**Song Generation (Voice Clone Synthesis)**

**Project Members:**
**21K-3387 Syed Zain Abbas**
**21K-4885 Muhammad Hasan**
**21K-4922 Ali Ahmed**

**Course: Deep Learning for Perception**

**Instructor: Miss Sumaiya Zahid**

## 1. Project Definition:

The project involves developing a computational model capable of synthesizing novel audio sequences (songs) in the style of a specific singer, given a text input of lyrics. The input data consists of a set of audio recordings of the target singer and corresponding textual transcriptions of the lyrics. The desired output is a generated audio waveform that exhibits the vocal characteristics and stylistic attributes of the target singer when given new lyric text.

## 2. Input Data Specifications:

- **Audio Data:**
  - Format: `.wav` audio files.
  - Language: Urdu language vocals.
  - Requirements:
    - Consistent sampling rate across all files (e.g., 16kHz, 22.05kHz).
    - Normalized amplitude levels.
    - Minimal background noise.
    - Clear vocal presence.
- **Lyrics:**
  - Format: Text files.
  - Script: Roman Urdu.
  - Requirements:

        - Accurate temporal alignment with the corresponding audio segments. This alignment must specify the precise time intervals within the audio that correspond to each textual unit (word, phrase, or phoneme).
        - Consistent text encoding.

**3. Data Preprocessing:**

- **Audio Preprocessing:**
    - **Resampling:** If necessary, resample all audio files to a uniform sampling rate.
    - **Normalization:** Normalize the amplitude of the audio signal to a specific range (e.g., [-1, 1]).
    - **Silence Removal:** Employ voice activity detection (VAD) algorithms to remove segments of silence at the beginning and end of audio files.
    - **Feature Extraction:** Convert the raw audio waveform into a time-frequency representation. Common options include:
        - Mel-spectrogram: A visual representation of the audio signal where the frequency scale is transformed to the Mel scale.
        - Mel-Frequency Cepstral Coefficients (MFCCs): Coefficients derived from the Mel-spectrogram, often used to represent the spectral envelope of the audio.
- **Text Preprocessing:**
    - **Normalization:** Apply text normalization procedures to ensure consistency in the Roman Urdu text (e.g., handling of capitalization, punctuation).
    - **Phonemization (Optional):** Convert the Roman Urdu text into a sequence of phonemes. This requires a grapheme-to-phoneme (G2P) conversion model or a phonetic dictionary for Urdu.

**4. Model Architecture Options:**

- **Tacotron 2:**
    - Components:
        - Encoder: Transforms the input text into a higher-level representation.
        - Attention Mechanism: Dynamically aligns the encoded text with the audio features.
        - Decoder: Generates a sequence of mel-spectrogram frames.
    - Output: Mel-spectrogram.

    - Vocoder Requirement: A separate neural vocoder is necessary to synthesize the final audio waveform from the generated mel-spectrogram.
- **FastSpeech/FastSpeech 2:**
    - Characteristics: Non-autoregressive models that generate speech in parallel, offering faster synthesis compared to autoregressive models like Tacotron 2.
    - Output: Mel-spectrogram.
    - Vocoder Requirement: Requires a separate neural vocoder.
- **End-to-End Models (e.g., VITS):**
    - Characteristics: Models that directly synthesize the audio waveform from the input text in a single stage.
    - Output: Audio waveform.
    - Vocoder Requirement: Integrated within the model.

**5. Model Training:**

- **Environment:**
    - Development environment with Python and deep learning libraries (e.g., PyTorch, TensorFlow).
    - GPU acceleration for training (e.g., NVIDIA T4 or P100).
- **Data Loading:**
    - Implementation of data loaders to efficiently feed preprocessed audio features and text sequences to the model during training.
- **Loss Function:**
    - Definition of appropriate loss functions to quantify the difference between the model's predictions and the ground truth.
- **Optimization:**
    - Use of optimization algorithms (e.g., Adam) to update the model's parameters based on the calculated loss.
- **Speaker Embedding:**
    - If voice cloning is implemented, the model should include a speaker embedding layer to capture speaker-specific characteristics.

**6. Audio Synthesis (Inference):**

- **Input:** New lyric text.
- **Processing:**
    - Text encoding.

    - Mel-spectrogram generation (if applicable).
    - Waveform synthesis (using a vocoder or directly in end-to-end models).
- **Output:** Synthesized audio waveform.