**Rasool Bux Palejo Neural Voice Resurrection Project**

***Phase-0 Master Technical Plan***

*(For DSP, Machine Learning, and Embedded Engineering Teams)*

**1. Executive Summary & Vision**

This project seeks to **digitally resurrect the voice and rhetorical cadence of Rasool Bux Palejo**, a legendary Sindhi political leader, writer, and orator, by scientifically modeling his linguistic, acoustic, and prosodic characteristics.  
The goal is to create an **AI-driven text-to-speech (TTS) system** that generates new spoken content in his authentic tone, rhythm, and emotional style.

The system will progress through three operational stages:

1. **Stage 1 – PC-based English TTS:**  
   Model training and inference on a workstation; user inputs text and receives downloadable audio in his voice.
2. **Stage 2 – Sindhi Long-Form Speech Generation:**  
   Adaptation of the model for Sindhi language, book-length narration, and rhetorical prosody control.
3. **Stage 3 – Embedded Offline Reader:**  
   Deployment of a pre-trained model on a small single-board computer (SBC) or edge device that plays pre-stored Sindhi/English texts using his synthetic voice.

Core modeling stack: **FastPitch + HiFi-GAN**, augmented with **prosody, duration, and style-embedding control**.

**2. End-to-End System Architecture**

**Input:** Text (English or Sindhi)  
**Processing Stages:**

1. **Text Normalization + G2P Conversion:**  
   Cleans text, expands numerals, and converts to phoneme sequence (using espeak-ng / phonemizer).
2. **Acoustic Model (FastPitch):**  
   Predicts mel-spectrograms from phonemes conditioned on F0 (pitch), energy, duration, speaker embedding, and style vector.
3. **Vocoder (HiFi-GAN):**  
   Converts mel-spectrograms to time-domain waveform.
4. **Post-Processing:**  
   Loudness normalization, silence insertion, metadata tagging.
5. **Output:** WAV/MP3 audio; downloadable or streamed.

**Training Data Path:**  
Audio → Denoising → Alignment → Phoneme/Prosody labeling → Dataset metadata.

**Deployment Path:**  
Trained model → ONNX export → Quantization → Runtime on PC/Web/Edge.

**3. Scientific Foundations**

**3.1 Linguistics & Dialectology**

* **Multilingual triad:** Sindhi ↔ Urdu ↔ English code-switching.
* **Lexical density:** High semantic load per phrase; necessitates rhythm-sensitive duration modeling.
* **Sindhi phonology:** Implosives /ɓ ɗ ʄ/, retroflexes /ʈ ɖ/, long vowels /ā ī ū/, nasalization—all central to identity.

**3.2 Phonetics & Prosody**

* **Pitch contour (F0):** rising-falling, slow decline at sentence ends.
* **Energy envelope:** deliberate crescendos for emotional peaks.
* **Temporal rhythm:** alternation between rapid logical sequences and reflective pauses.
* **Spectral balance:** low-mid warmth; harmonic richness typical of charismatic baritone voices.

**3.3 Cognitive & Rhetorical Science**

* **Ethos/Pathos/Logos balance:** measured assertiveness, emotional resonance, logical clarity.
* **Neural coupling:** cadence induces listener entrainment; must be reflected via F0-energy correlation modeling.

**4. Stage 1 – PC English Voice Cloning**

**4.1 Objective**

Create a controllable English-language TTS model reproducing his timbre, rhythm, and rhetorical style.

**4.2 Core Components**

* **Dataset:** ≥ 3 h cleaned English-dominant recordings.
* **Sampling:** 22.05 kHz / 16-bit mono.
* **Transcription:** Whisper-v3 + manual correction.
* **Alignment:** WhisperX or Montreal Forced Aligner.
* **Phonemization:** espeak-ng (en-gb).
* **Features:** F0, energy, duration, pause labels.

**4.3 Training Strategy**

* Fine-tune **FastPitch** on curated dataset.
* Use **ECAPA-TDNN speaker embedding** for timbre conditioning.
* Train **HiFi-GAN V1** vocoder on same voice.
* Validate via **MOS**, **F0 RMSE**, **spectral convergence**.

