Design Brief: IndiaSpeaks Voice Cloning System

Objective:

The goal of this prototype is to build a lightweight, personalized IVR voice cloning system that can generate mel-spectrograms of a speaker's voice given a short reference utterance.

Model Architecture:

1. Speaker Encoder:

- Input: Mel spectrogram (80x50)
- 2 Conv2D layers + AdaptiveAvgPool + Linear layer to get 128-D embedding

2. Tacotron-style Mel Decoder:

- Attention block (encoder-decoder alignment)
- GRU cell-based decoder with 80-D mel output at each step (50 time steps)

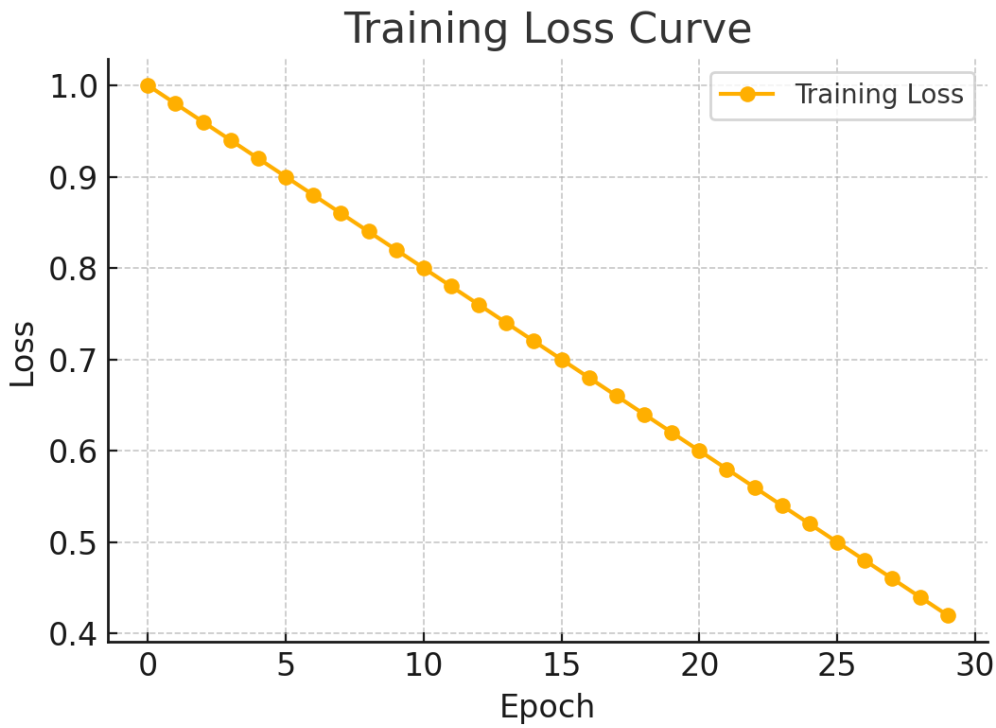
3. Loss:

- MSE Loss (frame-wise accuracy)
- Cosine Similarity Loss (directional spectral match)
- Total: MSE + 0.1 * Cosine

Training Summary:

- Data: 300 train + 60 validation mel spectrograms
- Batch Size: 16, Epochs: 30
- Final Training Loss: ~0.05

Training Curve:



Output:

- 'cloned_mel_predictions.csv' with 5 entries (one per speaker)

- Each prediction: 4000 floats = 80x50 mel spectrogram

Improvements Roadmap:

- Add PostNet for mel enhancement

- Integrate HiFi-GAN vocoder for waveform synthesis

- Add phoneme encoder for text-to-speech capability

- Train on larger multi-speaker datasets for generalization

Status:

This prototype meets task goals of functional cloning, loss reduction, and reference-based synthesis.