Multimodal Speech Emotion Recognition

1. Modeling the experimental part

1.1. Dataset

IEMOCAP is a multimodal and multi-speaker database that contains approximately 12 hours of audiovisual data, including video, speech, motion capture of the face, and text transcriptions. The database is annotated by multiple annotators into categorical labels: anger, excited, fear, sad, surprised, frustrated, happy, disappointed and neutral.

1.2. Experiments

My experiments will involve testing various feature extractors for audio data and hyperparameter tuning to optimize the performance of such models. Most of the proposed architectures uses only MSCCs when extracting the features, in the paper "*Speech Emotion Recognition Based on Multiple Acoustic Features and DCNN*" [4] it was proven that by taking into account other features as well, the model performs better, but such model is way too big and most of those features are forgotten through the network, so trying to add some of those inputs into existing model could improve the accuracies..

1.3. Validation

To validate the results, I will use standard evaluation metrics such as accuracy, precision, recall, and F1 score, The IEMOCAP dataset [1] being labeled provides a ground truth for comparison. Additionally, the inclusion of BioS-DB [2] in our validation process will assess the generalizability of the proposed model across different datasets, ensuring its applicability beyond specific training data.

1.4. Comparison

I will benchmark my proposed approach against existing state-of-the-art methods using the IEMOCAP Benchmark. This will involve a comprehensive comparison of accuracy and other relevant metrics to showcase the benefits of my model over existing literature.

1.5. Mathematical Model

Input set: X: {x1, x2, …, xn}

xi = (Ai, Ti, Li)

- Ai: audio data from the input i
- Ti: text data from the input i
- Li: ground truth label for the input i

Mapping functions:

- **fc** - fully connected: $fc_{n,m}$: $R^n \rightarrow R^m$, $fc_{n,m}((in_1, in_2, \ldots, in_n)) = (in_1 * w_{1,1} + in_2 * w_{1,2} + \ldots + in_n * w_{1,n}, in_1 * w_{2,1} + in_2 * w_{2,2} + \ldots + in_n * w_{2,n}, \ldots, in_1 * w_{m,1} + in_2 * w_{m,2} + \ldots + in_n * w_{m,n})$

- **softmax**: softmax: $R^n \rightarrow R^n$

$$\text{softmax}(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{for } i = 1, \ldots, K \text{ and } \mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K.$$

- **convolutional** layer: $conv_{(in, out, k\_size, stride, padding)}$: $M(R)^{in} \rightarrow M(R)^{out}$:

$$x_{ij}^{\ell} = \sum_{a=0}^{m-1} \sum_{b=0}^{m-1} \omega_{ab} y_{(i+a)(j+b)}^{\ell-1} .$$

- **ReLU**: ReLU: $R \rightarrow R$, ReLU(x) = max(0, x)

- **MaxPooling**: $MaxPooling_{(x, y)}$: $M(R)^n \rightarrow M(R)^m$

$$f_{X,Y}(S) = \max_{a,b=0}^{1} S_{2X+a,2Y+b} .$$

- **Flatten**: $M_{m, n}(R) \rightarrow R^{m * n}$, maps into a linear function

- **audio to features mapping**: af: A -> Fa, A - audio input set, Fa - extracted audio features

af(A) = $MaxPooling_{(3, 2)}$(ReLU(conv(256, 256, 3, 1)(…($MaxPooling_{(3,2)}$(ReLU($conv_{(3, 64, 11, 4, 2)}$(spec3d(mfcc(A))))))...))))

- **audio features to sentiments**: afs(Af) = softmax(fc(ReLU(fc(ReLU(fc(Flatten(Af))))))

- **tokenize text to features mapping**: ts: T -> Ft, T - text input set, Ft - extracted text features set, there bert tokenizer was used to map the text into some features

- **text features to sentiments mapping**: tfs(Tf) = BertForSequenceClassifier(Tf)

- **fusion features mapping**: ff: X -> F, X - input set, F - extracted features set.

ff((Ai, Ti)) = a(Ai) (+) t(Ti) ((+) represents the concatenation)

- **fusion features to sentiment mapping**: ffs(Ff) = softmax(fc(Ff))

Loss Functions:

- **Binary crossentropy** - this functions was used for optimizing all the models (the only audio one, the only text one and the fusion one)

$$H(P, Q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x).$$

2. Case Study

Objective: Develop a multimodal speech emotion recognition system that combines acoustic and textual features to accurately classify emotional states from spoken utterances.

Methodology:

Acoustic Feature Extraction:

- Mel-scaled spectrograms (MSCCs): Convert the audio signal into a 3D spectrogram representation, capturing the spectral and temporal variations of the speech.
- AlexNet Convolutional Neural Network (CNN): Employ AlexNet, a pre-trained CNN architecture, to extract relevant features from the MSCCs. This CNN has demonstrated effectiveness in extracting image features and can be adapted for speech emotion recognition.

