Exploring a Cutting-Edge Convolutional Neural Network for Speech Emotion Recognition

Navod neranjan thilakarathne
Faculty of Integrated Technologies
Universiti Brunei Darussalam
Brunei Darusalam
21h5104@ubd.edu.bn

Kasorn Galajit
NECTEC, National Science and Technology Development Agency,
Pathum Thani, Thailand
kasorn.galajit@nectec.or.th

Candy Olivia Mawalim
Japan Advanced Institute of Science and
Technology,Ishikawa, Japan
candylim@jaist.ac.jp

Hayati Yassin
Faculty of Integrated Technologies Universiti Brunei Darussalam
Brunei Darusalam
hayati.yassin@ubd.edu.bn

***Abstract*—In light of the ongoing expansion of human-computer interaction, advancements in the comprehension and interpretation of human emotions are of the utmost importance. SER, representing speech emotion recognition, is a critical element in this context as it enables computational systems to comprehend the emotions of humans. Throughout the years, SER has employed a variety of techniques, including well-established speech analysis and classification methods. However, in recent years, techniques powered by deep learning have been suggested as a viable substitute for conventional SER methods, owing to their more encouraging outcomes in comparison to the aforementioned methods. In this regard, by utilizing a novel Convolutional Neural Network (CNN) model designed to assess and categorize seven emotional states based on speech signals retrieved from three distinct databases, this research presents a novel approach to SER that yields 88.76% accuracy. The purpose of this research is to enhance the accuracy and efficiency of emotion identification, with the end goal of boosting applications in fields such as interactive voice response systems, mental health monitoring, and personalized digital assistants.**

Keywords—Speech emotion recognition, speech classification, convolution neural network, artificial intelligence, SER, deep learning, CNN

I. INTRODUCTION

In the current epoch of Human-Computer Interaction (HCI), the capacity of machines to perceive and react to human emotions carries significant implications for a variety of domains, encompassing transportation, entertainment, mental health diagnostics, customer service, and virtual assistants [1], [2]. When considering the diverse ways emotions can be conveyed, speech continues to be one of the most abundant and easily obtainable reservoirs of emotional indicators. Therefore, SER has emerged as a critical component in enabling intelligent systems to comprehend the meaning of human emotions [2]-[4]. To put it simply, SER is the process of extracting emotional states from spoken language and categorizing them. Its main goal is to interpret the intricate emotional nuances present in spoken language so that machines can recognize a variety of emotions, such as joy, grief, rage, and others [1]-[4].

The accurate identification of emotions in speech data continues to be a multifaceted and ever-changing obstacle, demanding techniques capable of capturing the subtle dynamics, patterns, and contextual information that are inherent in audio signals [3], [4]. On the other hand, human speech is a rich tapestry of information where the tone, pitch, and rhythm of our speech often carry more emotional weight than the words themselves. As per the current status of SER research, a range of strategies have been employed, for speech analysis and classification. However, Deep Learning (DL) powered techniques have been offered as a substitute for these traditional methods in recent times [4]-[6].

By analyzing emotions in human speech, SER improves various aspects of our lives, making it an indispensable technology in our increasingly interconnected world, necessitating cutting-edge technologies for SER that yield high accuracy [2]-[5]. Overall, the SER is comprised of two phases: feature extraction and feature classification, as depicted in Fig. 1 [1]-[3]. For the extraction of features, various technologies are used, such as acoustic feature extraction (involves analyzing various speech characteristics like pitch, tone, and speed through techniques like Mel-Frequency Cepstral Coefficients (MFCCs)), prosodic analysis (refers to the analysis of rhythm, stress, and intonation of speech) and spectral analysis (involves examining the frequency spectrum of speech through techniques like Fourier Transform and spectral band energies) [1]-[4]. Once the features are extracted, the next step involves classifying them into distinct emotions, which employs Machine Learning (ML) and DL techniques [1]-[4]. The commonly used machine learning algorithms include Support Vector Machine, K-Nearest Neighbor, and Decision Tree, where DL techniques, including Recurrent Neural Networks (RNNs) and CNNs, have shown remarkable proficiency in handling the complexities of speech data in recent years [1]-[4].

