Speech Emotion Recognition Using Machine Learning: A Comprehensive Evaluation

Soumya Karuturi
Department of CSE
SRM University-AP
Soumya_karuturi@srmap.edu.in

Jyotika Tammineedi
Department of CSE
SRM University-AP
jyotika_tammineedi@srmap.edu.in

Indra Narayana Korrakuti
Department of CSE
SRM University-AP
indranarayana_korrakuti@srmap.edu.in

Anupoju Tejaswi
Department of CSE,
SRM University-AP, India
tejaswi_anupoju@srmap.edu.in

Murali Krishna Enduri
Department of CSE
SRM University-AP
muralikrishna_e@srmap.edu.in

***Abstract*−The recognition of human emotions through machine-based systems presents a complex and challenging task. Machine learning models aim to automate this process, enabling systems to learn and adapt to emotional cues. Despite advancements in the field, achieving accurate emotion recognition from speech with high performance remains a significant challenge. In the field of human-computer interaction, the recognition and categorization of emotions play a vital role. Emotions are typically identified by examining behavioral cues such as facial expressions, voice tone, and body language. This research explores two feature extraction techniques to improve speech-based emotion recognition. Anger, disgust, fear, happiness, neutrality, surprise, and sorrow are seven emotional states they are investigated by using the open-source Audio-Visual Database of Emotional Speech (ADVES) delivered information for this research. For emotion classification, a multilayer perceptron (MLP) classifier was employed. The researchers compared their proposed model's performance to similar studies and evaluated the results. Using AVDES, the proposed model achieved 97.59% overall accuracy in classifying the seven emotions. Based on experimental data, the ensemble learning methodology is superior when compared with other state-of -the-art methods.**

Index Terms−Affective computing, Audible feature, Machine learning, MLP, Speech emotions

I. INTRODUCTION

Emotional identification systems are essential for understanding human feelings and behaviors [1]. Two prevalent methods for emotional expression and recognition are affective speech and facial expressions [2]. Recognizing emotions through facial expressions involves interpreting visual cues, such as facial muscle changes, to infer emotional states. However, these systems are limited as they depend solely on external indicators, which may not accurately represent internal emotions, often leading to misclassification [3]. Emotions significantly influence decision-making in individual, governmental, and commercial sectors [4]. For individuals, emotions shape preferences and choices. In governmental and commercial contexts, understanding collective emotions is crucial for policymaking, marketing, and customer engagement. The internet's vast information has highlighted the need for advanced analytical techniques to process data effectively. Social media platforms, where people share daily activities, reactions, and emotions, provide unstructured data reflecting real-time emotional states, presenting opportunities and challenges for emotion identification systems [5]. Utilizing such data, emotion recognition technologies can offer deeper insights into human behavior, enhancing decision-making across various fields.

Emotions significantly shape daily life, influencing perceptions and responses to various situations [6]. They arise from experiences or interactions and crucially affect human behavior and decision-making, especially in domains like e-commerce, restaurants, movies, and customer satisfaction. Emotions guide consumer choices and opinions on services or products. Social media platforms now include emotional expressions in their interfaces; for example, Facebook's reaction options ("angry," "happy," "love," and "surprise") enable users to express feelings about content [7]. This underscores the growing importance of emotions in digital communication and the need for systems to interpret emotional expressions. Emotion analysis, a subfield of sentiment analysis, interprets judgments, responses, and feelings from various data sources, including text, and is vital in data mining, web mining, and social media analytics [8]. The rise of social media has heightened relevance of emotion analysis, as users express opinions and feelings on diverse issues through text, yielding valuable insights. These advancements make emotion analysis crucial for interpreting human behavior in online interactions, with practical applications in customer service, marketing, and social behavior research. Accurate emotion analysis enables organizations to better understand and respond to their audiences, fostering meaningful interactions and informed decision-making.

