WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Speech Emotion Recognition Using Machine Learning & Deep Learning Techniques

Polireddi Indhumathi

B. Tech Student, Department of IT, GMR Institute of Technology, Rajam-532127, Andhra Pradesh, India
Email: 21341A1298@gmrit.edu.in

ABSTRACT

The goal of speech emotion recognition (SER) is to automatically infer a speaker's emotional state from their speech output. Furthermore, it involves recognizing emotions from different voice samples, which enhances human-computer connection. Since emotions are subjective, speaking is a good way to represent them. Important markers of emotions include tone, pitch, and expression. But because of its possible uses in contact centers, healthcare, and human-computer interface (HCI), SER has grown in significance recently. This area of study makes use of deep learning and machine learning methods to identify feelings including neutrality, joy, sadness, anger, fear, disgust, and surprise. In order to do this, pre-processing of speech data is applied in order to remove background noise and increase the prominence of high-frequency peak signals. The process of recognizing emotions in speech is difficult, and the complexity of the problem and the features in the dataset might affect how well an algorithm performs. Convolutional Recurrent Neural Nets, on the other hand, have demonstrated efficacy in speech emotion recognition tests (CRNN). Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are two powerful classification algorithms that may effectively train from labeled voice data to identify emotional states. The CRNN combines both characteristics for sequence modeling and feature extraction. With developments in deep learning, machine learning, and multimodal analysis creating new avenues for comprehending and addressing human emotions, the subject of SER is changing quickly. We may anticipate seeing SER technology become more and more ingrained in our daily lives as research advances, boosting human-computer interactions and strengthening our bonds with and support for one another.

Keywords: Speech Emotion, Recurrent Neural Network (RNN), Support Vector Machine (SVM), Convolutional Neural Network (CNN).

INTRODUCTION

Speech Emotion Recognition (SER) falls within the realm of affective computing, striving to automatically identify a speaker's emotional state solely from their speech. Various machine learning techniques, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Support Vector Machines (SVM), have been leveraged to tackle this objective. CNNs, acclaimed for image processing, can be repurposed for sequential data like speech signals in SER. They extract pertinent features from audio signal representations, particularly spectrograms depicting frequency spectrum changes over time. In this process, CNNs utilize convolutional layers to capture localized patterns, followed by pooling layers to condense information, and finally, fully connected layers for classification. RNNs, tailored for grasping temporal dependencies in sequential data, are well-suited for analyzing the dynamic nature of speech signals. They maintain a hidden state to capture information from prior time steps. Variants like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are commonly employed in SER to address temporal dependency challenges. SVM, a classical algorithm, has found utility in emotion recognition tasks, including SER. It classifies features derived from speech signals into distinct emotional categories by learning a hyperplane. The selection of an algorithm in SER depends on factors like dataset size, emotional complexity, and computational resources. Hybrid models, combining various algorithms—such as CNNs for feature extraction and RNNs for sequential modeling—show promise in augmenting overall SER performance.

