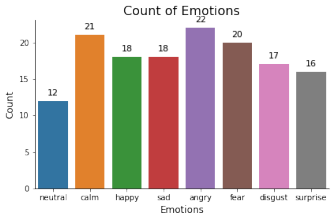

Figure 12. Testing dataset's emotions count

Data augmentation

Other than training dataset, data augmentation are applied to the training dataset for better training accuracy. The training dataset is augmented using white noise, time stretch, pitch scale, invert polarity, and random gain.

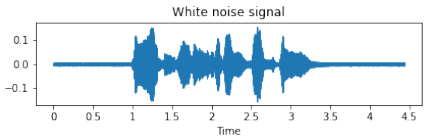

Figure 13. White noise waveform

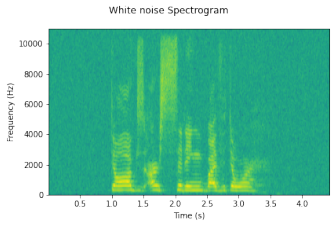

Figure 14. White noise spectrogram

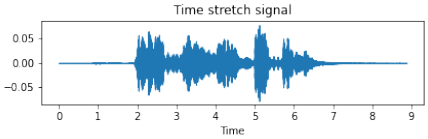

Figure 15. Time stretch waveform

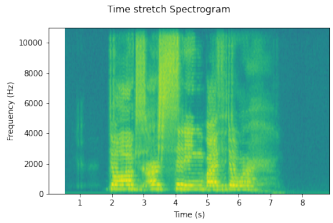

Figure 16. Time stretch spectrogram

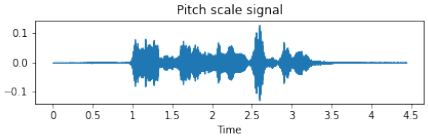

Figure 17. Pitch scale waveform

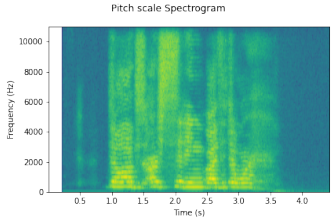

Figure 18. Pitch scale spectrogram

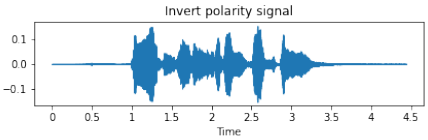

Figure 19. Invert polarity waveform

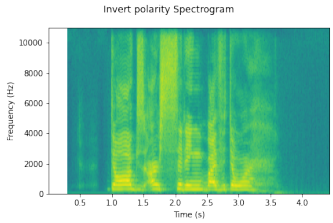

Figure 20. Invert polarity spectrogram

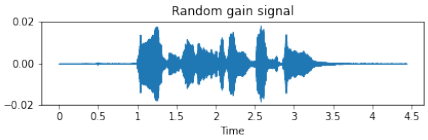

Figure 21. Random gain waveform

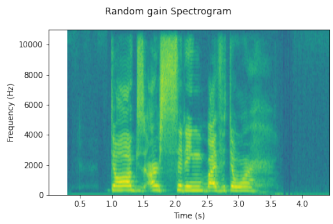

Figure 22. Random gain spectrogram

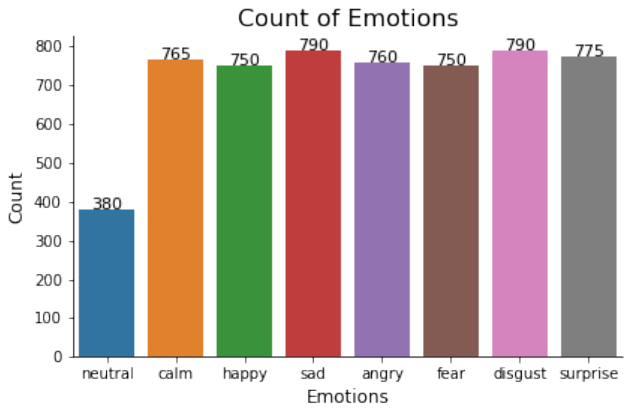

Figure 23. Augmented training dataset's emotions count

DEEP LEARNING TECHNIQUE

CNN model

The CNN model developed used four sequential convolutional blocks. Each convolutional blocks contains one convolutional layer (Conv1d), Rectified Linear Unit (ReLU), batch normalization over 2D or 3D input (BatchNorm1d) and max pooling (MaxPool1d). The only differences between each convolutional blocks are the input and output of the convolutional layer inside them.

