REST Journal on Data Analytics and Artificial Intelligence
Vol: 4(1), March 2025
REST Publisher; ISSN: 2583-5564
Website: http://restpublisher.com/journals/jdaai/
DOI: https://doi.org/10.46632/jdaai/4/1/90

Towards More Human Machines: Speech Emotion Recognition Using Mel-Spectrogram-Based Reconstruction

*** T. Sai Teja Sri Sanjana, K. Shivani, R. Kanishka, M. Madhavi**
Anurag University, Hyderabad, India.
*Corresponding Author Email: saitejasrisanjana@gmail.com

Abstract: *Speech Emotion Recognition is a field that aims to interpret human emotions from speech signals. By leveraging advanced machine learning and signal processing techniques, this project aims to create a model capable of detecting a range of emotions such as happiness, sadness, anger, fear, surprise, disgust, and neutrality by analyzing speech patterns. Specifically, the project focuses on implementing a self-supervised learning approach through the vector quantized masked autoencoder for speech (VQ-MAE-S), which is fine-tuned for optimal emotion recognition in various contexts.*

Key Words: *Speech Emotion Recognition (SER), VQ-MAE-S, autoencoders, Deep Learning, mel-spectrogram, RAVEDESS.*

1. INTRODUCTION

SER is an advanced topic of HCI (Human Computer Interaction) dealing with Speech Emotion Recognition analysis and interpretation of emotions conveyed through speech. As opposed to sentiment analysis in text, SER focuses on those aspects of communication that are non-verbal, instead using vocal features such as pitch, tone, rhythm, volume, and speech rate to make inferences about a speaker's emotional state. This allows SER systems to draw an immense array of emotions such as happiness, sadness, anger, fear, surprise, disgust, and neutral [1-4]. Importance of Speech Emotion Recognition (SER) Emotion plays a critical role in communication. Human interactions are often driven by feelings, and the ability to recognize and respond to these emotional cues is essential for understanding intent and facilitating meaningful conversations. Incorporating SER into technology allows machines to bridge this gap and create more intuitive interactions [6-9]. Enhancing Human-Computer Interaction (HCI) The capability of instilling emotion recognition capabilities complements the encounters between man and machine. For instance, in virtual assistants, SER could be applied to systems such as Siri or Alexa and would not only reply to a query but also give an appropriate response based on the emotional tone of the individual. It will create more personalized and empathetic interactions, therefore, making technology human-like [10-12]. SER in Customer Service and Marketing: In customer service, SER systems can detect frustration, dissatisfaction, or happiness in a customer's voice. Companies can use this insight to route calls more effectively, offer emotional support, or provide better-tailored solutions. For example, a customer service bot equipped with SER can detect an agitated customer and immediately transfer them to a human agent for quicker resolution. In marketing, SER aids in analyzing customer sentiment during product pitches, helping businesses craft messages that resonate emotionally with their target audience [13-17]. Adaptive Gaming and Entertainment: In terms of gaming, the SER can create more immersive and adaptive experiences. This helps games adjust and pace according to the player's emotional state. This is because games make experiences more dynamic when adjusted based on the emotional state of players. Thus, a system can automatically adjust difficulty levels or offer support if there is some frustration or disengagement. Thirdly, SER can make virtual reality environments richer, responding in real-time to the user's emotions-for example, in connection and immersion [18-20]. Thus, with the integration of high-performance techniques from deep learning, feature extraction, and signal processing, the SER system will operate under advanced machine learning settings: among the key methods used will include spectrogram

Copyright@ REST Publisher

analysis, pitch contour extraction, and decompositions of speech signals to extract features relevant for the classification of emotional status, then fed into a deep-learning model. The goal is to create a system that is usable on the ground in any application where sensitive, responsive emotion detection is needed. In a nutshell, the project is therefore to alter how machines perceive human emotions so that richer interactions in virtual assistants, customer services, marketing, games, and healthcare applications may be possible.

