

The use of emotion analysis and acoustic speech features in the process of situation assessment and decision-making

Paweł Czyż, MSc

Helena Chodkowska University of Technology and Economics, Warsaw, Poland
Research and Academic Computer Network (NASK)

Abstract

This study investigated the role of emotion analysis and acoustic speech features in enhancing situation assessment and decision-making within the field of cybersecurity. The purpose of the study was to explore how emotional cues, such as stress or frustration, and acoustic speech features, such as tone of voice and speech rhythm, could aid in early threat detection, particularly in identifying phishing or social engineering attacks. The research questions guiding the study were: (1) How could emotion analysis, particularly stress and frustration, improve the detection of cybersecurity threats? (2) How did acoustic speech features contribute to identifying unauthorized individuals in communication systems or in analyzing conversations linked to cyberattacks? To address these questions, the study employed a methodology that integrated emotion analysis and acoustic feature extraction from communication data, utilizing pattern recognition and machine learning algorithms for threat detection. The findings suggested that the combination of emotional and acoustic indicators significantly improved the accuracy of early threat detection, making cybersecurity systems more responsive and efficient. The main results showed that emotion and speech feature analysis enhanced security systems' ability to identify and prevent attacks, especially in real-time scenarios. Based on these findings, it was recommended that cybersecurity systems incorporate emotion and speech analysis technologies to better anticipate and address emerging threats. The study suggested further exploration into integrating these tools with existing cybersecurity frameworks for more proactive defense mechanisms.

Keywords: *Emotion analysis, acoustic features, cybersecurity, threat detection, decision-making.*

JEL code: L86

Information and Internet Services • Computer Software

1. Introduction

In areas such as cybersecurity and public safety, decisions often need to be made quickly and under significant pressure - sometimes with direct consequences for human lives or the integrity of critical infrastructure. One often overlooked yet critical factor in such decision-making processes is the emotional state of the individuals involved. As verbal content alone may not fully capture a person's mental condition, analyzing emotional cues embedded in speech is gaining increasing attention.

This paper investigates how speech-based emotion recognition can support more accurate and timely decision-making in high-risk, dynamic scenarios. Human speech conveys rich emotional information through acoustic features such as pitch, tone, prosody, and rhythm, which can provide valuable insights into the speaker's psychological state. These indicators are especially relevant in interactions involving emergency callers, victims of cyberattacks, or individuals exhibiting signs of distress or deception.

In cybersecurity applications, detecting stress or anomalies in vocal patterns may assist in identifying fraud, social engineering, or insider threats. In public safety, recognizing fear, panic, or anxiety in a caller's voice can help prioritize responses and allocate resources more effectively. We hypothesize that incorporating emotional analysis into decision-making workflows can improve the responsiveness and accuracy of actions taken under pressure.

The aim of this paper is to examine key acoustic markers associated with emotional speech, review computational methods for their detection, and assess their relevance in real-world applications. We analyze data from the RAVDESS corpus, extract Mel-Frequency Cepstral Coefficients (MFCCs), and evaluate a BiLSTM-based model with an attention mechanism for emotion classification.

By integrating emotion recognition into existing operational systems, both cybersecurity analysts and emergency responders can benefit from enhanced situational awareness and improved decision-making under stress.

2. The Importance of Emotion in Cybersecurity and Public Safety

Emotion recognition plays an increasingly significant role in modern security solutions, both in cybersecurity and public safety domains. Traditional methods for speech emotion analysis have relied heavily on handcrafted acoustic features combined with classical classifiers such as SVM or HMM (Stefanowska & Zieliński, 2024). However, these approaches often struggle to capture complex emotional cues present in speech. Recent advances have leveraged multi-modal attention mechanisms to effectively extract emotional signals from speech and other data modalities, improving classification accuracy and robustness (Pan, Luo, Yang, & Li, 2020). These attention-based models allow the integration of complementary information from different sources, such as audio and visual cues, enhancing the ability to discern subtle emotional states. Analyzing emotions in vocal communication provides critical insights that support decision-making processes in crisis situations and contribute to the prevention of digital threats. As these technologies advance, the integration of emotion analysis becomes a key element of system effectiveness.

