Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 216 (2023) 706–713

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Computer Science and Computational Intelligence 2022

Exploring deep learning algorithm to model emotions recognition from speech

Andry Chowanda[a,*], Irene Anindaputri Iswanto[b], Esther Widhi Andangsari[b]

[a]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480 Indonesia*
[b]*Psychology Department, Faculty of Humanities, Bina Nusantara University, Jakarta 11480 Indonesia*

Abstract

Recognising emotions from a conversation is a paramount task for a machine to understand the full context of the conversation. However, recognising emotions is not an easy task for a machine that cannot understand social context. This research aims to explore several conventional machine learning methods in emotion recognition modelling, such as Na¨ıve Bayes, K-Nearest Neighbour, Decision Trees and Random Forest. Moreover, this research also presents the exploration of deep learning architecture such as Vanilla Deep Neural Networks, Convolutional Neural Networks and Long-Short Term Memory. Moreover, five Mel-Frequency Cepstral Coefficients numbers are also explored (32, 40, 48, 64, 128) with two sampling techniques (Fast Fourier Transform and Kaiser Fast). The results demonstrate that the best training accuracy is achieved by several algorithms such as KNN, ANN and LSTM (100%). Moreover, the model trained with ANN (73%) achieves the best training accuracy.

Keywords: Emotions Recognition, Speech Signals, Machine Learning, Deep Learning

1. Introduction

Emotions are an essential part of social communication. When two interlocutors interact with each other in a social interaction setting, they display social signals or cues. The interlocutor captures those signals or cues, and then the interlocutor translates the signals or cues based on their mental model [1]. They then act based on the mental model and send signals or cues back to the other interlocutors. The loop will continue until the interaction is ended or broken. Hence, it is a paramount task to capture verbal communication in the social interaction between interlocutors and non-verbal cues or signals. One of the signals or cues that can be captured is emotions. The interlocutor's emotions can be captured through their gestures, facial expressions and speech patterns. Those signals or cues can be captured using several sensors, such as a camera (for visual gestures and facial expressions) and a microphone (for speech patterns).

[*] Corresponding author
E-mail address: andry.chowanda@binus.ac.id

10.1016/j.procs.2022.12.187

The captured signals or cues then be processed and modelled using a machine or deep learning algorithms. Several works in emotions modelling can be found in [2, 3] for visual gestures, [4, 5] for facial expressions and [6, 7] for speech patterns. There are a number of applications that implement speech recognition models, such as: medical application [8], entertainment and games [1], analytical tools [9] and business application [10]. Most emotions models that exist out there use either text [11] or facial expressions [5]. The research of emotions models from the speech is still limited as there are only a limited number of datasets in speech emotion recognition. The renowned dataset in speech emotions recognition that is usually used in the research are The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [12] and The Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database [13].

This research explores several algorithms, from machine learning algorithms to deep learning algorithms to model emotion recognition from speech modality or signals. This research aims to explore several conventional machine learning methods in emotion recognition modelling, such as Na¨ıve Bayes, K-Nearest Neighbour, Decision Trees and Random Forest. Moreover, this research also presents the exploration of deep learning architecture such as Vanilla Deep Neural Networks, Convolutional Neural Networks and Long-Short Term Memory. The best training accuracy was 100% and achieved by several models. The accuracy was achieved by KNN (all MFCCs numbers for both FFT and Kaiser Fast), ANN (MFCCs numbers = 40, 48, 64, 128 for FFR and MFCCs numbers = 40, 48, 64 for Kaiser Fast) and LSTM (all MFCCs numbers for both FFT and Kaiser Fast). The best test accuracy was 73% and achieved by the one trained with Vanilla Deep Artificial Neural Network with MFCCs numbers = 48 and FFT. The overview of the insights provided by this research can be implemented to build an affective system, particularly to recognise emotions from speech. As there is not much work done in this area, this should be the main contribution of the research. The rest of the paper is organised as follows: the next section, “Recent Work,” presents the work that has been done in the area of emotions modelling from speech modality. The research methodology is comprehensively described in the “Methodology Section” section. Moreover, the research results are comprehensively presented in the “Results and Discussion” section. Finally, the conclusion is drawn in the last section, the “Conclusion and Future Work”.

