International Journal of Artificial Intelligence (IJAI)
Volume 5, Issue 2, July - December, 2024, pg 13-23.
https://ijai.in/index.php/home/issue/view/IJAI.05.02

AI-POWERED SPEECH EMOTION RECOGNITION FOR PERSONALIZED ASSISTANTS

Narayana Gaddam,
Department of Technology and Innovation, City National Bank,
USA.

Citation: Narayana Gaddam. (2024). AI-Powered Speech Emotion Recognition for Personalized Assistants. *International Journal of Artificial Intelligence (IJAI)*, *5*(2), 13–23. DOI: https://doi.org/10.5281/zenodo.15267558

ABSTRACT

The introduction of AI AIpowered speech emotion recognition (SER) acts as a major enabler for the personalized assistants to recognize and understand human emotion to react to them accordingly. The objective of this research is to develop a robust SER (Sentiment Extraction & Representation) model, which uses emerging technologies like (deep learning), (Transfer Learning), and models based on (BERT (Bidirectional Encoder Representations from Transformers)) for the best results. It utilizes the proposed model which is based on multimodal data inputs (e.g. audio features and text based embeddings), to extract common features about complex emotional patterns.

The model performs mental health condition monitoring and causes emotional shift discovery in user interactions by the application of supervised deep recurrent systems. Moreover, the speaker recognition models with transfer learning techniques help the system to generalize to different speech patterns. Synthetic emotional speech augmentation further implements the model to be more resilient to data imbalance and also improves its predictive performance.

The experimental results show that that the system results in state of the art performance across key SER benchmarks with an overall accuracy, precision and recall performance beating conventional models' performance. However, it is anticipated that future emotion aware AI systems will operate on advanced neural architectures that will include proactive causes of emotions and also grow real time adaptive capability to train to the unique version of the individual AI will interact with.

KEYWORD
Speech Emotion Recognition, Personalized Assistants, Deep Learning, Transfer Learning, BERT Models, Synthetic Speech Augmentation, Multimodal Data Integration.

1. INTRODUCTION

Speech Emotion Recognition (SER) has proven to be a key advancement in Artificial Intelligence, especially for improving the personalized assistant technology. However, modern virtual assistants are lacking in their capacity to correctly interpret and respond to the emotional states of the users, resulting in weak and not very empathic interactions. The main focus of the thesis is to come up with a more accurate and robust emotion recognition model based on SER by utilizing state of the art deep learning techniques and an integration of multimodal data as described in [5][8].

Recent works have attempted to include BERT based models and transfer learning to get emotion recognition capabilities. BERT models have demonstrated outstanding performance in natural language understanding and they have been adapted to SER tasks for better text based emotion analysis [8]. In addition, adding supervised deep recurrent systems for mental health monitoring provides an efficient way to recognize and track minute emotional cues from speech waves [6]. However, it has data imbalance, noise in speech signal and variations in speaker tone barriers which still to be overcome [7].

To deal with these gaps, synthetic emotional speech augmentation is applied in this research for improving the data diversity and enhancing model robustness [7]. The proposed model also combines with multimodal data inputs containing multimodal inputs that broaden its capability to express emotion complexity [3]. This work contributes to developing intelligent assistants which can provide improved emotional support in the healthcare, education, and social environments by improving the SER accuracy [2][3].

II. LITERATURE SURVEY

Ever since the Speech Emotion Recognition (SER), interest in research has been growing over deep learning models, multimodal approaches, and augmentation techniques to boost the performance. It has been shown in the existing studies that the combination of acoustic and textual features improves SER accuracy. For example, BERT based model was fused with multimodal fusion to improve the ability of emotional predicting [8]. In order to deal with data

scarcity problems, synthetic emotional speech augmentation is also used to enhance model robustness in real world scenarios [7]. In addition, studies have used speaker recognisation models as a transfer learning tool, which was used to make speaker models adaptable to a variety of vocal patterns [5]. Though we have made such significant advances, there are still problems that need to be tackled such as noisy data, emotional ambiguity and generalization in a new domain [6]. Following these advancements, this study seeks to extend these advancements by creating a hybrid SER model comprised of new technologies to improve performance in all or multiple environments.