In recent years, DL has drawn more interest and is seen to be a developing subject of study in SER [4]–[6]. When compared to traditional methods, the application of DL in SER provides unique advantages. These advantages encompass the ability to identify intricate patterns and characteristics directly from raw data, eliminating the necessity for manual feature

extraction and adjustment. Additionally, DL methods tend to extract fundamental features from the input data and are adept at processing data that is not labelled. According to the state-of-the-art [2]-[4], CNN demonstrated better discriminative performance compared to other DL architectures employed and often outperformed traditional methods employed for SER.

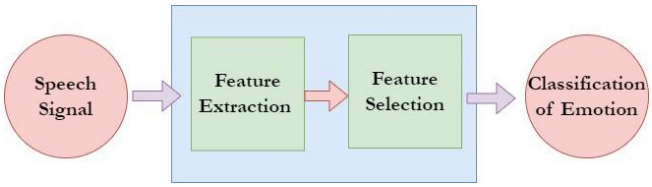

Fig. 1. Phases of SER

Thus, motivated by the fact that designing a more efficient way of identifying emotions through human speech, this study presents a novel CNN architecture for SER. By harnessing the capabilities of DL, this proposed architecture achieves exceptional accuracy in distinguishing subtle human emotions from speech data. By conducting an extensive investigation into this innovative CNN architecture, our objective is to propel the field of SER forward, presenting a potentially fruitful pathway for augmenting the emotional intelligence of forthcoming human-computer interactions. The key contributions of the study are outlined in the following.

1. Propose a novel CNN model for speech emotion recognition integrating CNN layers with Long Short-Term Memory (LSTM) layers, combining speech data from three distinct datasets for training the underlying model.
2. Employ data augmentation techniques to improve the generalization of the proposed CNN model and prevent overfitting.
3. Summarizes the state-of-the-art in order to give a broad picture of the present status of SER research.

The remainder of the study is organized in the following manner. Following the introduction, the second section highlights the background of the research, summarizing the latest literature. The methodology is highlighted in section three, and experimental results are highlighted in section four. Finally, the paper concludes with a conclusion.

II. Background and Related Work

Humans are able to better comprehend one another via the use of emotions, and SER is the process of analyzing spoken language in order to determine and categorize the emotional state or mood of the speaker. [1]-[3]. The goal of SER is to determine the underlying emotions expressed in speech, such as happiness, sadness, anger, fear, surprise, or neutrality [1]. This technology is a subset of the broader field of HCI, which focuses on recognizing and understanding human emotions and moods.

Overall, in recent times, there has been a significant surge in the utilization of ML-based approaches to tackle a wide array of SER tasks [4], [5]. However, identifying emotions from speech presents an intricate challenge due to the diverse ways people express their emotions, making it challenging to pinpoint the distinguishing features. In fact, even for human evaluators, this paralinguistic task proves to be quite demanding [4], [5]. Conventional methods employed to address this issue typically involve extracting fundamental descriptors from speech signals and then training ML models to learn from these features. However, the process of selecting optimal features for extraction is a formidable task, and the subsequent optimization is often a time-consuming endeavour [4]-[8]. Consequently, the traditional approach in speech analysis shifted towards harnessing robust semantic analysis strategies, frequently relying on model selection to fine-tune results [5]-[9]. As a result, DL appeared as a viable option that can overcome most of the challenges associated with ML [10]-[12].

As of now, DL techniques have risen to prominence as a remarkable approach within the domain of SER, and several compelling reasons underscore their significance: (1) ability to handle complex patterns, as speech signals are inherently complex and high-dimensional where DL algorithms, especially CNNs and RNNs, are adept at handling such data, extracting intricate patterns that simpler models might miss [4]-[8]. (2) ability to learn feature hierarchies: as DL models can automatically learn hierarchies of features in the context of SER, this means they can start by detecting simple patterns in the raw audio signal and progressively learn more abstract features, like emotional cues [4]-[8]. (3) higher performance: DL models have demonstrated higher accuracy in classifying emotions from speech compared to traditional ML methods, and DL models are better at generalizing from training data to new, unseen data, which is crucial for real-world applications [4]-[8]. (4) scalability with data: DL models generally improve as the amount of available data increases, making them well-suited for SER, where large datasets can be leveraged [4]-[8]. In summary, DL offers a suite of tools uniquely suited to the challenges of SER. Its ability to process complex, high-dimensional data, learn from large datasets, and adapt to the intricacies of human speech makes it a potent choice for advancing the state of the art in emotion recognition from speech.

