CSE 572 Data Mining
Final Project Literature Review

Project Title : Speech Emotion Recognition using Deep Learning

Team members :

Full Name	**ASU ID**
Aashvik Chennupati	1225259971
Ravi Tej Chaparala	1230035172
Jaya Shankar Maddipoti	1230911684
Hemanth Pasula	1225551795
Dinesh Koushik Deshapathi	1225344913

Research papers related to Speech Emotion Recognition:

Step 1 : Summary of the Related Work

[1] R. Y. Cherif, A. Moussaoui, N. Frahta and M. Berrimi, "Effective speech emotion recognition using deep learning approaches for Algerian dialect," 2021 International Conference of Women in Data Science at Taif University (WiDSTaif), Taif, Saudi Arabia, 2021, pp. 1-6, doi: 10.1109/WiDSTaif52235.2021.9430224. Link

Brief Summary:

- Created a new dataset of Algerian speech emotion recordings to address the scarcity of resources for under-represented languages in speech emotion recognition (SER).
- Applied various machine learning and deep learning models, including convolutional neural networks (CNNs), Long Short Term Memory (LSTM) networks, and Bidirectional LSTM (BLSTM), to the task of SER for Algerian dialect.
- The LSTM-CNN model achieved the highest classification accuracy of 93.34%, demonstrating the effectiveness of deep learning approaches for SER in Algerian dialect.

Strengths:

- Addresses the underrepresentation of languages in SER research by creating a new dataset for Algerian dialect.
- Employed a variety of machine learning and deep learning models, allowing for a comprehensive comparison of their performance.
- The LSTM-CNN model achieved a high classification accuracy, demonstrating the potential of deep learning for SER in under-resourced languages.

Limitations:

- Focuses on four primary emotions (happy, angry, neutral, sad), limiting the range of emotions recognized.

- The dataset size could be expanded to further enhance the generalizability of the findings.
- Study could be extended to explore other deep learning architectures and optimization techniques.

[2] H. Li, X. Zhang and M. -J. Wang, "Research on Speech Emotion Recognition Based on Deep Neural Network," 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 2021, pp. 795-799, doi: 10.1109/ICSIP52628.2021.9689043. Link

Brief Summary :

- Proposed a hybrid deep neural network (DNN) model for speech emotion recognition (SER) that combines convolutional neural network (CNN) and Long Short-Term Memory (LSTM) to extract both spectral and temporal information from speech spectrograms.
- The DNN model achieved a weighted accuracy of 61% and an unweighted accuracy of 56% on the IEMOCAP dataset, demonstrating its effectiveness in recognizing a range of emotions from speech spectrograms.
- The study highlights the potential of combining CNN and LSTM for SER and emphasizes the importance of feature representation using spectrograms.

Strengths:

- The proposed DNN model effectively combines CNN and LSTM to capture both spectral and temporal info from speech spectrograms, enhancing feature extraction for SER.
- The study provides a detailed analysis of the model's performance across different emotion categories, offering insights into its strengths and limitations.
- The research contributes to the exploration of deep learning techniques for SER and highlights the importance of feature representation using spectrograms.

Limitations:

- The overall accuracy of 61% (weighted) and 56% (unweighted) suggests room for improvement in the DNN model's performance.
- The study could be extended to investigate other deep learning architectures and optimization methods to further enhance SER performance.
- The research is limited to the IEMOCAP dataset, which, while valuable, is constrained in size and may not fully capture the variability of human emotions.

[3] Javier de Lope and Manuel Graña. 2023. An ongoing review of speech emotion recognition - ScienceDirect. *ScienceDirect*. Retrieved October 17, 2023. Link

Brief Summary:

- The paper reviews Speech Emotion Recognition (SER) in the context of Automated Emotion Recognition (AER) and highlights the importance of SER in applications like Human-Computer Interaction and Human-Robot Interaction.
- The paper provides an overview of various audio-visual databases commonly used in SER research and discusses conventional machine learning approaches and a wide range of deep learning approaches for SER.
- The paper emphasizes the challenges of limited data availability and suggests future research directions, including collaboration on existing corpora and advancements in deep learning.

