Group 7 Proposal – Campaign Speech Emotion Recognition

Chris, Samiksha, Faisal, Meng-Fei

1.1 Group Proposal

After you have selected a topic, a network, and a data set, submit a proposal of what you

plan to do for the project. The proposal should be a few hundred words, and should address the

following items.

• What problem did you select and why did you select it?

• What database/dataset will you use? Is it large enough to train a deep network?

• What deep network will you use? Will it be a standard form of the network, or will you

have to customize it?

• What framework will you use to implement the network? Why?

• What reference materials will you use to obtain sufficient background on applying the

chosen network to the specific problem that you selected?

• How will you judge the performance of the network? What metrics will you use?

• Provide a rough schedule for completing the project.

=========================================================================

=========================================================================

1. **Problem Selection and Reason for Selection**

**Speech emotion recognition (SER)** is the process of inferring human emotion from audio speech signals regardless of the semantic content. SER has been an area of research interest for more than two decades and continues to garner significant research interest with modern deep learning techniques. SER is increasingly relevant across various domains, including customer service, mental health, and human-computer interaction. By analyzing emotions conveyed through voice, SER can enhance interactions where non-verbal cues are absent, providing a deeper understanding of user needs and behaviors. This nuanced insight into emotional states makes SER a powerful tool for creating more empathetic and adaptive technology. (Mengfei)

Our group will develop a **speech emotion recognition model to analyze and infer the emotional tone** of campaign speeches of US presidential candidates. With this analysis we hope to gain insights into the emotional tone candidate speeches and better understanding of overall campaign strategies.

1. **Dataset**

Our group will explore different databases for training our model. We will start with the **Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D**) (1). CREMA-D is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities. Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral and Sad) and four different emotion levels (Low, Medium, High and Unspecified). This dataset is a popular dataset and has been cited in 360 papers.

1. **Methodology**

Our team will explore different deep learning networks to find the best option for this task. We will explore **Recurrent Neural Networks (RNNs), Convolution Neural Networks (CNNs), Transformers, and Hybrid models.** We plan to build our models in on a **Pytorch framework** to operate efficiently on GPUs.

SER is a popular topic of research, and our team is able to pull from existing meta-analyses and review studies of the field to implement the current state of the art models for our project (2) (3) (4).

We will use Accuracy, F1-Score, and Cohen Kappa as our key performance metrics when we build and test our models. Broadly speaking, based on survey studies of the field, accuracy is the most common metric reported. Given the standard dataset we will be using, we feel it will be useful to compare against existing models.

1. **Project Schedule**

Our rough project schedule below:

* Week 1: Data preprocessing and initial model selection
* Week 2-3: Model development and training
* Week 4: Model tuning and evaluation
* Week 5: Final testing and reporting

1. **Citations:**
2. 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.
3. Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors (Basel, Switzerland)*, *21*(4),
4. George, S. M., & Ilyas, P. M. (2023). A review on speech emotion recognition: a survey, recent advances, challenges, and the influence of noise. *Neurocomputing*, 127015.
5. Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning—A systematic review. *Intelligent systems with applications*, 200266.