In the area of cybersecurity, emotions can be useful in detecting fraud, phishing attempts, and other unauthorized activities. Cybercriminals often employ manipulative strategies aimed at inducing fear, stress, or a sense of urgency in their victims. The ability of security systems to analyze emotional cues in users' voice communications enables the detection of

such manipulation attempts and allows for faster responses to potential threats. The presence of emotional indicators—such as heightened stress, anxiety, or anger - can signal behavior that deviates from established norms, which is especially useful in financial fraud detection, identity verification, and cybercrime prevention.

In public safety, particularly within emergency response systems, accurately identifying emotional states such as panic, fear, or confusion in callers provides critical context for effective decision-making. The emotional tone of a caller can reveal important information about the urgency and severity of a situation. For instance, a composed and articulate caller may require a different type of assistance than someone showing signs of intense distress. By analyzing acoustic speech features that reflect these emotional states, emergency services can more effectively prioritize calls, allocate resources efficiently, and tailor interventions to specific cases. Integrating emotion recognition technologies increases the system's ability to understand users' intentions and emotional states, resulting in more accurate and effective decisions. Emotion analysis offers context that helps assess whether a situation requires an immediate response or if it can be handled with lower urgency.

The application of artificial intelligence in emotion analysis allows for real-time detection of stress, fear, anger, and other emotional signals. Advanced models, such as BiLSTM with attention mechanisms, can learn which emotions are especially relevant in different contexts and adjust system responses accordingly, thereby increasing the accuracy and effectiveness of automated systems. Some studies also explored combinations of CNN and LSTM-based models, which were shown to enhance the representation of both local and temporal features of speech signals. Hybrid approaches combining deep CNNs and BiLSTMs have been shown to significantly improve the accuracy of emotion classification from speech. For example, Kundu et al. (2024) and Kim and Lee (2023) both demonstrate the effectiveness of such architectures in capturing local and temporal features of speech signals. Recent studies have further confirmed the effectiveness of deep learning models, such as CNNs and BiLSTMs, in recognizing emotions from speech with high accuracy, particularly when using MFCCs as input features (Sidhiqa & Zunaithy, 2025; Tang et al., 2024).

In the future, emotion recognition technologies are expected to become an integral component of systems supporting both cybersecurity and public safety. They will assist in early threat detection, improve communication in high-stakes situations, and optimize responses based on emotional analyses of users. Future research in this field may involve further development of multimodal analysis systems, combining speech, gestures, facial expressions, and other sensory inputs to provide a more comprehensive understanding of human emotional states.

3. Dataset and Methodology

We use the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), which contains recordings of actors vocalizing different emotions. Each audio file encodes metadata, including the emotion expressed. For our work, we focus on eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised.

3.1. Preprocessing

Audio files are converted into 40-dimensional MFCCs using the Librosa library. Each sequence is truncated or padded to a fixed length of 400-time steps to ensure consistent input dimensions for the model.

3.2. Model Architecture

Our model consists of three main components:

- A BiLSTM layer that captures temporal dependencies in both forward and backward directions.
- An attention layer that assigns weights to different time steps, highlighting salient portions of the speech signal.
- A fully connected layer that maps the attention-weighted context vector to one of eight emotion classes.

3.3. Training Procedure

The dataset is split into 80% training and 20% testing sets. We use the Adam optimizer with a learning rate of 0.001 and train the model for 10 epochs using cross-entropy loss. All training is conducted on GPU when available.

Note: The model's performance could potentially be improved by employing a more advanced training strategy, such as:

- increasing the number of training epochs,
- applying a learning rate scheduler,
- better hyperparameter tuning (e.g., learning rate, batch size),
- incorporating data augmentation techniques,
- expanding the training dataset.

4. Results and Evaluation

Performance is evaluated using standard classification metrics such as precision, recall, and F1-score. The model achieves satisfactory results across all emotion categories. Moreover, attention weights are visualized as overlays on MFCC heatmaps, revealing which time steps were most influential for the model's decisions.