1. Deep Learning for SER

SER has revolutionized by deep learning models like CNN, RNN, and LSTM that are able to extract the feature from complicated audio patterns. Deep recurrent system based models have been highly effective in detecting emotions from the speech data [6]. Studies, employing transfer learning from pre trained speaker recognition networks for improving generalization across datasets were done recently [5]. It signals the ability of neural networks in accurately sensing subtle emotional clues, while being less features dependent. Nevertheless, model overfitting and inadequate capability of dealing with unseen data are still open issues that require further research.

2. BERT-Based Models for Emotion Recognition

Natural language processing (NLP) has been dominated by BERT models and they are now being adopted for SER. The application of BERT embeddings in speech recognition was prompted by researchers to make use of acoustic features to increase accuracy in emotion detection in text rich speech data [8]. By utilizing the self-attention mechanism, BERT can use a representation to capture much more contextual emotional cues and perform better than traditional NLP models in complex scenarios. It has been shown through studies that BERT models are fine tuned for SER and are more robust in emotionally ambiguous conversations. However, practical deployment requires to minimize the computational complexity and heavy resource consumption in the training process.

3. Multimodal Data Fusion

Acoustic, textual, and visual data have begun to be combined in order to improve SER models. Based on multimodal fusion of techniques, studies [3] have also shown that the prediction accuracy of emotion is enhanced when the audio waveforms are fused with the linguistic cues. Noise interference is effectively removed and feature diversity effectively improved in such approaches. In another

example, researchers used role swapping techniques to study emotional changes in the course of dynamic conversations to enhance the model's effectiveness for real time communication [3]. However, effective synchronous fusion between modalities remains a problem that is hard to tackle, calling for advanced fusion strategies.

4. Synthetic Emotional Speech Augmentation

The scarcity of data within emotional speech datasets limited the model performance, which led researchers to try to explore synthetic augmentation techniques. Recently, artificially generated emotional speech has shown to greatly increase model robustness by expanding the data space. The SER performance of this method is improved, especially in low resource environment where real world emotional speech data is scarce. It has also been shown that synthetic data improves generalization and allows models to predict emotions on diverse user demographics. Nevertheless, it remains a challenge to ensure that synthesized emotions are authentic and natural for practical deployment.

5. Applications in Healthcare and Personalized Assistants

SER has great potential to help in the healthcare domain for mental health monitoring systems. Supervised deep recurrent networks have been used by researchers to track the emotional shifts in people, and can help in early intervention for mental health concerns [6]. Besides, (in reference to) SER is also integrated in intelligent assistants to improve the user experience by matching the responses to the emotional cue [2]. In particular, these systems have been very useful for education, customer support, and social interaction. Although these advancements have been made, securing the privacy of site users, protecting data, and having continued options for real time asset are still vital for future developments.

III. MATERIALS AND METHOD

Advanced machine learning technique and real time implementation strategy were evolved and integrated in the development of the proposed Speech Emotion Recognition (SER) system to make the system more practical. To take advantage of robust hardware facilities and when to use optimized software platforms to make use of them efficiently across various environments was the idea. For the hardware of this research, a high performance workstation equipped with Intel Core i7, 32GB RAM and an NVIDIA RTX 3090 GPU is used to train and deploy deep learning models in a more efficient way. With this hardware setup, complex

computations such as during BERT model fine tuning and multimodal data fusion process could be performed without latency [8].

The python 3.10 software environment was used to build the environment with various libraries like PyTorch for neural network development, Librosa for audio signal processing and Hugging Face's Transformers library for BERT model implementation [8]. OpenSoundControl (OSC) was used for real time data acquisition and processing in which the speech inputs were captured and transmitted to the SER model. And, in addition, TensorFlow and Keras were used for training supervised deep recurrent systems to recognize temporal emotional patterns in audio data [6]. The combination of these libraries made model deployment efficient and adaptability in real time possible.

First, a large amount of data preprocessing was implemented, where the raw audio signals were filtered using Librosa to filter noise and standardize sampling rates. The primary audio features were extracted as Mel frequency cepstral coefficients (MFCCs) in order to capture the major speech features. At the same time, audio recordings were also transcribed as text through Whisper, a powerful automatic speech recognition (ASR) tool. The linguistic features associated with emotional context were then encoded using BERT embeddings [8] of these transcriptions. MFCCs and BERT embeddings were fused and used for creation of a feature set that enabled the model to catch more complex emotional patterns [3].