2. EXISTING SYSTEM

SER is an area within affective computing. This research interest grows in the detection of human emotions from speech signals. Work in the area has been documented for over two decades and has brought many developments among which is the use of Machine Learning techniques in extracting emotional cues from speech, an important development toward improving HCI. The para-linguistic features such as tone, pitch, and intensity of speech would refer to the emotional aspects and thus make speech an ideal modality for detecting emotion. One major challenge in SER is speaker independence. As the accents, gender, and speaking styles vary among different speakers, major variability is introduced into the speech data that actually affects the classification accuracy. In recent works, researchers have mainly targeted these problems by using state-of-the-art normalization strategies, such as cross-speaker histogram equalization along with robust feature extraction strategies. The other major challenge involved was low classification accuracy observed in the experiments of the speaker-independent setup, which was majorly caused due to the lack of generalization across diverse datasets. The quality of SER models mainly depends on features that are extracted from speech data. The features include classical features: pitch, energy, and zero-crossing rate. Traditional features also include spectral features such as Mel-Frequency Cepstral Coefficients (MFCC). Deep learning models allow the automatic extraction of deeper features to capture more complex patterns in speech. Wavelet transforms, and short-time Fourier transforms have also been used to provide time-frequency representations of speech signals to enhance the ability of the system to recognize emotions. There are many datasets that have been useful in furthering SER research. Among them, the datasets of RAVDESS and IEMOCAP are considered to be well-known for the evaluation of SER models because they possess rich emotional content and speaker diversity. These datasets include happiness, sadness, anger, and sometimes surprise, thus often coming into use, not just in traditional ML-based studies but also in DL-based SER studies. Existing SER systems primarily employ ML and DL techniques for emotion recognition from speech signals. These systems are essentially a pipeline of data pre-processing, feature extraction, and then the actual classification of emotion. Support Vector Machines (SVM), Gaussian Mixture Models (GMM) and k-nearest Neighbors (KNN) are some of the traditional ML algorithms that are applied to SER tasks. However, recent advancements show that deep learning approaches whereby models like CNNs, RNNs, or other types of autoencoders are used for the automatic learning of features from raw speech dataend. Most SER systems, including in this project, rely on well-known datasets, namely RAVDESS, IEMOCAP, Emo-DB, and SAVEE, providing a variety of labeled speech samples to train models, which actually covers many of these emotions: happiness, sadness, anger, and fear, and which helps to bench Markus around different studies. Various techniques, including machine learning and deep learning models, have been developed to improve the accuracy and robustness of the system. Numerous existing systems focused on using multimodal input, transfer learning, and CNN to classify emotions in real-time, especially for the dataset we have chosen i.e. RAVDESS. CNN-14 and bi-LSTM with attention mechanism, fine-tuning CNN-14 has improved emotion classification to 80.08% using the PANNs framework. ER uses the aural transformers and action units which combine acoustic features via Wav2Vec2.0 with facial units to improve recognition with an accuracy of 81.82%. The shallow CNNs outperformed the deep neural networks in noisy environments and on smaller datasets achieving 82.99% accuracy. Fine tuning of CNN-14 and AlexNet for speech recognition reaching an accuracy of 76.58%.

Copyright@ REST Publisher

3. LITERATURE REVIEW

TABLE 1. Literature Review

Year	RF.NO	Method	Dataset	Metric	Result
2023	1	VQ-MAE-S	RAVDESS – speech RAVDESS – song IEMOCAP EMODB	Accuracy, f1-score	84.1, 84.4 85.8, 85.7 66.4, 65.8 90.2, 89.1
2023	2	MFV and CNN	RAVDESS	Accuracy	92%
2021	19	Spatial Transformer Networks (STN) and a bi-LSTM with an attention mechanism	RAVDESS	Accuracy	80.08%
2021	20	fine-tuned xlsr-Wav2Vec2.0 transformer	RAVDESS	Accuracy	81.82%

4. METHODOLOGY

A SER system is proposed for speech signal processing. Therefore, the architecture considered is based on the self-supervised learning approach VQ-MAE-S. Further illustration of the architectures, algorithms, methods, and technical requirements of the system are provided: The proposed system for Speech Emotion Recognition is based on the VQ-MAE-S, which leverages self-supervised learning techniques to recognize emotions in speech signals. The architecture consists of the following components:

VQ-VAE Encoder: The VQ-VAE encoder maps raw speech audio to compact discrete latent representations, and thus flexibility is provided in feature extraction efficiently.