2. Recent Work

Automatic emotion recognition techniques have been evolving over these decades. Emotions from the interlocutors can be captured through their social signals, for example, their facial cues, body gestures, verbal cues and speech pattern. The social signals can be automatically captured via several sensors, such as a camera and microphone. The captured signals can then be automatically analysed and implemented in several affective systems. However, recognising emotions automatically is a daunting task for a social ignorant machine [1]. Moreover, there does not exist universal emotions model to be implemented in this area. Several models implement the Circumplex model, Vector model, Plutchik’s mode, and several others. Several primary problems and opportunities in the area of automatic emotion recognition are the datasets, the modality and the algorithms & architectures. There are limited datasets that are publically available to be used to model emotion recognition for several modalities. Moreover, most of the datasets that exist are dominated by datasets in English and performed by Caucasian interlocutors. Albeit emotions is a universal language worldwide, emotions’ expression and feeling are culturally different. Several datasets to train the emotion recognition model can be seen in this research [14, 12, 13]. Emotions can be perceived through several social signals or modalities. Several works in emotions modelling can be found in [2, 3] for visual gestures, [4, 5] for facial expressions and [6, 7] for speech patterns. Finally, modelling the emotions recognition system can be done using either machine learning or deep learning algorithms and architectures. Although, in general, deep learning is superior to machine learning, in some cases, machine learning still performs better than deep learning. This depends on the datasets, features extraction techniques and the case study.

Researchers over these decades have proposed several modalities and algorithms, and this research focuses on the algorithms and architectures used to model emotion recognition through speech modality. However, recognising emotions from speech modality has its own difficulties. First, the features should be extracted from the speech signals. Most researchers implement Mel-frequency cepstral coefficients (MFCCs) as one of the features representation of speech signals. Then, several steps are usually applied to process speech signals. The steps are framing, windowing, normalisation and noise reduction. Furthermore, the accuracy of the emotions recognition system is also ranging from ≈ 40% to≈ 80% depending on the dataset, and the algorithms Xu et al. [15] proposed Attention CNN to recognise Happy, Sad, Exited and Neutral using IEMOCAP dataset and achieved the best accuracy of 76.36%. Fahad et al.

[16] also proposed DNN - HMM to train the IEMOCAP dataset to recognise Happy, Sad, Exited and Neutral and achieved the best accuracy of 65.93%. Another researcher who implemented the IEMOCAP dataset to model emotion recognition via speech signals was Hujian et al. [17]. Hujian et al. proposed CNN - RNN to recognise Happy, Angry, Frustrated and Exited and achieved the best F1-Score of 46.73%. Another dataset that is generally used to model the emotions recognition system is the RAVDESS dataset. Khumbar et al. [18] proposed LSTM to recognise Happy, Calm, Angry, Surprise, Fearful, and Disgust in the RAVDESS dataset and achieved the best training accuracy of 80.81%. Moreover, Bharti et al. [19] proposed MVSM to recognise Happy, Said, Angry and Joy and achieved the best training accuracy of 97%. However, both Khumbar et al. nor Bharti et al. did not use all eight emotion labels in the dataset. Finally, Luna-Jimenez et al. [20] proposed CNN-14 of the PANNs framework to recognise all eight emotions in the RAVDESS dataset and achieved the best F1-Score of 80.08% a subject-wise 5-CV evaluation, 76.58% on CNN-14 architecure and 61.67% on Alex-Net architecture.