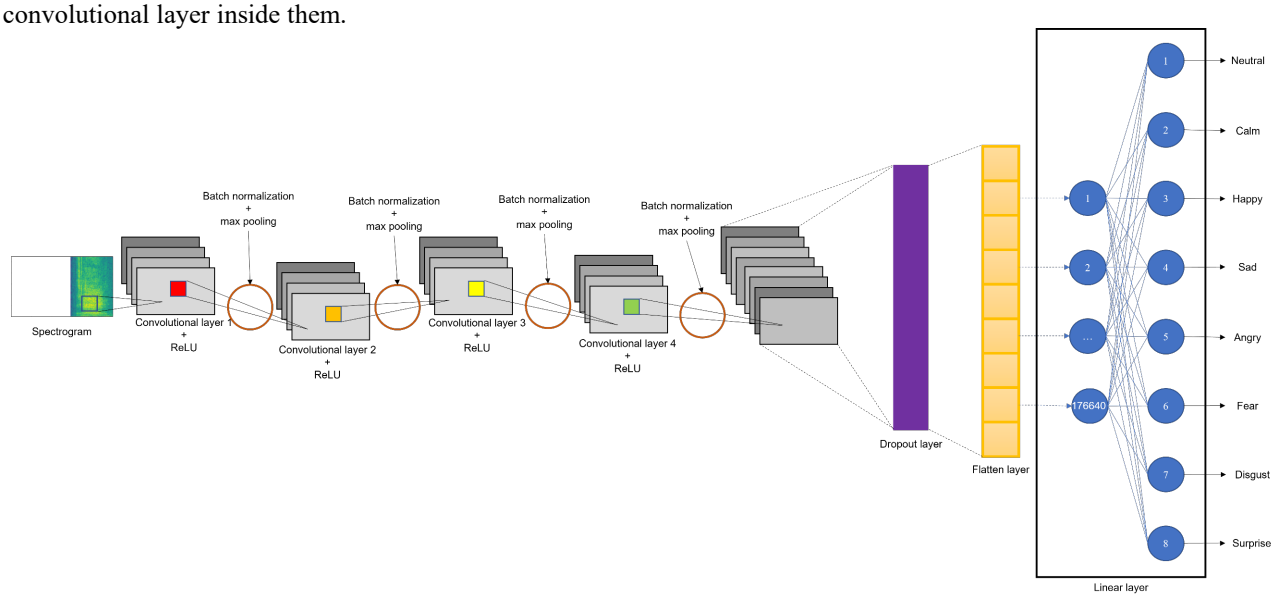

Figure 23. Overall view of CNN model

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv1d-1           [-1, 16, 22052]              64
              ReLU-2           [-1, 16, 22052]               0
       BatchNorm1d-3           [-1, 16, 22052]              32
         MaxPool1d-4           [-1, 16, 11026]               0
            Conv1d-5           [-1, 32, 11028]           1,568
              ReLU-6           [-1, 32, 11028]               0
       BatchNorm1d-7           [-1, 32, 11028]              64
         MaxPool1d-8            [-1, 32, 5514]               0
            Conv1d-9            [-1, 64, 5516]           6,208
             ReLU-10            [-1, 64, 5516]               0
      BatchNorm1d-11            [-1, 64, 5516]             128
        MaxPool1d-12            [-1, 64, 2758]               0
           Conv1d-13           [-1, 128, 2760]          24,704
             ReLU-14           [-1, 128, 2760]               0
      BatchNorm1d-15           [-1, 128, 2760]             256
        MaxPool1d-16           [-1, 128, 1380]               0
          Dropout-17           [-1, 128, 1380]               0
          Flatten-18              [-1, 176640]               0
           Linear-19                   [-1, 8]       1,413,128
================================================================
Total params: 1,446,152
Trainable params: 1,446,152
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.08
Forward/backward pass size (MB): 40.40
Params size (MB): 5.52
Estimated Total Size (MB): 46.00
----------------------------------------------------------------
```

Figure 24. Summary of CNN model

DNN model

For the DNN model, The CNN model's architecture is used as the base and replaced the last linear layer with sequential linear block. The linear block consists of four linear layers with different input and output features.

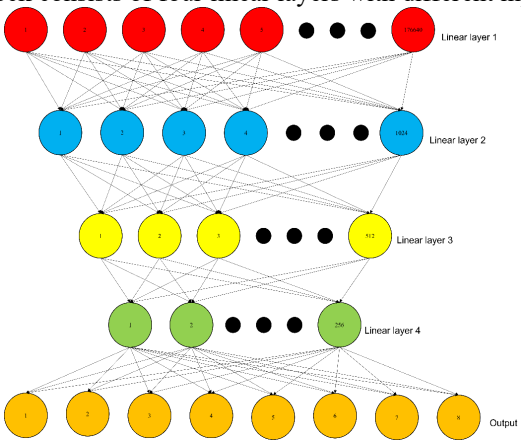

Figure 25. Visualize of linear layers inside of linear block

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv1d-1           [-1, 16, 22052]              64
              ReLU-2           [-1, 16, 22052]               0
       BatchNorm1d-3           [-1, 16, 22052]              32
         MaxPool1d-4           [-1, 16, 11026]               0
            Conv1d-5           [-1, 32, 11028]           1,568
              ReLU-6           [-1, 32, 11028]               0
       BatchNorm1d-7           [-1, 32, 11028]              64
         MaxPool1d-8            [-1, 32, 5514]               0
            Conv1d-9            [-1, 64, 5516]           6,208
             ReLU-10            [-1, 64, 5516]               0
      BatchNorm1d-11            [-1, 64, 5516]             128
        MaxPool1d-12            [-1, 64, 2758]               0
           Conv1d-13           [-1, 128, 2760]          24,704
             ReLU-14           [-1, 128, 2760]               0
      BatchNorm1d-15           [-1, 128, 2760]             256
        MaxPool1d-16           [-1, 128, 1380]               0
          Dropout-17           [-1, 128, 1380]               0
          Flatten-18              [-1, 176640]               0
           Linear-19                [-1, 1024]     180,880,384
           Linear-20                 [-1, 512]         524,800
           Linear-21                 [-1, 256]         131,328
           Linear-22                   [-1, 8]           2,056
================================================================
Total params: 181,571,592
Trainable params: 181,571,592
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.08
Forward/backward pass size (MB): 40.41
Params size (MB): 692.64
Estimated Total Size (MB): 733.14
----------------------------------------------------------------
```

