A Hybrid Approach to Low-Resource Text-to-Speech (TTS) Dataset Synthesis via Voice Cloning and ASR Datasets

1. Introduction

The development of high-quality **Text-to-Speech (TTS)** models requires large, meticulously curated datasets of audio-text pairs recorded by a target speaker. In scenarios where only a **low-resource donor voice** (limited audio samples) is available, generating a sufficient dataset for effective voice cloning and TTS training presents a significant challenge.

This whitepaper details a **hybrid methodology** that leverages an existing **large-scale, multi-speaker Automatic Speech Recognition (ASR) dataset** (e.g., Common Voice) and advanced **zero-shot voice cloning techniques** to synthesize a high-quality, synthetic TTS dataset in the target voice. The approach focuses on robust data preparation, quality filtering, and scalable synthesis.

---

2. Methodology: Synthetic Dataset Generation

The core idea is to first process the multi-speaker ASR dataset to isolate clean text and high-quality audio segments, ensuring precise text formatting and alignment. Then, a pre-trained, off-the-shelf voice cloning solution is used to **re-voice** the entire text corpus into the **target donor voice** using a brief audio sample as a reference.

2.1. Initial ASR Donor Dataset Preparation

A large, multi-speaker ASR dataset serves as the textual and temporal blueprint for the new dataset.

- **Format Conversion:** All audio files must be standardized to a consistent format, such as **WAV (e.g., 16 kHz, 16-bit mono)**, for uniform processing.
- **Audio Enhancement (Optional but Recommended):** Applying a **denoising** or **de-reverberation** process can improve the quality of the source audio segments.
- **Volume Normalization:** *Goal: Ensure consistent speaking loudness across all samples.* Audio files are normalized to a target loudness level, typically using the **Loudness Units Full Scale (LUFS)** or **dBFS** standard (e.g., target loudness of **-20.0 dBFS**). This reduces variability that could negatively impact subsequent modeling.

2.2. Advanced Textual and Temporal Preparation (Crucial for Raw ASR Data)

To maximize the quality of the synthetic TTS dataset, the raw text and audio from the ASR source often require deep cleanup and re-segmentation.

- **Punctuation Addition/Normalization:** Raw ASR transcripts often lack crucial punctuation (commas, periods, question marks). Punctuation must be added based on grammatical rules or through a **punction restoration model** to ensure the synthesized audio receives correct prosodic cues (pauses, inflection).
- **Sentence Chunking and Segment Cleaning:**
  - **Chunking:** Very long utterances (e.g., over $15-20$ seconds) must be split into smaller, grammatically complete sentences. Long utterances are challenging for TTS models to synthesize consistently.
  - **Text Normalization:** Numbers, abbreviations, and acronyms must be converted into their full written forms (e.g., '1995' to 'nineteen ninety-five').
- **Forced Alignment for Re-Segmentation:** *Goal: Align audio boundaries to precise sentence/phrase boundaries.* For files that are long or contain multiple sentences, a **Forced Aligner** (e.g., based on kaldi or MFA) is used. This process aligns the provided transcript text to the multi-speaker audio and generates precise start/end timestamps for each sentence segment. This is used to **split long audios into smaller, clean files** corresponding exactly to a single, punctuated sentence.

2.3. Quality and Temporal Filtering

Utterances are filtered post-segmentation to remove outliers based on speaking pace.

- **Silence Trimming (Post-Alignment):** *Goal: Remove remaining non-speech segments at segment boundaries.* A **Voice Activity Detection (VAD)** algorithm is used on the new, smaller segments to remove any leading and trailing silences not precisely captured by the Forced Alignment. A minimum final duration (e.g., **1.0 second**) is enforced.

  Example: Input duration reduced from 5.58s to 3.97s.

- **Words-Per-Minute (WPM) Calculation and Filtering:** WPM is calculated for each new, clean audio-text segment: $$WPM = \frac{\text{Number of Words}}{\text{Audio Duration in Minutes}}$$ Utterances are filtered using statistical outlier removal (e.g., keeping files within **± 2.0 standard deviations (σ)** of the mean WPM) to ensure the training data has a consistent and natural rhythm.

  Example: Filtering range of 88.24 WPM to 247.08 WPM.

2.4. Voice Cloning and Synthesis via Zero-Shot Revoicing

This stage generates the final audio data using the cleaned text , normalized audio and the target donor voice via modern zero-shot cloning.

1. **Zero-Shot Cloning Model Selection:** A modern, pre-trained **zero-shot voice cloning model** (e.g., an open-source solution like **Chatterbox**) is selected. These models replicate the **timbre and style** of a target speaker from a very **short reference clip** (the donor voice) without requiring new model training.
2. **Dataset Revoicing:** The **clean, segmented, and punctuated text transcripts** (from Sections 2.2 and 2.3) are synthesized using the zero-shot model, conditioned on the **donor voice's reference audio clip**. This process generates a large, monospeaker dataset.
3. **Post-Synthesis Cleanup:** All newly synthesized files undergo a **manual spot-check** for any pervasive artifacts introduced by the cloning process.

---

3. Training the Final TTS Model

The newly synthesized dataset (consisting of the cleaned text transcripts and the synthesized audio in the target voice) is now used for the final TTS model training.

- **Model Selection:** The synthetic dataset is suitable for training a wide range of end-to-end TTS models, such as **VITS** or other modern architectures.
- **Transfer Learning:** If an existing TTS model is available, the synthetic dataset can be used in a **fine-tuning** step, which often accelerates convergence and improves final quality.

---

4. Conclusion

This hybrid methodology offers a robust and scalable solution for creating large, high-quality TTS datasets from limited donor audio samples and readily available ASR resources. The reliance on **zero-shot voice cloning** makes the pipeline highly efficient for **low-resource TTS** scenarios.