International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 03 | March – 2025 DOI: 10.55041/ISJEM02527
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Survey on Speech Emotion Recognition with Expressive Speech Synthesis

Abhimanue S[1], **Dr. Jyothish K John**[2]

[1]Department of Computer Science and Engineering
Federal Institute of Science And Technology, Angamaly,
683 577, Kerala, India
[2]Department of Computer Science and Engineering
Federal Institute of Science And Technology, Angamaly,
683 577, Kerala, India

-- ***--

Abstract - Emotion plays a key role in identifying the state of a person, that is, whether they are angry, sad, happy, etc. The paper presents an integrated framework that recognizes emotions from speech, generates emotionally aware responses, and synchronizes facial expressions to provide an animated video response. The system provides real-time, empathetic interactions for emotional support. It focuses on identifying the emotion of the person, especially to know if the person is depressed or having a hard time, so that it can provide emotional support to them, to overcome the feeling of distress and isolation using emotion-incorporated synthesis. The proposed system provides responses like an actual human and keeps them company in the absence of an actual individual, who provides emotional support. It acts like a companion or friend in the time of need. The framework employs an integrated approach combining Wav2Vec for deep learning-based speech emotion recognition, Coqui TTS for emotion-aware voice synthesis, and Neuro Sync for synchronized facial animation visualization through Unreal Engine's Meta Human models. Unlike existing solutions that rely on handcrafted features or separate disconnected components, our framework creates an end-to-end solution that achieves 74.506% emotion recognition accuracy while generating contextually appropriate emotional responses with synchronized facial expressions in real-time. This multimodal approach distinguishes itself from traditional methods by seamlessly bridging emotion recognition and response generation, creating a more natural and empathetic human-computer inter- action experience specifically designed for emotional support applications.

Key Words: Emotion recognition, emotion-incorporated syn- thesis, emotional support agent

1. INTRODUCTION

Emotion is important in communication and understanding the state of a person. An appropriate response can only be given after the emotion has been identified. During states of distress, appropriate responses are necessary for the feeling to be overcome. The emotion of the person is identified by the proposed solution using a Wav2vec model for emotion classification identifying eight different emotions which includes calm, happy, sad, angry, fearful, disgusted, surprised, and neutral. The model is trained using four different datasets such as TESS, RAVDESS, CREMA-D, and SAVEE, by which around 12162 audio files are comprised. A probabilistic classification of emotions is produced by the Wav2vec classifier head. Once the emotional state of the person is identified by the system, a response is given back using an LLM, text-to-speech system with voice cloning to mimic a human and a Meta Human model which is animated to speak the voice response. Emotional sup- port is mainly focused on by the system for people who are lonely and depressed so that their depression can be overcome and a return to their normal state of life can be achieved.

In an increasingly digital world, emotional well-being and mental health have been made critical concerns, particularly for individuals by whom loneliness or depression is experienced. While significant advancements have been made by conversational AI systems, the ability to understand and respond to users' emotions effectively is often lacked by them, by which their ability to provide meaningful emotional support is limited. An intelligent system is needed by which the emotional state of a person can be accurately identified from their speech and emotionally appropriate responses can be delivered in real time. As a virtual companion, comfort and support should be offered by such a system, by which individuals are helped to over- come feelings of distress and regain emotional balance. This gap is addressed by the development of a speech-based emotion recognition system combined with emotion-incorporated synthesis to deliver empathetic, human-like responses.

2. LITERATURE REVIEW

Hu and Huang et al. [1] investigated how Emotion-Aware Conversational Agents incorporate emotional intelligence to enhance user experiences through real-time emotion detection and response. Their research examines the integration of speech emotion recognition and empathetic response generation to create more meaningful human-computer interactions. These intelligent systems employ Speech Emotion Recognition (SER) to identify various emotional states including happiness, sadness, anger, and surprise. Advanced models can even assess continuous emotional dimensions such as valence and arousal levels during conversations. By examining acoustic features in speech—including tonal qualities, pitch variations, and speaking cadence—these agents generate responses that correspond appropriately to the user's emotional state. This synchronization creates a more authentic and supportive conversational experience. For instance, emotion-aware agents in mental health contexts provide empathetic responses through validation, encouragement, or cognitive reframing techniques, offering emotional support during interactions. Their experimental study with 75 participants demonstrated that these systems not only enhanced perceived emotional intelligence but also contributed to stress reduction. This research under- scores the significant potential of emotion-aware agents across various fields including mental health support, educational environments, and interactive virtual assistants.

