Decoding Digital Emotions: Advancing Online Learning with Speech-Emotion Recognition Systems

Sherif Welsen(✉) and Yiyang Liu

The University of Nottingham Ningbo China, Ningbo, Zhejiang, China
Sherif.welsen@nottingham.edu.cn

Abstract. This chapter introduces a novel system tailored for emotion recognition in speech within online educational platforms. Developed using MATLAB, this system harnesses cutting-edge machine learning methodologies, employing datasets from the Berlin Database of Emotional Speech (EmoDB) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Precise detection and categorization of emotional expressions in speech are made possible by the hybrid model it employs, which combines Long Short-Term Memory (LSTM) networks with one- and two-dimensional convolutional neural networks. This system effectively improves the interpretation of student emotions in virtual learning contexts, achieving an impressive accuracy rate of 83.95%. However, it is important to note that this work also underscores the necessity for ongoing research to further refine the system's performance and dependability. This endeavour marks a crucial advancement in customizing online education, aiming to foster more empathetic and engaging virtual learning environments.

Keywords: Smart Campus · Speech Emotion Recognition (SER) · Convolution Neural Networks · Intelligent Tutoring · Long Short-Term Memory

1 Introduction

Speech is the fundamental basis of human contact, facilitating the seamless flow of thoughts and emotions [1]. Since the 1960s, significant strides have been made in speech recognition technology, enabling computers to convert spoken language into written text [1]. This technology has permeated various sectors, revolutionizing interactions through applications such as virtual assistants, voice-operated controls, real-time translation services, and customer support systems. Consequently, it has bolstered efficiency and paved the way for a more intelligent way of life [2].

Despite the many advancements in speech recognition technology, traditional systems primarily focus on the linguistic aspects of speech and often overlook the emotional nuances essential for human communication. This recognition of emotional under-tones has given rise to the emergence of Speech Emotion Recognition (SER), a field dedicated to interpreting emotions such as joy, anger, or sadness from vocal expressions. As a result,

E. C. K. Cheng (Ed.): AETS 2024, LNET, pp. 139–151, 2025.
https://doi.org/10.1007/978-981-96-4952-5_10

this enriches interactions by transcending mere words [3]. Modern Human-Computer Interaction (HCI) relies heavily on SER, which provides creative solutions like tracking drivers' emotional moods to avoid collisions [4]. SER goes beyond theoretical applications and has practical implications across various domains. It improves customer service by allowing representatives to assess caller emotions and respond empathetically [4].

By allowing tutors to identify and address students' emotional needs, SER technology significantly improves the academic experience in educational environments, especially those that are online [4]. Through real-time evaluation of emotional states and speech patterns, SER offers crucial insights that can greatly enhance the learning environment [5]. This technology extends its benefits to evaluating instructors' emotional stability, which directly influences the efficacy of their teaching and the quality of information delivery. Additionally, recognizing learners' emotions facilitates the customization of teaching strategies to accommodate individual learning preferences and emotional states, thereby fostering a supportive atmosphere that boosts student motivation and engagement [6, 7]. Moreover, SER allows for continually monitoring and adjusting teaching methods and resources based on real-time feedback regarding students' emotional conditions, empowering educators to refine their online instructional approaches [8]. However, despite its substantial potential, the successful implementation of SER within educational frameworks faces challenges that require ongoing innovation and adaptation to realize its benefits fully [3].

The work presented in this chapter aims to develop an innovative system for recognizing emotions in speech tailored explicitly for online education. This system, resulting from a final-year project-based initiative [9], is more than just an enhancement; it is an extra crucial tool designed to transform how personal tutors understand and respond to their students' emotions within virtual learning environments. The process of extracting emotions from the speech signal is carefully crafted to support online tutoring, making it an invaluable resource for students engaged in remote learning. Integrating an emotional extraction mechanism enables personal tutors to gain deeper insights into their students' perceptions and emotional states [10–12]. This capability is essential for maintaining understanding and pastoral care during personal tutoring sessions and adapting to the dynamic educational shifts brought about by global challenges [13].

