Speech Emotions Recognition for Online Education

Abdelaziz A. Abdelhamid[1, 2]

[1] Department of Computer Science, College of Computing and Information Technology, Shaqra University, Shaqra 11961, Saudi Arabia
[2] Department of Computer Science, Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566, Egypt
Emails: abdelaziz@su.edu.sa; abdelaziz@cis.asu.edu.eg

Abstract

The severe circumstances caused by COVID-19 make online education the best replacement for regular face-to-face education for continuing the education process. One year ago, and till now most schools adopted online learning during this pandemic shutdown, which indicates the applicability of this teaching methodology. However, the efficiency of this method needs to be improved to guarantee its effectiveness. Although face-to-face teaching has many advantages over online education, there is a chance to promote online learning by utilizing the recent techniques of artificial intelligence. From this perspective, we propose a framework to detect and recognize emotions in the speech of students during virtual classes to keep instructors updated with the feelings of students so and can behave accordingly. The approach of detecting emotions from the speech is much more helpful for cases when turning on the cameras at the student's side could be embarrassing. This case is very common, especially for schools in Middle East countries. The proposed framework can also be applied to other similar scenarios such as online meetings.

Keywords: Speech emotions; Online learning; Machine Learning.

1. Introduction

Speech signal contains much information about the speaker. Besides the main message said by the speaker, speech signal also contains information about the emotional state of the speakers regardless of their national borders, gender, and race [1-2]. As the emotional state of students is very important for their education, recent research pays attention to it to improve the online education process. The information about the emotional state can be used in many applications such as aided driving [4-5], health management [3], and other applications [6-9]. In the previous research on detecting and classifying human emotions, researchers indicated that humans could perceive emotions regardless of their cultural background. In addition, researchers defined human emotions as six categories namely, happy, angry, excited, frustrated, sad, and neutral. [10, 11]. Based on these categories, there is a set of speech datasets were developed by researchers to study the recent methodologies for this purpose [12].

There are previous attempts that are presented by authors to employ emotion detection and recognition in online education based on facial expressions. Although this approach is effective, sometimes it cannot be applied in online learning due to social constraints in some societies. In some countries in the Middle East, the case on turning on the web camera during online learning is considered to breach the privacy of students as they attend online classes from home. On the other hand, this is not the case for turning on the microphone and talking to their instructor as this is allowed. In this case, the application of emotion detection based on students' images is not applicable. Therefore, we present in this paper a framework for detecting the emotional state of a

Doi: https://doi.org/10.54216/FPA.100104
Received: June 19, 2022 Accepted: November 06, 2022

student from their voice. This framework is based on processing the voice of a student using a deep learning network which results in one of the types of emotions and then sends a notification about the emotional state to his/her instructor to behave accordingly.

Although the process of emotion recognition from speech signals seems complex when compared with emotion recognition from facial images, the features of the mel-spetrogram of the speech signal are rich in information that can be used for this purpose. In the literature, there are a lot of research efforts that utilized these features along with the recent progress in deep learning and could achieve noticeable recognition accuracy.

As online education provides a good alternative the regular education in a fast, economical, and convenient mode of learning. However, students cannot exchange and interact dynamically like in face-to-face learning due to the separation between students and teachers. This could result in isolation and loss of interest, loss of confidence, and isolation of students. In addition, sometimes students cannot express their feelings to the teachers, which may lead to laziness and confusion in their studies.

Affective computing is an active research field in artificial intelligence which is related to studying the computing-related effect on emotion [3]. This field is composed of a set of modules including communication, modeling, detection, and recognition that react to emotions [4]. The emotion recognition module is considered the vital and most effective part of this approach. This module is usually based on information from either audio or facial data or both.

Recently, there are several research efforts have been done to realize the module of speech emotion recognition. The methodologies used for this task include K-nearest neighbor, Kernel regression, Maximum likelihood Bays classifier, Neural networks, and deep learning [5]. In addition, the features used to train these classifiers include Mel-frequency cepstral coefficients, Mel-spectrograms, and other feature selection methods.

This paper is organized as follows. In section 2, the material and proposed method are presented and discussed. In addition, the achieved speech emotion recognition results are presented and discussed in the approach presented in section 3. Moreover, to emphasize the effectiveness of the proposed approach, a fair comparison between the proposed approach and another approach is presented in section 4. Finally, the discussion and conclusions are presented in section 5.

2. The Proposed Methodology

There is a shortage of methods that can be used for collecting effective data for learners. The current methods are limited to physiological sensors and questionnaires. These methods could capture limited information about learners and more efforts should pay to find methods that are more effective. The application of this information can be used to adapt the learning contents or the learning tasks which is the current approach for most of the current research. In this paper, we extend these research efforts by incorporating speech emotion recognition to improve the e-learning process and develop an effective tutoring system.