In the paper Speech Emotion Recognition with Multilayer Perceptron (MLP) [2], examines how MLP models can recognize emotions in speech. MLPs are widely used due to their simplicity and efficiency, making them well-suited for tasks involving features like Mel-frequency cepstral coefficients (MFCC), chroma features, and spectral properties. These features capture key emotional cues from speech, allowing MLP models to perform well on smaller datasets with relatively fast training times. Despite their advantages, MLPs struggle to model temporal dependencies—an important factor in emotion recognition, where the sequence and timing of sounds matter. This limitation makes MLPs less effective for tasks that require detailed sequence modeling, such as recognizing emotions in continuous speech.

The Hybrid Models Combining CNN and GRU [3], demonstrates the advantages of using a hybrid approach in SER. Convolutional Neural Networks (CNNs) extract local patterns from speech signals, capturing frequency and temporal variations that indicate emotional states. Meanwhile, Gated Recurrent Units (GRUs) process these features while maintaining temporal dependencies in speech data. By combining CNNs and GRUs, this approach effectively handles both spatial and temporal aspects of emotion in speech. CNN layers identify crucial local features, while GRU layers ensure the model retains con- text and sequential information. Studies have shown that this hybrid model outperforms traditional approaches in SER tasks, making it useful for real-time applications like virtual assistants, affective computing, and interactive gaming.

OpenSMILE for Feature Extraction [4], discusses the use of OpenSMILE as a powerful tool for extracting features in SER. OpenSMILE provides a standardized process for analyzing speech, capturing essential acoustic features such as energy, pitch, spectral characteristics, and formants. These features are crucial for detecting emotional expressions in speech. One key advantage of OpenSMILE is its consistency, which allows re- searchers to compare results across different datasets. Many studies have integrated OpenSMILE with deep learning models, leading to improved accuracy in emotion recognition tasks. The toolkit's ability to extract dynamic features over time also makes it particularly effective in capturing evolving emotional expressions during speech.

Lee and Williams [5], explore how attention mechanisms enhance emotion recognition models. Attention mechanisms help models focus on the most relevant parts of speech while filtering out irrelevant data. In SER, this technique allows models to identify specific speech segments where emotional cues are most prominent, leading to more accurate classifications. By emphasizing emotional highlights, attention mechanisms im- prove the recognition of subtle emotional variations, such as sarcasm or mixed emotions. Research has shown that adding attention layers to RNN-based models can increase accuracy by over 10%, especially in long and noisy audio sequences. These improvements make attention-based models particularly valuable in customer service applications, virtual assistants, and mental health support systems.

In Spectrogram-Based Emotion Recognition with Convolutional Transformer [6], researchers introduce a hybrid architecture that combines CNNs with Transformer layers for emotion recognition from audio spectrograms. CNN layers ex- tract spatial features, capturing localized emotional cues, while Transformer layers use attention mechanisms to learn temporal dependencies. This combination enhances the model's ability to understand emotional context in speech. Trained on the RAVDESS dataset using the Adam optimizer, the model achieved 87% accuracy, outperforming CNN-only models, which reached 78%. The addition of Transformer layers significantly improved temporal pattern recognition. Future research aims to extend this architecture to other audio classification tasks, making it applicable beyond emotion recognition.

The research Cross-Corpus Generalization [7], addresses the challenge of building SER models that perform well across different datasets. Many models excel on specific datasets but struggle with data from different sources due to variations in recording conditions and speaker demographics.To improve generalization, researchers have explored techniques like adversarial training and domain adaptation. These approaches help models learn features that remain consistent across datasets. Studies using adversarial training on RAVDESS, CREMA-D, and TESS datasets show that models trained with such techniques achieve better real-world performance. This research is particularly relevant for emotion recognition systems used in customer service and interactive voice assistants, where diverse speech inputs are common.

In Real-Time Processing for SER [8], focuses on reducing processing time in SER applications. Real-time recognition is essential for interactive applications like virtual assistants, gaming, and storytelling, where emotional responses must be detected quickly. To achieve low-latency processing, researchers have developed lightweight models, such as 1D- CNNs and optimized RNN variants, which reduce computational complexity. Some models can process speech in under 100 milliseconds, making them suitable for interactive systems. However, there is a trade-off between speed and accuracy, as complex models require more computation. Future work aims to balance these factors for optimal performance in real-time applications, especially on resource-constrained devices.