Strengths:

- Comprehensive coverage of SER, including databases and ML techniques.
- Provides an overview of key findings and approaches in the field.
- Emphasizes the importance of addressing data scarcity and deep learning methods.

Limitations:

- Lack of original research; serves as a review of existing work.
- Does not propose specific solutions to the challenges discussed.
- Offers a general outlook on future research but lacks concrete recommendations.

[4] J. Bhanbhro, S. Talpur and A. A. Memon, "Speech Emotion Recognition Using Deep Learning Hybrid Models," 2022 International Conference on Emerging Technologies in Electronics, Computing and Communication (ICETECC), Jamshoro, Sindh, Pakistan, 2022, pp. 1-5, doi: 10.1109/ICETECC56662.2022.10069212. Link

Brief Summary:

- The paper proposes a hybrid deep learning model for speech emotion recognition (SER) that combines a stacked convolutional neural network (CNN) and a long short-term memory (LSTM) network.
- The model achieves an accuracy of 93.9% in classifying emotions into one of eight categories on the RAVDESS dataset, which is a significant improvement over previous results.
- The model is likely to be generalizable to other datasets, as it is able to capture both spatial and temporal features of the speech signal.

Strengths:

- The paper highlights the significance of balanced datasets and the value of preprocessing, such as augmenting with AWGN, for improved model performance.
- The research emphasizes the advantages of using Mel spectrograms as input for deep learning models, reducing the need for manual feature engineering.
- The hybrid model's use of CNN and LSTM shows significant promise in improving SER, providing accurate emotion classification

Limitations:

- The research relies heavily on the RAVDNESS dataset, which may not fully represent the diversity of speech emotions in real-world scenarios.
- The paper lacks a detailed discussion of potential real-world applications of the SER model and its broader implications.
- The model is not able to handle noise in the speech signal.

[5] X. Ying and Z. Yizhe, "Design of Speech Emotion Recognition Algorithm Based on Deep Learning," 2021 IEEE 4th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), Shenyang, China, 2021, pp. 734-737, doi: 10.1109/AUTEEE52864.2021.9668689. Link

Brief summary:

- This paper addresses the essential technology of Speech Emotion Recognition (SER) and its significance in human-computer interaction, particularly in applications such as criminal investigation.

- The paper presents a deep learning-based SER algorithm, combining Convolutional Neural Network (CNN), Bi-directional Long- and Short-Term Memory (LSTM), and a multi-headed attention mechanism.
- The proposed model improves SER efficiency and capacity in human-computer interaction devices.

Strengths:
- Introduces a comprehensive deep learning model that combines different neural network components for SER.
- The multi-headed attention mechanism enhances the model's ability to focus on relevant speech features.
- The paper advances the field of SER, contributing to its application in critical areas such as criminal investigation.

Limitations:
- The computational resources required for the proposed model, including CNN and multi-headed attention, may be demanding.
- The discussion primarily focuses on applications in the English language, potentially limiting its generalizability to other languages.
- The applicability to purely audio-based scenarios may require further investigation.

[6] P. Tzirakis, J. Zhang and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 5089-5093, doi: 10.1109/ICASSP.2018.8462677. Link

Brief summary:

- Utilized a Convolutional Neural Network (CNN) to extract features from the raw signal.
- Employed a 2-layer Long Short-Term Memory (LSTM) network for considering contextual information.
- Achieved significant improvement in concordance correlation coefficient compared to state-of-the-art methods for the RECOLA database.

Strengths:

- Novel end-to-end model architecture for speech emotion recognition.
- Outperformed existing methods in terms of concordance correlation coefficient for both arousal and valence dimensions.
- Demonstrated the effectiveness of the model with the RECOLA database, which contains multimodal data.

Limitations:

- The model's performance for valence prediction might need further improvement.
- The study focuses primarily on audio input, whereas multimodal approaches might yield better results for valence prediction.
- While the model outperforms previous approaches, it's essential to consider the applicability of these results to other datasets or real-world scenarios.

[7] R. Ranjan, "Analysis of Speech Emotion Recognition and Detection using Deep Learning," 2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), Gwalior, India, 2022, pp. 1-5, doi: 10.1109/IATMSI56455.2022.10119297. link

Brief Summary:

- The paper introduces a deep learning approach utilizing LSTM networks for Speech Emotion Recognition (SER), achieving an average accuracy of 71% in recognizing seven different emotions from audio samples.
- It emphasizes feature extraction using Mel Frequency Cepstral Coefficients (MFCC) and explores further accuracy improvements through additional feature extraction techniques and larger datasets.

