**PROJECT DESCRIPTION (DSP-BASED SPEECH RECONSTRUCTION & DATASET ENGINEERING SYSTEM)**

**Project Name (Working Title):**

**RBP-DSP: A High-Precision Speech Signal Processing Pipeline for Clean Corpus Generation, Prosodic Feature Extraction, and AI-Ready Dataset Construction**

**Objective**

This project develops a **complete DSP pipeline** that transforms raw, noisy, variable-length audio recordings of Rasool Bux Palijo into a **fully processed, annotated, linguistically-aware, and AI-ready speech dataset**. The final dataset will be so thoroughly prepared that **future AI training (TTS, Voice-Cloning, Speech Synthesis, Prosody Modeling, or Speech-Driven Embodied Agents)** will require **no preprocessing at training time** — achieving a *“100% clean, aligned, structured, and scientifically enriched dataset.”*

**Core Vision**

The project’s long-term aim is to **digitally preserve and re-synthesize** the voice, speaking style, prosody, accent, and rhetorical identity of Rasool Bux Palijo. To do that, the system must:

1. **Extract the true voice signal** (without crowd noise, applause, environmental reverberation, or mic artifacts)
2. **Model linguistic identity** (Laari accent, discourse rhythm, phonetic realizations, breathing patterns)
3. **Model communication personality** (prosodic curvature, emotional crescendos, rhetorical pacing)
4. **Output a dataset** that mirrors **who he was as a *voice*** — not merely *what he said as text.*

**Input & Output**

| **Stage** | **Type** | **Description** |
| --- | --- | --- |
| **Input** | .wav, .mp3, .aac, etc. | Raw public speeches, interviews, TV talks, rallies, long-form recordings |
| **Intermediate** | DSP-processed signals | Denoised, dereverbed, source-separated, normalized speech |
| **Final Dataset** | AI-ready structured corpus | Clean speech + aligned transcripts + phonetic labels + prosody + features |
| **Verification Output** | Comparison plots & metrics | Objective DSP quality scores + side-by-side signal comparison |

**Folder Architecture**

📦 RBP-DSP/

├── raw\_data/ # unprocessed speech (any length, any number)

├── processed\_data/ # output of DSP pipeline

├── segments/ # utterance-level aligned clean chunks

├── features/ # MFCC, LogMel, Formants, Prosody, Glottal, etc.

├── annotations/ # TextGrid, phoneme tiers, word tiers, ToBI prosody tiers

├── metadata/ # timestamps, speaker notes, dialect flags, quality logs

├── verification/ # comparison plots, metrics, evaluation reports

└── scripts/ # DSP, alignment, feature extraction & verification code

**DSP PIPELINE (High-Precision, Industry-Level)**

Every audio clip (regardless of quality or duration) will pass through a **multi-stage DSP system**:

| **Stage** | **Algorithm(s)** | **Purpose** |
| --- | --- | --- |
| **1. Voice Activity Detection (VAD)** | WebRTC / GMM-HMM VAD / Neural VAD | Isolate only speech frames |
| **2. Source Separation** | Conv-TasNet / Demucs / DPCL++ | Remove chants, applause, background voices |
| **3. Dereverberation** | **WPE (Weighted Prediction Error)** | Suppress room reflections and echo tails |
| **4. Noise Reduction** | Hybrid: **MMSE-LSA + Wiener** or RNNoise | Remove stationary & non-stationary noise |
| **5. De-buzz / De-hum** | Adaptive notch + spectral tracking | Remove powerline hum & harmonics |
| **6. De-click / Dropout Repair** | Interpolative LPC-based restoration | Repair damaged frames |
| **7. Dynamic Range & Loudness Control** | Adaptive DRC + LUFS normalization | Produce consistent clean speech |
| **8. Final Speech Refinement** | GSS + Spectral Subtraction + Harmonic Enhancement | Retain natural timbre while maximizing clarity |

**Linguistic & Communication-Science Feature Extraction (Core Research Innovation)**

The dataset will **scientifically encode communication identity**, not just phonemes:

| **Category** | **Examples** |
| --- | --- |
| **Prosody** | F0 contour, energy contour, phrase boundaries, intonation tones (ToBI), rhythm indices (nPVI, %V), speech rate |
| **Phonetics** | VOT, formants (F1-F3), vowel space, consonant realization, Laari accent markers |
| **Voice Quality** | Jitter, Shimmer, HNR, CPP, Glottal parameters (IAIF/QUASI) |
| **Acoustic Features** | MFCC, PNCC, LPC, LogMel, Spectral flux, Zero-Crossing Rate |
| **Rhetorical Biometrics** | Pause strategy, emphasis peaks, climax crescendos, call-and-response timing |

**Scientific Verification System (Final Requirement)**

The project will include a **comparison program** that runs:

**Metrics:**

* SNR (before vs after)
* PESQ (speech quality)
* STOI (intelligibility)
* SDR/SI-SDR (source clarity)
* LSD (log-spectral distance)
* F0-contour similarity
* Envelope similarity (Hilbert)
* Spectrogram difference heat-maps
* Waveform overlay plots

**Output:**  
A PDF/HTML verification report for every run, comparing **raw vs processed** signals visually and numerically.

**Final Deliverable (Before AI Stage)**

A fully engineered dataset containing:

✅ clean speech  
✅ utterance segmentation  
✅ full transcripts  
✅ phoneme & syllable alignments  
✅ prosody (ToBI-LR tiers)  
✅ acoustic & glottal features  
✅ linguistic & rhetorical metadata  
✅ DSP verification metrics  
✅ scripts to re-run entire pipeline deterministically

This dataset will be **directly pluggable into Tacotron, VITS, FastPitch, WaveGlow, HiFi-GAN, or your own custom models**, with **zero preprocessing needed later**.