II. LITERATURE SURVEY

The advancement of (SER) systems has been driven by extensive worldwide research efforts. Initially, SER models primarily employed conventional machine learning approaches. For instance, Lee and Narayanan [9] applied statistical techniques like (HMMs) and (SVMs) to classify emotions using prosodic and spectral features. While these methods established a crucial foundation for SER by demonstrating the feasibility of emotion detection from speech, their dependence on manually crafted features restricted their effectiveness. Moreover, these traditional models struggled to generalize across diverse datasets and handle complex temporal patterns in speech. Recent progress has enabled the creation of models capable of automatically extracting relevant features from raw data, overcoming the

limitations of hand-crafted features. Trigeorgis et al. [10] showcased the potential of convolutional neural networks (CNNs) for SER using spectrogram-based inputs. For example, Fayek et al. [11] combined LSTMs with attention mechanisms, Transformer-based architectures have also made significant contributions to the field. Sawaf [12] demonstrated the efficacy of transformers in achieving state-of-the-art results on well-known datasets like IEMOCAP and RAVDESS, reporting test accuracies surpassing 90%. Another noteworthy development in SER is the emergence of multimodal approaches. Poria et al. [13] highlighted the advantages of integrating multiple data modalities, such as combining speech with textual information or visual cues like facial expressions. This multimodal fusion significantly enhances the robustness of emotion recognition, particularly in cases where audio data is noisy, ambiguous, or insufficient to accurately convey emotion. By leveraging complementary information from different modalities, these approaches have achieved superior performance compared to unimodal systems. Despite these advancements, data scarcity remains a persistent challenge in SER. The limited availability of high-quality labeled datasets hinders the training of complex models. These methods improve model generalization by simulating diverse audio conditions, thereby reducing the risk of overfitting.

The progression of (SER) systems from conventional machine learning techniques to sophisticated deep learning frameworks demonstrates continuous innovation in this domain. Researchers have substantially enhanced the precision and resilience of emotion detection by employing (CNNs), (LSTMs), transformer models, and multimodal strategies. The field's collaborative nature is evident in efforts to overcome challenges like limited data availability through transfer learning and data augmentation methods. These developments not only deepen our theoretical grasp of emotional speech but also facilitate practical implementations in areas such as human-machine interaction, mental health surveillance, and customer service analytics.

III. METHODOLOGY

In this section, The Speech Emotion Recognition (SER) initiative employs a systematic technical approach. Initially, a corpus of emotional speech recordings with labels for various emotions. The dataset's diversity is enhanced through data augmentation methods including pitch modification, temporal expansion, noise addition, and time displacement. The primary features extracted are (MFCCs), which capture crucial spectral characteristics for emotion identification. These features are subsequently formatted to serve as input for machine learning algorithms. The selected model is an LSTM-based neural network, comprising an LSTM layer for sequence handling, dense layers with ReLU activation for pattern detection, and dropout layers to mitigate overfitting. The model is optimized by using Adam and categorical cross entropy loss, with early stopping and model checkpointing implemented to enhance performance. The trained model's efficacy is assessed on a separate test set, yielding high accuracy (e.g., 97.59%). This comprehensive approach, combining effective data augmentation, robust feature extraction, and an advanced LSTM architecture, demonstrates exceptional efficiency in SER tasks.

The corpus utilized for this investigation comprises 2,800 voice recordings, categorized into seven emotions anger, happiness, sadness, neutrality, fear, disgust, and pleasant surprise, with each category containing 400 samples to maintain a balanced dataset. (MFCCs), a widely adopted feature extraction method in speech processing, were employed to capture essential audio characteristics, resulting in 40 features per sample. To expand and diversify the dataset, augmentation techniques such as time stretching (modifying audio speed without altering pitch), noise addition (introducing low-level random noise), and time shifting (slightly altering waveform timing) were applied, doubling the dataset to 5,600 samples. The distribution between original and augmented data was maintained at equal proportions, ensuring consistency for model training. The dataset is divided into 80% training and 20% testing sets. This configuration allowed for effective model tuning and performance monitoring during training. The model exhibited high effectiveness, achieving an accuracy of 97.59%, demonstrating its ability to generalize across unseen data. These results underscore its potential applications in domains such as human-computer interaction, where emotion detection enhances user experience, mental health monitoring for assessing emotional well-being, and call center analytics for automated sentiment analysis. Future research may explore advanced feature extraction methods, additional augmentation techniques, or transfer learning to further enhance the model's performance.