3. Methodology

Figure 1 illustrates the proposed research methodology in this research. To model the emotions recognition system from speech, the RAVDESS dataset was used. The RAVDESS dataset consists of 7,356 emotions data (from speech and song modality). The data was gathered from twenty-four actors. This research implements the data using speech modality. It has 4,948 speech data from twenty-four actors. The speech data were annotated into eight classes, and they are: Neutral, Calm, Happy, Sad, Angry, Fearful, Disgusted and Surprised. The data is divided into two subsets, the training (85%) and the testing subsets (15%). The next step is to extract the features from the speech signals. The first step of feature extraction is windowing. The second step is data normalisation, and the noise reduction process is applied to the signals. Finally, the features are represented by Mel-Frequency Cepstral Coefficients (MFCCs). This research implements the Librosa library to perform the windowing, normalisation and noise reduction process. The sampling rate implemented in this research is 22,050. The resample types explored in this research are Fast Fourier Transform (FFT) and Kaiser Fast. Moreover, the number of MFCCs representation explored in this research is 32, 40, 48, 64 and 128.

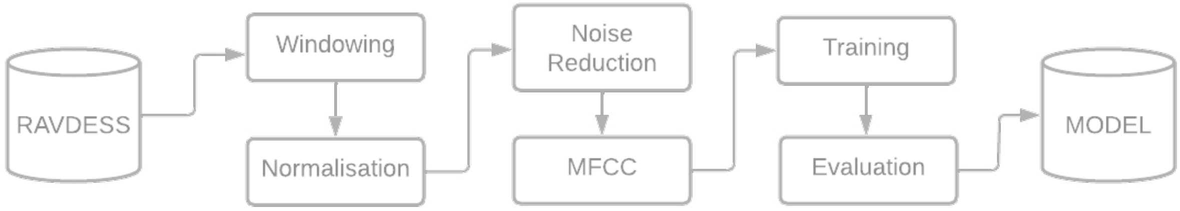
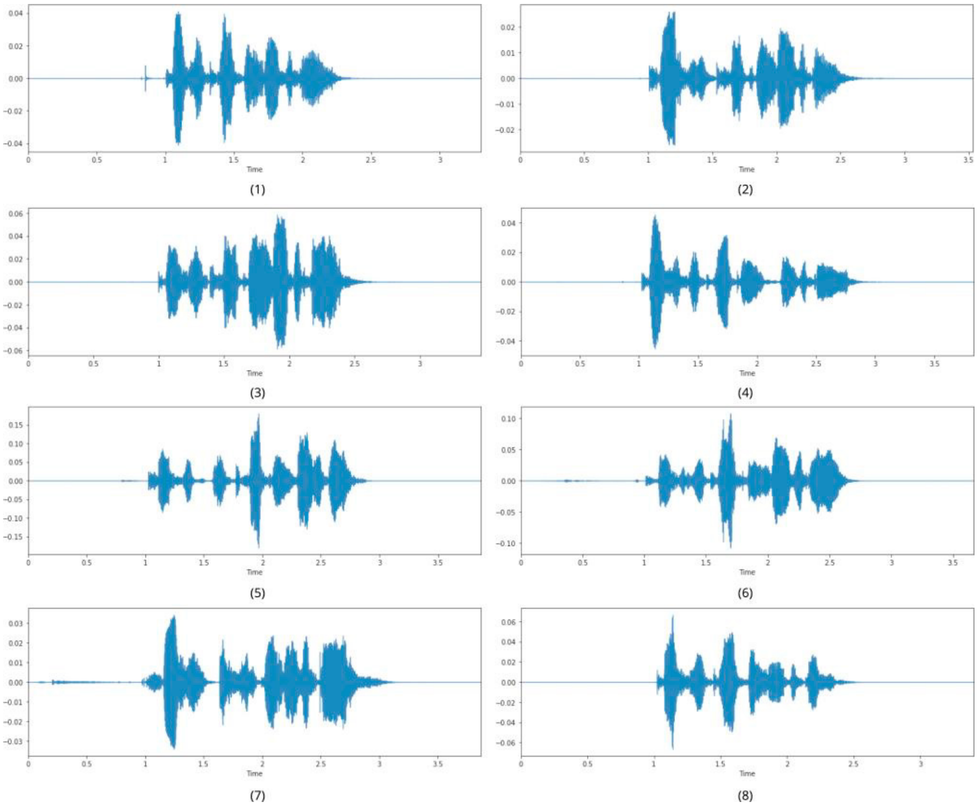


