**Project Proposal: Exploring the Adversarial Robustness of Speech-Command Recognition**

Team:

Nikhitha Kilari  
Jishnuvardhan Karpuram

The focus of our research will be on developing new adversarial attacks and defenses on 1D-CNN models trained on the Google Speech Commands v2 dataset containing 35 spoken commands. Initially, we will work on the audio clip preprocessing steps (resampling, framing, normalization) and devise a basic 1D convolutional network to get clean-accuracy baselines, the dataset is publicly available via… PyTorch and TensorFlow APIs, thus assuring seamless compliance with the assignment through-the-pipeline dataflow.

Using a mix of black-box and psychoacoustic approaches, we will create six unique attacks in the second stage, none of which were utilized in our assignments:

* **Simultaneous Perturbation Stochastic Approximation** (SPSA) Attack (gradient-approximation)
* **GenAttack** (evolutionary algorithms, optimization without gradients)
* **Attack on Spatial Transformation** (pitch-shifting, time-stretching disturbances)
* Using human auditory masking thresholds to conceal noise is known as a **psychoacoustic masking attack.**
* Enclosing hostile orders in harmless sounds is known as a "**hidden voice-command attack.**"
* **SimBA Audio Attack:** Iterative frequency-component substitution for a simple black-box attack

We'll adjust each strategy's attack budget (such as SPSA iterations, genetic-population size, and transformation magnitude) report top-1 accuracy drop, signal-to-noise ratio, and perceptual quality metrics to measure effectiveness and stealth.

In addition to the assignment coverage from prior sessions, we will assess the following three defenses:

* **Randomized Smoothing** (applying Gaussian noise followed by certified-radius analysis)
* **Feature Squeeze** (reducing audio bit-depth to median-filtered levels)
* **Defensive Distillation** (training a soft-label model which smoothens the decision boundaries of the primary model)

We will analyze clean, adversarial, and certified-robust accuracies within each defense budget and present the findings as accuracy-vs-robustness curves while considering deployment implications for practical speech systems. Every piece of code, plot, and analysis will be compiled into a singular Jupyter notebook alongside proper external implementation citations and structured for easy navigation.