The application of speech emotion recognition is performed in the previous research as a general-purpose software development and not dedicated to the e-learning process that requires special user modeling and user setting. In [13], authors developed a real-time standalone framework for social signal processing and recognition of speech signals. Similar work is presented by authors in [14]. In this work, the authors presented another tool for the same task, which is processing and recognizing the social signal.

Authors in [15] presented a provisioning system that offers feedback for human-robot interaction. Another approach is presented in [16], in which authors employed speech analysis of a sequence of the signal of fragments to enable affective learning in the setting of online learning. This approach is based on extracting the vocal intonations that are converted into an emotional state. This approach is then extended in another research article presented by authors in [17]. In this extension, the authors proposed a framework that employs affective computing for providing feedback to learners in an online learning framework.

Doi: https://doi.org/10.54216/FPA.100104
Received: June 19, 2022 Accepted: November 06, 2022

There are many studies in the literature that presented the process of speech emotion recognition [18]. The process usually starts with capturing the speech signal then dividing it into a sequence of frames then processing each frame to detect the emotional state of the speaker. However, this process is not trivial as each emotional state occurring in the human-computer interaction is spontaneous and usually becomes a complex task to distinguish these emotional states from the acoustic features. Meanwhile, the input speech signal might contain more than one emotion. This could also make this process more complex.

Authors in [19] presented the methodology of recognizing speech emotions for intelligent tutoring systems. In this research, authors categorized the tools used in the detection and recognition of speech emotions into motor-behavioral, physiological, and psychological. The motor-behavioral tools detect the emotional state using special software that uses cameras, a mouse, and a keyboard to record the behavioral movements of users. However, the detection of the emotional state using the physiological tools is based on recording the physiological state using a set of dedicated sensors designed for this purpose [20-25]. On the other hand, the detection of the emotional state based on psychological tools depends mainly on self-reporting tools that can record the subjective experience of users. All these methods can be used in many applications if they are enriched with effective computing. Therefore, we propose in this paper a framework for a computer application that can detect and recognize the emotional state of learners and thus can improve the system feedback without the involvement of the human teacher [26-30].

The proposed framework is depicted in Figure 1. As shown in the figure, when the learner interacts with the e-learning framework using voice, the framework captures the speech signal and encrypts it to be secured, then sends it to the cloud to be processed. The full operation of speech processing is performed in the cloud, as it requires computational capabilities, which sometimes are not available on the learners' local resources. Once the encrypted speech signal arrives in the cloud through the internet, it decrypts it to start the speech processing operation. The first step in the speech processing operation is feature extraction, in which the speech signal is converted into mel-spectrogram that is used as the main feature entities to be fed to the next step [31-34].

The extracted features are used in the deep learning model, which is already trained on a speech corpus containing a large set of speech signals labeled with the corresponding emotional state. This trained model is used to check the incoming signal against a set of emotions. Once the models detect the emotional state of the speaker, it sends this state to a rule-based engine that produces the appropriate feedback to the learner based on the detected emotional state of the current content of the educational material shown to the learning on the e-learning framework. This feedback is also encrypted before being sent to the learner to avoid any potential illegal breach.

The proposed approach is composed of three layers namely, the learner layer, network layer, and cloud layer. More details on these layers are presented in the next sections.

A. Learner layer

In this layer, the student as a learner has access to the e-learning management system and interacts with the course contents. In addition, this layer enables the user to perform an online examination and live interaction with the affective-enabled environment. As the learner intended in this work to study at his own time, place, and pace, he is considered a lifelong learner who is positively biased toward the informal learning paradigm.

In addition, this layer incorporates the internal hardware configuration of the learner's hardware, which includes the computer, laptop, or smart device that is used to access the effective learning environment. This environment is assumed to be integrated with a microphone to capture the speech signal from the learner. Moreover, it consists of a component called an affective computing tool that is responsible for the interaction between the learner and the e-learning system.

Doi: https://doi.org/10.54216/FPA.100104
Received: June 19, 2022 Accepted: November 06, 2022