Fig. 1. Methodology

The extracted features are then trained using several machine learning and deep learning algorithms. The machine learning algorithms used in this research are K-Nearest-Neighbors (KNN), Decision Trees (DT), Naïve Bayes and Random Forest. The deep learning algorithms implemented in this research are Deep Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN) and Long-Short Term Memory (LTSM). The KNN and DT algorithms implement the Grid Search technique to find the best hyper-parameters of each model. Moreover, all the data are normalised using a min-max scaler. The KNN hyper-parameters searched are the number of neighbours ($n_{n}eighbors \in [3, 5, 7, 9, 13, 15]$), the weights ($n_{n}eighbors \in [uniform, distance]$) and the leaf size ($lea\,f_{s}ize \in [10, 15, 20, 25]$). The DT hyper-parameters search are the criterion ($criterion \in [gini, entropy]$) and the maximum depth ($max_{d}epth \in [10, 11, 12, 13, 14, 15]$). The Naïve Bayes hyper-parameters implement the default settings for the algorithms. While, the Random Forest implements hyper-parameters as follows: criterion = Gini, maximum depth of trees = 20, maximum features = log^2, maximum leaf nodes = 50, minimum samples leaf - 2, minimum samples split = 15 and number of estimators 50,000.

The Artificial Neural Networks has seven layers consisting of an input layer (number of MFCCs), first dense layer (255) with 0.3 dropouts, second dense layer (128) with 0.3 dropouts, third dense layer (64) with 0.3 dropouts, flatten layer, fourth dense layer (64), fifth dense layer or output layer (8 or number of class). All layers except the fifth dense layer or output layer implement the RELU activation function, while the output layer uses the Softmax activation function. The Convolutional Neural Network has ten layers consisting of the input layer (number of MFCCs), first 1D convolutional layer (128) with 3x3 filter and 0.3 dropout layer, first max pooling layer with a pool size of 8, second 1D convolutional layer (128) with 3x3 filter and 0.3 dropout layer, second max pooling layer with a pool size of 8, third 1D convolutional layer (64) with 3x3 filter and 0.3 dropout layer, third max pooling layer with a pool size of 8, a flatten layer, first dense layer (64) and finally the second dense layer or output layer (8 or number of class). All layers except the second dense layer or output layer implement the RELU activation function, while the output layer uses the Softmax activation function. Finally, the Long-Short Term Memory (LSTM) architecture consists of four layers, and they are the input layer (number of MFCCs), the first LSTM layer (64), the second LSTM layer (64) and the first dense layer or output layer (8 or number of class) with Softmax activation function. The ANN, CNN, and LSTM architectures were trained in 2,000 epochs with a batch size of 16. Moreover, the ANN, CNN, LSTM architectures implements RMSProp optimiser with the best hyper-parameters, they are: learning rate = 0.001, $\rho = 0.9$, $\epsilon = 1e-07$ and decay rate = 0.0. Finally, several performance metrics are implemented as the model evaluation; they are training accuracy and testing accuracy (for all algorithms) and training loss and testing loss (for deep learning algorithms).

