**Individual Final Project Report**
**FAISAL ALQAHTANI**
**GROUP7**

## 1. Introduction

In my individual contribution to the group's Deep Learning project on Speech Emotion Recognition (SER), I focused on employing advanced pre-trained speech models to enhance the classification of emotions from audio. Initially, I integrated **Wav2Vec2** for feature extraction from raw waveforms. Later, I explored **HuBERT**, another cutting-edge self-supervised speech model, to see if alternative embeddings would yield performance gains. Both models are known for capturing rich acoustic features, which significantly simplifies the feature engineering process.

My goals were:

1. Leverage pre-trained self-supervised models (Wav2Vec2, HuBERT) to extract robust speech embeddings.

2. Train a neural network classifier on these embeddings to predict emotion classes.

3. Evaluate the model not only on the original training domain (CREMA-D dataset) but also test on external audio clips (e.g., political speech recordings) to assess generalization.

## 2. Description of My Work Using Pre-trained Models

### Wav2Vec2 Pipeline:

I began by processing the CREMA-D audio samples through Wav2Vec2, a model trained on large amounts of unlabeled speech data (Baevski et al., 2020). This generated context-rich embeddings. By averaging these embeddings over time, each utterance became a single fixed-size vector. I then fed these vectors into a feed-forward neural network with dropout layers to classify emotions such as Anger, Disgust, Fear, Happy, Neutral, and Sad.

### Switching to HuBERT:

To investigate if another self-supervised approach could improve results, I switched to HuBERT (Hsu et al., 2021). HuBERT uses a masked prediction objective over cluster assignments of acoustic units, potentially capturing different speech characteristics. I replicated the Wav2Vec2 pipeline with HuBERT: load raw audio, extract embeddings, average them, and input them into the same classifier. The results were on par with Wav2Vec2, solidifying the effectiveness of these embeddings.

## 3. Experimental Setup and Training

I trained the classification model on the **CREMA-D dataset** (Cao et al., 2014), which provides thousands of clips labeled with different emotions. I used a categorical cross-entropy loss, the Adam optimizer, and early stopping. Both Wav2Vec2 and HuBERT embeddings improved accuracy over traditional handcrafted features.

## 4. Testing on External Audio Clips

Beyond CREMA-D, I tested the final HuBERT-based model on external audio clips—specifically, segments from political speeches. Since the model expects audio as input, I used properly formatted WAV files (or converted existing audio to the correct sampling rate). The model produced predictions for these clips, demonstrating its flexibility. However, the predictions may not align with human interpretations of emotion in real-world political speech due to domain differences and the fact that the model never saw this kind of data during training.

This highlighted a key limitation: while the model performed well on the curated CREMA-D dataset, it struggled to generalize to new domains. To improve real-world applicability, I would need more domain-specific training data or employ domain adaptation techniques.

**5. Key Learnings**

- **Power of Self-Supervised Models:** Using Wav2Vec2 and HuBERT drastically simplified the SER pipeline, eliminating the need for manual feature extraction and often improving accuracy.
- **Domain Generalization is Challenging:** While the model excelled on the test set from CREMA-D, it faced difficulties when classifying emotions in more complex, real-world data such as political speeches.
- **Data Quality and Diversity:** For better generalization, the model would benefit from training on a more diverse range of speech samples, possibly fine-tuning on speech that matches the target domain more closely.

**6. Future Directions**

- **Domain Adaptation and Fine-Tuning:** Acquiring or fine-tuning on in-domain samples (political speeches with known emotional labels) could improve the model's reliability on real-world audio.
- **Multimodal Inputs:** Incorporating text transcripts alongside audio may yield more robust emotion recognition, as linguistic cues can complement acoustic signals.
- **More Advanced Architectures or Data Augmentation:** Experimenting with data augmentation and more sophisticated architectures could further improve performance.

**References**


- Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems (NeurIPS)*.
- Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing, 5(4),* 377–390.
- Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.