A. *Data preparation*

MFCC features were organized into arrays for model input, and labels (emotions) were one-hot encoded for classification. Post-augmentation, the dataset included 5,600 samples: 2,800 original and 2,800 augmented. An LSTM model, suitable for processing sequences like speech, was employed for emotion classification. The model structure was Input Layer: Processes 40 MFCC features for each audio sample. LSTM Layer: Captures MFCC sequence patterns with 256 neurons. Dense Layers: Three fully connected layers using ReLU activation for learning complex patterns. To optimize training: Model Checkpoints: Saved the best-performing model and Early Stopping: Halted training when validation accuracy plateaued to prevent overfitting. The model trained for up to 50 epochs with a batch size of 64, typically concluding earlier due to early stopping. The overall methodology of the study is presented in (Fig.1).

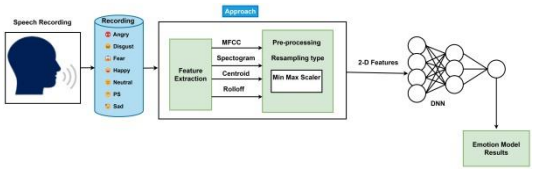

Fig. 1. The general workflow of the study.

B. *Preprocessing*

The model is trained up to 50 epochs with a batch size of 64, though early stopping often concluded training earlier. The

model was then tested on the dataset, achieving 97.59% accuracy, indicating its effectiveness in emotion recognition from speech. For this purpose, the LSTM model effectively captured the audio data patterns. A plot of training and validation accuracy across epochs showed consistent improvement in validation accuracy during early epochs, with minimal overfitting.

IV. RESULTS AND DISCUSSION

Fig. 2. Distribution of emotional label.

The model trained for speech emotion recognition achieved a test accuracy of approximately 97.59%. This result suggests that the model demonstrated high performance in classifying emotions based on the audio data utilized for training. Model distribution of emotion labels in dataset are presented in (Fig. 2). The training process encompassed multiple epochs, during which improvements in validation accuracy were observed. Early stopping was implemented to mitigate overfitting. Model layer details and parameter counts are presented in (Fig. 3 and Table I). In their research, Ma et al. [14, 15] used 12-dimensional statistical aspects of pitch and energy to construct a 70.9% precise rating of five emotions. Furthermore, multiple studies have differentiated between speech emotions, attaining precision with energy-pitch-related local characteristics [16-20]. The description of key features and specifications of the speech emotion recognition system is presented in Table II. The tables present

TABLE I. Model layer details and parameter counts

Layer (Type)	Output shape	Parameters
LSTM (lstm_11)	(None, 256)	264,192
Dense (dense_48)	(None, 128)	32,896
Dense (dense_49)	(None, 64)	8,256
Dense (dense_50)	(None, 64)	4,160
Dropout (dropout_35)	(None, 64)	0
Dense (dense_51)	(None, 32)	2,080
Dropout (dropout_36)	(None, 32)	0
Dense (dense_52)	(None, 7)	231

The provided tables offer an in-depth examination of a machine learning model's design and effectiveness recognizing emotions from vocal input. The dataset comprises 2800 audio samples, equally distributed among seven emotional states: anger, disgust, fear, happiness, neutrality, sadness, and surprise, with 400 examples per category. This equitable allocation prevents model bias towards any particular emotion, enhancing its overall performance across all emotional classes. To enhance the dataset's diversity and resilience, various data augmentation techniques were utilized. These methods included altering audio playback speed without pitch modification (time stretching), adjusting audio frequency to simulate different voice tones (pitch shifting), introducing background noise to improve real-world scenario handling (noise injection), and slightly modifying audio sample timing (time shifting) to create more varied inputs.

The model's structure is based on a (LSTM) network combined with dense layers and dropout regularization. The LSTM layer, containing 256 units with 264,192 parameters, is particularly adept at processing sequential data like audio signals, as it captures temporal dependencies crucial for understanding speech patterns. Following the LSTM layer are three dense layers, applying non-linear transformations to the input features. These layers employ ReLU activation functions to enhance the model's learning capacity. The output layer has dense layer with 7 units, uses softmax activation to generate probabilities for each emotion category, enabling multi-class classification.

The model underwent training for 50 epochs with early stopping implemented to prevent overfitting. This technique monitored validation accuracy and halted training when improvements ceased, ensuring the model maintained its generalization abilities. The model's total parameter count is approximately 330,000, striking a balance between computational efficiency and predictive power. It achieved a test accuracy of 97.59%, demonstrating exceptional performance in classifying emotions from speech, which reported 70.9% accuracy using pitch and energy features, and others that achieved accuracies of 77.8% and 86.8% using local and global energy-pitch features, respectively. The current model's performance underscores the benefits of advanced feature extraction, robust architecture, and effective regularization techniques.