According to the state of the art we have reviewed, it is evident that only a few studies have applied Artificial Intelligence (AI) techniques for SER. Among such studies (Khalil et al., 2019), [1] present an overview of DL techniques employed in SER and discuss recent literature where these methods are utilized for SER. (Akçay & Oğuz, 2020) , [2] discussed distinct areas of SER, provided a detailed survey of the current literature on each, and listed the current challenges. (Fayek et al., 2017) ,[3] presents state-of-the-art results on the IEMOCAP database for speaker-independent SER and presents quantitative and qualitative assessments of the model's performance. Also, the literature on SER systems and varied design components/methodologies is surveyed by (Wani et al., 2021),[4].The available literature on various databases, different features, and classifiers employed in SER has been reviewed by (Swain et al., 2018),[7] in their study. (Parthasarathy & Tashev, 2018) ,[14] have analyzed various CNN architectures used for SER, where they report performance on different frame-level features.

Having provided a brief overview of the background of SER and survey/review studies done, Table 1 summarizes the recent similar research work done in the domain of SER, where we highlight the algorithm /techniques used along with the scope of the study.

TABLE I. SUMMARY OF RECENT RELATED RESEARCH WORK

Reference	Algorithm/technique used	Scope of the study
[5], (Zheng et al., 2015)	CNN	This research presents a methodical approach to the implementation of an efficient emotion identification system that is built on deep CNNs.
[6], (Neumann & Vu, 2017)	CNN	The research involved using an attentive CNN with a multi-view learning objective to assess system performance. This assessment was based on varying the length of the input signal, experimenting with different acoustic feature sets, and analyzing various types of emotionally charged speech.
[8], (Badshah et al., 2017)	CNN	Presents a method for SER using spectrograms and deep CNN.
[9], (Huang et al., 2014)	CNN	The study introduces a method where a semi-CNN is used for learning features that are crucial for emotion detection in speech.
[10], (Lim et al., 2016)	CNN and RNN	The authors have designed a SER model based on CNN and RNN and evaluated on emotional speech data.
[12] , (Issa et al., 2020)	CNN	In order to identify emotions, the authors present a novel architecture that utilizes features extracted from sound files, including spectral contrast, chromagram, mel-frequency cepstral coefficients and tonnetz representation, and chromagram, as inputs for a one-dimensional CNN.
[13] , (Bertero & Fung, 2017)	CNN	The authors propose a real-time CNN for SER to recognize three emotional states in the study.
[15] ,(Tzirakis et al., 2018)	CNN and LSTM	The authors present a novel continuous SER model using CNN and LSTM.
[16] , (Mao et al., 2014)	CNN	The authors proposed using CNN to discover affect-salient features for SER. Their experimental findings on benchmark datasets demonstrate that the method employed produces consistent and dependable scene recognition performance.
[17] ,(Zhang et al., 2018)	CNN	The authors investigated how a deep CNN can be utilized to bridge the emotional gap in voice signals.

Overall, with the summarized research work, it is evident that CNN has become increasingly significant in SER in recent times. This is primarily attributed to their capacity to extract and learn features from unprocessed speech data, accommodate speech variability, and acquire hierarchical representations. Spatiotemporal and spectral aspects of speech, which are critical for identifying emotions, are both adeptly processed by CNNs. By integrating with additional neural network architectures such as RNNs and LSTMs, their capacity to capture the spatial and temporal dynamics of speech is significantly improved. Furthermore, the improved capability of CNNs to accurately identify emotions in speech, combined with developments in DL and the capacity to generalize across diverse datasets, has resulted in a decreased reliance on manual feature engineering. As a consequence, CNNs have become the preferred option for SER applications. On the other hand, commonly used features include MFCCs for spectral representation, spectral features like centroid and bandwidth for energy distribution analysis, and pitch for emotion and speaker identification. Additional features like formants for phoneme recognition, zero-crossing rate for distinguishing speech types, and temporal characteristics such as duration and silence are also used.