As a result, this chapter is dedicated to exploring the various challenges associated with the SER and evaluating both traditional and contemporary methodologies. The goal is to identify the most effective strategies for implementing SER in online teaching platforms, ultimately enhancing emotional connectivity and the overall responsiveness of virtual learning environments. The subsequent sections of this chapter are organized as follows: Sect. 2 provides a concise review of the literature concerning feature extraction, classification, and both machine and deep learning within the context of SER. Section 3 outlines the proposed system's workflow, and Sect. 4 offers a detailed overview of the proposed emotional extraction system. The findings and analytical discussions are explored in Sects. 5 and 6, respectively, with Sect. 7 concluding the study.

2 Literature Review

Identifying emotions through speech involves two key steps: feature extraction and feature classification [14]. In the first stage, a continuous signal, speech varies significantly across different utterances, each containing unique emotional cues. Researchers have extensively investigated various aspects of speech to capture this rich emotional content effectively. Many studies have analyzed acoustic and speech quality features to develop a comprehensive framework for representing emotions within speech [15–17]. Acoustic features typically include elements such as formants, which are concentrated energy bands in the vocal spectrum, as well as short-term energy and fundamental frequency reflecting pitch and intensity dynamics in speech [18]. On the other hand, speech quality features comprise characteristics like pitch variation, speech rate, and volume, each playing a vital role in conveying emotional states [19]. Illustratively depicted in Table 1, these diverse speech features are meticulously analyzed to align with corresponding emotional expressions, enabling the nuanced detection and classification of emotions. This systematic approach enhances emotion recognition accuracy and enriches the interface between humans and technology by providing a more intuitive understanding of emotional subtleties in spoken language.

Table 1. Emotion versus speech feature (recreated from [3])

Emotions	Pitch	Intensity	Speaking rate	Voice quality
Anger	abrupt on stress	much higher	marginally faster	Breathy, chest
Disgust	wide, downward inflexions	lower	much faster	irregular voicing
Fear	wide, normal	lower	much faster	irregular voicing
Happiness	much wider, upward inflexions	higher	faster/slower	breathy, blaring tone
Joy	High mean, wide range	Higher	Faster	Breathy, blaring timbre
Sadness	slightly narrower	downward inflections	lower	resonant

The next step involves creating classifiers divided into linear and non-linear types. These classifiers are essential for identifying emotional cues in speech [3]. Linear classifiers, such as Perceptrons, Linear Regression, and Logistic Regression, typically organize data along a straight line to categorize emotions [20]. On the other hand, non-linear classifiers use complex hyper-surfaces and handle multi-dimensional data, including Decision Trees, Gaussian kernel SVMs, and Gradient Boosting Decision Trees (GBDT) [20]. These non-linear classifiers are particularly effective in the SER domain because they can manage speech signals' inherent complexity and variability. Non-linear classifiers have a significant advantage over linear classifiers. They can be used in three-dimensional and multi-dimensional contexts, unlike linear classifiers. Different classification methods have been explored in the domain of SER, as outlined in Table 2. Linear classifiers

are limited in their ability to handle the complexity of speech signals, while non-linear classifiers can effectively address the intricacies and variations in speech. Studies have confirmed non-linear classifiers' superior efficacy and accuracy [3, 21, 22]. However, there is room for improving their overall performance and adaptability in practical settings [22]. For example, when an SVM classifier detects three emotions, it achieves an accuracy rate of about 90%. However, when the emotional range expands to seven, SVM integration with decision tree techniques is required to maintain high accuracy, achieving a peak accuracy of 82.9%. The effectiveness of these classifiers varies significantly across different languages and speech databases, with some studies recording accuracies as low as 56.25% [22]. This highlights the need for ongoing improvements in classifier performance across diverse linguistic contexts.

Table 2. Various Classifiers used for Speech Emotion Recognition (SER)

Classifiers	Linear or Non-linear	Reference
SVM classifier	Linear or Non-linear	[20]
PCA classifier	Linear or Non-linear	[23]
ELM classifier	Linear or Non-linear	[3]
HMM classifier	Non-linear	[24]
GMM classifier	Non-linear	[24]
Bayes classifier	Linear	[3]
K-Nearest Neighbor classifier	Linear	[3]

Recent advancements have shifted towards deep learning in SER, which eliminates the need for manual feature extraction and classification, unlike traditional machine learning methods [3]. Popular deep learning architectures like Recurrent Neural Networks (RNN), Long Short-Term Memory Networks (LSTM), and Convolutional Neural Networks (CNN) have proven highly effective in advancing SER capabilities [25]. Deep learning models have achieved remarkable recognition accuracies, reportedly as high as 97.1% for identifying four distinct emotions [22].