In Multimodal Emotion Recognition [9], integration of speech with other modalities—such as facial expressions, body language, and physiological signals—to enhance emotion recognition accuracy is being studied. While speech alone pro- vides valuable emotional information, combining it with visual and physiological data allows for a more comprehensive understanding of emotions. Studies show that merging audio with facial expression data improves accuracy by up to 15% com- pared to unimodal systems. This is particularly useful for recognizing complex emotions that may not be clearly expressed through speech alone. The use of wearable sensors and biometric monitoring also expands the potential of multimodal emotion recognition, particularly in healthcare applications where real-time emotion monitoring can improve patient care.

Emotional Speech Recognition Using Deep Neural Net- works [10], introduces a novel approach that combines CNN, CRNN, and GRU architectures for enhanced emotion recognition. Using Mel-spectrograms as input, CNN layers extract spatial features, while GRU layers capture temporal dependencies in emotional speech. The CRNN model integrates CNNs and RNNs, improving the processing of sequential data. Tested on the IEMOCAP[11] dataset, this model achieved an accuracy of 97.47%, significantly outperforming traditional RNN-based systems, which reached 88%. The use of dropout and batch normalization further improved generalization. Future research aims to integrate multimodal inputs—such as visual and physiological signals—to enhance cross-domain emotion recognition.

3. EMOTION-RECOGNIZED RESPONSE SYSTEM

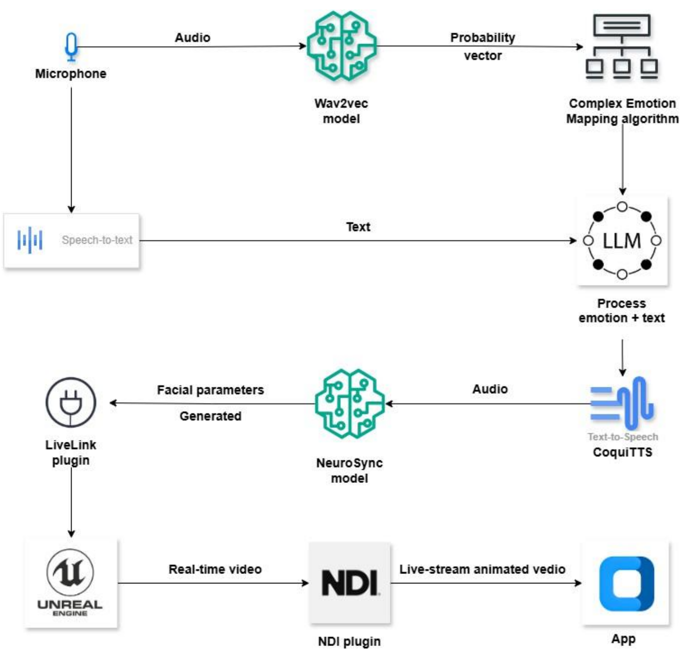

FIGURE I: SYSTEM ARCHITECTURE FOR THE EMOTION-RECOGNIZED RESPONSE SYSTEM.

This system processes user speech through a pipeline that begins with voice input capture, followed by Wav2vec model analysis of audio features to classify basic emotions probabilistically. A specialized algorithm then maps these basic emotions to more complex emotional states. The system generates text responses via Meta LLM[12], which are converted to natural-sounding speech using CoquiTTS[13] voice cloning technology. Simultaneously, Neuro Sync generates facial expression parameters that animate a Meta Human model in Unreal Engine. The animated output is live-streamed through NDI into a custom tkinter application, creating a complete emotion-aware conversational experience.

The Figure I represents the architecture of the system and its working. The major steps in the process are:

- Voice Input: – The system first captures raw speech from the user via a microphone or audio input device.
- Probabilistic classification: – The speech signal is processed by Wav2vec model, which extracts key audio features like MFCC (Mel Frequency Cepstral Coefficients), energy, pitch, and other prosodic features. These features provide a detailed representation of the emotional content embedded in the speech signal. The features are then used by the model to produce a probability vector of

basic emotions like happy, sad, angry etc.

- Complex emotion Classification using emotion mapping algorithm: – The probability vector is used by a complex emotion mapping algorithm which maps basic emotions to more complex emotions like depressed, melancholic, annoyed etc.
- Response Generation:- The text response for the voice in- put is generated using Meta LLM. The response is then converted to human like voice response by making use of voice cloning in CoquiTTS model.
- Facial parameter generation: - The facial parameters are generated using NeuroSync model, which is then passed to the Unreal Engine via livelink plugin.
- Emotion-Incorporated Synthesis using Unreal Engine and NDI SDK: – The Unreal Engine standalone uses the facial parameters passed by the NeuroSync model to animate a Meta Human model and the animation is live- streamed using NDI plugin into a custom tkinter application.