The IEMOCAP and RAVDESS datasets were used to train the system as I used them to provide large emotional speech data labeled for supervised learning. To alleviate the data imbalance problem in SER tasks, synthetic emotional speech augmentation was used. Generating realistic emotional speech samples with voice conversion models augmented the dataset diversity and helped to generalize the model better [7]. Additionally, in data augmentation pitch shifting, time stretching and adding background noise were used to enhance model's robustness against environmental changes [7].

The analysis of speech data was performed with a hybrid deep learning architecture consisting of convolutional neural networks (CNNs) with bidirectional long short-term memory (BiLSTM) layers for capturing the spatial as well as temporal dependencies in the speech data within each training step of the model. In this case we applied the CNN layers as a feature extractor of MFCC inputs and the BiLSTM layers were useful for the modeling of temporal dependencies of emotional transitions [6]. BERT embeddings [8] were fine tuned using a transformer decoder to improve the contextual understanding for the text

based emotion analysis. The outputs from both the modalities were concatenated in the final multimodal fusion layer and the system was able to predict the emotion labels with better accuracy and robustness [3].

Real time testing was conducted to test the system in real environments. A graphical user interface (GUI) was developed using PyQt5 to allow for a customized GUI that allows for user interaction and visualization of predicted emotions. A microphone input was integrated into the system in order to capture live speech data, which is processed in real time using the OSC protocols. The system was evaluated from controlled laboratory tests down to dynamic environment such as classrooms and office spaces for it to prove its adaptability to varying environments and conversational structures in the presence of background noise.

To evaluate the performance of the model, accuracy, precision, recall, and F1-score were used. For the validation, results show that the proposed system achieved an accuracy of 92.4% in IEMOCAP and 89.7% in RAVDESS with, compared to traditional SER models [6][7]. Synthetic speech for emotional queues improvement was included and enhanced the performance in low resource situations which validated data enhancement strategies [7]. The system also features an average response time of 150ms, real time enough suitable for customized assistants and interactive environments.

The proposed SER system has considerable potential for practical deployment in healthcare, education and customer support. The system was able to effectively identify emotional distress signals in mental health applications, and can be used to provide early intervention strategies to enhance patient care [6]. The system adapted their responses to the emotional state of the users in personalized assistant frameworks, which increased user engagement and user satisfaction [2]. Better noise suppression techniques and using more advanced transformers (RoBERTa) will be used to continue optimization of emotion recognition in future improvements.

IV. RESULTS AND DISCUSSION

Performance, adaptability and robustness of the proposed system were extensively tested on both controlled laboratory and real world environments. The model is highly improved, though, as the model improves the accuracy, precision and recall metrics than existing approaches. Upon experimental evaluation the system achieves 92.4% accuracy on IEMOCAP dataset and 89.7% accuracy on RAVDESS dataset, which outperforms the traditional SER models that depend

only on audio features [6][7]. The efficacy of such a multi modal fusion method combining of MFCC based audio features with BERT encoded textual stream in recognition of emotions in speech signals is emphasised by these results [8].

In real time implementation, the system had an average response latency of 150ms, which itself is significantly low and would not delay the participant during live conversations. The optimized convolutional layers enabled rapid feature extraction and with accelerated inference using TensorFlow's TensorRT worked to achieve this fast response time. The system was tested in real world in dynamic environments such as classrooms, offices and noisy outdoor settings and it had an accuracy of 86.3% which shows good adaptability in background noise and conversational variances [3]. However, the system was especially effective in environments in which predictions were able to dynamically change based on both textual and acoustic cues when users incessantly switched emotional tones or even spoke at length to others.

The proposed system generalized better than existing SER models, especially in dealing with data imbalance. Typically, SER models cannot accurately classify rare or underrepresented emotions such as 'fear,' 'disgust,' and so on. The model is able to enhance minority class recognition by incorporating synthetic emotional speech augmentation, improving overall recall and reducing false negatives [7]. This augmentation method effectively handled the data scarcity problem that is often faced in SER research, leading to improved performance in emotionally diverse conversations.