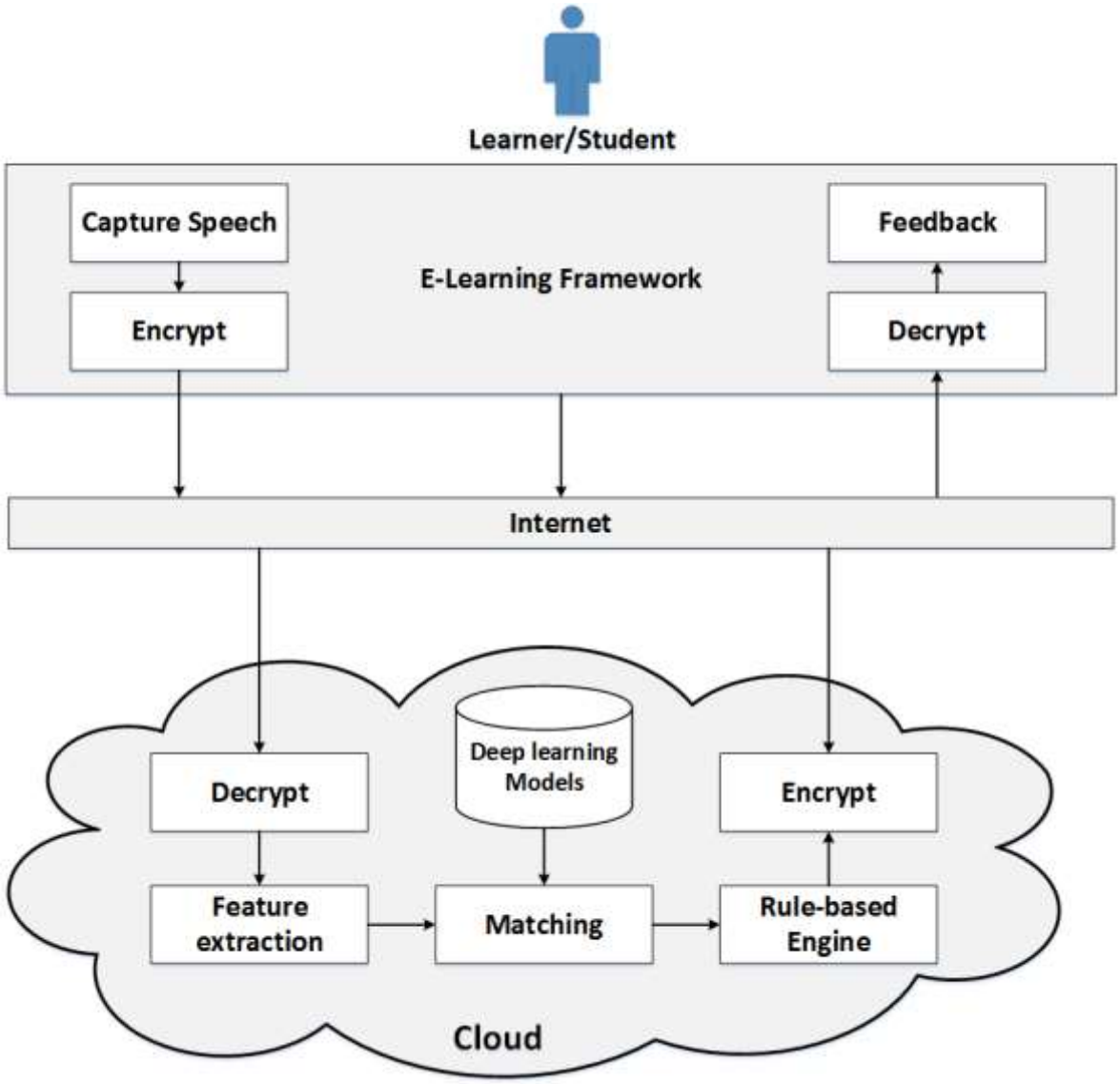

Figure 1: The proposed system architecture

- *Affective computing tool*

This tool is the main actor in the interaction process between the learner and the system. It is responsible for receiving the speech signal from the learner and applies encryption to cipher the speech content to be protected. Then, it sends the encrypted speech over the internet to the target cloud that performs the main operation of speech emotion recognition. In addition, this tool is responsible for receiving feedback came from the main engine on the cloud and then decrypting it as it comes in an encrypted format. Then based on this feedback it guides the learner in a way that improves his acceptance and satisfaction with the online learning material.

Doi: https://doi.org/10.54216/FPA.100104
Received: June 19, 2022 Accepted: November 06, 2022