4. Results and Discussion



(1) (2) (3) (4) (5) (6) (7) (8)

Fig. 2. The Speech Signals Visualisation

Figure 2 demonstrates the speech signals visualisation of a male actor (actor number one) saying the sentence “Kids are talking by the door” with eight emotions imbued to the sentence. From the top left image, image number 1 indicates the neutral emotion, number 2 shows the calm emotion, number 3 demonstrates the happy emotion, number 4 shows the sad emotion, number 5 demonstrates the angry emotion, number 6 illustrates the fearful emotion, number 7 demonstrates the disgust emotions, and number 8 indicates the surprised emotion. At a glance, some of the emotions look pretty similar; however, some patterns can be learned by the learning algorithms. A total of 70 experiments have been explored in this research. The experiments consist of seven algorithms (machine and deep learning), five settings of MFCCs number and two sampling approaches. The algorithms implemented in this research are K-Nearest Neighbour, Decision Trees, Na¨ıve Bayes, Random Forest, Vanilla Deep Neural Network, Convolutional Neural Networks and Long-Short Term Memory. The MFCCs number implemented in this research are 32, 40, 48, 64, and 128. MFCCs numbers lower than 32 and higher than 128 did not provide significant improvement to the model. Furthermore, the Fast Fourier Transform (FFT) and Kaiser Fast are the two sampling approaches.