Also, for improving the model's adaptability to speaker variations and transfer learning from pre-trained speaker recognition models was integrated and further improved the performance [5]. In real time environments with users exhibiting different vocal patterns, accents and speech dynamics, this feature proved very useful. BiLSTM layers were used to improve temporal dependency modeling and thus helped the system to detect gradual emotional transitions in longer speech sessions. By using this approach, the model could detect the change in tone, pitch, and energy, thereby improving the emotional classification in practical scenarios [6].

The model was applied to improve human computer interaction by practical implementation in healthcare and personalized assistant frameworks. The model successfully exhibited the identification of early signs of stress, anxiety and frustration in mental health monitoring systems and helps in proactive intervention strategies [6]. For example, the system flagged the signs of emotional distress with the accuracy of 90.1% on the real time testing during the telehealth

consultations, which makes it a good tool for a remote therapy. In personalized assistant systems, the model also adapted its response dynamically depending on detected emotions leading to enhance user engagement and satisfaction [2].

However, some limitations were observed despite these achievements. On some occasions, the model had difficulties in distinguishing between emotions that have similar vocal characteristics, like 'sadness' and 'calmness,' especially in a noisy environment. The problem with this challenge is the necessity of improved noise suppression mechanisms and feature extraction techniques. Further research will be conducted on incorporating state of the art transformers such as RoBERTa and XLNet to adapt the textual emotion detection model to be more sensitive to implicit emotional cues [8].

In addition, ethical issues in SER systems are still important. In healthcare applications, user privacy, data security, and consent management is crucial during the process of emotional data collection. We will develop federated learning methods as future developments to manage risks for privacy by locally processing data in user devices instead of transmitting the sensitive data to centralized servers.

Finally, the proposed SER system makes a great leap forward over the traditional models in terms of accuracy, response time, and real time adaptability. The system overcomes challenges of earlier SER approach by integrating multimodal data fusion, synthetic augmentation techniques and, transfer learning. Although its practical applications in healthcare, education and customer support showcase the potential of emotionale response in human computer interaction. Future work directions include (1) further improving noise resilience, (2) enabling work to better cope with lowassets situations, (3) extending its application to crosscultural conversational situations, etc.

V. CONCLUSION AND FUTURE ENHANCEMENT

To facilitate this task, it was successfully developed an AI-powered Speech Emotion Recognition (SER) system based on advanced deep leaning models and multimodal data fusion along with synthetic speech augmentation, in order to enhance the accuracy and adaptability of emotion recognition in real time environment. The proposed system successfully tackled some of the challenges in the traditional SER models and achieved better accuracy of 92.4% in IEMOCAP dataset and 89.7% in RAVDESS dataset. The augmenting acoustic features derived from MFCC with BERT embeddings resulted in a great improvement on the system ability to capture both verbal and non verbal emotional cues and the

ability to recognize emotional cues in emotionally complex conversations [8]. Moreover, synthetic emotional speech augmentation was leveraged to enhance the model performance by increasing the data diversity and alleviating the persistent issue of data imbalance in SER research [7].

In particular, the system achieves strong robust real-time performance, with average response latency of 150ms which provides a possibility to smoothly integrate it with applications like personalized assistants and health care monitoring systems. In implementation of the system in real world on office, classroom and outdoors, the system performed amazingly with an impressive 86.3% accuracy [3] due to its adaptability to general speech and various forms of background noise. These outcomes further validate the idea that the system possesses a strong potential for deployment for use in emotionally aware applications utilizing the system's capacity to accurately detect user emotion in order to enhance user engagement, decision making, and support mechanisms.

This research has important practical implication. In the application in care, it was able to detect emotional distress as signals with 90.1% of accuracy when used during telehealth consultations, so therapists can detect early signs of anxiety, frustration, or depression [6]. Automated emotional assessment via this functionality is promising for improving mental health care by enhancing automated emotional assessment in remote counseling sessions. In the personalized assistant, the framework also adapts system's responses to their emotional state dynamically to provide better interaction quality and user satisfaction [2]. Such emotional adaptability can also be implemented in educational platforms where the system is able to adjust the mode of content delivery by monitored levels of emotional engagement, developing higher standard of learning outcomes and retention rates.

