Samiksha Burkul

DATS 6303 Final Project – Individual Report

Introduction:

Speech is an important part of communication among human beings and is a way to express thoughts, feelings, and moods. Beyond just information being transferred, emotion plays an important role in speech communication and emotional speech is an effective way of conveying a message. Subtle changes in lexical or grammatical emphasis can change the meaning of a statement, with changing the words (1). This fact can be best demonstrated by the fact that a voice call is more informative than a text message.

Speech Emotion Recognition (SER) is a field of study which focuses on inferring human emotion from speech signals. SER is emerging as an important topic with applications in healthcare, machine-human interaction, education, intelligent assistance and many others (2). Many questions of SER are classification problems and can be addressed with supervised machine learning methods.

This project focuses on SER applied to political campaign speeches. By detecting emotions such as anger, happiness, fear, and neutrality from voice alone, the goal is to gain deeper insights into the emotional tenor of political messaging. Understanding these emotional cues can help decode communication strategies and the underlying sentiments intended to influence public opinion.

For this project, our group created several models including a CNN, LSTM, pretrained transformer model, a fine-tuned audio spectrogram transformer model, and a multimodal model. The overall goal of the project was to compare these model types and their performance for SER tasks and then apply the best performing model to political speech audio files.

Individual Contribution:

My primary contribution to this project was the development of a Custom Convolutional Neural Network (CNN) for emotion classification.

Mel Spectrograms

Mel spectrograms serve as a vital representation of audio signals in this project. They transform raw audio data into a time-frequency representation, using the Mel scale to mimic human auditory perception. By capturing how sound energy is distributed across frequencies over time, Mel spectrograms enable machine learning models to discern complex patterns in speech. Using the librosa.feature.melspectrogram() function, each audio file was converted into a fixed-size 2D image with 64 frequency bands and 128 time frames. These standardized spectrograms

were instrumental in enabling the CNN to extract features effectively and classify emotions accurately.

Model Process and Model Architecture:

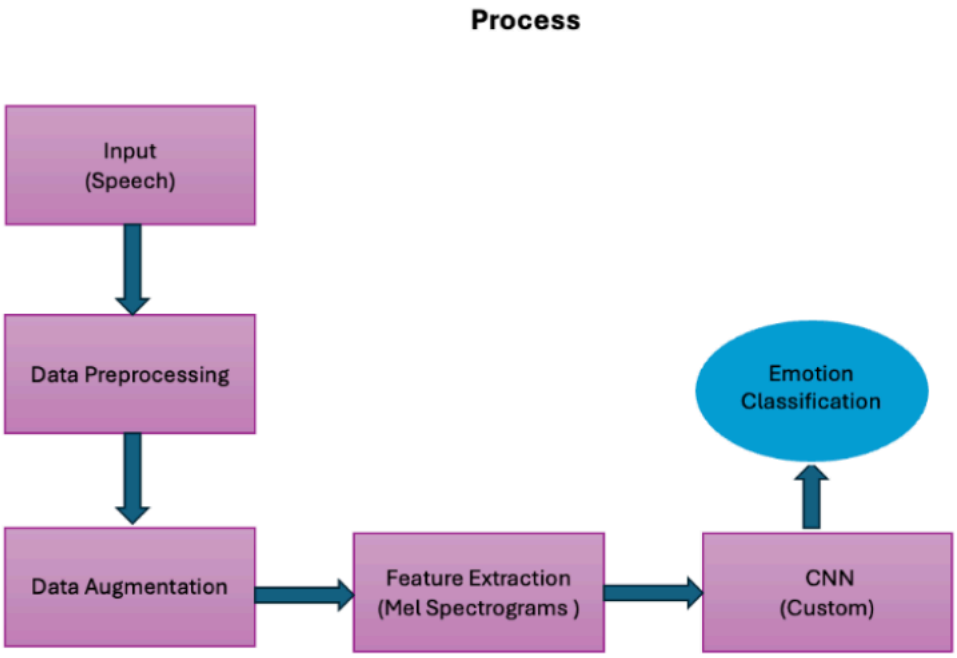

Figure 1. Model Process

- **Data Preprocessing**: The audio dataset was preprocessed by splitting it into training, validation, and test sets. This ensured a robust evaluation of the model and prevented overfitting.
- **Data Augmentation**: To increase the diversity of the training data and enhance the model's generalizability, various augmentation techniques were applied to the audio data:
  Time Shift: Randomly shifted audio signals in the time domain.
  Pitch Shift: Adjusted the pitch of audio by a random number of semitones.
  Noise Addition: Added random noise to simulate real-world audio variations.
- **Feature Extraction**: Since CNNs require fixed input sizes, audio files were converted into visual Mel spectrograms (2D time-frequency images). Each spectrogram was standardized to dimensions of 64 frequency bands and 128-time frames to ensure consistency.

Figure 2. Model Architecture

The CNN architecture was designed with six convolutional blocks to progressively extract spatial and temporal patterns from the input data. The number of filters increased from 32 to 512 across layers, allowing the network to capture low-level to high-level features. Batch normalization was applied after each convolutional layer to stabilize and accelerate the training process. Max-pooling layers were used to reduce spatial dimensions, improving computational

efficiency. Finally, two fully connected layers were employed to capture high-level abstract features, followed by a softmax activation function for multiclass emotion classification.