However, traditional machine learning still holds value, mainly due to its cost-effectiveness and simplicity, despite its lower precision compared to deep learning. Deep learning boasts higher scalability and potential accuracy but at a higher complexity and cost [3]. The ongoing challenge within SER is to find an optimal model or classifier that balances high accuracy with practical applicability, especially for online educational platforms. Moreover, the SER field grapples with limited datasets that lack real-time relevance, typically recorded under controlled conditions without considering environmental variables like background noise. This limitation presents a clear and distinct

difference when compared to other fields of machine learning, such as image processing or speech recognition, which have databases that consist of millions of samples.

3 Proposed System Workflow

The suggested system's workflow, as shown in Fig. 1, is divided into three main steps to enhance speech emotion recognition. The initial stage, Data Input, focuses on collecting and carefully observing speech data, ensuring its quality and relevance for subsequent analysis. Transitioning into the second stage, the process shifts to feature extraction and generating feature vectors suitable for machine learning. This involves assigning emotional labels to the speech samples, refining the dataset by removing or adjusting data based on emotional relevance, standardizing and preparing data for analysis, enhancing the dataset's size and variability through data augmentation, and extracting key features indicative of emotional states. The final stage, Machine Learning, utilizes the prepared data to train the model to recognize and classify emotions, followed by testing the model with a separate data set to verify its accuracy in identifying emotions across diverse scenarios. This streamlined workflow is designed to ensure that the system is both robust and reliable in real-world applications.

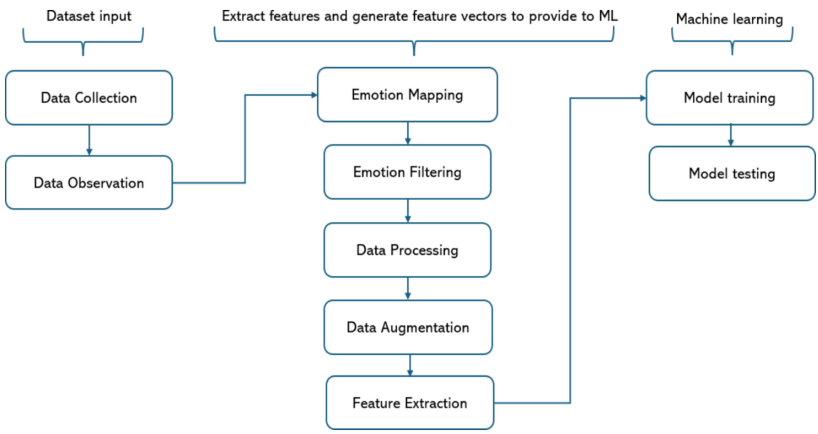

Fig. 1. Speech Emotion Recognition Workflow

The effectiveness of the model was validated using two established datasets renowned in the Speech Emotion Recognition (SER) arena: the German Emotional Speech (EmoDB) dataset and the Ryerson Audio-Visual Database of Emotional Speech and Song (Ravdess) dataset. These datasets are pivotal for enabling comparative studies with existing research. The EmoDB dataset features contributions from ten professional actors—equally divided between males and females—who recorded 535 spoken phrases depicting common conversational emotions, including anger, boredom, anxiety, happiness, sadness, disgust, and neutrality. Initially captured at a sampling rate of 48 kHz,

these recordings were subsequently down-sampled to 16 kHz to meet processing standards [26]. Similarly, the Ravdess dataset features contributions from twenty-four professional actors, evenly split between males and females, who performed eight distinct emotions. These emotions include neutral, calm, happy, sad, angry, fearful, disgusted, and surprised expressions. This dataset contains 1440 speech files, establishing a solid foundation for detailed emotional analysis throughout the project [27]. This comprehensive collection of emotional speech files significantly enhances the model's training and validation phases, ensuring a thorough assessment of its performance across varied emotional contexts. The distribution of emotion classes across the two datasets is illustrated in Fig. 2.