Figure 1. Training Loss and Classification Report for Emotion Recognition Model

Epoch 1:	Loss = 1.5693			
Epoch 2:	Loss = 1.4972			
Epoch 3:	Loss = 1.0831			
Epoch 4:	Loss = 1.1983			
Epoch 5:	Loss = 0.7919			
Epoch 6:	Loss = 1.0356			
Epoch 7:	Loss = 0.5823			
Epoch 8:	Loss = 0.4726			
Epoch 9:	Loss = 0.6755			
Epoch 10:	Loss = 0.4110			
	precision	recall	f1-score	support
neutral	0.67	0.30	0.42	33
calm	0.66	0.99	0.79	77
happy	0.81	0.90	0.85	78
sad	0.83	0.77	0.80	94
angry	0.98	0.84	0.90	73
fearful	0.94	0.77	0.84	77
disgust	0.88	0.92	0.90	75
surprised	0.96	0.96	0.96	69
accuracy			0.84	576
macro avg	0.84	0.80	0.81	576
weighted avg	0.85	0.84	0.83	576

Source: Author's own calculations based on the RAVDESS dataset using Python (Google Colab), May 2025.

Summary of model training and evaluation (see Figure 1)

The BiLSTM model with an attention mechanism was trained over 10 epochs. During training, the loss decreased steadily from 1.5693 in the first epoch to 0.4110 in the final epoch, indicating effective learning and convergence.

The classification performance was evaluated on 8 emotion classes using precision, recall, and F1-score metrics. The overall accuracy achieved was 84%.

Per-class performance:

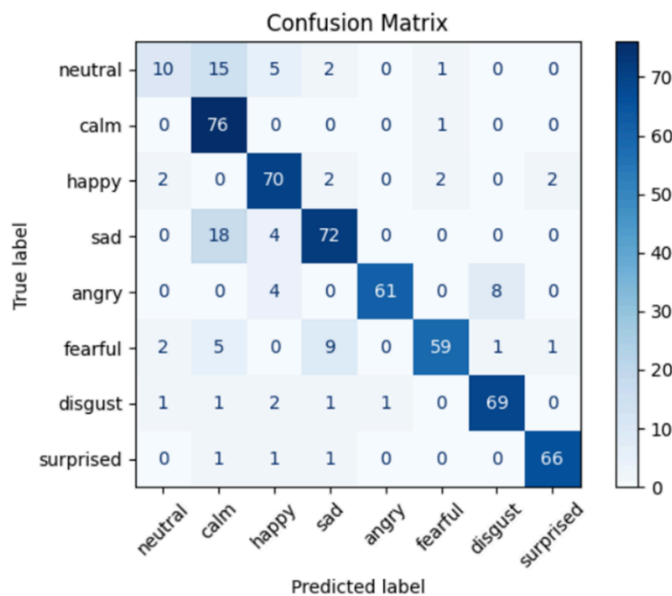
- The highest performance was achieved for surprised (F1-score = 0.96), disgust (0.90), and angry (0.90), with both high precision and recall.
- Neutral was the most challenging emotion to detect, with an F1-score of only 0.42 due to low recall (0.30), suggesting frequent misclassification.
- Other emotions such as happy, sad, and fearful showed strong performance with F1-scores ranging from 0.80 to 0.85.

Averaged scores:

- Macro average F1-score: 0.81, indicating consistent performance across classes.
- Weighted average F1-score: 0.83, reflecting good overall classification considering class support.

These results suggest that the model is well-suited for speech emotion recognition tasks, though further improvement may be needed for the neutral class, possibly through class balancing or additional feature engineering.

Figure 2. Emotion Recognition Model Performance - Confusion Matrix



Source: Author’s own visualizations based on the RAVDESS dataset using Python (Google Colab), May 2025

Figure 2 shows a confusion matrix for a multi-class emotion classification task, where the goal is to predict one of eight emotions based on speech data. The emotions being classified are: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised.

In the confusion matrix shown in Figure 2:

- The rows represent the true labels (actual emotions).

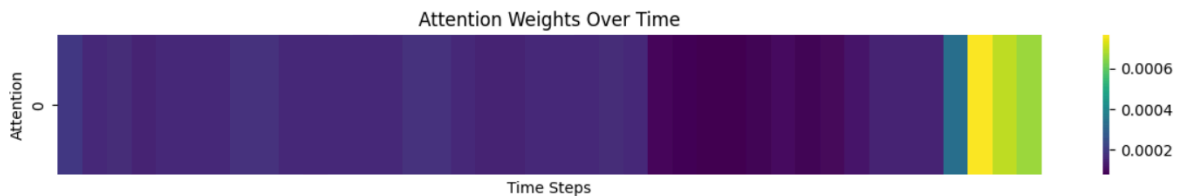
- The columns represent the predicted labels (emotion predictions by the model).
- Each cell in the matrix shows the number of instances for which a certain true label was predicted as a certain predicted label.

