Emotions Classification from Speech with Deep Learning

Andry Chowanda
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480

Yohan Muliono
Cyber Security Program
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480

***Abstract*—Emotions are the essential parts that convey meaning to the interlocutors during social interactions. Hence, recognising emotions is paramount in building a good and natural affective system that can naturally interact with the human interlocutors. However, recognising emotions from social interactions require temporal information in order to classify the emotions correctly. This research aims to propose an architecture that extracts temporal information using the Temporal model of Convolutional Neural Network (CNN) and combined with the Long Short Term Memory (LSTM) architecture from the Speech modality. Several combinations and settings of the architectures were explored and presented in the paper. The results show that the best classifier achieved by the model trained with four layers of CNN combined with one layer of Bidirectional LSTM. Furthermore, the model was trained with an augmented training dataset with seven times more data than the original training dataset. The best model resulted in 94.25%, 57.07%, 0.2577 and 1.1678 for training accuracy, validation accuracy, training loss and validation loss, respectively. Moreover, Neutral (Calm) and Happy are the easiest classes to be recognised, while Angry is the hardest to be classified.**

***Keywords*—Emotions recognition; speech modality; temporal information; affective system**

I. Introduction

Emotions are one of the essential communication factor during the social interactions. They provide additional meanings to verbal communication. Most of the conversation meaning can be captured mostly via non-verbal channels (e.g. speech prosody, body gestures and facial expressions) [1], [2], [3]. Hence, capturing emotions during social interactions between interlocutors is essential to building a system that can interact with humans effectively, efficiently, and naturally. Several efforts have been made to build models that can automatically classify emotions from non-verbal cues in the conversation. Some researchers aim to model the emotions classifier from image or video modality (e.g. Facial Expression Recognition and Hand and Body Gesture). The others use speech and text modality to recognise emotions from the conversation. Generally, the emotions are classified into six basic emotions plus neutral [4]. Recognising emotions from the conversation is a cumbersome task to a social ignorant computer [3]. Several problems exist in building good emotions classifier model from social conversation. First is the dataset; most datasets exist to model the emotions recognition are not balanced in the emotions class. This is due to not all emotions being expressed equally. The second problem is that not all the emotions recognition models have good performance to recognise emotions from the conversation. The results depend on the implemented machine or deep learning algorithms, the dataset used, pre-processing applied, and the modality used (video, image, text or speech). This research proposes and explores several deep learning architectures based on Temporal Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM) to extract features and classify emotions from speech. Most of the emotions recognition required temporal information to improve the model performance. Hence, this research proposes a combination of Temporal CNN to extract the features from the speech signals with LSTM to extract the features further and classify the emotions. The results have shown that MODEL-5 achieved the best model with the training accuracy score of 99.92%, validation accuracy of 78.22%, training loss of 0.0144 and validation loss of 0.8432. The rest of the sections in this paper are organised as follows: The next section illustrates the related work and state of art of emotions recognition from speech. The next section, Emotions Recognition from Speech, demonstrates the proposed framework to model emotions recognition from speech signals. The details of the experiment's settings are also shown in this section. The results are comprehensively presented and discussed in section Results and Discussion. Finally, the last section demonstrates the conclusion and future research direction of this research.

II. Related Work

A. Emotion Detection / Recognition

Emotions are one of the essential parts of social interactions. Emotions convey more than 80% meanings during the social interactions between interlocutors [2], [1]. Hence, detecting or recognising emotions is a paramount task to build a good and natural affective system. Emotion Detection / Recognition is a classification method that can bring up an important feature, namely the emotion contained in an input used for various uses [5]. The input used consists of various forms such as: speech [6], text [7] and visual [8], [9] cues. Most emotions detection/recognition tasks implement machine or deep learning (e.g. convolutional based, attention-based, recurrent based and transformer-based) to model the detector or recogniser. Analysing emotions can help in various fields, one of which is human and computer interaction which can later make computers better decisions for their users. some research regarding emotion detection has many variations,

such as: Sarcasm Detection [10], Mood Prediction [11] and Personality Detection [12].