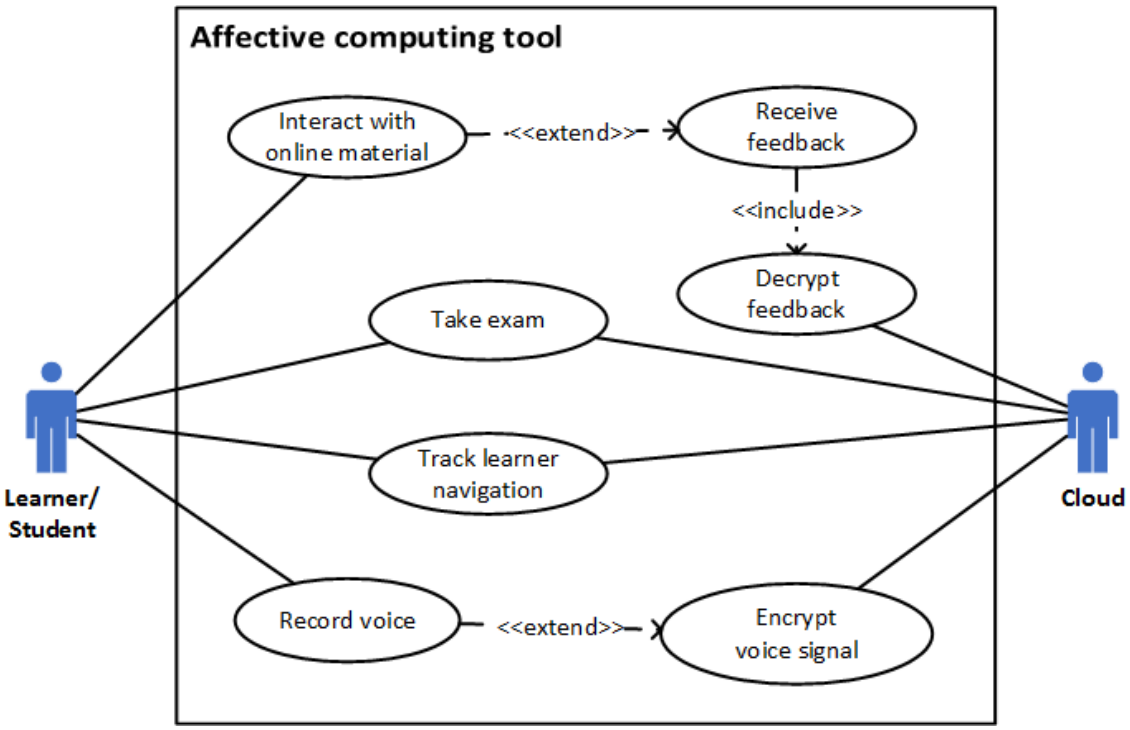

Figure 2: Use case diagram of the effective computing tool

The use case of the affective computing tool is depicted in Figure 2. In this figure, the learner interacts with online material by browsing the online lectures or watching a video through the online learning management system. This interaction is usually tracked using the affective computing tool. Based on this interaction, the affective computing tool may request feedback from the rule-based engine to boost and enrich the educational process. In addition, the effective computing tool supports the examination process, as it guides students in explaining and clarifying the exam questions.

B. Network layer

This layer refers to the communication between the e-learning management system and the cloud layer that performs speech-emotion recognition. This layer is responsible for receiving the encrypted voice and transmitting it to the cloud server to process it. On the other hand, it also receives encrypted feedback from the cloud server and transmits it to the learner's layer to present the learner in an appropriate manner.

C. Cloud layer

This layer represents the core of the proposed approach. It consists of the main operation of speech signal processing and emotion detection and recognition. The process starts with decrypting the incoming speech signal, then performs feature extraction using mel-spectrogram based on the fast Fourier transform. Then, the already-trained deep learning models are used to match the extracted features with respect to the six types of speech emotions under consideration. Once the emotion is detected and recognized, it is passed to the rule-based engine to produce the appropriate feedback.

Doi: https://doi.org/10.54216/FPA.100104
Received: June 19, 2022 Accepted: November 06, 2022