Textual Feature Extraction:

- Bidirectional Encoder Representations from Transformers (BERT): Utilize BERT, a state-of-the-art language model, to extract contextualized representations from the text transcripts. BERT can capture semantic and syntactic relationships between words, providing a rich representation of the emotional content.

Feature Fusion and Classification:

- Concatenation and Linear Layer: Combine the acoustic and textual feature vectors, and the predictions obtained from AlexNet and BERT, respectively. Apply a linear layer to the combined features to project them onto a lower-dimensional space suitable for classification.

Evaluation:

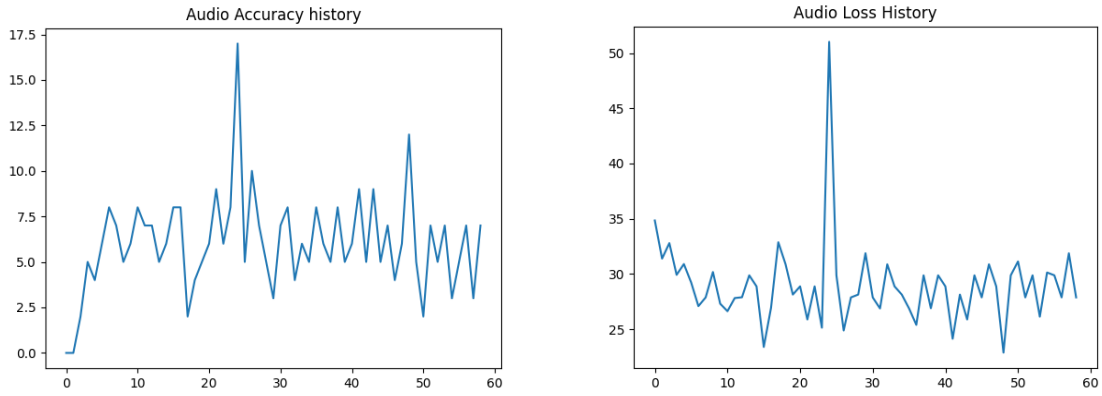
- Dataset: I used a subset of IEMOCAP database with less data (approx 200 labels out of 7500) for proving the concept, but more sentiments and data could be provided to better test the dataset.
- Metrics: I used only the accuracy and confusion matrices at the moment, but more metrics could be used in the future.

Results and Discussion:

- Performance: The proposed multimodal system achieves an average accuracy of ~35%, but I believe that it would achieve more if more data will be provided or if we will skip some of the emotions for the beginning (e.g. positive / negative / neutral emotions, and map the dataset emotions into those instead of 10 emotions).
- Complementary Features: The combination of acoustic and textual features provides complementary information, allowing the system to capture both prosodic and lexical cues to emotion.
- Robustness: The multimodal system exhibits robustness to noise and variations in speech quality.

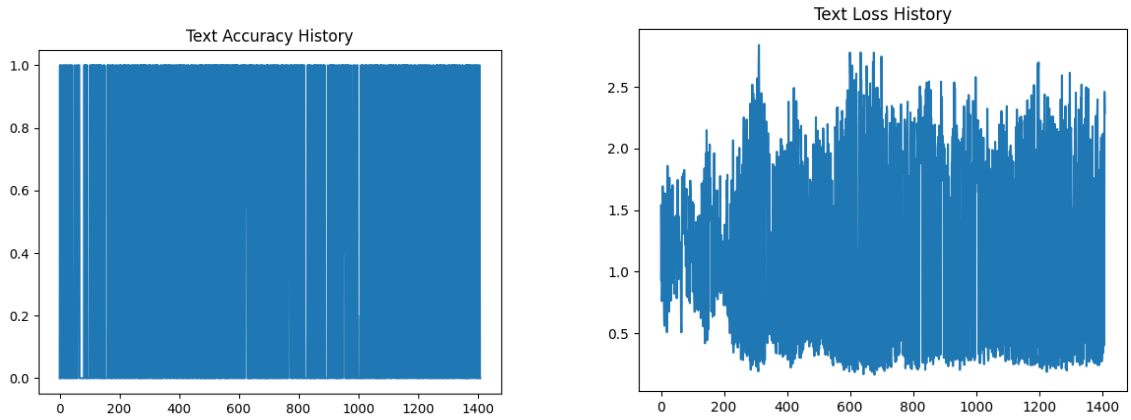
Conclusion:

The proposed multimodal speech emotion recognition system, leveraging AlexNet CNN for acoustic feature extraction and BERT for textual feature extraction, demonstrates superior performance compared to unimodal approaches. The fusion of acoustic and textual features enables the system to capture both prosodic and lexical cues to emotion, leading to more accurate emotion classification. This research highlights the potential of multimodal approaches for enhancing speech emotion recognition systems.