B. Speech Emotion Recognition

Speech Emotion Recognition (SER) is a method for mapping the features of a speech into the emotions contained in the speech. SER is not a new field of study [13]. However, along with the development of technologies, several methodological developments can be applied to SER. Thus, making research in the SER field more varied and complex to achieve more optimal results. SER usually utilises a classification algorithm to map input in a speech to output in the form of emotion classification. In general, the pipeline for SER is data pre-processing, features extraction and model training + evaluation. The data pre-processing generally involves data augmentation as well as data framing and windowing. Features extraction techniques are implemented to the data after the data is being pre-processed. The features can be extracted in the form of Spectral features, Prosodic features and the combination of both Spectral and Prosodic features. Finally, the features are then trained and evaluated using machine (or deep) learning algorithms.

C. Convolutional Neural Network in Speech Emotion Recognition

Although CNN is well-designed for Image Recognition it could be extended to Natural Language Processing and Speech Processing [14][15][16] Research regarding CNN for Speech Emotion Recognition conducted in 2016 [17] and 2018 [18] using RECOLA datasets [19]. by combining Convolutional Neural Network and Long Short-Term Memory which resulted in an outperformed model compared to traditional approaches on signal processing techniques. Then in 2017 [6] conducted research regarding Speech Emotion Recognition using Deep Convolutional Neural Network (DCNN) and Discriminant. Temporal Pyramid Matching (DTPM) to classify speaker's emotion resulting in a good model for automatic feature learning on speech emotion recognition tasks. The research concludes that DCNN is not only effective for image recognition but also in Speech Emotion Recognition. In 2020 [20] conducted research about CNN based Framework for Enhancing Audio Signal Processing for Speech Emotion Recognition proposing a framework that utilises a discriminative CNN using spectrogram which according to the author, the spectrogram has many features that texts or phonemes cannot represent.

III. Emotions Recognition from Speech

This research proposes the five best architectures by combining Temporal Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) to extract and classify emotions from speech signals. The dataset used in this research is The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [21]. The dataset contains more than 7000 audio files of emotional speech acted by twenty-four professional actors. This research only uses the emotional song dataset, where there are 920 audio data files and four basic emotions (i.e. Angry, Fear, Happy and Sad) plus Neutral. The RAVDESS dataset encodes the audio information (e.g. label and gender) in the filename. Hence, some text sub-string methods were applied to extract the label of the files. In this research, gender information is not used. Moreover, several pre-processing techniques were implemented to the dataset to enhance the quality of the dataset. First, the sample rate of the dataset was set to 16 KHz, and to normalise the speech time, the signal audio was padded to a maximum of 3 seconds. The dataset was split into two sets of data train and test with the ratio of 80%:20% (736:184). Moreover, the training data then were augmented to improve the quality of the data. This research proposed two settings of the data augmentation: seven times of the training data (5,152) and three times of the training data (2,208). Table I illustrates the emotions class distribution on each augmentation setting. The column Train Aug 1 denotes the augmentation with three times of the training data, while column Train Aug 2 refers to the augmentation with seven times of the training data. The dataset has an imbalanced dataset, where the Sad class is the majority class, and the Angry class is the minority class. Fig. 1 illustrates the example of the speech signal. The X-axis indicates the time is second (s), while the Y-axis indicates the amplitude of the signals in Decibel (dB). The left side of the image illustrates the original speech signal and the right side of the image demonstrates the augmented speech signal with noise.

TABLE I. Dataset Distribution

No	Label	Train Aug 1	Train Aug 2	Test
1	SAD	465	1120	46
2	NEUTRAL	450	1036	39
3	HAPPY	444	1015	39
4	FEAR	441	1015	36
5	ANGRY	408	966	24
6	**TOTAL**	**2,208**	**5,152**	**184**

Feature extraction methods using Short-time Fourier Transform (STFT) and Mel Frequencies were applied to generate the Mel-Spectrogram representation on each audio file. The features were extracted using several parameters, such as: the hop length of 512, the window of 256. To normalise the features, all the vector then padded with zeros up to 2,048 to match the Fast Fourier Transform input. The next step was to generate Mel-Spectrogram from the Mel frequencies generated from Mel bins of 128 and the maximum frequency of 4.0 KHz. Finally, the features were framed with a window step of 128 and a window size of 64. Fig. 1 and Fig. 2 illustrate the examples of the Mel Spectrogram features visualisation from a happy female audio (left side of the image) and a happy male audio (right side of the image). The features extracted are then used in training with the proposed architectures. There are five best architectures proposed in this research.

