Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL)
ISSN P: 2303-5528 | E-ISSN: 2303-6397
Vol. 13 (2) 2026

Deep Learning-Based Emotion Recognition for Better Human-Computer Interaction

Adrian Kovalen
Danubia Technical University, Austria

Submission : 17/10/2025 Acceptance : 20/07/2026 Published : 11/08/2026

Abstract:
In order for systems to comprehend and respond more appropriately to user emotions, emotion recognition from speech is an essential component in enhancing human-computer interaction (HCI). An in-depth examination of deep learning methods for voice emotion recognition, focusing on convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid designs that integrate feature extraction and temporal analysis. We examine the description of features using spectrograms and mel-frequency cepstral coefficients (MFCCs), as well as the impact of acoustic features on the model's ability to classify emotions thru pitch, tone, and energy. We evaluate various deep learning algorithms that utilize large datasets and robust model designs to accurately identify emotions such as happiness, sadness, anger, and neutrality. Cultural differences, speaker unpredictability, and ambient noise are among the challenges to emotion perception that this study addresses. understanding how to utilize deep learning to create emotion recognition systems that are far more accurate and responsive, and how to design human-computer interfaces that are more empathetic and responsive to users' needs.

keywords Emotion Recognition, Speech Analysis, Deep Learning, Human-Computer , Interaction (HCI)

Introduction

Emotion recognition from speech is an important part of developing HCI since it helps computers to comprehend and respond correctly to user emotions. Virtual assistants, contact centers, and interactive learning platforms are examples of voice-based systems that are rapidly growing in popularity. Improving user experience, empathy, and efficacy in these areas relies heavily on the ability to accurately discern emotions. One of the best methods to express ourselves verbally is by subtle changes in pitch, tone, and volume. By associating human emotional displays with automated answers, emotion recognition makes use of these audio cues to create interactions that appear more real and meaningful. Emotion recognition has been greatly enhanced by deep learning, which trains models on vast amounts of audio data. Many speech analysis tasks have demonstrated promise for both Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs are great at extracting features from spectrograms, while RNNs are good at capturing temporal correlations in audio sequences. For accurate emotion classification, a hybrid architecture combining RNNs and CNNs is the way to go because it takes advantage of both spatial and temporal analysis. Spectrograms and mel-frequency cepstral coefficients (MFCCs) are acoustic features that these models use to pick up on the nuanced vocal expressions of emotions including happiness, sadness, rage, and

neutrality. Acquiring trustworthy emotion identification is challenging because to factors including cultural differences in emotional expressiveness, individual speech pattern variability, and real-world noise. We require models that can learn from diverse datasets and generalize well to multiple contexts and speaker characteristics if we are to overcome these challenges. Since emotion detection systems are finding more and more real-world applications, it is crucial that they are accurate enough to function in loud environments, such as outdoors or with background noise. investigating the performance of various model structures and acoustic feature representations for emotion detection from speech using deep learning methods. By examining how different deep learning algorithms handle emotional signals, this study aims to assist designers in creating HCI systems that are more sympathetic and reactive. Strong emotion recognition models can improve many other applications, including entertainment, virtual assistance, customer service, and mental health support, leading to more adaptive and user-centered human-computer interaction (HCI) experiences.

Empathy Recognition with Deep Learning

Deep learning has substantially enhanced emotion recognition from speech by enabling systems to identify complex emotional patterns that are frequently overlooked by traditional approaches. These algorithms can take audio data and extract spatial and temporal information using a mix of deep learning architectures like RNNs, CNNs, and hybrid models. Because these models learn to link certain emotional states with temporal and auditory elements like pitch, tone, and energy, they are excellent resources for creating more sensitive and sympathetic HCI systems. The following deep learning methods are mainly employed for emotion recognition from speech.

1. Convolutional Neural Networks (CNNs) for Feature Extraction

For processing and feature extraction from spectrograms and mel-frequency cepstral coefficients (MFCCs), two visual representations of audio signals, convolutional neural networks (CNNs) find extensive application in emotion recognition. In order for CNNs to detect changes in emotional tone, they apply convolutional filters to these representations. These features include frequency patterns and energy shifts.

- **Spectrograms as Input**: Spectrograms, which show the frequency and amplitude changes of an audio stream over time, work very well with CNNs. In the frequency domain, emotions such as melancholy and rage may exhibit distinct patterns that convolutional neural networks (CNNs) can train to identify.
- **MFCCs for Feature Richness**: Emotional speech can be more accurately identified with the use of convolutional neural networks (CNNs) trained on MFCC recordings of the short-term power spectrum. When it comes to emotion categorization, CNNs trained on MFCC representations perform better.

2. Recurrent Neural Networks (RNNs) for Temporal Analysis

Emotions expressed through speech are dynamic; they develop with time. Analysis of temporal patterns in speech is ideally suited to RNNs, particularly LSTM networks and GRUs, due to their exceptional ability to capture sequential relationships. RNNs take in data in a sequential fashion, remembering details from earlier frames; this allows them to comprehend emotional development and context.

- **LSTM Networks for Long-Term Dependencies**: LSTMs can detect the development of emotional cues throughout an audio recording since they are good at capturing long-term dependencies. An LSTM network can tell the difference between, say, a steady eruption of rage and a level-headed, prolonged display of indifference.
- **GRUs for Computational Efficiency**: GRUs can still capture temporal patterns in speech, but they do it with fewer parameters, making them more computationally efficient. This feature is similar to LSTMs.