**4.4 Outputs**

* Interactive **web-based UI** (FastAPI + HTML + download).
* Model checkpoints, prosody statistics, and baseline evaluation report.

**5. Stage 2 – Sindhi Long-Form Speech & Book Cadence**

**5.1 Objective**

Extend the model to handle full-book narration in Sindhi, preserving natural phrasing, sentence prosody, and dialectal features.

**5.2 Data & Linguistics**

* Collect 3–6 h of Sindhi speeches and readings.
* Normalize Arabic-script Sindhi (Unicode NFC; harmonize yā/kaaf/heh variants).
* Build **Sindhi lexicon** and **phoneme mapping** for espeak-ng.
* Annotate pauses, rhetorical markers, and code-switch points.

**5.3 Model Adaptation**

* Continue fine-tuning FastPitch using multilingual phoneme embeddings.
* Integrate **language-condition tag (Sindhi = sd)**.
* Extend HiFi-GAN with Sindhi data for acoustic continuity.
* Implement **prosody memory module** to maintain context across long segments.

**5.4 Narration Control**

* Sentence segmentation via punctuation; insert micro-pauses (150–300 ms).
* Chapter boundary tones modeled through prosody vector shifts.
* Optional chime or breath noise insertion for realism.

**6. Stage 3 – Embedded Offline Speech Device**

**6.1 Hardware Targets**

* **SBC:** Raspberry Pi 5, Jetson Orin Nano, or equivalent.
* **Audio Path:** I²S DAC → Class-D amp → speaker.
* **UI:** Buttons + small TFT/OLED display + SD card.

**6.2 Software Stack**

* **Runtime:** ONNX Runtime (ARM build).
* **Language front-end:** C++/Python text normalizer + phonemizer.
* **TTS daemon:** text → phoneme → mel (ONNX FastPitch) → audio (ONNX HiFi-GAN).
* **Double-buffer streaming** for continuous playback.

**6.3 Optimization**

* **Quantization:** INT8 dynamic for mel model; FP16 for vocoder.
* **Target RTF:** ≤ 0.5 (on Pi 5).
* **Power budget:** < 6 W average.

**7. Data Engineering & Corpus Development**

| **Step** | **Process** | **Tools** |
| --- | --- | --- |
| 1 | Denoising, dereverb, loudness normalization | RNNoise, ffmpeg, librosa |
| 2 | Silence & breath trimming | SoX, pydub |
| 3 | Forced alignment | WhisperX, MFA |
| 4 | Phoneme extraction | phonemizer |
| 5 | Prosody extraction | RMVPE (F0), energy envelope |
| 6 | Metadata assembly | `utt\_id |

**8. Signal Processing Strategy**

**8.1 Acoustic Representation**

* 22.05 kHz sample rate, 1024-point FFT, 256-hop, 80 mel bins.
* Hann windowing; pre-emphasis = 0.97.
* Mel-spectrogram normalized to [0, 1].

**8.2 Feature Extraction**

* **F0:** RMVPE algorithm.
* **Energy:** frame-wise RMS.
* **Duration:** from forced alignments.
* **Pause & phrasing tags:** silence > 200 ms = phrase boundary.

**8.3 DSP Metrics**

* **SNR, LSD, F0 RMSE, spectral convergence** during training.
* **Perceptual metrics:** STOI, PESQ, SRMR.

**9. Machine & Deep Learning Strategy**

| **Layer** | **Model** | **Notes** |
| --- | --- | --- |
| Text → Mel | FastPitch | Duration/pitch/energy predictors, prosody tokens |
| Vocoder | HiFi-GAN V1 | Trained on same speaker for coherence |
| Speaker Encoder | ECAPA-TDNN | Produces 256-D embedding for conditioning |
| Style Encoder | GST/Reference Encoder | Captures emotional tone & rhetorical mode |

**Loss Functions**

* (MSE + MAE)
* , ,
* for HiFi-GAN
* Total = weighted sum (λ₁–λ₄).

**Optimization:** AdamW, cosine annealing, grad clip 1.0.  
**Batch:** 16–32; training epochs: 200 (≈ 2 weeks on RTX 3080).