Figure 26. Summary of DNN model

Training script

The criterion used as loss function is cross entropy loss, which is a metric used to measure the accuracy of a classification model. Adam optimizer was used to keep track and adjust the weights and learning rate to reduce the loss function. Both the learning rate and weight decay rate is set as $1e^{-4}$.

RESULTS

Table 1. Accuracy of both models

Epochs	CNN model			DNN model		
	Training accuracy	Validation accuracy	Testing accuracy	Training accuracy	Validation accuracy	Testing accuracy
1	0.215278	0.111111	0.104167	0.224971	0.166667	0.125
2	0.414352	0.173611	0.125	0.416522	0.145833	0.145833
3	0.54456	0.138889	0.152778	0.529659	0.215278	0.236111
4	0.642072	0.152778	0.173611	0.634983	0.104167	0.131944
5	0.695747	0.152778	0.125	0.722946	0.097222	0.152778
6	0.765046	0.173611	0.145833	0.754196	0.104167	0.125

The training accuracy of CNN model increases each epoch. The model stopped training at the 6th epochs because the validation loss is higher than the lowest validation loss for five epochs passed consecutively. For validation accuracy, it begins with fluctuated results, and then it increases to 17.36% at the last epoch. In testing accuracy's perspective, its

all-time high accuracy (17.36%) is achieved at 4th epoch, then drops to 12.50%. It then drops to 12.50%, which is lower than both training and validation accuracy. At the end of training, the testing dataset only managed to reach 14.58% accuracy only.

In DNN model, the training has early stopped at 6th epoch. The training accuracy increased from the lowest (22.50%) to the highest (75.42%) throughout the epoch passed. Meanwhile in the case of validation, although the accuracy is decreased from 1st epoch (16.67%) to 2nd epoch (14.58%), it increased to 21.53% at the 3rd epoch. The accuracy floats between 10.42% and 9.72% from 4th epoch to 6th epoch. The testing accuracy are increased at the beginning of the training. At 3rd epoch, it managed to reach 23.61%, which is the highest accuracy. Then, it tends to drop from 13.19% at 4th epoch to 12.50% at 6th epoch, with a sudden drop at 5th epoch (15.28%).

