
# Voice Cloning for Song Generation

Course: Deep Learning for Perception

Instructor: Miss Sumaiyah Zahid

21K-3387 (Syed Zain Abbas)

21K-4885 (Muhammad Hasan)

21K-4922 (Ali Ahmed)

## Objective

The objective of this project is to develop a voice cloning model that can generate new audio tracks in the voice of a specific artist—such as Atif Aslam—by training the model on a limited set of the artist's existing songs along with the corresponding Roman Urdu lyrics. Once trained, the model should be capable of synthesizing new songs from any given Roman Urdu lyrics, producing audio that mimics the singing voice and style of the original artist.

## Problem Statement

Most current voice cloning and singing synthesis models are designed primarily for English or other high-resource languages, often using large datasets and pre-trained voices. There is a lack of publicly available tools and methodologies for generating singing audio in Urdu—especially using Roman Urdu transcriptions, which are common in informal digital communication.

Additionally, training a model to capture both the voice and the expressive style of a singer is a complex task that typically requires extensive and high-quality aligned datasets. In this project, we aim to overcome these challenges by building a system that can work with:

- Limited data (around 10 songs)
- Roman Urdu text (rather than Urdu script)
- Raw `.wav` audio files with minimal preprocessing
- Moderate computational resources (Kaggle's P100 or T4 GPUs)

The ultimate goal is not high accuracy, but rather a functional proof-of-concept that demonstrates the ability to generate intelligible and stylistically consistent audio from new lyrics.

## Methodology

The project is carried out in the following key stages:

1. **Data Preparation**
    a. Collect `.wav` files of songs by the target artist.
    b. Align each audio file with its corresponding Roman Urdu lyrics, ensuring that lyrics match the audio in content and sequence.
    c. Preprocess audio by normalizing sample rates (16kHz), trimming silence, and converting to mono channel.
2. **Text-Audio Preprocessing**
    a. Prepare metadata files required for training (e.g., paths to audio files and transcriptions).
    b. Tokenize or clean Roman Urdu text to improve model compatibility.
    c. Optionally apply phoneme conversion or forced alignment techniques.
3. **Model Training**
    a. Use a Text-to-Speech (TTS) model such as Tacotron2 to convert Roman Urdu text to mel spectrograms.
    b. Train a vocoder model (e.g., HiFi-GAN or WaveGlow) to convert mel spectrograms into `.wav` audio files.
    c. Combine these stages into a full TTS pipeline customized for singing voice synthesis.
4. **Voice Synthesis / Generation**
    a. Provide new Roman Urdu lyrics as input to the trained model.
    b. Generate corresponding audio that mimics the trained singer's voice.
    c. Evaluate output subjectively for vocal quality and style retention.