Can AI Recognise Pilot's Vocal Emotional Expression under Emergency Situations?

Wen-Chin Li, Kuang-Lin Hsieh, Jeremia Pramudya & Declan Saunders

Safety and Accident Investigation Centre, Cranfield University

SUMMARY

The development of Artificial Intelligence (AI) in the field of voice recognition has prompted interest in the field of emotional voice recognition (EVR); EVR is now one of the key challenges in the applications of Natural Language Processing (NLP). When conducting an accident investigation, the voice data from the cockpit voice recorder (CVR) usually provide significant evidence of the pilot's mental state, which can support some hypotheses on occurrences by investigators, especially the verbal speech between pilots and air traffic controllers. In the past, emotion analysis mainly relies on image and text analysis technology. With the development of large language models (LLM) and AI; such as ChatGPT, Llama, and Perplexity, EVR has become possible. This research aims to explore the potential of using a Recurrent Neural Network (RNN) and the open-source dataset, Toronto Emotional Speech Set (TESS), to identify the pilot's speech and emotions in emergencies. Further research may combine voice with physiological data, and with facial expressions to serve the purposes of operational and safety monitoring.

KEYWORDS

Artificial intelligence, Deep Learning, Emotional Voice Recognition, Large Language Model, Machine Learning

Introduction

The human ability to perceive nuances in verbal communication is extremely sophisticated; with accuracy and capacity achieved over years of evolution to communicate consciousness and perception (Alharbi et al., 2021). Recent studies into EVR, using artificial intelligence (AI), have demonstrated a capacity for feature extraction, such as pitch, tone, and speech rate. These features correlate with human emotion: a high pitch and speech rate could indicate excitement, while a slower speech rate and tone could indicate sadness (Cowen & Keltner, 2017). In aviation, pilot decision-making during emergencies is crucial to ensuring the safety of passengers and crew members. When faced with unexpected events, pilots must quickly respond and make decisions based on limited information and high-pressure conditions. These situations reduce the pilot's ability to fully utilise the resources and information available, increasing the chance of human error. Emotional responses, such as anxiety, stress, or fear, can impair a pilot's attention, memory, and judgment, thus affecting their choices during critical moments. Emotional interference with cognitive functions may lead to delays in decision-making or poor judgment, which could severely threaten flight safety (Nocak & Mrazova, 2015).

This research aims to use Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) models to recognise the emotion in verbal communications in high-stress environments. Deep Learning (DL) has emerged as a key research area in Machine Learning (ML), gaining attention for its ability to process raw data without manual feature extraction. In recognising data, DL surpasses traditional methods by automatically detecting complex patterns and handling

unlabeled data. DL has made significant advancements in natural language processing (NLP), achieving breakthroughs in applications like speech recognition and other language-related tasks (Barhoumi & BenAyed, 2025). A Recurrent Neural Network (RNN) is a type of artificial neural network with directed cycles, allowing information to persist. Unlike feedforward networks, RNNs retain past inputs, making each output dependent on both current and previous inputs (Sherstinsky, 2020). Consider a general recurrent network with input $\{x^{(t)}\}$, output $\{z^{(t)}\}$, and target sequence $\{y^{(t)}\}$, where $z^{(t)} \in \mathbb{R}^p$ and $y^{(t)}a \in \mathbb{R}^p$. The data-generating process (DGP) is modeled as:

$y^{(t)} = z^{(t)} + \varepsilon^{(t)}$, *for* $t \in \mathbb{Z}$,

Where $\{\varepsilon^{(t)}\}$ is a sequence of independent and identically distributed (*i.i.d.*) random vectors. This additive error assumption is widely employed in statistical modeling and aligns with commonly used loss functions designed to quantify the discrepancy between $y^{(t)}$ and $z^{(t)}$ (Zhao et al., 2020). Python has become a dominant programming language in EVR due to its extensive libraries and frameworks that facilitate ML and DL applications. The librosa library is widely used for feature extraction from audio signals, enabling researchers to analyse MEL-frequency cepstral coefficients (MFCCs), pitch, and energy levels essential for emotion detection (Barhoumi & BenAyed, 2025). DL frameworks like TensorFlow and PyTorch provide robust tools for building RNN, LSTM, and gated recurrent unit (GRU) models, which are effective in processing sequential voice data (Abdel-Hamid et al., 2014). These models learn temporal dependencies in speech, allowing the system to recognise emotions from voice modulation (Luo et al., 2017).

Methods

3.1 Data collection and loading:

This research utilises the open-source database, TESS (Toronto Emotional Speech Set). It is then randomly divided into two groups (80% of training data and 20% of testing data) to build the accuracy of the model TESS (Toronto Emotional Speech Set):

TESS is a speech dataset that contains speech samples from two female actresses, aged 29 and 64, using a set of 200 target words expressing seven different emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. The dataset includes 2,800 samples in total.