4. TECHNOLOGICAL STACK

The system integrates a variety of state-of-the-art technologies to perform speech emotion recognition and emotion- incorporated synthesis. Below is a breakdown of the key technologies used in the development of the system:

- **Wav2vec** - Wav2vec is a self-supervised learning model for speech representation learning. It is designed to process raw audio waveforms and learn meaningful speech representations and is mainly used for speech recognition and other audio-based tasks. The wav2vec model is fine-tuned using the dataset and the wav2vec classifier head is used for probabilistic classification.
- **Meta Llama API** - Meta's Llama 3 is the latest iteration in their series of large language models (LLMs), de- signed to understand and generate human-like text. The Llama API along with prompt engineering is used to generate human like response.
- **CoquiTTS** - is an open-source text-to-speech (TTS) framework that allows you to train and use deep learning models for speech synthesis. It's known for its flexibility, high-quality speech generation, and support for multiple languages and voices.
- **NeuroSync** - It is a transformer-based sequence-to-sequence neural network that generates real-time facial animations from audio input. It produces facial blend- shape coefficients which can be integrated with Unreal Engine using LiveLink for real-time animation.
- **Unreal Engine** - It is a powerful real-time 3D creation tool developed by Epic Games. It is widely used for game development, film production, architectural visualization, virtual reality, and simulation applications.
- **Python** - The system is implemented using Python, a versatile programming language that supports the integration of various machine learning libraries and frame- works.
- **TensorFlow and PyTorch** - These deep learning frame- works are utilized for different components of the system. PyTorch is used for implementing and fine-tuning the Wav2Vec model for speech emotion recognition, leveraging its dynamic computation graph for efficient training. TensorFlow is used to manage and optimize the Coqui TTS model for voice synthesis, ensuring high- quality emotion-aware speech generation. Both frame- works enable seamless integration of deep learning techniques within the pipeline, from emotion classification to real-time speech synthesis and animation.

Algorithm 1 Emotion Recognition and Response Generation

```
Input: Audio input from the user
Output: Emotion-incorporated voice response and ani- mated response
Step 1: Probabilistic Classification
The Wav2vec model is used for probabilistic classification by training with raw audio data.
Use RAVDESS, CREMA-D, TESS, and SAVEE datasets to train and test the model
The output is a probability vector of 8 emotions.
Step 2: Complex Emotion Mapping
The probability vector is used for complex emotion map- ping from among the 8 emotions. The complex emotions include frustrated, anxious, depressed, worried, amused, etc.
Step 3: Emotion-Incorporated Speech Synthesis
Generate a text response based on the detected emotion.
Pass the text response into CoquiTTS to produce a voice response.
The voice response is fed to the NeuroSync model to pro- duce facial parameters.
Unreal Engine animates a MetaHuman model for lip synchronization.
The animated response is live-streamed to a custom Tkinter application.
Step 4: End Process
The process is now complete, and the user hears the generated emotion-aware speech response.
```