III. METHODOLOGY

The methodology followed in the study encompasses six steps, as depicted in Fig. 2.

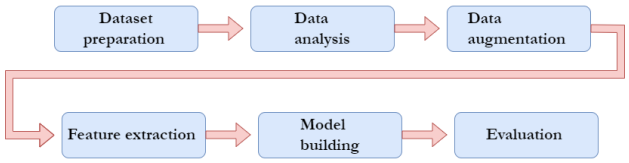

Fig. 2. Methodology of the research.

As illustrated in Fig. 2, the initial stage entails dataset preparation. The dataset was constructed through the concatenation of data extracted from three distinct databases: the toronto emotional speech set [19], the crowd-sourced emotional multimodal actors dataset [18], and the surrey audio-visual expressed emotion database [20]. The prepared dataset encompasses short voice messages, which include English phrases voiced by professional actors (including both male and female voices), and the data was in the .wav file format. From the three datasets, data were extracted in a way to contain seven types of emotions: angry, happy, disgust, surprised, sad, happy and fear. Overall, the final concatenated dataset contains 10722 voice records.

A. *Data Analysis*

After the dataset has been prepared, an exploratory data analysis is conducted to examine the dataset's underlying data. Fig. 3 showcases the number of .wav voice files with each emotion, where it is evident that the emotion surprise was the lowest recorded.

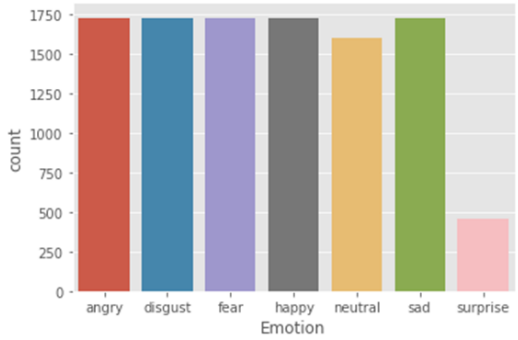

Fig. 3. Count of emotions in the dataset.

To have a glimpse of the underlying data contained in the dataset, two types of emotions, fear and angry, are illustrated in the form of wave plots in Fig. 4, and their respective spectrograms are depicted in Fig. 5.

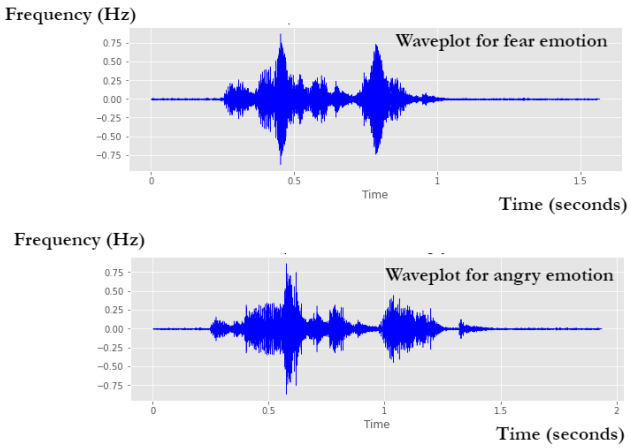

Fig. 4. Wave plots pertaining to fear and angry emotions.

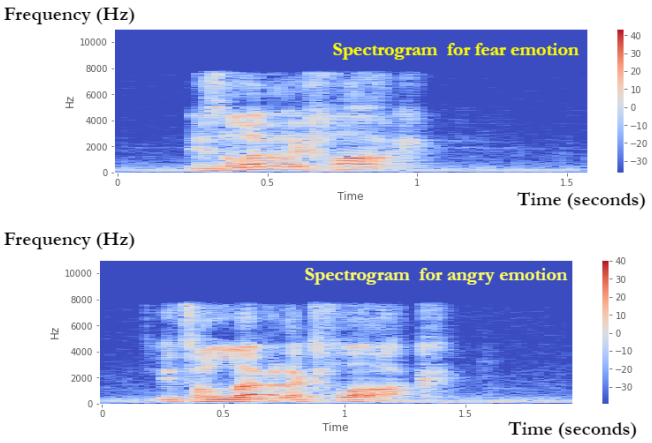

Fig. 5. Spectrogram pertaining to fear and angry emotions.