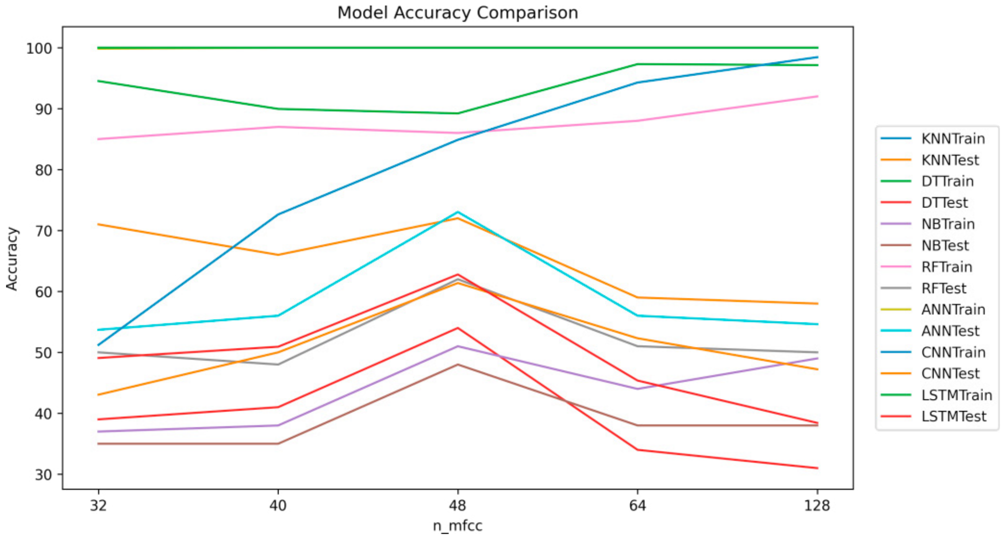


Fig. 3. Results - Accuracy Comparison - FFT

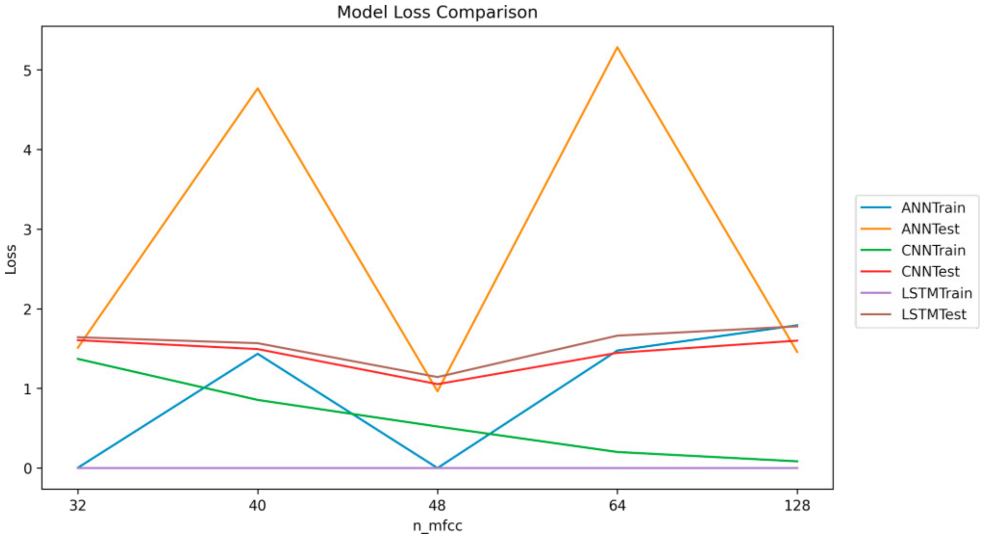


Fig. 4. Results - Loss Comparison - FFT

Figure 3 illustrates the overall accuracy comparison of all models trained with FFT. At the same time, Figure 4 demonstrates the overall loss comparison of models trained with FFT and deep learning. In addition, Figure 5 illustrates the overall accuracy comparison of all models trained with Kaiser Fast. In comparison, Figure 6 demonstrates the overall loss comparison of models trained with Kaiser Fast and deep learning. Overall, the best training accuracy was 100% and achieved by KNN (all MFCCs numbers for both FFT and Kaiser Fast), ANN (MFCCs numbers = 40, 48, 64, 128 for FFR and MFCCs numbers = 40, 48, 64 for Kaiser Fast) and LSTM (all MFCCs numbers for both FFT and Kaiser Fast). The worst training accuracy was Na¨ıve Bayes with MFCCs numbers = 32 and Kaiser Fast (36%). The best test accuracy was 73% and achieved by the one trained with Vanilla Deep Artificial Neural Network with MFCCs numbers = 48 and FFT. In contrast, the worst test accuracy was achieved by the Decision Tree with MFCCs numbers = 128 and FFT.