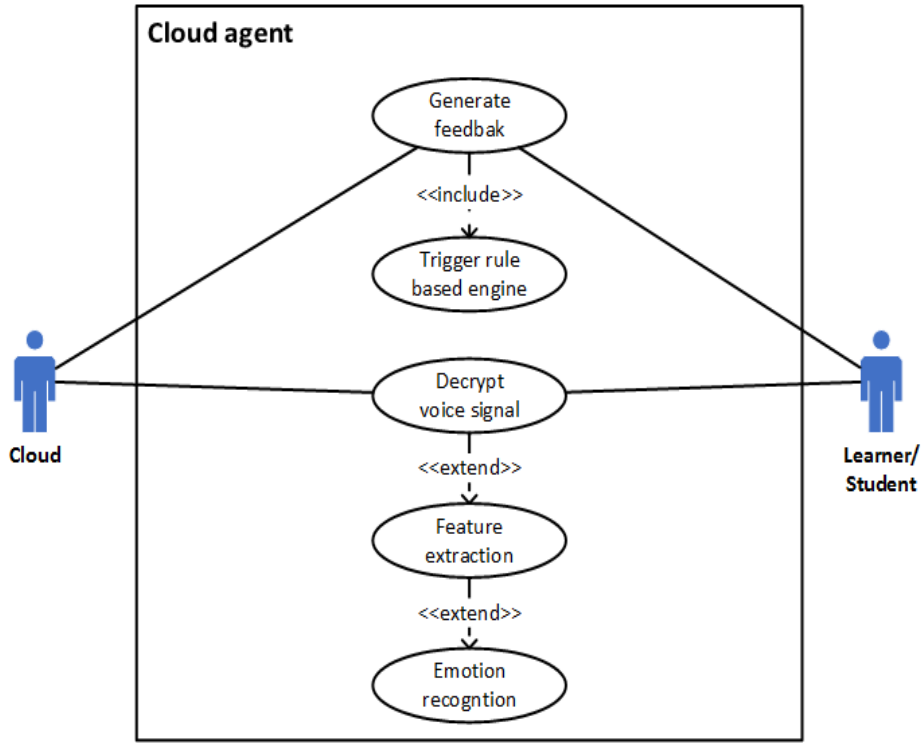

Figure 3: Use case diagram of the cloud agent

Figure 3 depicts the use case diagram of the cloud agent. As shown in this figure, the main tasks of the cloud agents are speech emotion recognition and the generation of feedback. The process of feedback generation may be triggered by the result of the speech emotion recognition process or by a request from the affective computing tool based on the learner navigations in the course material.

3. Results and Discussion

The proposed speech emotion recognition approach is described in this section. This system is based on deep learning and uses four layers of feature learning blocks. These blocks are employed for learning the embedded information in the extracted me-spectrogram features. These learning blocks are composed of a convolutional neural network (CNN) along with a regularization layer to normalize the intermediate values calculated in the learning block. The architecture of the proposed emotion recognition system is shown in Figure 4.

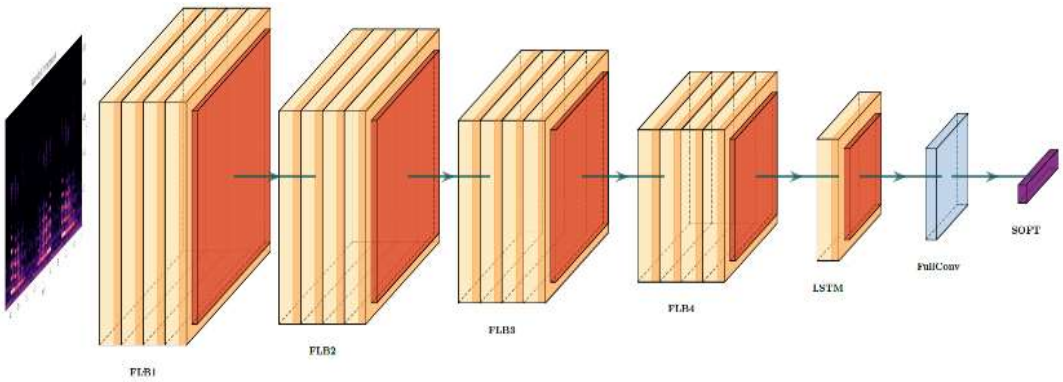

Figure 4: The proposed deep learning network used in the speech emotion recognition task.

Doi: https://doi.org/10.54216/FPA.100104
Received: June 19, 2022 Accepted: November 06, 2022

As shown in Figure 4, the deep learning network takes the image of the features of the input speech signal and outputs one of the six speech emotions. The resulting emotion is then fed to the rule-based engine for further processing.

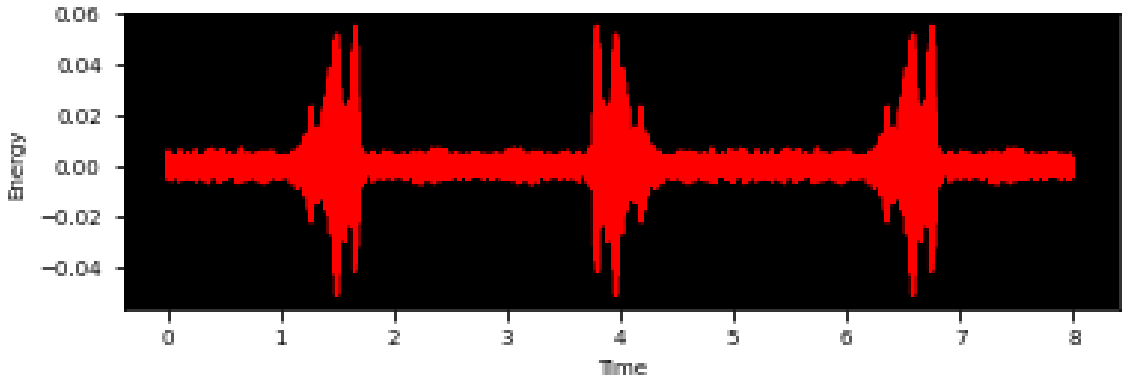

Figure 5: Sample speech signal

Figure 5 shows a sample speech signal from the IEMOCAP dataset. As shown in this figure, the most dominant information in this signal is the message said in this signal. However, the extraction of emotion is considered a challenge that we can manage using the recent approaches in deep learning.