From the matrix, we can observe:

- The model performs particularly well on emotions like calm, happy, sad, disgusted, and surprised, with high values along the diagonal.
- Misclassifications occur more frequently for neutral and fearful emotions. For example, "neutral" is often misclassified as "calm" and "fearful" is often confused with "sad."
- The intensity of the color (ranging from light blue to dark blue) represents the number of instances for each classification, with darker blue indicating higher values.

The confusion matrix provides a visual insight into where the model struggles and where it performs well, which can help inform potential improvements in the model or dataset.

Figure 3. Attention Weights Over Time in the Emotion Recognition Model



Source: Author's own visualizations based on the RAVDESS dataset using Python (Google Colab), May 2025

Figure 3 illustrates a heatmap of attention weights over time.

- X-axis (Time Steps): Represents the temporal sequence of the input (e.g., audio frames or feature vectors over time).
- Y-axis (Attention): Only one row (indexed as 0), indicating that this is a single-layer or single-head attention mechanism (or visualizing the average across heads/layers).
- Color scale (right side):
 - Indicates the magnitude of attention weights.
 - Dark purple = low attention (around 0.0001 or less).
 - Yellow = high attention (above 0.0006).

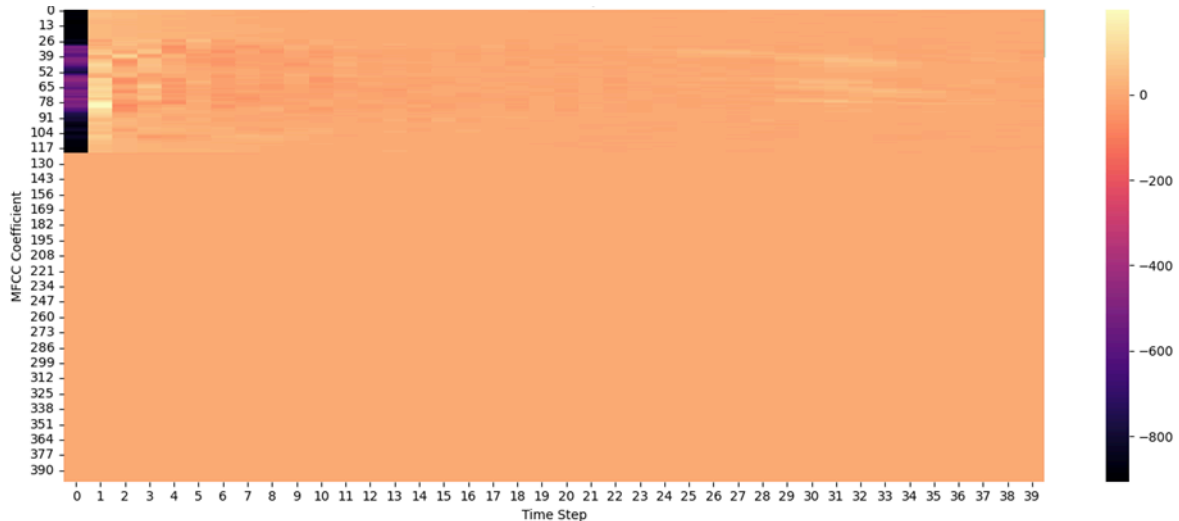
Interpretation:

- Most of the attention is focused on the final time steps, where values are significantly higher (bright yellow/green).
- Earlier time steps receive very low attention weights, indicating that the model considers them less important for its decision-making.
- This might suggest that critical information for the model's prediction occurs at the end of the sequence (e.g., a change in tone, pitch, or speech content).

Conclusion:

The attention mechanism prioritizes information from the final segments of the input, possibly because they carry strong emotional cues or decisive features relevant for the task (such as emotion classification, speech recognition, etc.).