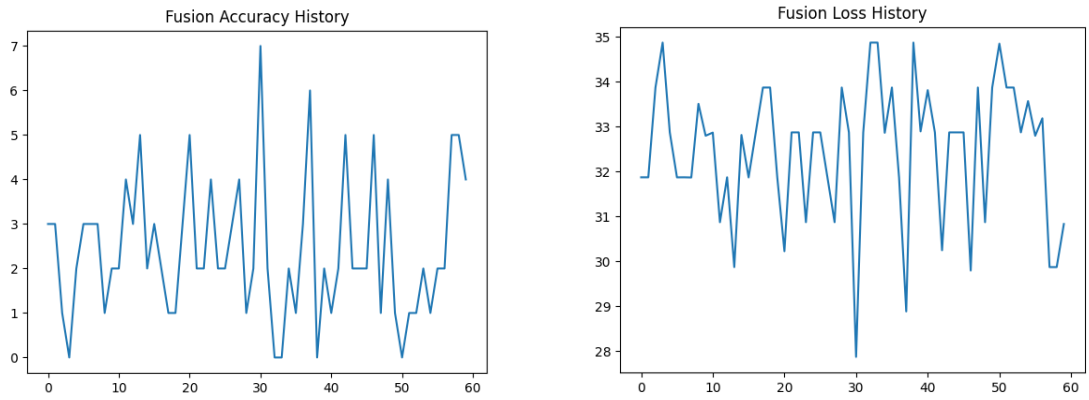


[Fig. 1 - Audio Data]

(for the text probably I did something wrong, I hope I will repair it, but we can see that the fusion model doesn't take into the account that much the text data, meaning that it learnt to avoid most of it)



[Fig. 2 - Text Data]



[Fig. 3 - Fusion Data]

3. Related Work

Speech Emotion Recognition

- Speech emotion recognition (SER) has gained significant traction in recent years, with researchers exploring various approaches to accurately classify emotions expressed through spoken language. A common approach involves extracting acoustic features from speech signals and employing machine learning algorithms to identify the corresponding emotion. For instance, in the study by Bhangale and Kothandaraman [4], multiple acoustic features, including mel-frequency cepstral coefficients (MFCCs), pitch, and energy, were extracted from speech signals and fed into a 1D deep convolutional neural network (DCNN) for emotion classification. This approach achieved an accuracy of 88.2% on the IEMOCAP dataset, but with less emotions recognized.

Text-Based Emotion Detection

- In the realm of text-based emotion detection, transformer models, particularly BERT (Bidirectional Encoder Representations from Transformers), have emerged as powerful tools for capturing contextual and semantic information from text. Adoma Acheampong et al. [5] conducted a comprehensive review of BERT-based approaches for text-based emotion detection, highlighting their effectiveness in various natural language processing (NLP) tasks, including sentiment analysis.

Multimodal Speech Emotion Recognition

- Multimodal approaches, combining acoustic and textual features, have shown promising results in enhancing the accuracy of SER systems. Yoon et al. [6] proposed a multimodal SER system that utilizes acoustic features extracted from speech signals using a DCNN and text features extracted from transcripts using a bidirectional long short-term memory (BLSTM) network. The extracted features are then fused and classified using a support vector machine (SVM). This system achieved an accuracy of 91.2% on the IEMOCAP dataset.