Fig. 3 demonstrates the blueprint of the proposed architectures. The architecture blueprint consists of four parts: the CNN block, The Flatten layer, the LSTM block and the Softmax Layer. The Temporal CNN block consists of one temporal convolutional layer (3x3 filter), one batch normalisation layer, one Exponential Linear Unit (ELU) activation layer, one Max Pooling layer (3x3 filter) and one Dropout layer. The CNN block can have two to four blocks of layers in the proposed architecture (see Table II. The Flatten layer aims to flatten all the extracted layers with temporal features. Moreover, the LSTM block consists of 128 units of LSTM layers. The LSTM block has one to two LSTM layers plus one bi-directional layer in the proposed architecture (see Table II. Finally, the

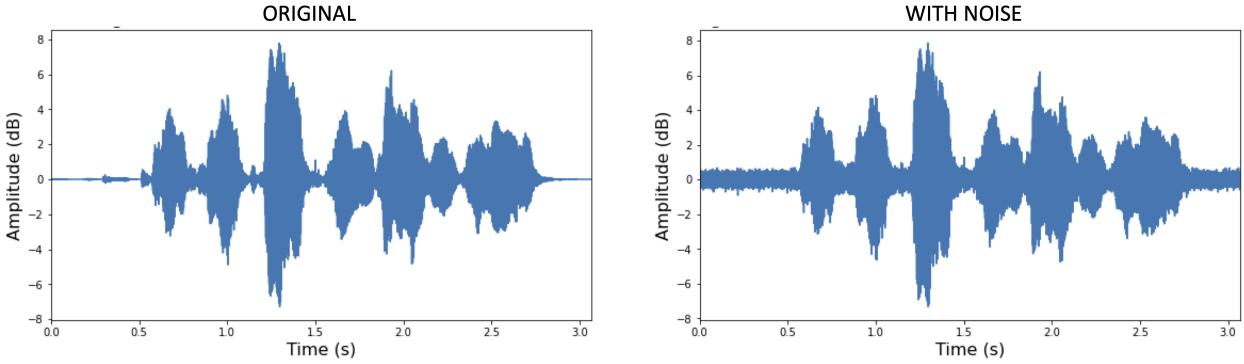

Fig. 1. Speech Signals: Original (left), Augmented with Noise (right).

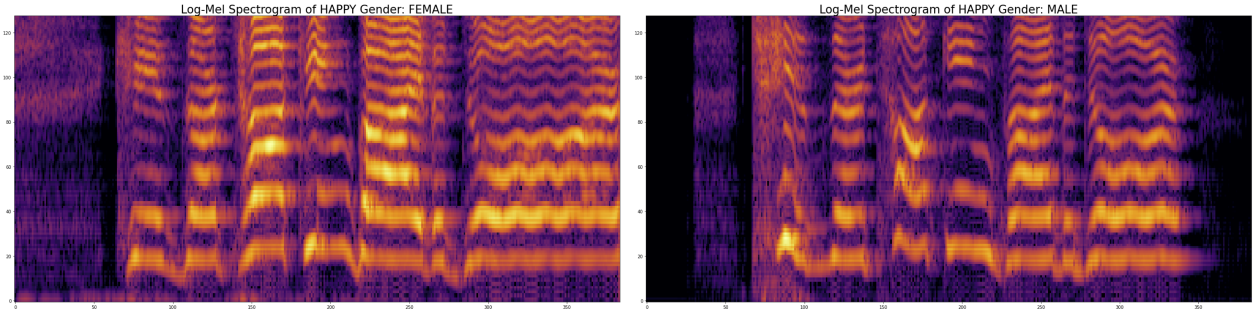

Fig. 2. Mel Spectogram Features: Happy Female (left), Happy Male (right).

features extracted then are classified in the Softmax activation layer. Table II shows the overall settings of the proposed architectures. The CNN column indicated the number of the CNN blocks (e.g. 2 indicates that there are two blocks of CNN layers in the model), while the LSTM column shows the number of the LSTM blocks, where the asterisk mark (*) indicates a Bidirectional LSTM layer. The Dropout implemented in this research is between 0.3 to 0.4. Finally, the column Aug indicates the augmented training dataset used, and three indicates three times the original training dataset and seven indicates seven times the original training dataset.