3. Hybrid Architectures for Emotion Recognition

Hybrid models capture spatial features and temporal dependencies within a single framework by combining CNNs and RNNs, leveraging the capabilities of both architectures. The use of convolutional neural networks (CNNs) for feature extraction and recurrent neural networks (RNNs) for temporal dynamics enables robust processing of complicated audio data.

- **CNN-LSTM Models**: Spectrograms or MFCCs are used by these models' CNN layers to extract spatial information, which are subsequently fed into LSTM layers, which capture the speech signal's sequential structure. By offering a more comprehensive view of the auditory data, this method has been demonstrated to enhance performance in emotion recognition tasks.
- **Attention Mechanisms**: To improve the model's capacity to detect nuanced emotional signals, some hybrid models use attention layers to zero in on specific parts of speech that are more suggestive of certain emotions.

4. Autoencoders for Unsupervised Emotion Recognition

One application of autoencoders in emotion recognition is unsupervised feature learning; these systems learn to compress and rebuild data. A model can develop latent representations of audio signals that can highlight patterns relevant for emotion classification by training autoencoders on unlabeled speech data.

- **Variational Autoencoders (VAEs)**: VAEs add a stochastic component to autoencoding, which aids the model in producing latent features that represent the speech data's underlying emotional changes. When there is a lack of labelled emotional data, VAEs can help with transfer learning.
- **Denoising Autoencoders**: Emotion identification applications benefit greatly from these autoencoders since they are trained to recover clean audio data from inputs that contain noise.

5. Transfer Learning and Pretrained Models

Emotion identification systems can adapt to various tasks with minimal further training by leveraging pre-trained deep learning models on big audio datasets through transfer learning. Emotion recognition from voice is one area where transfer learning shines because it enhances model performance in real-world circumstances and decreases the requirement for big annotated datasets.

- **Pretrained CNNs and RNNs**: To improve training speed and generalizability, models that have already been trained on large-scale audio datasets for general audio classification can be fine-tuned for emotion recognition tasks.

- **Multimodal Transfer Learning**: Emotion identification can be made even more robust by combining pretrained models from several modalities, such face recognition and text-based sentiment analysis, which each offer unique insights into the emotional environment.

Using deep learning techniques like CNNs, RNNs, and hybrid architectures, emotion recognition from speech has been substantially enhanced. Emotion recognition relies on spatial and temporal aspects, which these methods capture. These techniques, together with autoencoders and transfer learning, have enhanced systems' ability to understand emotional signals, leading to more adaptable and sympathetic responses from computers, which in turn has improved human-computer interaction. Integrating these methods with attention mechanisms and multimodal inputs will undoubtedly lead to the development of better and more context-aware emotion recognition models in future studies.

Conclusion

Emotion recognition from voice, powered by deep learning, is a groundbreaking advancement in human-computer interaction. By enabling systems to recognize and comprehend emotional signals, deep learning-based emotion identification paves the way for a more personalized, empathetic, and responsive user experience. Systems that capture both the spatial and temporal components of speech, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid designs, significantly improve the accuracy of emotion classification. These algorithms use acoustic cues such as pitch, tone, and energy to distinguish between emotional states such as happiness, sadness, anger, and neutrality. Beyond the potential, there are challenges to overcome in emotion recognition in speech, including differences in speaker variability, cultural standards about emotional expressiveness, and background noise. To manage this complexity, we need strong models that can adapt to human vocal inflections and generalize to varied contexts. By incorporating approaches such as attention mechanisms and transfer learning, systems can be enhanced in their resilience and contextual awareness, leading to even better performance. More natural and flexible human-computer interaction (HCI) applications enabled by deep learning-based emotion recognition could have far-reaching benefits in many domains, including customer service, mental health support, education, and entertainment. Better understanding of multimodal emotion recognition and ethical concerns is required before we can fully use emotion-aware technologies to enhance our everyday interactions with computers.

Bibliography

- Author. (2024). AI-Driven Anomaly Detection in Network Security: A Comparative Study of Machine Learning Algorithms. *Journal of Quantum Science and Technology*, *1*(3), 94–98. https://doi.org/10.36676/jqst.v1.i3.31
- Gupta, R. (2024). Cybersecurity Threats in E-Commerce: Trends and Mitigation Strategies. *Journal of Advanced Management Studies*, *1*(3), 1–10. https://doi.org/10.36676/jams.v1.i3.13
- Lippon Kumar Choudhury. (2022). STUDY ON LOGIC AND ARTIFICIAL INTELLIGENCE SUBSETS OF ARTIFICIAL INTELLIGENCE. *Innovative*

Research Thoughts, *8*(1), 127–134. Retrieved from https://irt.shodhsagar.com/index.php/j/article/view/1114

- Singh , R. (2024). Artificial Intelligence and the Future of Legal Practice: Opportunities and Ethical Challenges. *Indian Journal of Law*, *2*(4), 57–61. https://doi.org/10.36676/ijl.v2.i4.41
- Singh, S. (2024). Advancements in Targeted Cancer Therapies: A Review of Recent Medical Research. *Shodh Sagar Journal for Medical Research Advancement*, *1*(3), 1–5. https://doi.org/10.36676/ssjmra.v1.i3.19