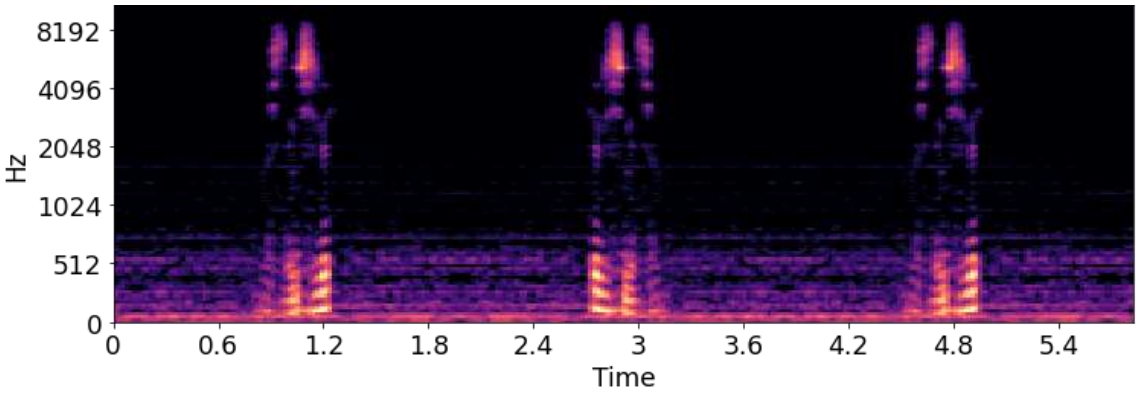

Figure 6: Mel-spectrogram of the sample speech signal

Figure 6 presents the mel-spectrogram of the above speech signal. This figure contains a great deal of information related to the speech, speaker, and message content. Therefore, this type of feature is preferred in the task of speech emotion recognition.

The proposed approach is evaluated in terms of the standard speech emotion dataset namely, IEOCAP [12]. This dataset is well prepared for the task of speech emotion recognition. It contains a set of speech signals recorded by speakers of different ages males and females. The established experiment is performed using the proposed deep learning framework, which is trained and tested using mel-spectrogram features.

The results of recognizing the six speech emotions in the dataset are shown in Figure 7. In this figure, it can be clearly shown that the performance of the proposed approach is very promising and can be used in the live scenarios of speech emotion recognition. The process of live detection and recognition of speech emotions is kept in mind and planned to be achieved in the future work of this research.

Doi: https://doi.org/10.54216/FPA.100104
Received: June 19, 2022 Accepted: November 06, 2022

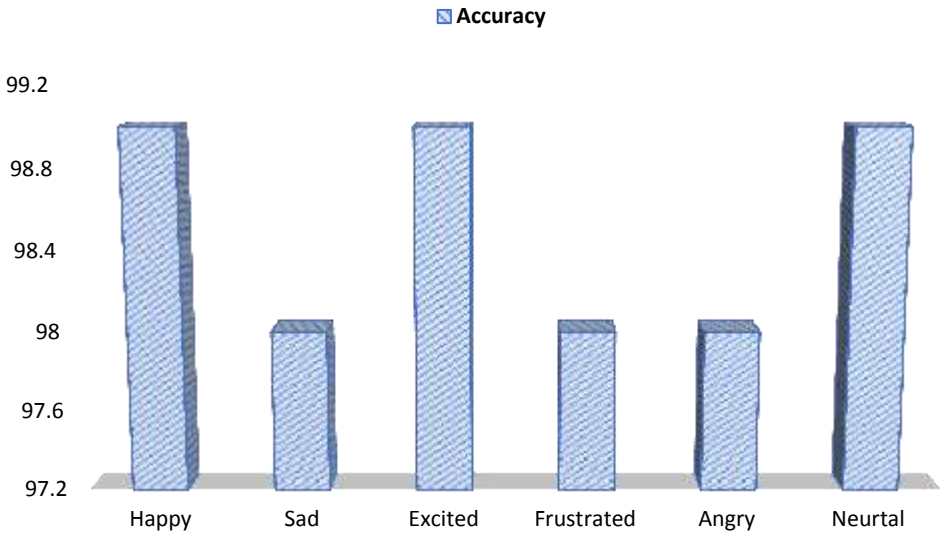

Figure 7: Speech emotion recognition accuracy

As shown in Figure 7, the accuracy of the proposed system is at least 98% for all types of speech emotions. These results are obtained when the system is trained on the speaker-independent samples of the IEMOCAP dataset and tested on a subset of 20% of the speaker-independent samples.