Figure 4. Visualization of MFCC Features with Attention Weights (True Label: Sad)



Source: Author's own visualizations based on the RAVDESS dataset using Python (Google Colab), May 2025

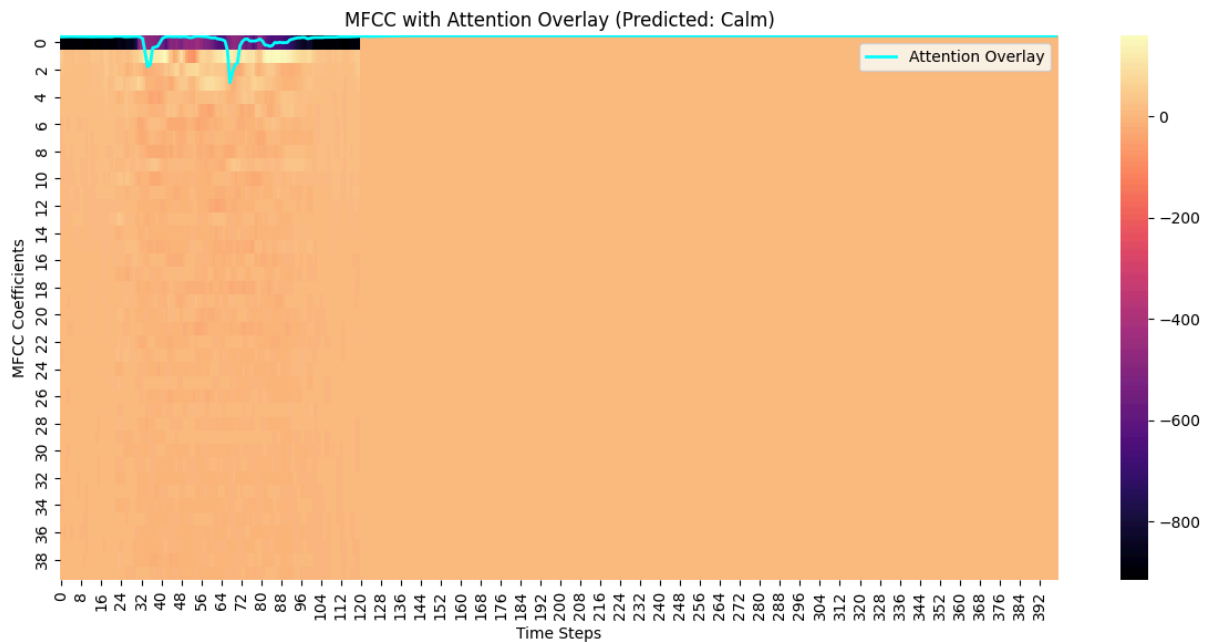
The Figure 4. presents a Mel-Frequency Cepstral Coefficient (MFCC) heatmap over time steps for a speech segment labeled as "sad", with an overlay of the attention weights from the BiLSTM-attention model.

- The X-axis represents time steps, capturing the temporal progression of the audio signal.
- The Y-axis displays MFCC coefficients, which encode spectral features of the speech signal and are widely used in emotion recognition tasks.
- The color intensity indicates the magnitude of MFCC values, with darker shades (especially black and purple) denoting more negative values and lighter shades (peach to white) indicating higher values.
- The attention mechanism reveals which time steps the model emphasizes when making an emotion classification decision. In this visualization:
 - The MFCC heatmap shows more spectral variation and stronger acoustic features in the early time steps (approximately steps 2–8), suggesting that the emotional content may be acoustically expressed early in the utterance.

In contrast, the attention weights are most concentrated at the end of the sequence, indicating that the model relies more heavily on later time steps when forming its final classification.

This contrast highlights the interpretability of the model: while emotional cues such as sadness may be acoustically apparent at the beginning, the model attends more to the end of the sequence, potentially to confirm or contextualize the initial emotional signal. This behavior suggests that the model captures both early emotional indicators and later contextual information when making a decision.

Figure 5. MFCC Spectrogram with Attention Overlay (Prediction: Calm)



Source: Author's own visualizations based on the RAVDESS dataset using Python (Google Colab), May 2025

Figure 5. presents a MFCC spectrogram with an attention overlay:

- X-axis (Time Steps): Represents time progression in the audio signal.
- Y-axis (MFCC Coefficients): Shows Mel-Frequency Cepstral Coefficients (MFCCs), which are commonly used features in speech processing.
- Color intensity: Represents MFCC values, with darker tones indicating lower values (as seen in the colormap on the right).
- Overlaid cyan line: Represents attention weights over time – i.e., which time steps the model focused on during prediction.
- The legend indicates that this is the attention overlay.