B. *Data Augmentation*

The next step involves augmenting the data, which is a technique commonly used in computer vision tasks, such as image classification, to increase the amount of training data and improve the generalization of DL models whilst preventing overfitting. As a part of augmenting the data, random noise was added to the voice samples. During this, we specified the noise amplitude, which determines how loud the noise will be in the final signal. A random noise was then generated with the same length as the speech signal. Afterward, generated noise was added to the original speech signal to create a noisy version. Overall, adding controlled random noise to the training data can make the model more robust to these variations. It essentially trains the model to recognize the target speech even when partially masked by noise.

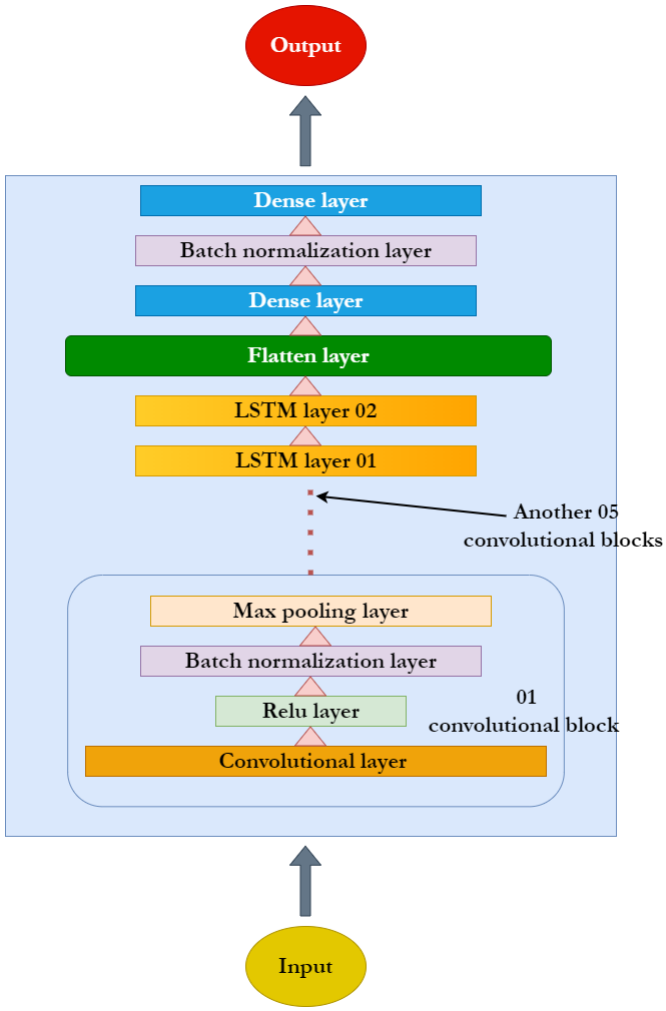

Fig. 6. The architecture of the proposed CNN model

C. *Feature Extraction*

For the extraction of the features, the Librosa library in Python was used where we have used two main features: Zero Crossing Rate (the rate of sign-changes of the signal during the duration of a particular frame) and MFCC form a cepstral representation where the frequency bands are not linear but distributed according to the Mel-scale. Once the features were

extracted, empty values were removed, and NaN values were filled with 0. Next, label encoding was done to convert emotional states into numerical form, and finally, all the data were standardized using the Python Standard Scaler method. This has been done to ensure that all features have the same scale and mean, which can help improve the performance of the DL model. Afterward, as the final step, the dataset was divided into 90% for training and 10% for testing.

D. Model Building

The final model's input layer consisted of an initial convolution layer that was composed of 512 filters and a kernel size of 3. Following the input layer is a convolutional block consisting of a batch normalization and max pooling layer. Five more blocks are included in the final model, which is subsequently composed of two LSTM layers, a flattened layer, a dense layer with batch normalization, and a dense layer employing a SoftMax activation function. In Fig. 6, the architecture of the proposed CNN model is highlighted which has been finalized after several trials with varied combinations of hyperparameters that yield the best results.

The detailed summary of the layers of the proposed model is depicted in Table 2.