In order to evaluate the effectiveness of the proposed approach, it is compared with other systems. However, to the best of our knowledge, there are a few attempts in the literature that tried to exploit speech emotion recognition in online education and in the form of an integrated framework. Therefore, we compared the proposed framework with the framework presented by the authors in [17].

Table1: Comparison between the proposed approach and another approach in the literature

Factor	**The approach in [17]**	**Proposed Approach**
Architecture	Localized on the learner's device	Hosted on a cloud server
Emotion recognition methodology	Sequential minimum optimization and WEKA	Deep learning CNN
Supports site navigation tracking	No	Yes
Require high computational resources on the user side	Yes (The full operation is run locally)	No (The full operation is run on the cloud)
Evaluation methodology	Using questionnaires	Using standard speech emotion dataset

Listed in Table 1 are the main factors included in the comparison. As shown in this table, the proposed approach is distinguished from the approach presented in [17] in terms of the computational resources required by the system. In [17], these computational resources are assumed available on the learner's side (on his own computer). This case is not always guaranteed, and thus affects the applicability of this approach. However, the proposed approach depends on hosting the full process of speech emotion recognition as well as the rule-based engine on the cloud. This cloud hosting can easily be accomplished by the organization responsible for the learning management system.

On the other hand, the speech emotion recognition methodology along with the evaluation methodology is performed in terms of the deep neural network which could achieve high recognition

Doi: https://doi.org/10.54216/FPA.100104
Received: June 19, 2022 Accepted: November 06, 2022

accuracy. However, these operations are performed on the classical approach which may suffer from less degree of accuracy.

4. Conclusions and Future Perspectives

In this paper, we presented a novel framework for integrating speech emotion recognition in the e-learning process to improve students' performance and satisfaction. The presented framework is based on a set of layers that interact together through message passing. This message is passed to/from layers in an encrypted format to secure it from outside breaching. The main layer in the proposed framework is the cloud layer, which is responsible for the main process of speech emotion recognition. The concept of cloud is adopted as it offers more powerful computing resources that afford the computation needed in detecting and recognizing the embedded emotions of the learner. The speech emotion methodology proposed in this paper is based on deep learning applied to the mel-spectrogram as an effective feature representing the speech contents. In addition, the paper presents an analysis of the proposed approach from the perspective of software engineering. Therefore, a set of diagrams have been presented and discussed throughout the text. Moreover, the proposed approach is compared with a similar approach presented in the literature to emphasize the significance of this approach.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] C. Darwin and P. Prodger, 'e Expression of the Emotions in Man and Animals, Oxford University Press, Oxford, MA, USA, 1998.
[2] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 97–115, 2001.
[3] R. Donoso, C. San Mart´ın, and G. Hermosilla, "Reduced isothermal feature set for long wave infrared (LWIR) face recognition," Infrared Physics&Technology, vol. 83, pp. 114 –123, 2017.
[4] T. Liu, H. Liu, Z. Chen et al., "FBRDLR: fast blind reconstruction approach with dictionary learning regularization for infrared microscopy spectra," Infrared Physics & Technology, vol. 90, pp. 101–109, 2018.
[5] Z. Huang, H. Fang, Q. Li et al., "Optical remote sensing image enhancement with weak structure preservation via spatially adaptive gamma correction," Infrared Physics & Technology, vol. 94, pp. 38–47, 2018.
[6] Y. Bi, M. Lv, Y. Wei, N. Guan, and W. Yi, "Multi-feature fusion for thermal face recognition," Infrared Physics & Technology, vol. 77, pp. 366–374, 2016.
[7] H. Liu, Z. Zhang, S. Liu, J. Shu, T. Liu, and T. Zhang, "Blind spectrum reconstruction algorithm with L0-sparse representation,"Measurement Science and Technology, vol. 26, no. 8, pp. 085501–085507, 2015.
[8] H. Wu, Y. Liu, L. Qiu, and Y. Liu, "Online judge system and its applications in C language teaching," in Proceedings of the International Symposium on Educational Technology (ISET),
[9] pp. 57–60, Beijing, China, July 2016.
[10] T. Liu, Z. Chen, H. Liu, Z. Zhang, and Y. Chen, "Multi-modal hand gesture designing in multi-screen touchable teaching system for human-computer interaction," in Proceedings of the Second International Conference on Advances in Image Processing, pp. 100–109, Chengdu China, June 2018.
[11] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," Journal of Personality and Social Psychology, vol. 17, no. 2, pp. 124–129, 1971.
[12] P. Ekman, "Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique," Psychological Bulletin, vol. 115, no. 2, pp. 268–287, 1994.
[13] Samarth Tripathi and Sarthak Tripathi and Homayoon Beigi, " ulti-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning," arXiv, 1804.05788, 2019.