How it does:

- Input: The raw audio signal is taken first, and processing will be done to form a power spectrogram using tools and techniques such as Short-Time Fourier Transform, which are provided through libraries like Librosa.
- Latent Space Representation: The encoder, through vector quantization, maps the spectrogram to a latent space. This takes continuous values from the spectrogram into a discrete number of limited values - codes. This captures meaningful patterns while reducing the dimensionality, which is foremost essential for the learning efficiency of the model.

MAE:

- Patch-Based Masking: Under the MAE approach, masking happens on most of the quantized speech tokens, meaning they are not observed by the model. That leads to training by forcing the model to be able to reconstruct the masked parts from the remaining visible parts of the input, which encourages the understanding of deeper patterns in data.
- Encoder: The encoder uses a transformer-based architecture which is particularly excellent at learning contextual relationships. It processes the visible tokens (unmasked portions) to generate a robust representation of the speech input with pertinent emotional cues.
- Decoder: The decoder is tasked to reconstruct the masked tokens from the encoded output. This makes the decoder fill gaps that would improve the model's ability to cope up with partially available data. Consequently, this improves its robustness for noise and variability in speech inputs.

Emotion Classifier

The system is fine-tuned on labeled emotional speech datasets like RAVDESS and IEMOCAP to specialize in emotion detection after pre-training with a large corpus.

Classification Approaches:

- Linear Layer: A shallower version where the learned features are subjected to a linear transformation for emotion classification.
- Query2Emo: A stronger version that enhances emotion understanding by forcing specific features related to emotions in the learned representations. This way, more contextual understanding, and better performance can be achieved by the classifier.

Training Phases

In its developmental stage, the model will undergo two major phases:

- Pre-training Phase: Learning generalized speech representations from an extremely large, unsupervised dataset like VoxCeleb2. This pre-training phase is very important because it will allow the model to gain a bit of very basic knowledge of speech characteristics in various settings. Pre-training is conducted in a self-supervised learning regime where the model learns to predict masked tokens, thus providing strong feature representations adaptable to various tasks including emotion recognition.
- Fine-tuning Phase: Adapt learned representations to be utilized for emotion classification tasks with supervised datasets like RAVDESS and IEMOCAP. In this stage, the model is trained on emotions data, that specializes in the recognition and classification of various emotions. The fine-tuning process refines feature representations, thereby increasing the accuracy of the model and generalization to unseen data.