Table 2. Loss function of both models

Epochs	CNN model			DNN model		
	Training loss	Validation loss	Testing loss	Training loss	Validation loss	Testing loss
1	0.020872	0.031819	0.030526	0.022013	0.036057	0.034124
2	0.013607	0.045381	0.055937	0.012863	0.040116	0.037291
3	0.01026	0.079158	0.095206	0.010786	0.060016	0.046448
4	0.008446	0.066742	0.087654	0.00856	0.372092	0.588932
5	0.007632	0.187434	0.173247	0.007119	0.139436	0.142207
6	0.006029	0.111054	0.086643	0.006001	0.610483	0.629148

In CNN model, the validation loss did not decrease. The validation loss tends to continue increase for each epoch, and achieved the highest validation loss (0.1111) at the last epoch. In another hand, the training loss keeps decreasing for each epoch. At the last epoch, it reached 0.006029, making it the lowest loss function between the three dataset's loss function. For the testing loss, it increased from 1st epoch (0.03053) to 3rd epoch (0.09521). Although it dropped a little at the next epoch (0.08765), it floats up and down for the next two epochs. It increased to 0.1732 at 5th epoch, then dropped to 0.08664.

The loss function in training dataset of DNN model is decreasing dramatically from 0.02201 to 0.006001. On another hand, the loss function in validation dataset keeps increasing from 1st epoch (0.03606) to 4th epoch (0.3721). The validation loss dropped to 0.1394 at 5th epoch, but still not reaching the lowest validation loss. At the last epoch, the validation loss comes to its highest (0.6105). The testing loss does not decrease, but rather increase without dropping at all. The huge gap difference is from 0.03412 at 1st epoch raised up to 0.6291 at 6th epoch.

CONCLUSION

This paper reports the performance of CNN model and DNN model on RAVDESS Emotional speech audio. In summary, the perfomance of CNN model is better than DNN model, as it have higher overall accuracy and lower loss function than DNN model. Even though the final training accuracy of both model is above 75%, the testing accuracy is lower than 15%. This is also known as data overfitting. Hence, several improvement should be made to overcome this issue in the future studies, such as divide training, validation, and testing datasets equally through the emotions category, increase or decrease the convolutional layers or tune the learning rate and dropout layers' probability.

ACKNOWLEDGEMENT

First and foremost, I appreciate all my friends who aid me as a guide throughout this thesis. I am also grateful for the support from my supervisor, Dr. Ismail Bin Mohd Khairuddin and all the lecturers in Innovative Manufacturing, Mechatronic, and Sports Laboratory (iMAMS Lab) for lending me a hand during the study of this thesis. Their expertise in the mechatronics field has guided me along with the study, and their valuable advice and motivation have driven me to complete this thesis successfully.

I am grateful to all the Pytorch forum users who asked the questions I have gone through. They helped me save the waiting time for expertise to reply with solutions. Thank you to those who made videos on YouTube with detailed explanations. They helped me understand more about deep learning's concepts and techniques.

I also want to say thank you to my beloved family members, as they have given me encouragement, support and determination for this thesis. Last but not least, I would like to say thank you to all of those who I have met but did not mention in this words that lent me a hand throughout this thesis.

REFERENCES

[1] Rahmani, M. H., Almasganj, F., & Seyyedsalehi, S. A. (2018). Audio-visual feature fusion via deep neural networks for automatic speech recognition. Digital Signal Processing, 82, 54–63. https://doi.org/10.1016/j.dsp.2018.06.004

[2] Rebai, I., BenAyed, Y., Mahdi, W., & Lorré, J.-P. (2017). Improving speech recognition using data augmentation and

acoustic model fusion. Procedia Computer Science, 112, 316–322. https://doi.org/10.1016/j.procs.2017.08.003
[3] Su, R., Wang, L., & Liu, X. (2017). Multimodal learning using 3D audio-visual data for audio-visual speech recognition. 2017 International Conference on Asian Language Processing (IALP). https://doi.org/10.1109/ialp.2017.8300541
[4] Schoneveld, L., Othmani, A., & Abdelkawy, H. (2021). Leveraging recent advances in deep learning for audio-visual emotion recognition. Pattern Recognition Letters, 146, 1–7. https://doi.org/10.1016/j.patrec.2021.03.007
[5] Hossain, M. S., & Muhammad, G. (2019). Emotion recognition using deep learning approach from audio–visual emotional big data. Information Fusion, 49, 69–78. https://doi.org/10.1016/j.inffus.2018.09.008
[6] Hao, M., Cao, W.-H., Liu, Z.-T., Wu, M., & Xiao, P. (2020). Visual-audio emotion recognition based on multi-task and Ensemble Learning with multiple features. Neurocomputing, 391, 42–51. https://doi.org/10.1016/j.neucom.2020.01.048
[7] Farhoudi, Z., & Setayeshi, S. (2021). Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition. Speech Communication, 127, 92–103. https://doi.org/10.1016/j.specom.2020.12.001