Comparative Analysis Report: Hindi vs English TTS Synthesis

1. Objective

This report presents an analysis of Hindi and English text-to-speech (TTS) outputs using acoustic features.
The goal is to identify synthesis quality disparities between Indian and non-Indian languages using state-of-the-art models.

The Hindi synthesis was performed using **AI4Bharat's** `ai4bharat/indic-tts-hin-bark` model, while the English synthesis likely used a more mature TTS model such as from SpeechBrain or Coqui.

---

2. Acoustic Feature Comparison

Feature	English	Hindi	Relative Difference	Interpretation
Duration (sec)	2.62	2.85	Hindi +0.23 sec	Hindi speech is slightly longer
Spectral Centroid (Hz)	651.06	1592.82	~2.45× higher in Hindi	Hindi has more high-frequency energy; may sound sharper
RMS Energy	0.26	0.15	Hindi ~42% lower	Hindi is softer in loudness; may need normalization
Zero Crossing Rate	0.04	0.08	English is 50% lower	Hindi has more abrupt signal transitions (possible artifacts)

---

3. Key Observations

- **Longer duration** of Hindi suggests pacing/misalignment issues in synthesis.
- **Higher spectral centroid** in Hindi points to sharper or potentially distorted output, possibly due to vocoder tuning mismatches.

- **Lower RMS energy** in Hindi makes the output sound flatter or quieter.
- **Higher zero crossing rate** in Hindi may indicate unnatural waveform transitions or noisy synthesis.

---

4. Likely Causes & Limitations in Hindi TTS

The findings reflect systemic weaknesses in Indian language synthesis:

Tokenization & G2P Conversion

- English-centric models ignore features like aspiration, retroflexion, nasalization, and schwa deletion.
- Hindi's phonology demands tailored grapheme-to-phoneme mapping.

Tacotron2 or Sequence-to-Sequence Modeling

- Models like Tacotron2 often misalign longer or complex syllabic structures in Hindi, causing stretched or skipped phonemes.

Vocoder (e.g., HiFi-GAN)

- Trained mainly on English, vocoders often produce noisy or sharp Hindi speech — contributing to high spectral centroid and ZCR.

Lack of Prosody & Expressiveness

- Flat energy curves and emotionless intonation stem from insufficient Indian expressive corpora.

---

5. Conclusion

These findings empirically confirm that Indian languages — specifically Hindi — face limitations in current TTS pipelines.
To improve output quality and naturalness:

- Models must be trained with larger, diverse Indian speech datasets.
- Phoneme-level modeling and duration prediction need customization for Indian scripts.
- Vocoders should be retrained or fine-tuned with Indian acoustic characteristics.
- Incorporating pitch, energy, and prosody modeling is crucial for expressiveness.

This analysis contributes toward developing a **more inclusive, robust, and linguistically faithful speech synthesis pipeline** for Indian languages.