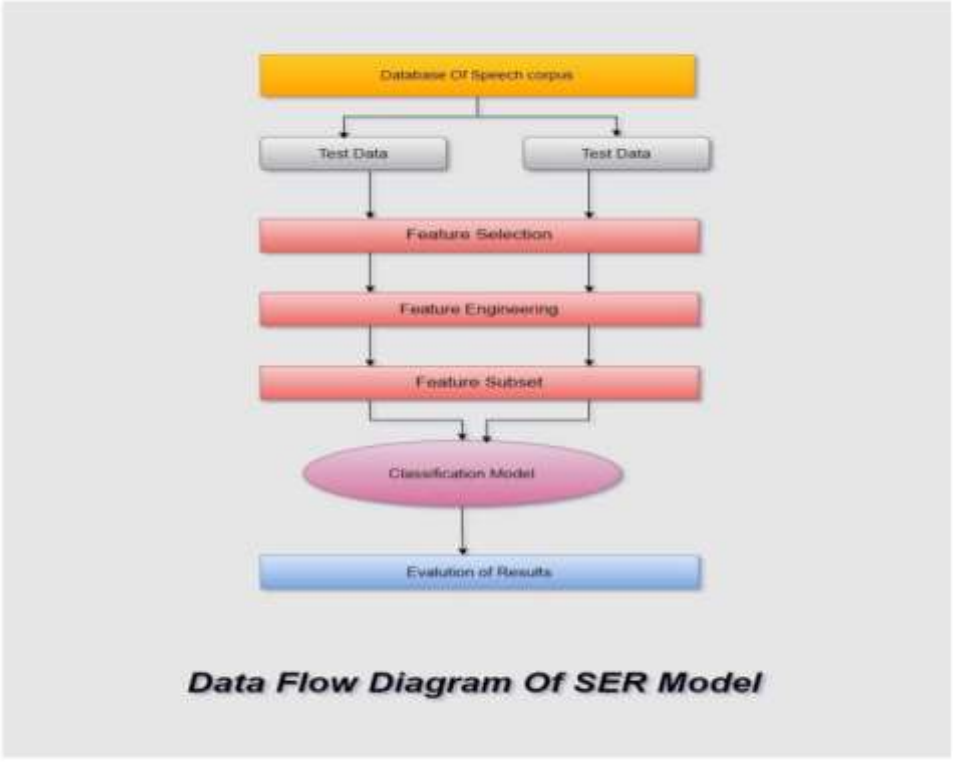
RESEARCH APPROACH

Speech Emotion Recognition (SER) involves analyzing speech recordings to detect and classify various emotions conveyed by speakers. Several databases cater to this purpose, such as Emo-DB, BESD, and RAVDESS. These collections feature recordings with annotated emotions, aiding in model training. To discern emotions from speech, SER commonly employs diverse features. Pitch, formant frequencies, energy levels, MFCCs, and prosodic features contribute crucial information about voice characteristics. Feature engineering steps like normalization, feature combination, and selection refine these attributes to enhance model performance. Recursive Feature Elimination (RFE) algorithms assist in feature subset selection by iteratively removing less contributory features.

Various classification models serve SER tasks, including Support Vector Machines (SVMs), Random Forests, and Neural Networks. SVMs discern classes by defining a hyperplane in the feature space, while Random Forests, comprising decision trees, ensure robustness and accuracy. Neural Networks, mimicking the human brain's structure, handle diverse tasks like classification and language processing. Evaluating SER models involves testing their performance on separate datasets. The unweighted average recall (UAR), a common metric, measures the model's ability to predict emotions across all speakers accurately. Higher UAR values signify the model's proficiency in recognizing diverse emotional states exhibited in speech.

METHODOLOGY:

Speech Emotion Recognition (SER) involves utilizing speech corpora like Emo-DB, BESD, and RAVDESS, which contain recordings of individuals expressing various emotions through different phrases or sentences. These corporated are annotated by human listeners to label emotional content.

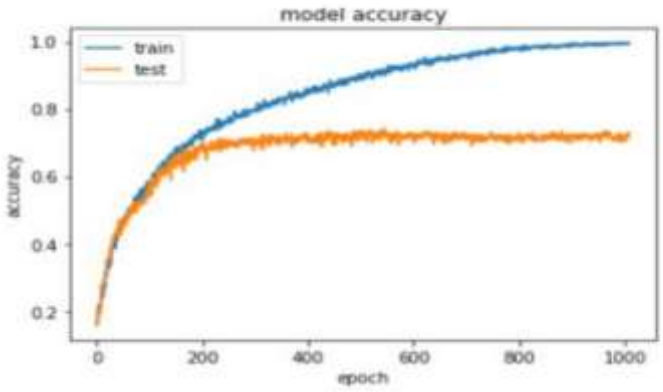


Data Flow Diagram Of SER Model

Features commonly employed in SER encompass pitch, formant frequencies, energy levels, Mel-frequency cepstral coefficients (MFCCs), and prosodic elements such as intonation and rhythm. Feature engineering techniques like normalization, feature combination, and selection aim to enhance SER performance by refining and optimizing these features. Selecting relevant features can be achieved using methods like Recursive Feature Elimination (RFE), which systematically removes less contributing features for better model performance.

For classification, SER deploys diverse models including Support Vector Machines (SVMs), Random Forests, and Neural Networks. SVMs delineate different classes by creating a hyperplane in the feature space, while Random Forests, consisting of decision trees, are known for their robustness. Neural Networks, inspired by the brain's structure, are versatile and applicable for various tasks, including SER. Evaluating the SER model involves assessing its performance on a separate test set. The unweighted average recall (UAR) serves as a common metric, measuring the model's ability to predict emotions across all speakers. A higher UAR signifies the model's accuracy in predicting diverse emotional states exhibited in speech.

RESULTS



CONCLUSION

The convergence of machine learning (ML) and deep learning (DL) has propelled speech emotion recognition (SER) to unprecedented levels of accuracy and sophistication. ML algorithms provide a robust foundation, extracting meaningful features from audio signals to discern subtle nuances in emotional expression within speech. Complementing this, DL models, particularly neural networks, introduce complexity and abstraction, facilitating the automatic learning of hierarchical representations. This collaborative integration enhances prediction accuracy, fostering more nuanced models capable of capturing the intricacies of diverse emotional expressions. The unified approach holds far-reaching implications, extending from refining human-computer interactions to revolutionizing mental health assessments through automated emotion monitoring. The ongoing evolution of ML and DL in SER promises continuous advancements, with the refinement of algorithms leveraging the strengths of both paradigms, paving the way for innovative solutions and the development of emotion-aware technologies.