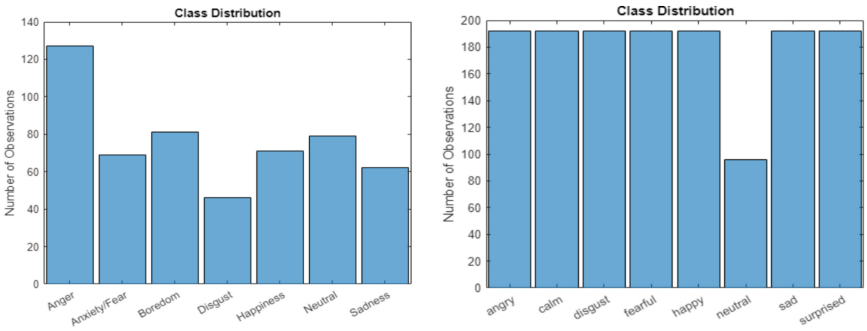

Fig. 2. Emotion classes distribution of EmoDB dataset (left) and Ravdess dataset (right)

4 Proposed Emotion Detection System Development

The emotions in the two datasets were initially focused on everyday communication contexts. However, due to the educational focus of this project, it was necessary to adapt these datasets to suit academic settings better. This adaptation was crucial for enhancing the project's relevance and improving the model's applicability in educational scenarios. As a result, the emotional spectrum was expanded to include specific categories: Disgust was redefined as Dislike and Anxiety or Fear was redefined as Confusion. A more detailed range of Curiosity to Happiness, while Anger was redefined as Frustration [28–30]. This expansion re-fined the specificity of emotion recognition and ensured a more nuanced understanding of emotional cues in educational contexts. Each audio file in the dataset was carefully tagged with the appropriate sentiment to align the emotional content with the label property of the audio datastore object, thereby enhancing the system's accuracy.

In order to maintain consistency between the two datasets, the duration of the audio snippets was standardized to four seconds by either clipping or padding the files. Only the data from the initial channel was utilized for the audio files in the Ravdess dataset, which were recorded using binaural technology. Although efforts were made, the datasets still presented issues due to their small size and uneven distribution of emotion categories. In order to address this issue, a total of 100 different data augmentation approaches

were implemented on each audio file, therefore improving the model's capacity to generalize [31]. These augmentations involved normalizing the audio to reduce amplitude variations, thus making model training more stable. Additionally, parallel processing was used to more efficiently handle data partitions, which helped speed up the overall processing time.

The audio clips are segmented into frames that last 128 ms each, with a 32-ms overlap between frames. Each frame is processed using a Hamming window function to reduce the impact of signal boundary effects. The features selected include pitch, intensity, and Mel spectrogram. Pitch values are extracted for each frame, creating a sequence that varies over time. Similarly, intensity values are calculated for each frame, generating another dynamic sequence. The pitch and intensity data are structured as one-dimensional arrays. Conversely, the Mel-Spectrum is represented as a matrix with two dimensions. This matrix is then averaged over time to create a vector with only one row. The vectors are merged to create a comprehensive feature vector for machine learning, guaranteeing uniform length across all feature vectors.

This study utilizes a 1D Convolutional Neural Network (CNN), which manages structured data and is particularly effective in extracting temporal details from audio signals [31]. The network architecture includes several key components:

Input Layer: Receives time series data structured into 240 features.

Convolutional Layers: This model comprises two 1D convolutional layers. The initial layer contains thirty-two filters of size five, utilizing causal padding to align outputs with inputs and prevent the leakage of future information. The second layer increases to 64 filters, maintaining the size and causal padding.

Activation and Normalization: Each convolutional layer is followed by a ReLU activation function to add nonlinearity and a normalization layer to accelerate training and stabilize the model.

Pooling Layer: This layer employs global average pooling (GlobalAveragePooling1D) across the time dimension, reducing the feature set to one dimension, which minimizes parameters and combats overfitting.

Fully Connected and Output Layers: The network progresses to a fully connected layer that connects to a classification layer. The number of categories within the training set determines the size of the classification layer. The network employs a softmax function to provide a probabilistic distribution of the predicted categories.