TABLE II. SUMMARY OF LAYERS OF THE PROPOSED MODEL

Layer
Convolutional layer 1
Batch normalization layer 1
Max pooling layer 1
Convolutional layer 2
Batch normalization layer 2
Max pooling layer 2
Convolutional layer 3
Batch normalization layer 3
Max pooling layer 3
Convolutional layer 4
Batch normalization layer 4
Max pooling layer 4
Convolutional layer 5
Batch normalization layer 5
Max pooling layer 5
Convolutional layer 6
Batch normalization layer 6
Max pooling layer 6
Flatten layer
Dense layer
Batch normalization layer 7
Dense layer

E. Evaluation

An accuracy, precision, recall, and F1 score based on True Positives (accurately predicted positives), True Negatives (accurately predicted negatives), False Positives (erroneously predicted positives), and False Negatives (erroneously predicted negatives) were employed to assess the model's performance once model development was complete [21]-[24]. In general, accuracy signifies the comprehensive validity of a model's predictions, whereas recall evaluates the model's capacity to detect all pertinent instances and precision signifies the accuracy of positive predictions [21]-[24]. The F1 score integrates recall and precision into a solitary metric, striving for a compromise between the minimization of false positives and false negatives [21]-[24].

IV. EXPERIMENTAL RESULTS

The experiment was carried out on a personal computer with the following specifications: AMD Ryzen 9 5900x 12-Core Processor 3.70 GHz, 64GB RAM, GeForce GTX 1660 326GB VGA Memory. The hyperparameters employed in the model training are highlighted in Table 3. Overall, we have tried varied combinations of hyperparameters, where the highlighted configurations in Table 3 yield the best results.

TABLE III. HYPERPARAMETERS PERTAINING TO THE PROPOSED MODEL.

Parameter	Value
Optimizer	Adam
Batch size	64
Epochs	34 (early stopping call back used)
Learning rate	0.001
Loss function	Categorical cross-entropy

A callback function was implemented to terminate model training once the validation set reached the minimum loss threshold, where training loss decreased and showcased a linear loss after the 20th epoch, as depicted in Fig. 7. The execution of the model was terminated at the 34th epoch. After model training, it was evaluated on a separate test set, and it is evident that it reached a testing accuracy of 88.76 % on test data. The obtained performance evaluation metrics are highlighted in Table 4.

TABLE IV. PERFORMANCE EVALUATION METRICS

Parameters	Accuracy	Precision	Recall	F1 score
Values (%)	88.76	88.43	88.56	87.43

Fig. 7 illustrates how the training and validation loss varies with the number of epochs, whereas Fig. 8 demonstrates how the training and validation loss varies with the number of epochs.

Fig. 7. Training and testing loss over the number of epochs

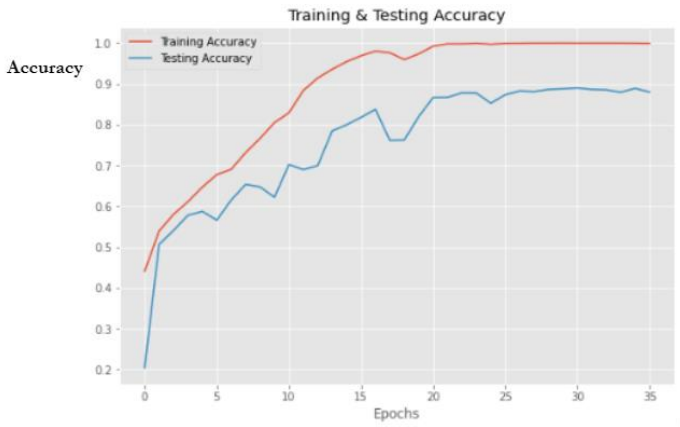

Fig. 8. Training and testing accuracy over the number of epochs

Fig. 9 showcases the confusion matrix obtained based on the final classified results, where it is evident that the majority of the emotional states are correctly classified and that minor misclassifications are there.