Results:

CNN Classification Report:

	precision	recall	f1-score	support
disgust	0.6334841628959276	0.5511811023622047	0.5894736842105263	254.0
happy	0.44543429844098	0.7843137254901961	0.5681818181818182	255.0
sad	0.5543859649122806	0.6220472440944882	0.5862708719851577	254.0
fear	0.6601941747572816	0.2677165354330709	0.38095238095238093	254.0
angry	0.7513513513513513	0.547244094488189	0.6332574031890661	254.0
neutral	0.6097560975609756	0.6880733944954128	0.646551724137931	218.0
accuracy	0.5742108797850907	0.5742108797850907	0.5742108797850907	0.5742108797850907
macro avg	0.6091010083197995	0.5767626827272603	0.5674479804428133	1489.0
weighted avg	0.6089752528311755	0.5742108797850907	0.56553595832375	1489.0

Figure 3. Classification Report

The CNN model was trained for 30 epochs using mel spectrograms as input features. The model achieved a final training accuracy of approximately 97%, indicating successful learning of the training data. However, the validation accuracy fluctuated throughout training, reaching a peak of 60.37% but ultimately settling at a lower level. This discrepancy suggests that the model may have overfit to the training data, failing to generalize well to unseen data.

The validation loss remained significantly higher than the training loss, further supporting the overfitting hypothesis. The model's performance on the test set was moderate, achieving a test accuracy of 57.89% and a macro F1-score of 0.56. These results indicate that the model was able to distinguish between emotions to a certain extent, but its performance was limited by the overfitting issue.

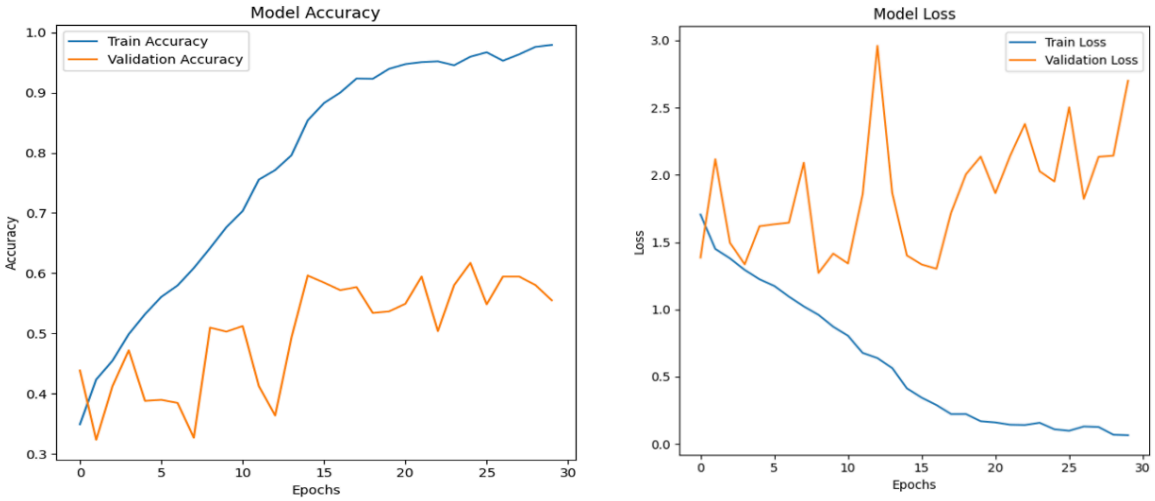

Figure 4. CNN accuracy

Summary and Conclusion:

The CNN model for emotion classification, though promising, faced challenges like overfitting and achieved a maximum F1-score of 57%. While this model can identify emotions from audio, it is insufficient for complex speech recognition tasks that require accurate transcription and understanding of spoken language.

Given the limitations of the custom CNN model, future work will explore the use of pre-trained models, such as those based on Transformer architectures, which have demonstrated superior performance in speech recognition tasks. By leveraging the knowledge and representations learned from large-scale language and speech datasets, these pre-trained models can potentially improve the accuracy and efficiency of our emotion classification system.

Code Usage:

The code primarily utilized Python libraries like TensorFlow/Keras, NumPy, and LibROSA. While leveraging existing codebases for core functionalities, approximately 10% of the code was custom-developed, including data preprocessing pipelines, CNN architecture design, and model training/evaluation. Future work will explore the potential of pre-trained models for improved performance.

Future work:

Advanced Architectures: Investigating more sophisticated architectures like Transformer-based models to capture long-range dependencies in audio data.

Data Augmentation: Implementing advanced data augmentation techniques to increase data diversity and improve model generalization.

Transfer Learning: Leveraging pre-trained models like Wav2Vec2 to incorporate prior knowledge and improve model performance.

Hyperparameter Tuning: Conducting a thorough hyperparameter tuning process to optimize the model's performance.

Ensemble Methods: Combining multiple models to improve overall accuracy and robustness.

Pre-trained Models: Utilizing pre-trained models as a starting point for training, leveraging their learned features and representations to accelerate training and improve performance.

References:

[1]. Alzubaidi, L., Zhang, J., Humaidi, A.J. *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* **8**, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8

[2]. S. Chu, S. Narayanan and C. J. Kuo, "Environmental Sound Recognition with Time–Frequency Audio Features," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 6, pp. 1142-1158, Aug. 2009.

[3.] Cnn, C.N.N., 2020. Speech emotion recognition using convolutional neural network (CNN). *International Journal of Psychosocial Rehabilitation*, *24*(8), pp.1-20.

[4]. Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE transactions on multimedia*, *16*(8), 2203-2213.

[5]. Hajarolasvadi, Noushin, and Hasan Demirel. "3D CNN-based speech emotion recognition using k-means clustering and spectrograms." *Entropy* 21, no. 5 (2019): 479.

[6]. V. Sowmya, A. Rajeswari, "Detection of Emotion cues from Tamil Speech signals", *2024 10th International Conference on Communication and Signal Processing (ICCSP)*, pp.328-333, 2024.