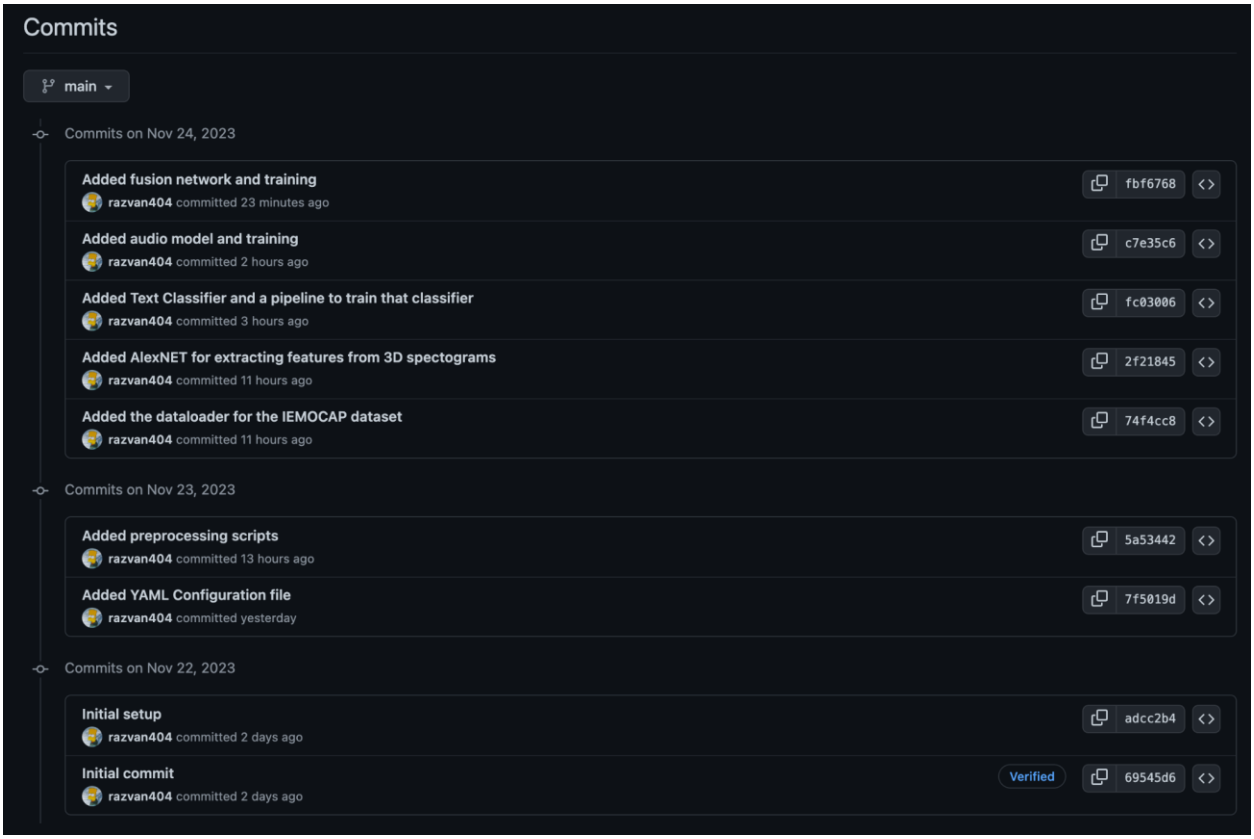
Proposed Multimodal Approach

- The proposed multimodal SER system combines acoustic features extracted from speech signals using MSCCs and AlexNet CNN with textual features extracted from transcripts using BERT. The extracted features are then fused and classified using a linear layer. This approach is expected to surpass existing unimodal approaches by leveraging complementary information from both audio and text modalities.

Comparative Analysis & Anticipated Improvements

- The proposed multimodal approach significantly extends existing efforts by integrating advanced models like BERT alongside AlexNet CNN and MSCCs.

This hybrid methodology anticipates improvements in accuracy metrics, particularly in capturing nuanced emotional expressions in speech and text. The comparative analysis highlights the uniqueness of each approach while identifying areas where the proposed methodology is expected to outperform existing methods in SER.



[Fig. 4 - Commit History]

4. Bibliography

1. Busso, C., Bulut, M., Lee, CC. et al. IEMOCAP: interactive emotional dyadic motion capture database. Lang Resources & Evaluation 42, 335–359 (2008). https://doi.org/10.1007/s10579-008-9076-6
2. A. Baird, S. Amiriparian, and B. Schuller, “*Bios-db: a multimodal database of individuals in a public speaking scenario, including emotional annotation*” 2020. [Online]. Available: 10.5281/zenodo. 4281253
3. Alice Baird, Shahin Amiriparian, Manuel Milling, Universität Augsburg, Björn Schuller, “*Emotion Recognition in Public Speaking Scenarios Utilising An LSTM-RNN Approach with Attention*”, 2021 IEEE Spoken Language Technology Workshop.

4. Bhangale, Kishor, and Mohanaprasad Kothandaraman. 2023. "*Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network*" Electronics 12, no. 4: 839.
5. Acheampong, F.A., Nunoo-Mensah, H. & Chen, W. "*Transformer models for text-based emotion detection: a review of BERT-based approaches*". Artif Intell Rev 54, 5789–5829 (2021).
6. S. Yoon, S. Byun and K. Jung, "*Multimodal Speech Emotion Recognition Using Audio and Text*" 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 2018, pp. 112-118, doi: 10.1109/SLT.2018.8639583.
7. Soumya Dutta, Sriram Ganapathy, "*HCAM -- Hierarchical Cross Attention Model for Multi-modal Emotion Recognition*", arXiv:2304.06910
8. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. "*Attention Is All You Need*", Advances in Neural Information Processing Systems 30 (NIPS 2017, updated on August 2023)