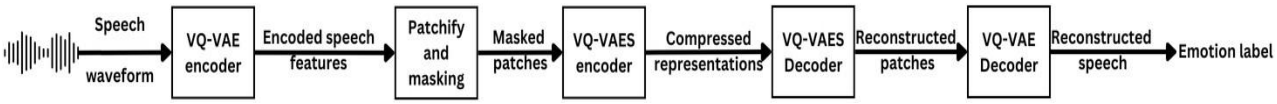

FIGURE 1: Architecture

Module Design and Organization: Speech Waveform: The waveform of speech is basically the raw audio signal that has the representation of sound, generally time-domain, where the amplitude of the sound wave is plotted over time. This waveform is captured via microphones or audio recording devices and is usually stored in formats like WAV or MP3. The waveform of the audio signal captures the raw features, including intonation, rhythm, and emotional tones that speech consumes. It's sampled at a particular frequency, usually 16 kHz or 44.1 kHz. This discretizes the continuous wave of sound into a discrete signal. VQ-VAE Encoder: The VQ-VAE encoder takes in the raw waveform of speech and compresses it into a more compact latent representation. Learn a meaningful compressed representation of audio, preserving key features with the noise sneaking out from there. The encoder then takes in the raw waveform its spectrogram representation and transforms this into a set of quantized vectors in a latent space. It accomplishes this by passing audio data through convolutional layers that extract hierarchical features. The output of the encoder will be a sequence of discrete codes that represent the audio signal in reduced form. Such codes are possible to capture patterns and trends in the speech signal relevant to further processing. Patchify and Masking: At this stage, the encoded representation is split into small patches and presented for masking. Patchify: The output of the encoder in VQ-VAE is split into small patches or segments. Each patch or segment would contain part of the audio signal. Dividing in this manner helps control the data size while letting the model zoom in on smaller areas of information that could be worked with. Masking: One section of the patches is masked randomly, meaning they are invisible to the model. Therefore, only a few patches are viewable for training while others are masked out. The principle of masking is to make the model learn robust features that can be generalized well even if some of the input data are missing. This also enhances the model's ability to infer missing information based on the context given by the visible patches. VQ-VAE Decoder: The VQ-VAE decoder reconstructs the original audio representation from the encoded latent representation obtained in the former steps. To learn to recover the masked (hidden) parts of the audio conditioned on its visible parts. The decoder takes the encoded representation and tries to reconstruct the complete patches set, including the masked ones. The missing information can then be reconstructed by the decoder, allowing the model to learn how to perform well with incomplete and noisy data. Rebuilt Speech: The output corresponding to the reconstructed audio waveform is obtained from the decoder of VQ-VAE. Reconstruct the original speech signal along with masked information. The encoded/decoded staged learned features generate the reconstructed audio as close to the original speech wave as ideally possible, and this reconstructed audio captures the details and the nuances required in emotion detection. Emotion Classification: This is the final step of the overall technique after speech reconstruction wherein the emotions occurring in an audio signal are classified. Identify and classify the emotional status of the speaker based on features extracted throughout the architecture. This reconstructed speech is fed to a classification model for emotions. The

proposed model could be a simple linear layer or a more advanced method, like Query2Emo, that uses learned features to classify emotions like happiness, sadness, anger, etc. The model analyzes the characteristics of the speech-for example, the tone, pitch, and rhythm-and determines the underlying emotion, and gives out the corresponding classification.