The findings suggest that the developed model demonstrates high efficacy in detecting emotions from audio signals, consistent with existing research in this domain. Subsequent studies could explore the model's efficacy across various languages, dialects, and acoustic settings to assess its broader applicability. Present results conform to several previous inferences in the speech emotions [21-24] Moreover, incorporating diverse input modalities, such as visual cues from facial expressions or physiological measurements, might improve its capacity to identify emotions in multifaceted scenarios. Assessing the model's performance in real-time, noisy environments would further validate its suitability for practical applications, including human-machine interaction, mental health surveillance, and

customer service analytics. This research lays a robust groundwork for advancing emotion recognition technologies and opens avenues for future innovations in the field.

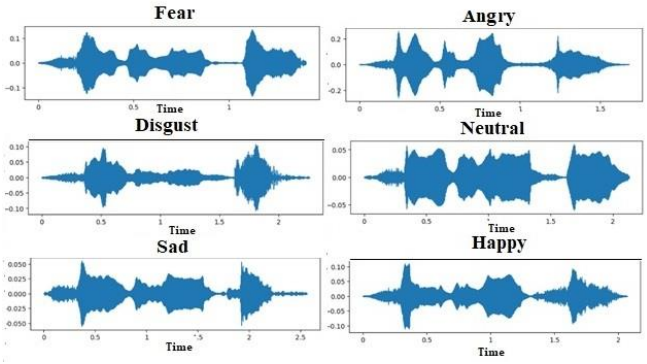

Fig. 3. Waveform analysis of emotional speech signals.

TABLE II. Key Features and specifications of the speech emotion recognition system

Aspect	Detail description
Dataset Size	2800 audio files
Emotions Included	Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise
Samples per emotion	400 samples per emotion
Features extracted	Mel-Frequency Cepstral Coefficients (MFCCs), 40 features
Data augmentation	Time stretching, pitch shifting, noise injection, time shifting
Model architecture	LSTM with Dense layers, Dropout for regularization
Model layers	LSTM (256 units), Dense (128, 64, 32 units), Output Dense (7 units)
Activation functions	ReLU (hidden layers), Softmax (output layer)
Optimizer	Adam
Loss Function	Categorical Crossentropy
Training Epochs	50 (with early stopping based on validation accuracy)
Test accuracy	97.59%

V. CONCLUSION

This model effectively captures speech's temporal dynamics, which are essential for emotion detection. To boost the model's resilience and improve its adaptability to various conditions, multiple data augmentation techniques were implemented. The SER system attained a remarkable test accuracy of 97.59%, The model's high accuracy makes it well-suited for applications in human-computer interaction, where understanding user emotions is crucial. Moreover, the system has potential uses in multimedia analysis, such as enhancing emotion-based content recommendations or improving video indexing and retrieval systems. Including audio samples with varying levels of background noise, different accents, and diverse linguistic contexts can improve the system's robustness and adaptability to real-world scenarios. Another area for improvement involves exploring more sophisticated architectures. For example, combining speech data with facial expressions or textual cues can provide a multimodal perspective on emotions, making the system more resilient to ambiguities in individual modalities. Multimodal systems are particularly advantageous in noisy environments or situations where a single modality may not fully convey the emotional state.

REFERENCES

[1] A. Albahri, M. Lech and E. Cheng, “Effect of speech compression on the automatic recognition of emotions,” Int. J. Signal Process. Syst. 4, p. 55–61. doi: 10.12720/ijsps.4.1.55-61. 2016.

[2] M.B. Akçay and K., Oğuz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” Speech Communication, 116, p. 56−76. 2020.

[3] F. Eyben, “Real-time Speech and Music Classification by Large Audio Feature Space Extraction,” (Springer, 2016). (2016). https://doi.org/10.1007/978-3-319-27299-3.

[4] K. R. Scherer, “Vocal communication of emotion: a review of research paradigms,SpeechCommunication,40,p.227–256.doi:10.1016/S0167-6393(02)00084-5. 2003.

[5] M.B. Mustafa, M.A. Yusoof, Z.M. Don and M. Malekzadeh, “Speech emotion recognition research: An analysis of research focus”. International Journal of Speech Technology, 21, p. 137−156. 2018.