3.2 Feature Extraction:

Mel Frequency Cepstral Coefficients (MFCC): Designed on the Mel frequency scale, MFCC attempts to simulate how the human ear perceives and processes speech, making it one of the most commonly used feature extraction methods in speech processing (Barhoumi & BenAyed, 2025). The MFCC will extract audio from a WAV file format, and apply coefficients in a 13-dimensional space, one dimension being the loudness and frequency of the voice (Ittichaichareon et al., 2012). Figure 2 shows the MFCC creating a vector consisting of loudness, pitch, frequency, and speech sounds including tone, rhythm, and articulation respectively.

3.3 Training Model with neural network:

Features extracted will be fed into the RNN and LSTM network to capture the temporal dependencies inherent in speech signals. LSTM is a specialised type of RNN designed to overcome the vanishing gradient problem that happens in RNN. LSTM contains a memory cell that can store information over long periods and has several gates (input, forget, and output gates) that control the flow of information in and out of the cell. These gates enable the LSTM to decide which information to retain, which to forget, and which to output. This allows LSTM to capture long-term dependencies in sequences. LSTM performs better than

RNN in tasks requiring the model to retain information for longer durations, as they can maintain and modify the internal state based on both the current input and previous context.

3.4 Applying the audio into the system:

This research aims to analyse the emotion with the system trained via DL, thus, the audio of the communication between the pilot of Flight 1549 and ATCOs for 90 seconds released by the Federal Aviation Authority (FAA) is applied. The proposed methodology for the EVR analysis consists of the following steps seen in Figure 1 below.

Figure 1 Flowchart of the overall architecture

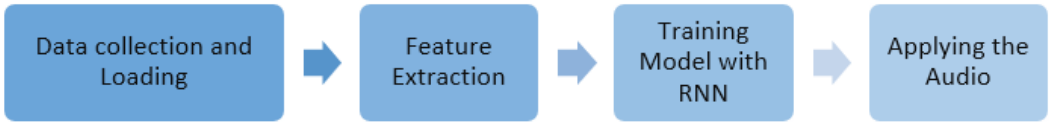

Figure 2. An example of the anger extracted audio

```
Sample Rate: 48000 Hz, Signal Length: 166567 samples
MFCC Features: [ 1.04644124 -6.68054316 -0.48840813 -0.60830217 -0.27202627  0.30199612
 -0.94593789  0.12241271 -0.50382968 -0.02061761 -0.28371644 -0.29715581
  0.39338679]
```

Results

In this study, we employ Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) for emotion recognition in cockpit dialogues. The model achieved an accuracy of 89.54% on the training data and 85.26% on the test data, shown in Table 1, demonstrating strong performance in identifying emotions. The pilot had 8 speeches to communicate with ATCO in 90 seconds, as shown in Table 2. The result demonstrated that AI recognised the pilot's emotion of fear (97.0%) at the beginning of the bird strike, followed by anger (2.8%), and finally disgust and pleasant surprise (0.1%). In those 90 seconds of communication, the emotion was also detected as sadness (51.1%), fear (47.7%) and neutral (1.2%).

Table 1: The accuracy of the model

Data (Source: TESS)	Method	Feature Extraction	Accuracy
Training Data	RNN + LSTM	MFCC	89.54%
Test Data			85.26%

Table 2: The emotion of the pilot predicted from the model

emotions / audio	angry	fear	disgust	happy	pleasant surprise	sad	neutral
Pilot's 1st speech: 00:00 – 00:06	2.8%	97.0%	0.1%	0.0%	0.1%	0.0%	0.0%
Pilot's 2nd speech 00:10 – 00:10	0.7%	85.8%	0.0%	0.0%	0.0%	2.4%	11.1%
Pilot's 3rd speech 00:33 – 00:35	7.6%	88.9%	0.0%	0.5%	0.0%	0.5%	2.5%
Pilot's 4th speech 00:39 – 00:39	0.1%	70.4%	0.1%	0.0%	0.0%	27.7%	1.7%
Pilot's 5th speech 00:48 – 00:52	0.2%	99.8%	0.0%	0.0%	0.0%	0.0%	0.0%
Pilot's 6th speech 00:58 – 00:58	0.1%	42.1%	0.0%	0.0%	0.0%	54.0%	3.8%
Pilot's 7th speech 01:20 – 01:21	51.3%	31.7%	5.1%	4.8%	0.0%	7.1%	0.0%
Pilot's 8th speech 01:23 – 01:24	0.0%	47.7%	0.0%	0.0%	0.0%	51.1%	1.2%

