

Comparative Analysis Report: English vs Hindi Text-to-Speech (TTS) Synthesis

Project Overview

The goal of this project is to compare the performance of Text-to-Speech (TTS) models for English and Hindi languages. This comparison aims to evaluate the synthesis quality, acoustic feature richness, and technical robustness of pre-trained models using the `TTS`, `SpeechBrain`, and `indic-nlp-library` frameworks.

We systematically executed speech synthesis for one English and one Hindi sentence using high-performing models from the Coqui TTS framework and then analyzed them via acoustic metrics and visualization techniques.

Tools & Libraries Used

- `TTS` (`coqui-ai/TTS`) for speech synthesis
 - `SpeechBrain` for model loading and speech pipelines
 - `Indic NLP Library` for possible Hindi tokenization and normalization (limited use)
 - `librosa` and `matplotlib` for feature extraction and visualization
 - `IPython.display.Audio` for audio preview in notebook
-

Experimental Setup

English

- **Model:** `tts_models/en/ljspeech/tacotron2-DDC`
- **Dataset:** `LJSpeech`
- **Text Used:** `Your order will be delivered`

Hindi

- **Model:** `tts_models/hi/eksteps/fastspeech2-vits`
 - **Dataset:** `EkStep Indic Corpus` (less documentation)
 - **Text Used:** `आपका ऑर्डर कल डिलीवर किया जाएगा`
-

Feature-Based Analysis

Extracted Metrics

Metric	English Output	Hindi Output	Interpretation
Duration	~2.74 seconds	2.4 seconds	Similar
RMS (Energy)	0.20-0.30	0.35 (Higher)	Hindi louder but not necessarily clearer
Zero Crossing Rate (ZCR)	~0.042	0.019	Hindi has smoother waveform; possibly muffled
Spectral Centroid	~640 Hz	223 Hz	Hindi spectrum concentrated in lower frequencies


Interpretation

- **Higher spectral centroid** in English indicates brighter, more intelligible sound.
- **Lower ZCR** in Hindi may suggest more voicing but less clarity.
- **Higher RMS** in Hindi could indicate louder sound but not better quality.

Observed Issues in Hindi Model

1. **Initial Feature Extraction Failure:**
2. Hindi audio failed to produce `np.ndarray`, suggesting internal format issues or low energy.
3. **Silent or Low-quality Audio:**
4. Hindi synthesis sounded more robotic, possibly due to poor training corpus or G2P mapping.
5. **No Access to Preprocessing:**
6. English text underwent proper normalization and phonemization.
7. Hindi used internal preprocessing; user couldn't inject or inspect G2P output.

Code-Level Observations

Step	English	Hindi
Audio Feature Extraction	Successful	 Initially failed
Phoneme Support	Full G2P & normalization	Black-box
Audio Quality	Natural and clear	Robotic and dull
Visualization Readiness	Complete plots and charts	Later fixed after debugging

Visual Analysis

Bar charts were generated for each metric (RMS, Spectral Centroid, ZCR). English consistently scored higher in spectral clarity and waveform complexity, indicating more natural speech.

However, these were only visible after careful debugging, such as ensuring audio data format and sample rate compatibility.

Conclusion

Based on the extracted metrics, audio playback, and debugging process, we conclude:

Hindi TTS still lags behind English TTS

- **Lower acoustic richness** (low spectral centroid, low ZCR)
- **More robotic quality** (subjective and objective)
- **Limited preprocessing transparency**

The results emphasize the need for: – Larger, cleaner datasets for Hindi – Better Grapheme-to-Phoneme models – Open phoneme-level control and normalization for Indic languages
