IndiaSpeaks Data Standards

This document outlines the normalization standards, allowable characters, digit expansion policies, and handling of non-verbal content for training TTS models across Indian languages.

Language-wise Normalization Rules

hi: Use lowercase Latin transliteration, remove diacritics, normalize 'aa'/'a', 'ee'/'i', etc.

bn: Use WX-style transliteration, strip tone marks, normalize conjuncts.

ta: Latin-encoded Tamil with consistent vowel markers, expand compound forms.

te: Telugu characters transliterated with consistent CV structure; collapse repetitive markers.

gu: Strip diacritic marks, normalize 'sh', 'kh', etc. to consistent grapheme forms.

kn: Use ISO standard transliteration; remove nukta forms.

ml: Normalize chillu characters, use consistent anusvara/visarga replacement.

mr: Latin-encoded Marathi; collapse variations like 'shh', 'rrh' to canonical.

pa: Strip tonal marks; normalize 'kh', 'gh' clusters.

ur: Roman Urdu; replace Arabic digits with Latin; remove decorative symbols.

en: Lowercase, remove all diacritics, preserve only basic punctuation.

Allowable Punctuation

Only the following characters are allowed in the normalized text corpus:

- Period (.)
- Comma (,)
- Question Mark (?)
- Exclamation Mark (!)
- Apostrophe (')
- Hyphen (-)
- Colon (:)
- Semicolon (;)

Digit Expansion Tables

All numeric expressions must be expanded into full words per language. Examples:

Hindi:
- 2024 → do hazaar chaubees
- 100 → ek sau

Tamil:
- 2024 → இரண்டாயிரம் இருபத்திநான்கு
- 100 → நூறு

English:
- 2024 → two thousand twenty-four
- 100 → one hundred

Non-verbal Token Policy

Non-verbal tokens such as [laugh], [breath], [noise], etc. must be marked explicitly in square brackets. During training, these tokens should be retained in transcripts but ignored in mel prediction targets. Each token must be consistently used and well-defined in the tokenizer.

Allowed Non-verbal Tokens:

- [laugh]
- [breath]
- [cough]
- [pause]
- [noise]

Page reserved for notes or additional standards.

(Empty)

Page reserved for notes or additional standards.

(Empty)

Page reserved for notes or additional standards.

(Empty)