5. DISTINCTION FROM EXISTING SYSTEMS

Aspect	Emotion-Recognized Response System	Existing Systems
Feature Extraction	Uses Wav2Vec for self-supervised feature extraction from raw audio, capturing detailed speech representations without requiring handcrafted features.	**Open SMILE for Feature Extraction**: Extracts traditional features like MFCC, pitch, and energy but lacks deep contextual understanding.
Emotion Classification	Wav2Vec's classifier head predicts emotions directly from audio, leveraging pre- trained speech representations for robust recognition.	**MLP and CNN-GRU models**: Typically used for SER, but may require extensive training and fine-tuning to generalize well.
Emotion-Incorporated Synthesis	Uses Coqui TTS for high-quality voice cloning to synthesize emotional speech from text, preserving speaker identity.	**Tacotron 2 and WaveGlow**: Commonly used for speech synthesis but may lack personalized voice cloning capabilities.
Facial Animation	Neuro Sync maps emotional speech to facial animation, generating realistic lip-sync and expressions.	**Deep Learning-Based Facial Animation**: Existing methods use CNNs or GANs for lip-syncing but may not fully capture fine-grained emotional expressions.
Real-Time Emotion Recognition and Animation	Enables real-time processing by integrating Wav2Vec (SER), Coqui TTS (voice synthesis), and Unreal Engine (animation via Live Link).	**Lightweight 1D-CNNs for Real-Time SER**: Some models prioritize speed over accuracy, while existing real-time animation tools may rely on predefined gestures.
Streaming and Application Integration	Uses the NDI plugin to stream animation from Unreal Engine to a Tkinter-based application, enabling seamless live interaction.	**Standard Streaming Pipelines**: Other systems may rely on OpenCV or game engine exports, which can introduce latency or require additional processing.
Generalization Across Datasets	Wav2Vec's self-supervised training allows better generalization across datasets, reducing reliance on labeled data.	**Domain Adaptation in SER**: Existing methods often require fine-tuning for different datasets, limiting cross-corpus performance.
Multimodal Emotion Recognition (Future Scope)	Currently focuses on speech-based emotion recognition but can integrate with vision-based emotion analysis in future up- dates.	**Multimodal Emotion Recognition**: Many systems use audio-visual fusion, achieving higher accuracy but requiring more computational resources.
Explainability (Future Scope)	Could incorporate explainability techniques to provide insights into Wav2Vec's emotion predictions for improved trust- worthiness.	**Explainable AI in SER**: Existing methods attempt to visualize model decisions, but this is still an evolving area.

4.CONCLUSION

Significant progress in speech emotion recognition (SER) has been made with the integration of deep learning models, by which both accuracy and real-time processing have been enhanced. The proposed system is designed with Wav2Vec for feature extraction and classification, through which the need for handcrafted features is eliminated while deep contextual speech representations are captured. More robust and efficient emotion detection compared to traditional methods is ensured by this approach. High-quality voice cloning that preserves the speaker's identity and emotional tone is enabled by Coqui TTS, which is employed to synthesize emotion-aware speech. The generated speech is then mapped to realistic facial animations using Neuro Sync, by which synchronized lip movements and expressive reactions are created. A Meta Human model is animated through the integration of Unreal Engine with Live Link, and seamless live streaming to a Tkinter based application is facilitated by the NDI plugin, through which real-time interaction is allowed.

Valuable applications in virtual assistants, interactive AI, and mental health monitoring can be found in the ability to recognize emotions from speech and generate synchronized emotional responses. Accuracy and realism could be further enhanced as future improvements by expanding the system to incorporate multimodal emotion recognition—where speech is combined with visual and physiological cues. Additionally, the framework could be made more suitable for real-time applications by optimizing for low-latency processing on edge devices. The development of emotionally intelligent systems is enabled by the integration of Wav2Vec for SER, Coqui TTS for voice synthesis, Neuro Sync for animation, and Unreal Engine for visualization. Human-computer interactions are enhanced by this combination, through which the way is paved for AI-driven ap- plications that can perceive, interpret, and respond to emotions in a more human-like manner.

REFERENCES

[1] J. Hu, "The acoustically emotion-aware conversational agent with speech emotion recognition and empathetic responses," *IEEE Transactions on Affective Computing*, 2023.

[2] A. A. Alnuaim, "Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier," 2022.

[3] S. Chamishka and T. Jayasinghe, "Speech emotion recognition using cnn and bilstm models," *Multimedia Tools and Applications*, vol. 81, 2022.

[4] E. Nachmani, "Spoken question answering and speech continuation using spectrogram-powered llm," 2024.

[5] M. Lee and S. Williams, "Real-time emotion detection using recurrent neural networks with gru and attention," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022.

[6] D. Clark and L. Davis, "Spectrogram-based emotion recognition with convolutional transformer," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[7] S. Zhang, Y. Wang, and H. Li, "Spontaneous speech emotion recognition using multiscale deep convolutional lstm," *IEEE Transactions on Affective Computing*, 2022.

[8] S. Chamishka, "A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling," 2022.

[9] J. Harris and O. Garcia, "Bimodal emotion recognition using cnn and lstm for audio-visual fusion," *IEEE Trans- actions on Affective Computing*, vol. 13, 2022.

[10] J. Hu and L. Xie, "Emotional speech recognition using deep neural networks," *Sensors*, vol. 22, 2022.

[11] C. Busso and C.-C. Lee, "Iemocap: Interactive emotional dyadic motion capture database," 2007.

[12] C. Cummins and V. Seeker, "Meta large language model compiler: Foundation models of compiler optimization," 2024.

[13] E. Go¨lge, "Coquitts an open-source text-to-speech framework," 2021.