Strengths:

- The use of LSTM networks to identify emotions is quite innovative, and it performs well with an average accuracy of 71%.
- The paper includes easy-to-understand graphs and images that help us grasp how emotions are detected.
- The paper offers detailed code examples for preprocessing data, extracting features, and creating models, making it practical and reproducible.

Limitations:

- The paper doesn't dive into the existing research in this field or discuss how this technology might be used in real-life applications.
- There could be more information on how the LSTM model is designed and how audio data is prepared.
- With only 2800 samples, the study's findings might not be widely applicable, and there's also a lack of information on the computing resources needed for real-world use.

[8] Y. Zhang, J. Du, Z. Wang, J. Zhang and Y. Tu, "Attention Based Fully Convolutional Network for Speech Emotion Recognition," 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 2018, pp. 1771-1775, doi: 10.23919/APSIPA.2018.8659587. Link

Brief Summary:

- The paper introduces an attention-based fully convolutional network for speech emotion recognition.
- Utilizes a fully convolutional network to handle variable-length speech without segmentation.
- Incorporates transfer learning with a pre-trained model, achieving state-of-the-art results on IEMOCAP with a weighted accuracy of 70.4% and unweighted accuracy of 63.9%.

Strengths:

- Demonstrates the adaptation of convolutional neural network architectures designed for visual recognition to the speech emotion recognition task.
- Shows that transfer learning, especially using a pre-trained model on natural scene images, enhances the performance of speech emotion recognition.
- Proposes an attention-based fully convolutional network that can handle variable-length speech and effectively identify emotion-relevant regions in the spectrogram

Limitations:

- The paper does not provide an in-depth analysis of the interpretability of the attention mechanism.
- It remains unclear how the proposed model would perform on datasets beyond the IEMOCAP corpus.
- The paper lacks a comprehensive discussion on potential challenges and future directions for improving speech emotion recognition systems.

[9] M. Saloumi et al., "Speech Emotion Recognition Using One-Dimensional Convolutional Neural Networks," 2023 46th International Conference on Telecommunications and Signal Processing (TSP), Prague, Czech Republic, 2023, pp. 212-216, doi: 10.1109/TSP59544.2023.10197766. Link

Brief Summary:

- One-Dimensional Convolutional Neural Network (1D-CNN) applied to recognize emotions in short voice messages (lasting <3 seconds).
- Utilizes the Ravee dataset voiced by professional actors for training.
- Achieves a recognition accuracy of up to 83%, employing Mel-frequency cepstral coefficients and short-time Fourier transform for feature extraction.

Strengths:

- Efficient utilization of 1D-CNN for emotion recognition in short voice messages.
- Incorporates data augmentation techniques, improving testing accuracy by increasing the number of training samples.
- Employs Mel-frequency cepstral coefficients and short-time Fourier transform, providing a comprehensive approach to feature extraction.

Limitations:

- The dataset used (RAVDESS) may have limited diversity and may not represent real-world scenarios.
- While achieving good accuracy, the model's performance on specific emotions may vary.
- The study does not extensively compare the proposed model with other state-of-the-art emotion recognition models.

[10] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124. Link

Brief summary:

- The paper discusses the importance of Speech Emotion Recognition (SER) in Human-Computer Interaction (HCI) and the transition to real-world applications.
- Deep learning techniques, such as Deep Boltzmann Machine (DBM), Recurrent Neural Network (RNN), Recursive Neural Network (RvNN), Deep Belief Network (DBN), Convolutional Neural Network (CNN), and Auto Encoder (AE), are highlighted as promising approaches for SER.
- The paper mentions that deep learning methods outperform traditional approaches. Where the combined accuracy of SVM and DBN is as high as 94.6%.

Strengths:

- The paper provides a comprehensive overview of traditional techniques and deep learning methods for SER.
- It highlights the advantages of deep learning in improving emotion recognition and the specific deep learning algorithms used.

- The paper acknowledges the significance of deep learning in addressing real-world emotion recognition challenges.