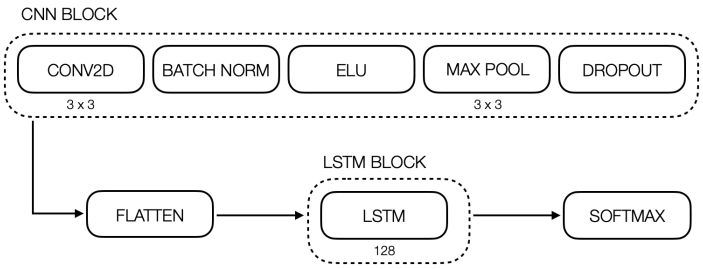

Fig. 3. Proposed Architecture.

IV. Results and Discussion

Five architectures with two settings of augmentation data were explored in this research. The architectures combine Temporal CNN and LSTM (or Bidirectional LSTM) to extract and classify the emotions. Fig. 3 illustrates the baseline of the proposed architectures and Table II demonstrates the proposed

TABLE II. Proposed Architecture Settings

NO.	NAME	CNN	LSTM	DROPOUT	AUG
1	MODEL-1	2	1	0.3	3
2	MODEL-2	2	1	0.4	3
3	MODEL-3	2	2	0.3	3
4	MODEL-4	3	1*	0.3	7
5	MODEL-5	4	1*	0.3	7

architecture settings. This research explores several settings of deep learning architectures and results in the five best architectures that provide the best results. The best result was achieved by the MODEL-5 that consisted of four CNN blocks combined with one bidirectional LSTM layer with a dropout value of 0.3. MODEL-5 also implemented data augmentation seven times bigger than the original training data. The MODEL-5 provides 99.92% and 78.22% of training accuracy and validation accuracy scores, respectively. Moreover, the MODEL-5 provides 0.0144 and 0.8432 of training loss and validation loss, respectively. The dropout value of 0.4 did not significantly improve the model compared to the dropout value of 0.3.

TABLE III. Overall Results

NO.	NAME	Train Acc	Val Acc	Train Loss	Val Loss
1	MODEL-1	98.73%	58.15%	0.1396	1.1680
2	MODEL-2	94.25%	57.07%	0.2577	1.1678
3	MODEL-3	96.01%	55.98%	0.2098	1.2720
4	MODEL-4	98.84%	69.85%	0.0625	0.9312
5	MODEL-5	**99.92%**	**78.22%**	**0.0144**	**0.8432**

Table III illustrates the overall results of the experiments.

The results have shown that the models trained with seven times training data augmentation perform better than the models trained with three times training data augmentation. Overall, there are no significant differences in the training accuracy score of models trained with three times training data augmentation compared to the models trained with seven times training data augmentation. However, the seven times training data augmentation model provides higher validation accuracy and lower validation loss. Moreover, the models trained with three times training data augmentation suffer from over-fitting despite batch normalisation and dropout were applied to the CNN and LSTM architectures. MODEL-1 that implemented 2 CNN blocks and 1 LSTM block with 0.3 dropouts trained with three times training data augmentation resulted in 98.73%, 58.15%, 0.1396 and 1.1680 in training accuracy, validation accuracy, training loss and validation loss, respectively. Moreover, MODEL-4 implemented three blocks CNN, one bidirectional LSTM with dropout value of 0.3 and trained with seven times training data augmentation resulted in 98.84%, 69.85%, 0.0625 and 0.9312 in training accuracy, validation accuracy, training loss and validation loss, respectively. Training with two layers of LSTM did not improve the validation accuracy, albeit two layers of LSTM improved the training accuracy. The results show that the model trained with two layers CNN and one layer LSTM (MODEL-2) provides 94.25%, 57.07%, 0.2577 and 1.1678 for training accuracy, validation accuracy, training loss and validation loss, respectively. Moreover, the model trained with two layers CNN and two layers LSTM (MODEL-3) provides 96/01%, 55.98%, 0.2098 and 1.2720 for training accuracy, validation accuracy, training loss and validation loss, respectively. Fig. 4 illustrates the confusion matrix for each classes in the best model (i.e. MODEL-5). The result shows that Neutral (Calm) and Happy emotions are the easiest emotions to classify from the given speech dataset. Moreover, the Angry emotion is the hardest emotion to classify compared to the other classes. The Angry emotion is also mostly miss-classified as the false positive in the other classes. Most likely, it is due to the number of the Angry class in both the training and testing dataset. Finally, the Adam and SGD optimiser do not provide a significant difference to the training accuracy, validation accuracy, training loss and validation loss.