**10. Linguistic & Phonetic Engineering**

* **Sindhi phoneme set:** derive ≈ 45 phonemes incl. implosives, retroflex, nasals.
* **Custom lexicon:** human-verified word → phoneme dictionary.
* **Phoneme embedding:** 128 D learned vectors.
* **Text normalization:** expand numbers, dates, units; standardize punctuation.
* **Prosody markup:** XML/JSON tags (<pause>, <emph>, <rise>).

**11. Prosody & Rhetorical Cadence Modeling**

* Extract **style vectors** from archetypal speeches (rally, elegy, reflection).
* Implement **prosody scaling factors:**
  + shift ± 20 cents for emphasis.
  + Duration stretch 1.1× for solemn tone.
* **Pause insertion rules:** semantic + syntactic boundaries.
* Encode **speech modes:** RALLY, REFLECTIVE, NARRATIVE, LAMENT.

**12. Software Architecture & Interface Planning**

* **Backend services:** FastAPI REST endpoints (/synthesize, /status, /download).
* **Frontend:** HTML/JS UI with text box, language selector, download button.
* **Model I/O:** PyTorch (checkpoint) → ONNX → Runtime.
* **Configuration:** YAML project config for paths and hyperparameters.
* **Logging & Analytics:** TensorBoard + CSV logs for prosody statistics.

**13. Evaluation Metrics & QA Protocol**

| **Category** | **Metric** | **Target** |
| --- | --- | --- |
| **Naturalness** | MOS (1–5) | ≥ 4.2 |
| **Speaker Similarity** | Cosine similarity (ECAPA embeddings) | ≥ 0.75 |
| **Intelligibility** | Re-ASR WER | ≤ 5 % |
| **Prosody Accuracy** | F0 RMSE < 20 Hz; Duration RMSE < 40 ms |  |
| **DSP Quality** | PESQ > 3.5; STOI > 0.9 |  |
| **Edge Performance** | RTF ≤ 0.5; RAM ≤ 1 GB |  |

**Listening tests:** double-blind 10-listener panel covering emotion, clarity, authenticity.

**14. Risk Analysis & Mitigations**

| **Risk** | **Mitigation** |
| --- | --- |
| Limited Sindhi corpus | Semi-supervised augmentation; cross-lingual transfer |
| Noise in recordings | Spectral subtraction + RNNoise pre-processing |
| Prosody instability (long text) | Context-window prosody memory + sentence-level chunking |
| Edge compute limits | INT8 quantization, FP16 fallback for vocoder |
| Ethical misuse | Audible disclaimer + watermarking module in output |

**15. Engineering Milestones & Deliverables**

| **Phase** | **Duration** | **Deliverables** |
| --- | --- | --- |
| **P-1: Corpus Creation** | 4 weeks | Cleaned audio + metadata + Sindhi lexicon |
| **P-2: Stage 1 Training** | 6 weeks | FastPitch/HiFi-GAN models + English web demo |
| **P-3: Prosody Control + Evaluation** | 4 weeks | Style vectors + MOS report |
| **P-4: Stage 2 Adaptation** | 6 weeks | Sindhi model + book narration demo |
| **P-5: Stage 3 Edge Integration** | 8 weeks | ONNX runtime demo device |
| **P-6: Documentation & Launch** | 2 weeks | User manual, ethical statement, research paper draft |

**Closing Statement**

This plan unites **communication science, phonetics, digital signal processing, and deep learning** into a practical engineering framework.  
The system will not merely synthesize voice; it will **reconstruct the prosodic and rhetorical spirit** of Rasool Bux Palejo — preserving his intellectual and emotional resonance for future generations.

Each subsystem (DSP → Acoustics → Linguistics → ML → Edge Deployment) has been aligned scientifically and technically to ensure feasibility, reproducibility, and fidelity.

**Prepared for:** Multidisciplinary Engineering Team (DSP / ML / Linguistics / Embedded)  
**Core Lead Model:** FastPitch + HiFi-GAN with Prosody Enhancement  
**Target Platforms:** Windows (Training & Demo) → Linux/Edge (Deployment)  
**Primary Languages:** English / Sindhi  
**Document Version:** Phase-0 Master Plan v1.0