Classification Layer: This final layer calculates the loss and predicts classes according to the softmax output.

Figure 3 shows the model's architecture. The training regime included the Adam optimizer for up to 200 iterations with an initial learning rate of 0.01, employing left-padding to standardize input lengths and a mini-batch size of 27. For the training of a 2D CNN, the complete Mel-Spectrum is utilized as input, acknowledging its critical role in enhancing classification accuracy. Enhancements to this hybrid model include reducing the maximum number of batch sizes to thirty-two and an adjusted initial rate of learning to 0.005 after the first two training cycles, alongside the integration of L2 regularization to curb complexity and prevent overfitting. Training adjustments are monitored using a validation dataset during each epoch to ensure optimal model performance. This approach is incredibly potent in handling the dynamic nature of audio data, capitalizing

on LSTM networks to capture extensive temporal dependencies and localized features crucial for effective emotion recognition.

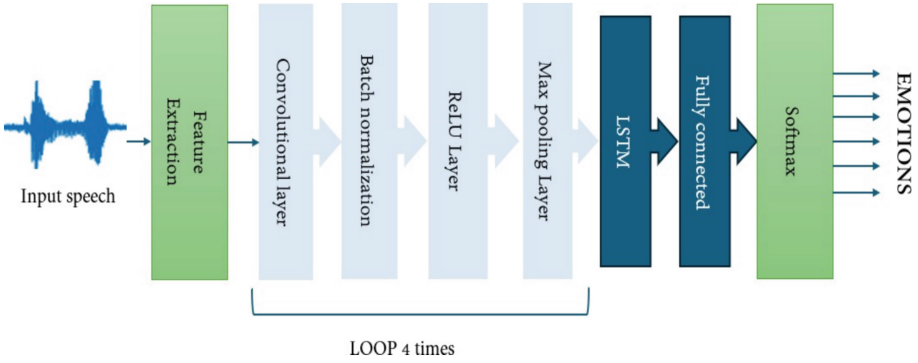

Fig. 3. The proposed 2DCNN LSTM model architecture

5 Results

The results presented in this study involved systematically assigning emotion codes to label audio files from the dataset through manual verification. Additionally, the number of files within different emotional categories was meticulously counted and cross-verified against the dataset specifications to confirm the precision of the emotion mapping process. An audio file for each emotion was randomly selected during the pre-processing phase to generate graphs in the time domain, frequency domain, and Mel-Spectrum. For the 1D Convolutional Neural Networks (CNNs), the pitch, intensity, and Mel-Spectrum vectors must be aligned as row vectors with uniform column counts. This project utilizes two main types of features for network training: the feature matrix and the label vector. The feature matrix contains the attributes of the audio files, while the label vector categorically encodes the emotional labels for each audio sample, providing a clear emotional context. The feature matrix is structured as an n x 1 cell array, with each cell containing a double-type matrix. On the other hand, the label vector is organized as an n x 1 categorical array, with each row representing an emotional category. The input requirements differ between the two models: two-dimensional for a 1D CNN and three-dimensional for a 2D CNN. This setup ensures precise alignment of the feature matrix and label vector, as illustrated in Table 3.

The integrated datasets were segmented into training, validation, and test sets in ratios of 70%, 15%, and 15%, respectively. The analysis utilized a confusion matrix to assess the model's performance against actual test data labels, focusing on accuracy, precision, recall, and F1 scores as key metrics. Initially, the best CNN 1D model achieved approximately 70% accuracy. By adopting a more sophisticated model and optimizing training through increased epochs and early stopping, the validation accuracy improved to 83.95%, with overall model accuracy reaching 78%, illustrated in Tables 4 and 5. The data of the confusion matrix is illustrated in the form of the scores of Precisions, Recall, and F1 performance indicators, as described in Table 6.