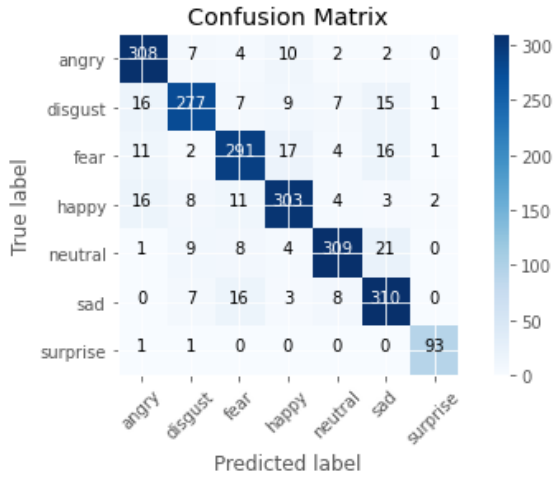

Fig. 9. Confusion matrix

Table 5 provides a brief comparison of our research findings with similar research to differentiate our work from theirs.

TABLE V. PERFORMANCE COMPARISON OF SIMILAR RESEARCH

Reference	Employed classifier	Features employed	Dataset/(s) used	Accuracy (%)
[5]	CNN (02 convolution and 02 pooling layers)	Mel spectrogram	Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset	40.00
[8]	CNN (03 convolutional and 03 fully connected layers)	Mel spectrogram	Berlin dataset	84.30
[10]	CNN + LSTM (04 convolutional layers with the LSTM network)	Short-time fourier transform	Berlin dataset	88.01 (precision only)
[13]	CNN (one convolutional layer)	Mel spectrogram	TEDLIUM v2 corpus dataset	66.10
[17]	AlexNet pretrained CNN	Mel spectrogram	Berlin database, RML audio-visual dataset, eNTERFACE05 audio-visual dataset and the BAUM-1s audio-visual dataset	87.31
Our proposed model	Modified CNN	Mel spectrogram	Crowd-sourced emotional multimodal actors dataset, toronto emotional speech dataset, and surrey audio-visual expressed emotion dataset	88.76

Overall, it is evident that our proposed novel CNN model outperforms all other models employed in recent similar research, though the methodology and the dataset employed are different. Nonetheless, even though most of the research tests their models on individual datasets or individually on different datasets, our proposed model was tested on data combined from three distinct datasets utilizing data augmenting techniques to improve the generalization of the proposed model. On the other hand, with our research, it is also evident that the choice of architecture, the number of convolutional layers, and the size of the filter kernels, can determine the CNN's ability to capture intricate features within the data and hyperparameters like the learning rate and batch size, play crucial roles in how efficiently the network converges during training and whether it generalizes well to unseen data. Thus, finding the right balance between these factors is essential for achieving optimal CNN performance which can further improve the proposed SER model performance. On the other hand, the added LSTM layers capture the temporal dependencies in the data for speech emotion recognition where, adding LSTM layers to a CNN allows the model to leverage the strengths of both convolutional and recurrent architectures, making it capable of handling complex tasks that involve both spatial and temporal information.

V. CONCLUSION

In summary, the increasing significance attributed to comprehending and interpreting human emotions is propelling the swift progression of the domain of HCI. Within this particular domain, SER has surfaced as an indispensable component, enabling computational systems to interpret human emotions accurately. Despite the significant impact that SER has had on conventional methods, promising results have been noted in recent times due to the emergence of DL-driven methodologies. By presenting a novel CNN architecture that classifies and evaluates seven distinct emotional states from speech signals, this study makes a valuable contribution to the field of SER. The noteworthy accomplishment of attaining an accuracy rate of 88.76 % serves as evidence of the potential inherent in this pioneering methodology. The key objective of

this study is to optimize the accuracy and effectiveness of emotion recognition, with wider implications in consideration, such as interactive voice response systems, mental health surveillance, and customized digital assistants. In an era where human-computer interactions must become more emotionally intelligent, the results of this study mark a substantial advancement in utilizing DL and CNNs to respond to and comprehend human emotions in a variety of technological contexts. Enhanced through continued development and refinement, this innovative SER approach has the potential to facilitate the creation of digital systems that are more responsive and empathetic, thereby elevating the standard of human-computer interactions across various domains.

ACKNOWLEDGMENT

The ASEAN IVO (http://www.nict.go.jp/en/aseanivo/index.html) project, spoof detection for automatic speaker verification, was involved in the production of the contents of this presentation and financially supported by NICT (http://www.nict.go.jp/en/index.html).