The study also conducted human evaluation on the same 90 seconds of audio transcript, to evaluate and compare the emotional recognition conducted by the AI. The evaluation was conducted as either an emotion was present or was not present, instead of a percentage of emotion in each speech. Based on an analysis with 34 subject matter experts (SME), it was concluded that the highest emotion detected in each transcript was (1) fear (8 out of 8 speeches), followed by (2) neutral (8 out of 8 speeches), then (3) anger (5 out of 8 speeches) and finally (4) disgust (4 out of 8 speeches). The emotional recognition of the human analysis revealed a significant deviation from the AI model, which demonstrates the complexity and diverse nature of EVR.

Discussion

This research successfully applied EVR modeling to analyse the cockpit dialogue of pilots and ATCOs in emergencies, effectively identifying various emotional states. Table 2 contains the weighting of each emotion based on the features of the audio input, which were different from those collected by the SME's. The variance is likely the result of nuances in human speech that are hard to define categorically and are highly subjective to the individual. Despite this variation, the model predicted its accuracy to the speech at 89.54%, indicating a strong performance to the training dataset based on the emotional parameters defined. However, when conducted on the test dataset, accuracy in emotional recognition decreased to 85.26%. This could be a result of the male voice in the test dataset where the training was conducted with a female voice. Further explanation could be the relative tone of the pilot during the emergency, duration of the communications, and background/environmental noise during each communication. Future research would consider how to further optimise the EVR model to perform more precise and accurate analyses using a larger dataset of emotional parameters and diversity among the actors for the data collection phase. This research has applications in accident investigation roles, predominantly the identification of stressors and other hidden signs in verbal communications that can help investigators understand why the pilots made the decisions they did. There is also an application as a real-time support tool between ATCOs and pilots during high-stress and high-workload scenarios. Finally, in flight simulators, emotion recognition can be used to track how trainee pilots respond emotionally to stress, providing valuable insights into their emotional resilience and decision-making under pressure. This feedback can be used to adjust training programs, helping to improve how pilots manage their emotions and perform under critical conditions.

Conclusion

This study utilised the TESS dataset to achieve a successful EVR on the transcript between the pilot and ATCO during Flight 1549, able to capture several different emotions conveyed within a singular speech input. While the accuracy of the detection is widely disproportionate to the human-evaluated EVR, the model was able to take a highly challenging transcript and self-disseminate the emotion conveyed by the pilot. Moving forward, the study will aim to expand the TESS dataset to include a wider diversity, which will aid in the detection and classification of the AI to the emotion as represented by the human. This technology can support accident investigation during CVR transcript analyses, when identifying the pilots' cognitive and mental state before their decision-making; letting the investigator into the mind of the flight crew. Further applications can support real-time assistance during abnormal or emergencies, which could include ATCO alerting, voice-to-text conversion, and speech personalisation for high-workload environments. This area of study can also open new optimisation for future Single Pilot Operations (SiPO), providing an emotional support network between the pilot and ground network, to promote more comprehensive decision-making during emergencies, when time limitations are present.

References

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *22*(10), 1533–1545. https://doi.org/10.1109/TASLP.2014.2339736

Alharbi, S., Yahya, I., Alsarhan, A., Sedeeq, E., Kareem, A., & Hussein, M. (2021). Automatic speech recognition: Systematic literature review. *IEEE Access, 9,* 131858–131876. https://doi.org/10.1109/ACCESS.2021.3112535

Barhoumi, C., BenAyed, Y. (2025). Real-time speech emotion recognition using deep learning and data augmentation. *Artificial Intelligence Review*, *58*(2). https://doi.org/10.1007/s10462-024-11065-x

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences, 114*(38), E7900–E7909. https://doi.org/10.1073/pnas.1702247114

Ittichaichareon, C., Suksri, S., & Yingthawornsuk, T. (2012). Speech Recognition using MFCC. *International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012)*. https://www.academia.edu/download/85360099/9_20712576.pdf

Luo, F., Guo, W., Yu, Y., & Chen, G. (2017). A multi-label classification algorithm based on kernel extreme learning machine. *Neurocomputing*, *260*, 313–320. https://doi.org/10.1016/j.neucom.2017.04.052

Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*. https://doi.org/10.1016/j.physd.2019.132306

Zhao, J., Huang, F., Lv, J., Duan, Y., Qin, Z., Li, G., & Tian, G. (2020). Do RNN and LSTM have Long Memory? In H. D. III & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning* (Vol. 119, pp. 11365–11375). PMLR. https://proceedings.mlr.press/v119/zhao20c.html