Limitations:

- The paper lacks specific numerical results or percentages to quantify the performance improvements achieved by deep learning techniques.
- While it introduces various deep learning methods, it does not offer a detailed comparison of their strengths and weaknesses.
- The paper mentions the limitations of deep learning, but it does not delve deeply into potential solutions or future research directions.

Step 2: Organization of relevant work

In the field of Speech Emotion Recognition (SER), the integration of neural networks and spectrograms plays a pivotal role in revolutionizing the accuracy and efficiency of emotional analysis in spoken language [1], [5], [6]. Neural networks, particularly convolutional neural networks (CNNs). These papers showcase the growing significance of deep learning in improving emotion recognition accuracy. Recurrent neural networks (RNNs) like Long Short-Term Memory (LSTM) and multi-headed attention mechanism, have emerged as powerful tools for their ability to automatically learn complex features and temporal dependencies within speech data and focus on relevant speech features. Complementing this, spectrograms provide a visual representation of the frequency and temporal content of audio signals, allowing neural networks to extract both spectral and temporal information [2], which is crucial for understanding the nuances of emotional expression in speech. This synergy between neural networks and spectrograms empowers SER systems to capture and interpret emotional cues with a higher degree of accuracy, offering a promising avenue for enhancing human-computer interaction, sentiment analysis, and a wide range of applications where understanding emotional states is critical.[5] improves the speech recognition by replacing LSTM with Bi-LSTM and usage of attention mechanism with multi-headed mechanism.

[1] addresses the underrepresentation of languages in SER by creating a new dataset for Algerian dialect. Diversity in data is crucial for building more robust SER models. We can consider similar strategies for underrepresented languages or dialects. As outlined in [1], the LSTM-CNN model obtained an impressive classification accuracy of 93.34%, underscoring the efficacy of deep learning techniques in the context of Speech Emotion Recognition (SER) for the Algerian dialect. The study [2] introduced a hybrid deep neural network (DNN) model for Speech Emotion Recognition (SER), utilizing a combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) to capture both spectral and temporal information from speech spectrograms. This DNN model demonstrated effectiveness by achieving a weighted accuracy of 61% and an unweighted accuracy of 56% when tested on the IEMOCAP dataset, showcasing its capability to recognize a variety of emotions from speech spectrograms. The research emphasized the potential of CNN-LSTM combinations for SER and highlighted the significance of spectrogram-based feature representation in this context. [2], [4], [6], [8], [9] emphasize the use of feature extraction methods, such as Mel-frequency cepstral coefficients (MFCC) and spectrograms, for capturing emotional cues in speech data. These features are essential for the effective modeling of emotions.

[5], [8] introduces an attention-based model that can focus on relevant speech features. This approach can be explored in our project to improve the identification of emotion-relevant regions

in audio data. [6] suggests the potential for multimodal approaches that combine audio and visual data for enhanced SER. Investigating such approaches may be beneficial for our project.

To address the limitations observed in these groups, our project can aim to expand on the strengths of existing approaches. For instance, if the literature review highlights limitations related to dataset size, we can strive to gather a larger and more diverse dataset for training and evaluation. If the literature review indicates a need for improving specific aspects of deep learning models, we can focus on optimizing model architectures, conducting hyperparameter tuning, and applying data augmentation techniques. Additionally, we can explore the transfer learning potential, as mentioned in [1], [4], [8], [9] to leverage knowledge from larger datasets for better SER performance. Also, computational resources required for models like multi-headed attention mechanisms would be demanding, as mentioned in [5]. Our project will aim to address these limitations to advance the field of Speech Emotion Recognition. To achieve this, we propose a comprehensive approach that encompasses various aspects highlighted in the literature. This includes broadening the range of recognized emotions to capture complex and subtle expressions [1] [4] [9], curating a more diverse dataset that encompasses languages and dialects [1], exploring various deep learning architectures and conducting performance comparisons [1] [2] [4] [6] [8], developing noise reduction techniques, incorporating multimodal data [6], and discussing potential real-world applications [5] [8] and future research directions [3] [10]. By addressing these limitations and capitalizing on the strengths of existing approaches, our project aims to make substantial contributions to the field of SER.