Table 3. Feature matrix and Label vector for 1D CNN and 2D CNN

	1D CNN		2D CNN	
#	Feature	Label	Feature	Label
1	200 × 1 double	7 Neutral	128 × 1 × 56 double	7 Neutral
2	200 × 1 double	1 Frustration	128 × 1 × 56 double	7 Neutral
3	200 × 1 double	4 Confused	128 × 1 × 56 double	7 Neutral
4	200 × 1 double	7 Neutral	128 × 1 × 56 double	7 Neutral
5	200 × 1 double	6 Sadness	128 × 1 × 56 double	4 Confused
…	…	…	…	…
n	200 × 1 double	7 Neutral	128 × 1 × 56 double	6 Sadness

Table 4. Confusion Matrix - CNN 2d LSTM; No. DA = 100; Epochs = 3;

	Frus.	Bore.	Disl.	Conf.	Happ.	Sadn.	Neut.
Frus.	87.3%	0%	0.0%	0.0%	5.3%	0.0%	7.4%
Bore.	0.0%	67.6%	8.3%	0.0%	8.3%	0.0%	15.8%
Disl.	0.0%	2.5%	65.0%	0%	15.0%	0.0%	17.5%
Conf.	9.3%	0.0%	10.0%	90.4%	0.0%	0.0%	0.0%
Happ.	19.1%	0.0%	0.0%	0.0%	80.9%	0.0%	0.0%
Sadn.	0.0%	11.1%	11.1%	11.1%	0.0%	55.6%	11.1%
Neut.	0.0%	31.7%	0.0%	0.0%	0.0%	0.0%	68.3%
Acc. 74.4%							

Table 5. Confusion Matrix - CNN 2d LSTM; No.DA = 100; Epochs = 10; EarlyStop

	Frus.	Bore.	Disl.	Conf.	Happ.	Sadn.	Neut.
Frus.	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Bore.	0.0%	83.3%	8.3%	0.0%	8.3%	0.0%	0.0%
Disl.	14.3%	0.0%	85.7%	0.0%	0.0%	0.0%	0.0%
Conf.	10.0%	0.0%	0.0%	80.0%	10.0%	0.0%	0.0%
Happ.	0.0%	0.0%	27.3%	0.0%	72.7%	0.0%	0.0%
Sadn.	11.1%	0.0%	0.0%	0.0%	0.0%	88.9%	0.0%
Neut.	33.3%	0.0%	0.0%	0.0%	0.0%	0.0%	66.7%
Acc. 78%							

Table 6. Precision, Recall, and F1 Scores

Class	Precision	Recall	F1 Score
Frustration	59.3%	100.0%	74.4%
Boredom	100.0%	83.4%	90.9%
Dislike	70.7%	85.7%	77.5%
Confusion	100%	80.0%	88.9%
Happiness	79.9%	72.7%	76.1%
Sadness	100.0%	88.9%	94.1%
Neutral	100.0%	66.7%	80.0%

6 Discussion

Figure 4 illustrates the training progression chart linked with the confusion matrix detailed in Table 4. Over approximately 3500 iterations in a single GPU setup, the model underwent three comprehensive training cycles. The training accuracy, depicted by a blue curve, notably escalates from about 10% to a stable rate near 80%, with a smoothed accuracy curve aiding in tracking these consistent upward trends. This demonstrates the model's capacity to learn from the data incrementally. Conversely, the verification accuracy begins higher and remains relatively constant, peaking at 81.48%, suggesting effective prevention of overfitting.

The training loss experiences a sharp decrease initially, levelling outpost the first epoch consistently below 0.5, indicative of a robust model fit. Notably, training loss consistently stays below validation loss, reflecting the model's solid generalization to new data. Eventually, the loss plateaus, hinting at convergence to a potentially optimal solution. The model incorporated early stopping and batch normalization strategies to mitigate overfitting, though a slight gap between training and validation accuracies suggests some overfitting. Enhancements in early stop-ping routines and expanded use of cross-validation methods could further bolster model generalization [32].

Turning to specific emotion recognition, as depicted in Table 5, the model demonstrates 100% precision in identifying "boredom," confidently distinguishing it from other emotions. However, with an 83.4% recall rate, some "bored" in-stances are mislabeled, often confused with "dislike" and "happiness," underscoring the overlapping vocal characteristics of these emotions. Despite these challenges, the model's F1 score of 90.9% underscores a strong balance between precision and recall, reflecting reliability. Future enhancements could focus on improving recall to boost this F1 score, optimizing both accuracy and consistency in real-world scenarios. Comparative studies, like those by Wu [33] and Huang [34], which achieved accuracies of 75.5% and 85.2% using the EmoDB dataset, further contextualize this model's performance within the field.