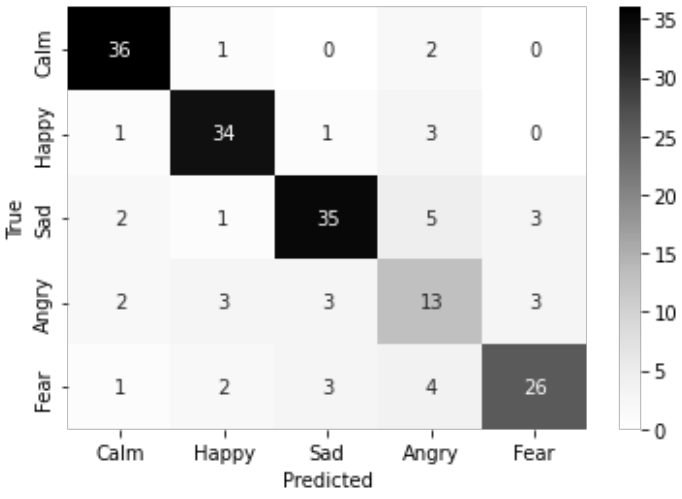

Fig. 4. Confusion Matrix for MODEL-5.

V. Conclusion and Future Work

Five settings of architectures with the combination of CNN and LSTM (or Bidirectional LSTM), number of dropouts and the data augmentation settings were explored in this research. The architectures were implemented to train the emotions recognition models using The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset. The dataset was pre-processed and augmented with two data augmentation settings (i.e. three times and seven times of the original data). The results show that the best model was achieved by MODEL-2, which provides 94.25%, 57.07%, 0.2577 and 1.1678 for training accuracy, validation accuracy, training loss and validation loss, respectively. Moreover, Neutral (Calm) and Happy emotions are the easiest emotions to classify from the given speech dataset, while the Angry emotion is the hardest emotion to classify compared to the other classes. This is due to the number of data in the Angry class in both the training and testing dataset.

For future direction research, more combinations of the architectures, such as the attention architectures and Transformer based architectures, will be explored to increase the recogniser model performances. Moreover, the multi-modal features can also be explored to increase the accuracy and tackle the over-fitting problem. Furthermore, the features from videos (e.g. facial expressions and body gestures), speech and text, can be explored to build a better model for emotion recognition. Finally, the emotions recogniser model that has been trained can be implemented to the more complex affective system such as virtual humans, where recognising emotions can be one of the tools to extract non-verbal meanings from the human interlocutors.

References

[1] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, 2011.

[2] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.

[3] Andry Chowanda and Alan Darmasaputra Chowanda. Recurrent neural network to deep learn conversation in indonesian. *Procedia computer science*, 116:579–586, 2017.

[4] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

[5] Yoones A Sekhavat, Milad Jafari Sisi, and Samad Roohi. Affective interaction: Using emotions as a user interface in games. *Multimedia Tools and Applications*, 80(4):5225–5253, 2021.

[6] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, 20(6):1576–1590, 2017.

[7] Boaz Shmueli and Lun-Wei Ku. Socialnlp emotionx 2019 challenge overview: Predicting emotions in spoken dialogues and chats. *arXiv preprint arXiv:1909.07734*, 2019.

[8] Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):1–26, 2018.

[9] Andry Chowanda. Separable convolutional neural networks for facial expressions recognition. *Journal of Big Data*, 8(1):1–17, 2021.

[10] Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22, 2017.

[11] Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam. Music mood detection based on audio and lyrics with deep neural net. *arXiv preprint arXiv:1809.07276*, 2018.

[12] Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, 53(4):2313–2339, 2020.

[13] Mehmet Berkehan Akçay and Kaya Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76, 2020.

[14] A Rakhlin. Convolutional neural networks for sentence classification. *GitHub*, 2016.

[15] WQ Zheng, JS Yu, and YX Zou. An experimental study of speech emotion recognition based on deep convolutional neural networks. In *2015 international conference on affective computing and intelligent interaction (ACII)*, pages 827–831. IEEE, 2015.

[16] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345, 2019.

[17] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204. IEEE, 2016.

[18] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5089–5093. IEEE, 2018.

[19] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013.

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

[21] 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.