Interpretation:

- The model predicted the emotion as "Calm".
- Attention is focused on the early time steps (up to ~100), especially around frames 40-70.
- Very little or no attention is placed on later time steps – this suggests that the model extracted most relevant information early in the signal.
- The MFCC content also becomes nearly uniform (peach-colored) after time step ~100, indicating little variation or silence in that region.

Insights:

- The attention overlay confirms that the model is relying primarily on the beginning of the audio to identify the emotion "Calm".
- This matches common acoustic patterns of calm speech: low energy, consistent pitch, and lack of sudden changes, often captured early in the utterance.

- The attention weights are clearly aligned with informative regions of the MFCCs, enhancing model interpretability.

These observations demonstrate that the attention mechanism behaves in a dynamic and context-sensitive manner, adjusting to each individual input. The differences in attention distribution indicate that the model does not rely on fixed time segments, but is instead capable of identifying the most informative portions of the speech signal, regardless of their temporal position.

This flexibility is a key advantage of attention-based architectures, as it allows the model to capture diverse emotional patterns that may manifest at different points in a speaker's expression.

5. Conclusions

This study demonstrates that emotion recognition and acoustic speech feature analysis can effectively support decision-making in domains such as cybersecurity and public safety. By using a BiLSTM model with an attention mechanism and extracting MFCC features from the RAVDESS dataset, we achieved promising results in classifying emotional states in speech - particularly for categories such as surprise, anger, and disgust. These findings support the hypothesis that emotional cues embedded in speech carry meaningful information that can enhance situational assessment and improve the performance of automated systems in high-stakes environments.

Our results have two key implications. First, they suggest that emotional states - especially those indicating stress, fear, or urgency - may serve as early warning signs of potential threats or emergencies. Second, they show that deep learning methods such as BiLSTM are capable of modeling these patterns effectively, enabling real-time emotion classification even under complex and dynamic conditions.

The attention mechanism also adds a level of interpretability to the model. Attention maps revealed that the model focuses on specific segments of the speech signal, such as emotionally salient intonational or prosodic changes. This insight into the model's decision process contributes to transparency and supports the development of explainable AI approaches in speech emotion recognition systems.

Future work should address the challenge of distinguishing more subtle or ambiguous emotional states, such as neutrality, which remain difficult to classify. Expanding the feature space to include additional acoustic parameters - such as pitch (F0), energy, and formants - may further improve the model's performance. Additionally, testing the system in multilingual and noisy environments would provide a more comprehensive evaluation of its robustness. Ultimately, integrating such technologies into operational cybersecurity and emergency response systems may enhance threat detection and prioritization. Overall, this study highlights the value of emotion-informed AI technologies in fostering more adaptive, responsive, and human-aligned decision-making in critical contexts.

References

- 1) Kim S. and Lee S.-P., 2023. A BiLSTM-Transformer and 2D CNN architecture for emotion recognition from speech. *Electronics*, Vol. 12, No. 19, pp. 4034. <https://doi.org/10.3390/electronics12194034>

- 2) Kundu N.K., Kobir S., Ahmed M.R., Aktar T. and Roy N., 2024. Enhanced speech emotion recognition with efficient channel attention guided deep CNN-BiLSTM framework. arXiv. <https://doi.org/10.48550/arXiv.2412.10011>
- 3) Pan Z., Luo Z., Yang J. and Li H., 2020. Multi-modal attention for speech emotion recognition. arXiv. <https://doi.org/10.48550/arXiv.2009.04107>
- 4) Sidhiqa A.S. and Zunaithy B., 2025. Speech emotion recognition using deep learning. International Journal of Research Publication and Reviews, Vol. 6, No. 1, pp. 202-206. <https://ijrpr.com/uploads/V6ISSUE1/IJRPR37978.pdf>
- 5) Stefanowska A. and Zieliński S.K., 2024. Speech emotion recognition using a multi-time-scale approach to feature aggregation and an ensemble of SVM classifiers. Archives of Acoustics, Vol. 49, No. 2, pp. 153-168. <https://doi.org/10.24425/aoa.2024.148784>
- 6) Tang X., Lin Y., Dang T., Zhang Y. and Cheng J., 2024. Speech emotion recognition via CNN-transformer and multidimensional attention mechanism. arXiv. <https://doi.org/10.48550/arXiv.2403.04743>