5. FUTURE DIRECTION

As we look into the future, several paths become apparent which would be avenues for further development based on this project to improve the capabilities and applications of the developed SER system. Several of these avenues address current limitations but also introduce SER technology into new and potentially impactful areas. Language and Accent Recognition: Among the biggest challenges that will be present in emotion detection is variability due to different languages and regional accents. Future work focuses on extending the SER system to recognize emotions in different languages and harder accents: data curation and incorporation of multilingual datasets containing varied dialects and accents for effective generalization. Transfer Learning: Utilizing the techniques of transfer learning to leverage this trained model in one language to others for comprehension and processing thereby strengthening its ability. Testing and Fine-Tuning: Further testing and fine-tuning with extremely high accuracy in languages appropriately to grasp cultural sensitivities in the expression of emotions. Enhancement for Noisy Worlds: Though the current model was noise-resistant, there was still room for improvement. This next attempt will involve Advanced Noise Reduction Techniques: Utilizing advanced noise cancellation algorithms for pre-processing audio input to reduce interference of background noises on emotion detection. Adversarial Training: The training employs adversarial training. For this purpose, the adverse noise environment is allowed to surface while training. This improves a model's ability to identify emotional cues even if there is interference present. Real-World Testing: Detailed field testing in the environment of different application settings for developing and improving model performance on real-world, noisy data. Building Practical Applications: SER technology has matured into a degree of technological readiness that would open many doors to practical applications. Future work will focus on: Mental health evaluation tools: A collaboration with mental health experts in developing applications that make use of SER, to assess the emotional well-being of a patient will provide valuable information to the therapists in observing the states. Customer service enhancement application development: As customers interact, an application can decode the interactions in real-time, hence it helps a business to respond accordingly to the emotions received leading to high customer satisfaction and retention. Sentiment Analysis Across Industries: The growth of SER technology in finance, marketing, and education to analyze customer feedback as a way to better understand how ways to improve the user experience improve engagement. Integration with Other Modalities: To provide an even more multifaceted understanding of human emotions, future research will continue to integrate SER with other modalities: Combining Audio Emotion Detection with Facial Expression Analysis to Increase Precise Accuracy and Compensate for the Emotional Context It will be achieved through face recognition. Different contexts can lead to more subtle perceptions of emotional states. Body Language Analysis: How Nonverbal Behavior like Posture and Gesture Can Complement Speech Emotion Detection Should Understand Body Language to Discover Emotions Better for the System Even in Face-to-Face Interactions. Multimodal Systems: Design multimodal systems that integrate knowledge from sources such as audio, visual, and kinesthetic to reconstruct a comprehensive view of human emotions in interaction with computers that will enrich the depth of human-computer interaction. Some channels for future work undertaken by the SER project to venture are to overcome the current limitations of the platform and further its applicability to various contexts. This will not only increase the strength of the technology but also see it find applications in various disciplines, gradually taking it to a more profound understanding of human emotions and interaction with machines that behave more empathetically. The potential positive benefits from integrating SER with other modalities promise a radical influence on areas such as mental health, customer service, emotional intelligence in technology, and so on.

6. CONCLUSION

In conclusion, this project significantly advances Speech Emotion Recognition by integrating cutting-edge techniques such as vector quantization, autoencoder models, and self-supervised learning. The model's ability to accurately classify emotions in real-time, even in noisy environments, showcases its robustness and potential for real-world applications. With a focus on multimodal data integration and diverse emotional state recognition, the system enhances human-computer interaction, offering promising implications in fields like mental health, customer service, and entertainment. This work paves the way for more empathetic, intuitive, and efficient technologies, positioning itself as a foundation for future innovations in emotional intelligence for AI.

Copyright@ REST Publisher

REFERENCES

[1]. Samir Sadok, Simon Leglaive, Renaud Séguier. "A Vector Quantized Masked Autoencoder for Speech Emotion Recognition." CentraleSupélec, IETR UMR CNRS 6164, France, 2023. [arXiv:2304.11117v1]

[2]. Chandrasekhara Reddy, T., Sirisha, G., Reddy, A.M. Smart Healthcare Analysis and Therapy for Voice Disorder using Cloud and Edge Computing, Proceedings of the 4th International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2018, 2018, pp. 103–106, 9001280

[3]. Chandrasekhara Reddy, T., Srivani, V., Mallikarjuna Reddy, A., Vishnu Murthy, G. Test case optimization and prioritization using improved cuckoo search and particle swarm optimization algorithm, International Journal of Engineering and Technology(UAE), 2018, 7(4.6 Special Issue 6), pp. 275–278

[4]. Dayaker, P., Chandrasekhara Reddy, T., Honey Diana, P., Mallikarjun Reddy, A. Recent trends in security and privacy of sensitive data in cloud computing, International Journal of Engineering and Technology(UAE), 2018, 7(4.6 Special Issue 6), pp. 241–245

[5]. Mallikarjuna Reddy, A., Venkata Krishna, V., Sumalatha, L. Face recognition approaches: A survey, International Journal of Engineering and Technology(UAE), 2018, 7(4.6 Special Issue 6), pp. 117–121

[6]. Obulesh, A., Vamshi Krishna, P., Mallikarjuna Reddy, A. Classification of lung patterns based on transition neighborhood pattern (TNP), Journal of Advanced Research in Dynamical and Control Systems, 2018, 10(15), pp. 35–41