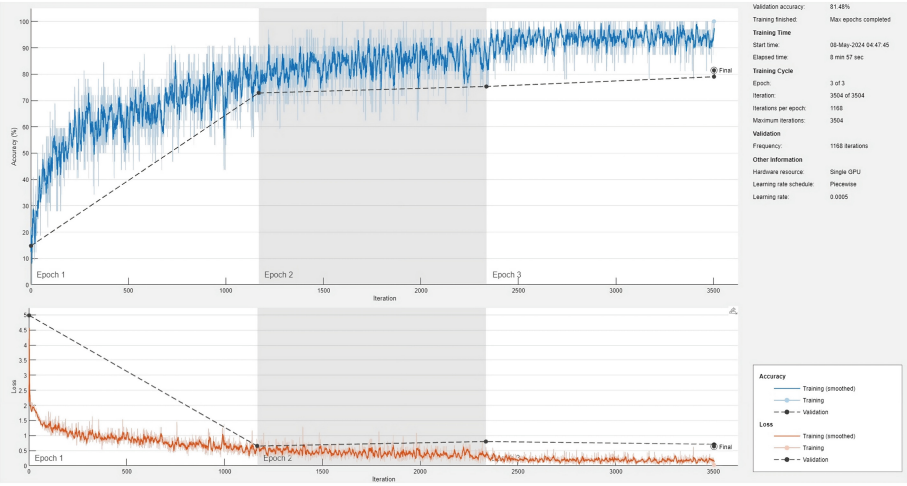

Fig. 4. The training progress chart related to Table 4

7 Conclusion

This chapter concludes a final-year project-based module that effectively deployed one-dimensional convolutional neural networks alongside advanced deep learning structures combining CNNs and LSTM networks to develop SER systems in online educational environments. Utilizing well-known datasets such as EmoDB and Ravdess, the project achieved a validation accuracy of up to 83.95%. The results underline the potential of these technologies to identify complex emotional states and adapt to real-world settings accurately. However, the advancement of SER systems must carefully address ethical and privacy issues, particularly as they handle sensitive personal data that could be misused if protections are inadequate. Ensuring robust data protection and clear data handling policies is essential.

Additionally, the possibility of inherent biases must be considered. Without diverse training datasets—including variations in race, gender, linguistic accents, and cultural and emotional expressions—SER systems risk perpetuating existing social biases and delivering discriminatory outcomes. Future initiatives might incorporate unsupervised learning to leverage unlabeled data, thus enhancing SER systems' versatility across various languages and cultural backgrounds. Expanding the range of emotional states in datasets and advancing real-time processing capabilities could further refine the practicality and impact of SER technologies.

References

1. Gaikwad, S.K., Gawali, B.W., Yannawar, P.: A review on speech recognition technique. Int. J. Comput. Appl. **10**(3), 16–24 (2010)
2. Yu, Y.: Research on speech recognition technology and its application. In: 2012 International Conference on Computer Science and Electronics Engineering. IEEE (2012)