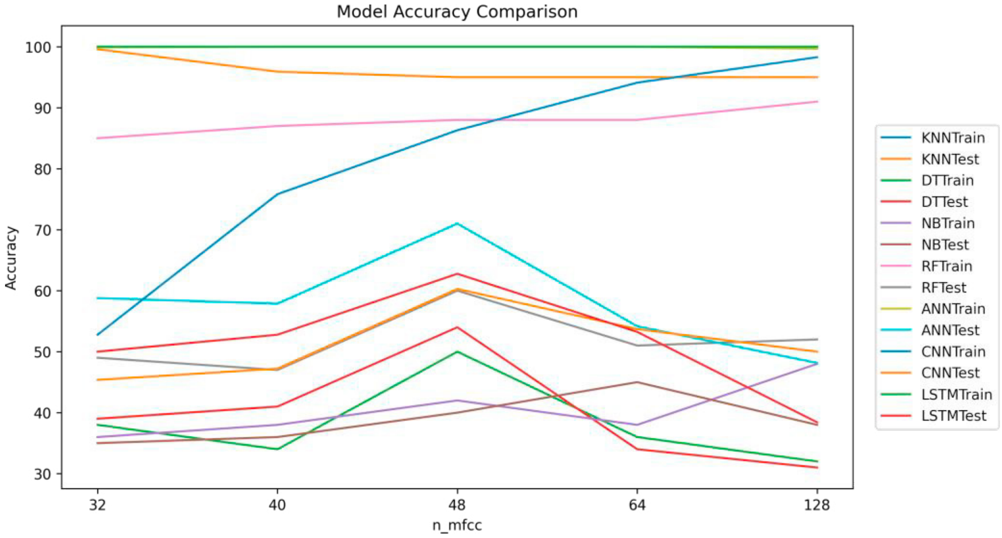


Fig. 5. Results - Accuracy Comparison - Kaiser Fast

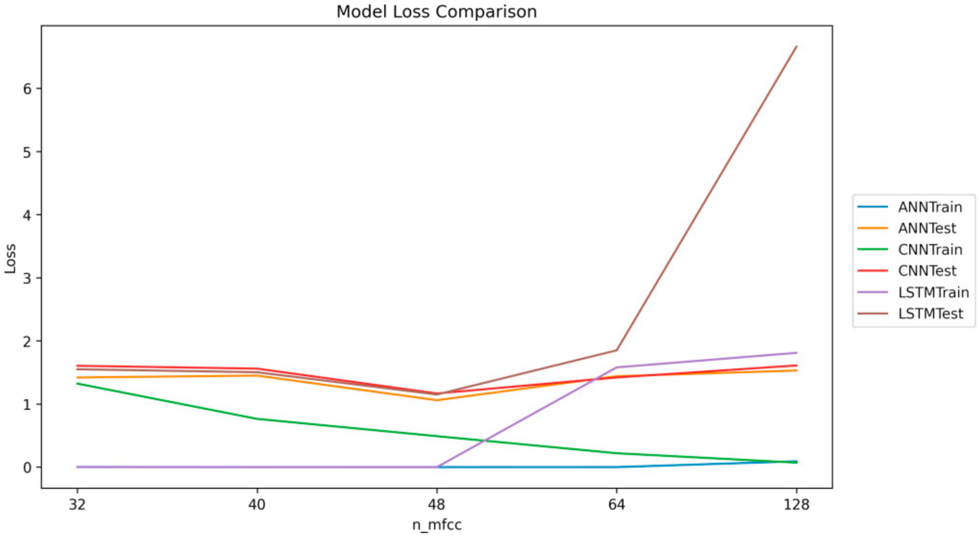


Fig. 6. Results - Loss Comparison - Kaiser Fast

In general, there are no significant differences between the model trained with Kaiser Fast and FFT. Moreover, the models trained with MFCCs numbers = 48 more superior to the other models. However, most models are overfitted, albeit the dropout value has been increased to 0.3. Increasing the value to more than 0.3 is evidently not providing any improvement to the overfitting problem. This is due to the nature of the dataset, where most of the research done using the RAVDESS dataset and implementing all the eight emotions suffers from an overfitting problem. To summarise, the training and testing accuracy achieved by this research are higher than several models from previous research that used the RAVDESS dataset. However, the performances of the models from this research are still inferior compared to the best model from Luna-Jimenez et al. [20].

5. Conclusion and Future Work

A total of 70 research settings have been explored, resulting in 70 models of emotion recognition from speech. The models were trained with seven machine and deep learning algorithms, five MFCCs settings and two sampling settings. The best training accuracy was 100% and achieved by several models. The accuracy was achieved by KNN (all MFCCs numbers for both FFT and Kaiser Fast), ANN (MFCCs numbers = 40, 48, 64, 128 for FFR and MFCCs numbers = 40, 48, 64 for Kaiser Fast) and LSTM (all MFCCs numbers for both FFT and Kaiser Fast). The best test accuracy was 73% and achieved by the one trained with Vanilla Deep Artificial Neural Network with MFCCs numbers = 48 and FFT. Most models suffer from an overfitting problem despite applying several techniques to deal with overfitting. The results achieved by the models trained in this research are superior to the models from previous research using the RAVDESS dataset. However, the models are still inferior compared to the best model from Luna-Jimenez et al. [20]. For the future research direction, the combination of several features other than MFCCs can be implemented in the training process. Moreover, multi-modal signals (e.g. text, facial expression and speech) can be explored to increase performance. The following research agenda can also explore attention algorithms and transformer-based architectures.

Acknowledgements

This work is supported by DIPA, Directorat Riset, Teknologi, dan Pengabdian Kepada Masyarakat, Direktorat Jendral Pendidikan Tinggi, Riset, dan Teknologi. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. Year of 2022 No: 410/LL3/AK.04/2022.

.

