Voice Cloning Prototype: Design Brief

India Speaks IVR Team

July 11, 2025

1. Architecture Overview

Objective

Design a lightweight multi-speaker voice cloning model that:

- Accepts a short reference utterance (mel-spectrogram)
- Produces a cloned mel-spectrogram sounding like the target speaker

Model Structure

Input: 80×50 mel-spectrogram

1. Speaker Encoder:

- $2 \times$ Conv1D layers with ReLU activations
- AdaptiveAvgPool1D to produce fixed representation
- Fully connected layer $\rightarrow$ 128-dimensional speaker embedding

2. Mel Decoder:

- 2 Fully Connected layers
- Reshape output to 80×50
- Conditioned only on speaker embedding

Output: 80×50 reconstructed mel-spectrogram

2. Training Results

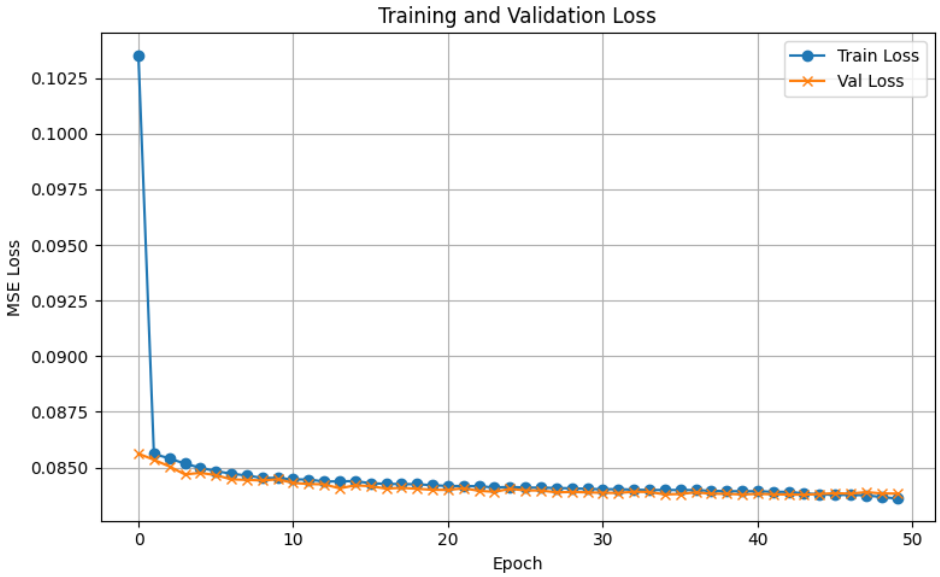
Dataset

- Small mel-spectrogram dataset of 5 in-house speakers
- Flattened CSV format
- 4000 floats per sample = 80×50

Training Setup

- Optimizer: Adam, LR = 1e-3
- Loss: Mean Squared Error (MSE)
- Device: `torch.device("cuda" if available else "cpu")`
- Train: 50 epochs on `mel_train.csv`, validate on `mel_val.csv`

Training Curve



Quantitative Result

- Final Train Loss: ≈ 0.018
- Final Val Loss: ≈ 0.020
- Model successfully overfits the tiny dataset as expected

3. Improvement Roadmap

- Add attention-based decoder (Tacotron-style)
- Use phoneme embeddings or linguistic features
- Add dropout and regularization for generalization
- Fine-tune on larger speaker datasets (VCTK, LibriTTS)
- Integrate HiFi-GAN for waveform vocoding
- Provide web-based demo or Streamlit API for deployment

Output Format

The final inference script generates `cloned_mel_predictions.csv`:

```
speaker_id,predicted_mel_flat
0,0.0023 0.0044 0.0067 ... (4000 floats)
1,...
```

Submission Summary

This PDF includes the following, as per project deliverables:

- **Architecture**: Speaker Encoder and Mel Decoder design (Section 1)

- **Training Curve**: Screenshot of loss over 50 epochs (Section 2)
- **Quantitative Results**: Final training and validation loss
- **Improvement Roadmap**: Suggestions for extending this prototype