References

[1]. Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Ali Mahjoub, M., & Cleder, C. (2020). Automatic Speech Emotion Recognition Using Machine Learning. IntechOpen. doi: 10.5772/intechopen.84856

[2]. Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H*+., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, *7*, 117327-117345.

[3]. Costantini, G., Parada-Cabaleiro, E., Casali, D., & Cesarini, V. (2022). The emotion probe: On the universality of cross-linguistic and cross-gender speech emotion recognition via machine learning. *Sensors*, *22*(7), 2461.

[4]. Alluhaidan, A. S., Saidani, O., Jahangir, R., Nauman, M. A., & Neffati, O. S. (2023). Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network. *Applied Sciences*, *13*(8), 4750.

[5]. Chaudhary, A. N. K. U. S. H., Sharma, A. K., Dalal, J., & Choukiker, L. (2015). Speech emotion recognition. *J Emerg Technol Innov Res*, *2*(4), 1169-1171.

[6]. G. Wen, H. Li, J. Huang, D. Li, and E. Xun, ‘‘Random deep belief networks for recognizing emotions from speech signals,’’ Comput. Intell. Neurosci., vol. 2017, Mar. 2017, Art. no. 1945630.

[7] A. Mohamed, G. E. Dahl, and G. Hinton, ‘‘Acoustic modeling using deep belief networks,’’ IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 1, pp. 14–22, Jan. 2012.

[8] C. Huang, W. Gong, W. Fu, and D. Feng, ‘‘A research of speech emotion recognition based on deep belief network and SVM,’’ Math. Problems Eng., vol. 2014, Aug. 2014, Art. no. 749604.

[9] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, ‘‘Greedy layerwise training of deep networks,’’ in Proc. Adv. Neural Inf. Process. Syst., 2007, pp. 153–160.

[10] W. Q. Zheng, J. S. Yu, and Y. X. Zou, ‘‘An experimental study of speech emotion recognition based on deep convolutional neural networks,’’ in Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII), Sep. 2015, pp. 827–831.

[11] Y. Kim, ‘‘Convolutional neural networks for sentence classification,’’ 2014, arXiv:1408.5882. [Online]. Available: https://arxiv.org/ abs/1408.5882

[12] W. Fei, X. Ye, Z. Sun, Y. Huang, X. Zhang, and S. Shang, ‘‘Research on speech emotion recognition based on deep auto-encoder,’’ in Proc. IEEE Int. Conf. Cyber Technol. Automat., Control, Intell. Syst. (CYBER), Jun. 2016, pp. 308–312.

[13] J. Deng, R. Xia, Z. Zhang, Y. Liu, and B. Schuller, ‘‘Introducing sharedhidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2014, pp. 4818–4822.

[14] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, ‘‘Sparse autoencoderbased feature transfer learning for speech emotion recognition,’’ in Proc. IEEE Humaine Assoc. Conf. Affect. Comput. Intell. Interact., Sep. 2013, pp. 511–516.

[15] L. Cen, W. Ser, and Z. L. Yu, ‘‘Speech emotion recognition using canonical correlation analysis and probabilistic neural network,’’ in Proc. 7th Int. Conf. Mach. Learn. Appl. (ICMLA), Dec. 2008, pp. 859–862.

[16] X. Zhou, J. Guo, and R. Bie, ‘‘Deep learning based affective model for speech emotion recognition,’’ in Proc. Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People, Smart World Congr. (UIC/ATC/ScalCom/CBDCom/IoP/ SmartWorld), Jul. 2016, pp. 841–846.

[17] S. E. Eskimez, Z. Duan, and W. Heinzelman, ‘‘Unsupervised learning approach to feature analysis for automatic speech emotion recognition,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp. 5099–5103.

[18] J. Zhao, X. Mao, and L. Chen, ‘‘Learning deep features to recognise speech emotion using merged deep CNN,’’ IET Signal Process., vol. 12, no. 6, pp. 713–721, 2018.

[19] W. Zhang, D. Zhao, Z. Chai, L. T. Yang, X. Liu, F. Gong, and S. Yang, ‘‘Deep learning and SVM-based emotion recognition from Chinese speech for smart affective services,’’ Softw., Pract. Exper., vol. 47, no. 8, pp. 1127–1138, 2017.

[20] L. Zhu, L. Chen, D. Zhao, J. Zhou, and W. Zhang, ‘‘Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN,’’ Sensors, vol. 17, no. 7, p. 1694, 2017