Doi: https://doi.org/10.54216/FPA.100104
Received: June 19, 2022 Accepted: November 06, 2022

[14] Wagner, J., Lingenfelser, F., & Andre, E.. The Social Signal Interpretation Framework (SSI) for Real Time Signal Processing and Recognitions, In Proceedings of INTERSPEECH, Florence, Italy, 2011.
[15] Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., & Andre, E. The Social Signal Interpretation (SSI) Framework: Multimodal Signal Processing and Recognition in Real-time. Proceedings of the 21st ACM International Conference on Multimedia, MM '13. Barcelona, Spain. 831–834, 2013.
[16] Jo Jianhua, T., Tieniu, T., & RosalindW, P. Affective computing: a review. Affective computing and intelligent interaction. Springer Berlin Heidelberg, 3784, 981–995, 2005.
[17] Jones, C., & Sutherland, J. Acoustic emotion recognition for affective computer gaming. In C. Peter & R. Beale (Eds.), Affect and Emotion in Human-Computer Interaction. LNCS. 4868. Heidelberg: Springer, 2008.
[18] Bahreini, K., Nadolski, R. & Westera, W. Towards real-time speech emotion recognition for affective e-learning. Educ Inf Technol 21, 1367–1386 (2016).
[19] Beale, R., & Creed, C. Affective interaction: How emotional agents affect users. International Journal of Human-Computer Studies, 67(9), 755–776, 2009.
[20] Feidakis, M., Daradoumis, T., & Caballe, S. Emotion Measurement in Intelligent Tutoring Systems: What, When and How to Measure. Third International Conference on Intelligent Networking and Collaborative Systems, 807–812. 2011.
[21] Huhnel, I., Fölster, M., Werheid, K., & Hess, U. Empathic reactions of younger and older adults: no age related decline in affective responding. Journal of Experimental Social Psychology, 50, 136–143, 2014.
[22] Nwe, T., Foo, S., & De Silva, L. Speech emotion recognition using hidden Markov models. Speech Communication, 41, 603–623, 2003.
[23] Pfister, T., & Robinson, P. Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis. IEEE Transactions on Affective Computing, 2(2), 66–78, 2011.
[24] Chen, L., Mao, X., Xue, Y., & Cheng, L. L. Speech emotion recognition: features and classification models. Digital Signal Processing, 22(6), 1154–1160, 2012.
[25] Bahreini, K., Nadolski, R., & Westera, W. FLITWAM and Voice Emotion Recognition. Games and Learning Alliance (GaLA) Conference. Paris, France, 23–25, 2013.
[26] Beale, R., & Creed, C. Affective interaction: How emotional agents affect users. International Journal f Human-Computer Studies, 67(9), 755–776, 2009.
[27] Ben Ammar, M., Neji, M., Alimi, A. M., & Gouardères, G. The affective tutoring system. Expert Systems with Applications, 37(4), 3013–3023, 2010.
[28] Chen, L., Mao, X., Xue, Y., & Cheng, L. L. Speech emotion recognition: features and classification models. Digital Signal Processing, 22(6), 1154–1160, 2012.
[29] Happy, S. L. Dasgupta, A., Patnaik, P., Routray, A. Automated Alertness and Emotion Detection for Empathic Feedback during e-Learning. IEEE Fifth International Conference on Technology for Education (T4E). 47–50, 2013.
[30] López-Cózar, R., Silovsky, J., & Kroul, M. Enhancement of emotion detection in spoken dialogue systems by combining several information sources. Speech Communication, 53(9–10), 1210–1228, 2011.
[31] Pekrun, R.. The impact of emotions on learning and achievement: towards a theory of cognitive/ motivational mediators. Journal of Applied Psychology, 41, 359–376, 1992..
[32] Hamzah A. Alsayadi, Abdelaziz A. Abdelhamid, Islam Hegazy, and Zaki T. Fayed: Arabic speech recognition using end-to-end deep learning. IET Signal Process. 15(8), 521– 534 (2021).
[33] A. Abdelhamid, E.-S. M. El-kenawy, B. Alotaibi, M. Abdelkader, A. Ibrahim et al., Robust speech emotion recognition using CNN+LSTM based on stochastic fractal search optimization algorithm, IEEE Access, vol. 10, pp. 49265-49284, 2022.
[34] A. Abdelhamid, W. Abdulla, B. Macdonald. WFST-Based Large Vocabulary Continuous Speech Decoder for Service Robots. ACTA Press, 2012. www.actapress.com, https://doi.org/10.2316/P.2012.771-009

Doi: https://doi.org/10.54216/FPA.100104
Received: June 19, 2022 Accepted: November 06, 2022