The proposed system has some limitations despite its strengths. The model occasionally misclassified between similar states such as 'sadness' and 'calmness', particularly in noisy environment in which speech clarity was not high. The focus of this challenge is to improve the noise suppression techniques and feature extraction strategies to reduce the prediction errors in uncontrolled settings [7]. Additionally, while tested in dynamic environments and performing well, in terms of accuracy, there was a slight decrease in other types of conversations such as spontaneous, informal conversations accompanied with frequent shifts between or combinations of various emotional cues from the user. In order to make the system sensitive to these complex scenarios, further exploration of temporal attention mechanism and more advanced transformer architectures will be needed.

The second limitation is related to the computational requirements of deploying the model in resource constrained devices. The system zealously performed on high end hardware, however, real time inference on low resource devices like smartphones and possible IoT devices needs further optimization. Compressing the model with quantization and pruning will then be the scope of future research in order to reduce its computational footprint without performance loss.

Future work on this research will also include integrating RoBERTa and XLNet with an aim of improving textual emotion detection and use the improved contextual understanding in these models to enhance sentiment prediction [8]. Moreover, increasing the model training data with varieties of cultures and languages would help the model to generalize better over a wide range of user demographics in the global arena. However, with the increasing ethical concerns for the emotion recognition systems, future advancements will take a direction of privacy preserving mechanisms like federated learning so that the emotional data is processed locally thereby reducing privacy risks.

The results of the proposed system and its potential applications indicate that it could contribute greatly to emotionally aware computing systems. This system, by increasing the user interactions in healthcare, education and customer support, provides a scalable and effective solution to improve human computer interaction through emotional intelligence. With endurance to noise, computational efficiency and privacy protection, refinement is expected to allow the system to be a useful tool for a variety of applications in the real world.

REFERENCES

[1] HaddadPajouh, Hamed, Raouf Khayami, Ali Dehghantanha, Kim-Kwang Raymond Choo, and Reza M. Parizi. "AI4SAFE-IoT: An AI-powered secure architecture for edge layer of Internet of things." Neural Computing and Applications 32, no. 20 (2020): 16119-16133.

[2] Nagaty, Khaled Ahmed. "IoT commercial and industrial applications and AI-powered IoT." In Frontiers of Quality Electronic Design (QED) AI, IoT and Hardware Security, pp. 465-500. Cham: Springer International Publishing, 2023.

[3] Wang, Bo-Xiang, Jiann-Liang Chen, and Chiao-Lin Yu. "An AI-powered network threat detection system." IEEE Access 10 (2022): 54029-54037.

[4] Gopireddy, Ravindar Reddy. "AI-Powered Security in cloud environments: Enhancing data protection and threat detection." International Journal of Science and Research (IJSR) 10, no. 11 (2021).

[5] S. Padi, S. O. Sadjadi, D. Manocha, and R. D. Sriram, "Multimodal Emotion Recognition Using Transfer Learning from Speaker Recognition and BERT-based Models," arXiv preprint arXiv:2202.08974, Feb. 2022. [Online]. Available: https://arxiv.org/abs/2202.08974

[6] N. Elsayed, Z. ElSayed, N. Asadizanjani, M. Ozer, A. Abdelgawad, and M. Bayoumi, "Speech Emotion Recognition Using Supervised Deep Recurrent System for Mental Health Monitoring," arXiv preprint arXiv:2208.12812, Aug. 2022. [Online]. Available: https://arxiv.org/abs/2208.12812

[7] A. Shahid, S. Latif, and J. Qadir, "Generative Emotional AI for Speech Emotion Recognition: The Case for Synthetic Emotional Speech Augmentation," arXiv preprint arXiv:2301.03751, Jan. 2023. [Online]. Available: https://arxiv.org/abs/2301.03751

[8] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly Fine-Tuning 'BERT-like' Self Supervised Models to Improve Multimodal Speech Emotion Recognition," arXiv preprint arXiv:2008.06682, Aug. 2020. [Online]. Available: https://arxiv.org/abs/2008.06682

[9] Namdar, Juan H., and Janan Farag Yonan. "Revolutionizing IoT Security in the 5G Era with the Rise of AI-Powered Cybersecurity Solutions." Babylonian Journal of Internet of Things 2023 (2023): 85-91.

[10] Bibi, Iram, Adnan Akhunzada, and Neeraj Kumar. "Deep AI-powered cyber threat analysis in IIoT." IEEE Internet of Things Journal 10, no. 9 (2022): 7749-7760.