[7]. Manoranjan Dash, N.D. Londhe, S. Ghosh, et al., "Hybrid Seeker Optimization Algorithm-based Accurate Image Clustering for Automatic Psoriasis Lesion Detection", Artificial Intelligence for Healthcare (Taylor & Francis), 2022, ISBN: 9781003241409

[8]. Manoranjan Dash, Design of Finite Impulse Response Filters Using Evolutionary Techniques - An Efficient Computation, ICTACT Journal on Communication Technology, March 2020, Volume: 11, Issue: 01

[9]. Manoranjan Dash, "Modified VGG-16 model for COVID-19 chest X-ray images: optimal binary severity assessment," International Journal of Data Mining and Bioinformatics, vol. 1, no. 1, Jan. 2025, doi: 10.1504/ijdmb.2025.10065665.

[10]. Manoranjan Dash et al.," Effective Automated Medical Image Segmentation Using Hybrid Computational Intelligence Technique", Blockchain and IoT Based Smart Healthcare Systems, Bentham Science Publishers, Pp. 174-182,2024

[11]. Manoranjan Dash et al.," Detection of Psychological Stability Status Using Machine Learning Algorithms", International Conference on Intelligent Systems and Machine Learning, Springer Nature Switzerland, Pp.44-51, 2022.

[12]. Samriya, J. K., Chakraborty, C., Sharma, A., Kumar, M., & Ramakuri, S. K. (2023). Adversarial ML-based secured cloud architecture for consumer Internet of Things of smart healthcare. IEEE Transactions on Consumer Electronics, 70(1), 2058-2065.

[13]. Ramakuri, S. K., Prasad, M., Sathiyanarayanan, M., Harika, K., Rohit, K., & Jaina, G. (2025). 6 Smart Paralysis. Smart Devices for Medical 4.0 Technologies, 112.

[14]. Kumar, R.S., Nalamachu, A., Burhan, S.W., Reddy, V.S. (2024). A Considerative Analysis of the Current Classification and Application Trends of Brain–Computer Interface. In: Kumar Jain, P., Nath Singh, Y., Gollapalli, R.P., Singh, S.P. (eds) Advances in Signal Processing and Communication Engineering. ICASPACE 2023. Lecture Notes in Electrical Engineering, vol 1157. Springer, Singapore. https://doi.org/10.1007/978-981-97-0562-7_46.

[15]. R. S. Kumar, K. K. Srinivas, A. Peddi and P. A. H. Vardhini, "Artificial Intelligence based Human Attention Detection through Brain Computer Interface for Health Care Monitoring," 2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON), Dhaka, Bangladesh, 2021, pp. 42-45, doi: 10.1109/BECITHCON54710.2021.9893646.

[16]. Vytla, V., Ramakuri, S. K., Peddi, A., Srinivas, K. K., & Ragav, N. N. (2021, February). Mathematical models for predicting COVID-19 pandemic: a review. In Journal of Physics: Conference Series (Vol. 1797, No. 1, p. 012009). IOP Publishing.

[17]. S. K. Ramakuri, C. Chakraborty, S. Ghosh and B. Gupta, "Performance analysis of eye-state charecterization through single electrode EEG device for medical application," 2017 Global Wireless Summit (GWS), Cape Town, South Africa, 2017, pp. 1-6, doi:10.1109/GWS.2017.8300494.

[18]. Gogu S, Sathe S (2022) autofpr: an efficient automatic approach for facial paralysis recognition using facial features. Int J Artif Intell Tools. https://doi.org/10.1142/S0218213023400055

[19]. Rao, N.K., and G. S. Reddy. "Discovery of Preliminary Centroids Using Improved K-Means Clustering Algorithm", International Journal of Computer Science and Information Technologies, Vol. 3 (3), 2012, 4558-4561.

[20]. Gogu, S. R., & Sathe, S. R. (2024). Ensemble stacking for grading facial paralysis through statistical analysis of facial features. Traitement du Signal, 41(2), 225–240.