References

1. Chowanda, A., Blanchfield, P., Flintham, M., Valstar, M.. Erisa: Building emotionally realistic social game-agents companions. In: *International conference on intelligent virtual agents*. Springer; 2014, p. 134–143.
2. Bhattacharya, U., Rewkowski, N., Banerjee, A., Guhan, P., Bera, A., Manocha, D.. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In: *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE; 2021, p. 1–10.
3. Wu, J., Zhang, Y., Sun, S., Li, Q., Zhao, X.. Generalized zero-shot emotion recognition from body gestures. *Applied Intelligence* 2022; **52**(8):8616–8634.
4. Li, B., Lima, D.. Facial expression recognition via resnet-50. *International Journal of Cognitive Computing in Engineering* 2021;**2**:57–64.
5. Chowanda, A.. Separable convolutional neural networks for facial expressions recognition. *Journal of Big Data* 2021;**8**(1):1–17.
6. Carl, M., Icht, M., Ben-David, B.M.. A cross-linguistic validation of the test for rating emotions in speech: Acoustic analyses of emotional sentences in english, german, and hebrew. Tech. Rep.; ASHA; 2022.
7. Sun, L., Zou, B., Fu, S., Chen, J., Wang, F.. Speech emotion recognition based on dnn-decision tree svm model. *Speech Communication* 2019;**115**:29–37.
8. Rahman, M.M., Sarkar, A.K., Hossain, M.A., Hossain, M.S., Islam, M.R., Hossain, M.B., et al. Recognition of human emotions using eeg signals: A review. *Computers in Biology and Medicine* 2021;**136**:104696.
9. Schneider, J., Sandoz, V., Equey, L., Williams-Smith, J., Horsch, A., Graz, M.B.. The role of face masks in the recognition of emotions by preschool children. *JAMA pediatrics* 2022;**176**(1):96–98.
10. Lin, C.K., Shen, K.S.. Exploring the o2o (online to offline) marketing design of electric vehicles based on consumers' emotions. *SN Applied Sciences* 2022;**4(**8):1–14.
11. Chowanda, A., Chowanda, A.D.. Recurrent neural network to deep learn conversation in indonesian. *Procedia computer science* 2017; **116**:579–586.

12. Livingstone, S.R., Russo, F.A.. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). 2018. doi:\bibinfo{doi}{10.5281/zenodo.1188976}. Funding Information Natural Sciences and Engineering Research Council of Canada: 2012-341583 Hear the world research chair in music and emotional speech from Phonak; URL https://doi.org/10.5281/zenodo.1188976.
13. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., et al. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 2008;**42**(4):335–359.
14. Suryani, D., Ekaputra, V., Chowanda, A.. Multi-modal asian conversation mobile video dataset for recognition task. *International Journal of Electrical and Computer Engineering (IJECE)* 2018;**8**(5):4042–4046.
15. Xu, M., Zhang, F., Khan, S.U.. Improve accuracy of speech emotion recognition with attention head fusion. In: *2020 10th annual computing and communication workshop and conference (CCWC)*. IEEE; 2020, p. 1058–1064.
16. Fahad, M., Deepak, A., Pradhan, G., Yadav, J., et al. Dnn-hmm-based speaker-adaptive emotion recognition using mfcc and epoch-based features. *Circuits, Systems, and Signal Processing* 2021;**40**(1):466–489.
17. Huijuan, Z., Ning, Y., Ruchuan, W.. Coarse-to-fine speech emotion recognition based on multi-task learning. *Journal of Signal Processing Systems* 2021;**93**(2):299–308.
18. Kumbhar, H.S., Bhandari, S.U.. Speech emotion recognition using mfcc features and lstm network. In: *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*. IEEE; 2019, p. 1–3.
19. Bharti, D., Kukana, P.. A hybrid machine learning model for emotion recognition from speech signals. In: *2020 International Conference on Smart Electronics and Communication (ICOSEC)*. IEEE; 2020, p. 491–496.
20. Luna-Jime´nez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J.M., Ferna´ndez-Mart´ınez, F.. Multimodal emotion recognition on ravdess dataset using transfer learning. *Sensors* 2021;**21**(22):7665.