3. Khalil, R.A., et al.: Speech emotion recognition using deep learning techniques: a review. IEEE Access **7**, 117327–117345 (2019)
4. Schuller, B., Rigoll, G., Lang, M.: Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE (2004)
5. Alkhamali, E.A., Allinjawi, A., Ashari, R.B.: Combining transformer, convolutional neural network, and long short-term memory architectures: a novel ensemble learning technique that leverages multi-acoustic features for speech emotion recognition in distance education classrooms. Appl. Sci. **14**(12), 5050 (2024)
6. Lin, M., et al.: A review of emotion recognition of learners for online education. Control Decis. **39**(4), 1057–1074 (2024)
7. Dehbozorgi, N., Kunuku, M.T.: Exploring the influence of emotional states in peer interactions on students' academic performance. IEEE Trans. Educ. (2023)
8. Huang, Y.: Real-time application and effect evaluation of multimodal emotion recognition model in online learning. In Proceedings of the 2024 10th International Conference on Computing and Data Engineering (2024)
9. Welsen, S., Zhang, M., Chu, Y.: Project-based network simulation of campus remote seat booking system. In: 2022 IEEE 2nd International Conference on Educational Technology (ICET). IEEE (2022)
10. Wang, C.-H., Lin, H.-C.K.: Constructing an affective tutoring system for designing course learning and evaluation. J. Educ. Comput. Res. **55**(8), 1111–1128 (2018)
11. Alyuz, N., et al.: Semi-supervised model personalization for improved detection of learner's emotional engagement. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction (2016)
12. Hayashi, Y. Detecting collaborative learning through emotions: An investigation using facial expression recognition. In: Intelligent Tutoring Systems: 15th International Conference, ITS 2019, Kingston, Jamaica, June 3–7, 2019, Proceedings 15. Springer (2019)
13. Keerthika, M., et al.: Emotional AI: computationally intelligent devices for education. In: Emotional AI and Human-AI Interactions in Social Networking, pp. 87–99. Elsevier (2024)
14. Koolagudi, S.G., Rao, K.S.: Emotion recognition from speech: a review. Int. J. Speech Technol. **15**, 99–117 (2012)
15. Li, X., et al.: Music theory-inspired acoustic representation for speech emotion recognition. IEEE/ACM Trans. Audio, Speech, Lang. Process. **31**, 2534–2547 (2023)
16. Li, D., et al.: Exploiting the potentialities of features for speech emotion recognition. Inf. Sci. **548**, 328–343 (2021)
17. Er, M.B.: A novel approach for classification of speech emotions based on deep and acoustic features. IEEE Access **8**, 221640–221653 (2020)
18. Vijayan, D.M., et al.: Development and analysis of convolutional neural network based accurate speech emotion recognition models. In: 2022 IEEE 19th India Council International Conference (INDICON). IEEE (2022)
19. Lee, C.M., Narayanan, S.S.: Toward detecting emotions in spoken dialogs. IEEE Trans. Speech Audio Process. **13**(2), 293–303 (2005)
20. Tianyang, Y.: Study on using scenarios of linear and non-linear classifiers. In: 2020 International Conference on Computing and Data Science (CDS). IEEE (2020)
21. Abbaschian, B.J., Sierra-Sosa, D., Elmaghraby, A.: Deep learning techniques for speech emotion recognition, from databases to models. Sensors **21**(4), 1249 (2021)
22. Wani, T.M., et al.: A comprehensive review of speech emotion recognition systems. IEEE Access **9**, 47795–47814 (2021)

23. Wang, S., et al.: Speech emotion recognition based on principal component analysis and back propagation neural network. In: 2010 international conference on measuring technology and mechatronics automation. 2010. IEEE
24. Dileep, A.D., Sekhar, C.C.: HMM based intermediate matching kernel for classification of sequential patterns of speech using support vector machines. IEEE Trans. Audio Speech Lang. Process. **21**(12), 2570–2582 (2013)
25. Lim, W., Jang, D., Lee, T.: Speech emotion recognition using convolutional and recurrent neural networks. In: 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). IEEE (2016)
26. Burkhardt, F., et al.: A database of German emotional speech. In: Interspeech (2005)
27. Livingstone, S.R., Russo, F.A.: 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)
28. Cowen, A.S., Keltner, D.: Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proc. Natl. Acad. Sci. **114**(38), E7900–E7909 (2017)
29. Bondu, A., Lemaire, V.: Adaptive curiosity for emotions detection in speech. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE (2008)
30. Curious vs encouraging tone: meanings & examples (2019). https://www.studysmarter.co.uk/explanations/english/prosody/curious-vs-engaging-tone/
31. Ahmed, M.R., et al.: An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition. Expert Syst. Appl. **218**, 119633 (2023)
32. Ying, X.: An overview of overfitting and its solutions. J. Phys. Conference series. IOP Publishing (2019)
33. Wu, S., Falk, T.H., Chan, W.-Y.: Automatic speech emotion recognition using modulation spectral features. Speech Commun. **53**(5), 768–785 (2011)
34. Huang, Z., et al.: Speech emotion recognition using CNN. In: Proceedings of the 22nd ACM International Conference on Multimedia (2014)