[6] A.A.A. Zamil, S. Hasan, S.MD. Jannatul Baki, J.MD. Adam, I. Zaman,“Emotion Detection from Speech Signals using Voting Mechanism on Classified Frames." International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). IEEE, 201. 2019.

[4] H.M. Bui, M. Lech, E. Cheng, K. Neville, and I. Burnett, “Object recognition using deep convolutional features transformed by a recursive network structure,” IEEE Access 4, 10059–10066. doi: 10.1109/ACCESS.2016.2639543. 2017.

[5] G. Hinton, L. Deng, D. Yu, G. Dahl, A.R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” IEEE Signal Process. Mag. 29, p. 82–97, 2012.

[6] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. Ali Mahjoub and C. Cleder. “Automatic Speech Emotion Recognition Using Machine Learning”. Social Media and Machine Learning. 2019.

[7] S. Mekruksavanich, A. Jitpattanakul and N. Hnoohom, “Negative emotion recognition using deep learning for Thai language,” Joint international conference on digital arts, media, and technology with ECTI northern section conference on electrical, electronics, computer and telecommunications engineering (ECTI DAMT & NCON), IEEE, p. 71−74. 2020.

[8] Z.T. Liu, P. Xiao, D.Y. Li, and M. Hao, “Speaker-independent speech emotion recognition based on CNN-BLSTM and multiple SVMs,” In: Yu, H., Liu, J., Liu, L., Ju, Z., Liu, Y., Zhou, D. (eds) Intelligent Robotics and Applications. ICIRA 2019. Lecture Notes in Computer Science, p. 11742. Springer, Cham. https://doi.org/10.1007/978-3-030-27535-8_43. 2019.

[9] C.M. Lee and S.S. Narayanan, “Toward Detecting Emotions in Spoken Dialogs,” IEEE Transactions on Speech and Audio Processing, 13, p. 293−303, 2005.

[10] G. Trigeorgis, F. Ringeval, R. Brueckner and E. Marchi. “Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network” IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE p. 5200−5204. 2016.

[11] H.M. Fayek, M. Lech and L. Cavedon, “Evaluating deep learning architectures for Speech Emotion Recognition,” Neural Networks,92, 60−68, 2017.

[12] H. Sawaf, “Automatic speech recognition and hybrid machine translation for high-quality closed-captioning and subtitling for video broadcast,” In Proceedings of Association for Machine Translation in the Americas, p. 1–5. San Diego, California, USA: AMTA. 2012.

[13] S. Poria, D. Hazarika, N. Majumder, G. Naik, E and Cambria, R. Mihalcea, “MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations,” In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2, p.527–536, 2019.

[14] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng and L. Cai, Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms. In Proceedings of the Inter speech 2018, Hyderabad, India, 2–6, 2018.

[15] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in Proceedings of the 6th international conference on Multimodal interfaces. ACM, p. 205–211, 2004.

[16] S. Mirsamadi, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," In IEEE ICASSP. 2017.

[17] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera and G. Anbarjafari, "Audio-visual emotion recognition in video clips," IEEE Transactions on Affective Computing. 2017.

[18] K.S. Rao, T.P. Kumar, K. Anusha, B. Leela, I. Bhavana and S.V.S.K. Gowtham, "Emotion Recognition from Speech," (IJCSIT) International Journal of Computer Science and Information Technologies, 3 (2), p. 3603−3607, 2012.

[19] M. Sarma, P. Ghahremani, D. Povey, N.K. Goel, K.K. Sarma and N. Dehak, "Emotion Identification from Raw Speech Signals Using DNNs," In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September, p. 3097–3101. 2018.

[20] M. Schröder, "Emotional speech synthesis: a review," in Seventh European Conference on Speech Communication and Technology," (Aalborg), p. 1–4. 2001.

[21] B. Schuller, G. Rigoll and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," In IEEE ICASSP. 2004.

[22] R. Xia and Y. Liu, "A Multi-Task learning framework for emotion recognition using 2D continuous space," IEEE Trans. Affect. Comput. 8, p. 3–14. 2017.

[23] M. Xu, F. Zhang and W. Zhang, "Head Fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset," IEEE Access, 9, p. 74539–74549. 2021.

[24] S.Yoon, S. Byun and K. Jung, "Multimodal speech emotion recognition using audio and text," In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December, p. 112–118, 2018.