REFERENCES

[1] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," IEEE Access, vol. 7, pp. 117327–117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[2] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," Speech Communication, vol. 116, pp. 56–76, Jan. 2020, doi: 10.1016/j.specom.2019.12.001.

[3] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," Neural Networks, vol. 92, pp. 60–68, Aug. 2017, doi: 10.1016/j.neunet.2017.02.013.

[4] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," IEEE Access, vol. 9, pp. 47795–47814, 2021, doi: 10.1109/ACCESS.2021.3068045.

[5] W. Q. Zheng, J. S. Yu, and Y. X. 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), Xi'an, China: IEEE, Sep. 2015, pp. 827–831. doi: 10.1109/ACII.2015.7344669.

[6] M. Neumann and N. T. Vu, "Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech," 2017, doi: 10.48550/ARXIV.1706.00612.

[7] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," Int J Speech Technol, vol. 21, no. 1, pp. 93–120, Mar. 2018, doi: 10.1007/s10772-018-9491-z.

[8] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," in 2017 International Conference on Platform Technology and Service (PlatCon), Busan, South Korea: IEEE, Feb. 2017, pp. 1–5. doi: 10.1109/PlatCon.2017.7883728.

[9] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech Emotion Recognition Using CNN," in Proceedings of the 22nd ACM international conference on Multimedia, Orlando Florida USA: ACM, Nov. 2014, pp. 801–804. doi: 10.1145/2647868.2654984.

[10] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," in 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, South Korea: IEEE, Dec. 2016, pp. 1–4. doi: 10.1109/APSIPA.2016.7820699.

[11] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," Speech Communication, vol. 41, no. 4, pp. 603–623, Nov. 2003, doi: 10.1016/S0167-6393(03)00099-2.

[12] D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," Biomedical Signal Processing and Control, vol. 59, p. 101894, May 2020, doi: 10.1016/j.bspc.2020.101894.

[13] D. Bertero and P. Fung, "A first look into a Convolutional Neural Network for speech emotion detection," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA: IEEE, Mar. 2017, pp. 5115–5119. doi: 10.1109/ICASSP.2017.7953131.

[14] S. Parthasarathy and I. Tashev, "Convolutional Neural Network Techniques for Speech Emotion Recognition," in 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Tokyo: IEEE, Sep. 2018, pp. 121–125. doi: 10.1109/IWAENC.2018.8521333.

[15] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB: IEEE, Apr. 2018, pp. 5089–5093. doi: 10.1109/ICASSP.2018.8462677.

[16] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks," IEEE Trans. Multimedia, vol. 16, no. 8, pp. 2203–2213, Dec. 2014, doi: 10.1109/TMM.2014.2360798.

[17] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," IEEE Trans. Multimedia, vol. 20, no. 6, pp. 1576–1590, Jun. 2018, doi: 10.1109/TMM.2017.2766843.

[18] "CheyneyComputerScience/CREMA-D." Cheyney Computer Science, Dec. 20, 2023. Accessed: Dec. 21, 2023. [Online]. Available: https://github.com/CheyneyComputerScience/CREMA-D

[19] "Toronto emotional speech set (TESS)." Accessed: Dec. 21, 2023. [Online]. Available: https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess

[20] "Surrey Audio-Visual Expressed Emotion (SAVEE) Database." Accessed: Dec. 21, 2023. [Online]. Available: http://kahlan.eps.surrey.ac.uk/savee/Database.html

[21] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in Interspeech 2014, ISCA, Sep. 2014, pp. 223–227. doi: 10.21437/Interspeech.2014-57.

[22] Yi-Lin Lin and Gang Wei, "Speech emotion recognition based on HMM and SVM," in 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China: IEEE, 2005, pp. 4898-4901 Vol. 8. doi: 10.1109/ICMLC.2005.1527805.

[23] M. Jain et al., "Speech Emotion Recognition using Support Vector Machine," 2020, doi: 10.48550/ARXIV.2002.07590.

[24] S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh, "Speech emotion recognition," in 2014 International Conference on Advances in Electronics Computers and Communications, Bangalore, India: IEEE, Oct. 2014, pp. 1–4. doi: 10.1109/ICAECC.2014.7002390.