# Cross-Lingual TTS for Indian Languages: A Benchmark and Inference-Time Duration Estimation Strategies

**Adarsh Arigala*** **Arjun Gangwar*** **Srinivasan Umesh**
SPRING Lab, Indian Institute of Technology Madras, India
arigalaadarsh780@gmail.com arjungangwar@gmail.com
umeshs@ee.iitm.ac.in

## Abstract

Recent alignment-free Text-to-Speech (TTS) models achieve high synthesis quality by removing explicit alignment modules and simplifying training. However, they often struggle in cross-lingual scenarios due to differences in character durations across languages, leading to duration mismatch and degraded speech quality. Existing approaches typically address this issue using explicit duration predictors, increasing model complexity. Systematic evaluation of TTS performance for Indian languages in cross-lingual settings remains under-explored due to the lack of standardized evaluation benchmarks. In this work, we construct a cross-lingual evaluation benchmark covering English and nine Indian languages and propose a language-aware inference-time duration estimation strategy for alignment-free TTS that incorporates language-specific duration-per character statistics without retraining or architectural modifications. We further develop heuristic formulations that refine duration estimation based on language-dependent duration variations. Experiments on bilingual and multilingual A-DMA models demonstrate improved speech naturalness, intelligibility, and speaker similarity while reducing duration mismatch. [1]

## 1 Introduction

Text-to-Speech (TTS) systems have achieved significant progress with the emergence of neural architectures, enabling natural and expressive speech synthesis. Recent alignment-free models such as E2-TTS (Eskimez et al., 2024), F5-TTS (Chen et al., 2025) and A-DMA (Choi et al., 2025), simplify training by removing explicit alignment modules while maintaining high synthesis quality. Although these models perform well when reference and target speech is in the same language, they often struggle in cross-lingual scenarios where reference and target languages differ. This happens due to differences in writing systems, orthographic characteristics, and language-specific duration-per character variations across languages. Moreover, in F5-TTS, duration estimation is based on the number of UTF-8 bytes obtained by encoding the input text, rather than the actual number of characters. Since characters from different writing systems occupy different numbers of UTF-8 bytes and languages exhibit different duration-per-character patterns, the estimated duration can become inaccurate, leading to duration mismatch and degraded synthesis quality.

To address these challenges, recent works have explored various strategies to improve duration modeling and generation quality in alignment-free TTS. Cross-Lingual F5-TTS (Liu et al., 2026) removes the dependency on reference transcripts in flow-matching TTS and enables cross-lingual voice cloning through forced alignment during training. However, it introduces additional training complexity by requiring MMS-based alignment (Pratap et al., 2024) preprocessing and separately trained speaking-rate predictors, making the overall pipeline more complex than standard alignment-free TTS systems.

Another work, PFluxTTS (Pankov et al., 2026) proposes a hybrid flow-matching framework that combines a duration-guided model and an alignment-free model through inference-time fusion, improving the trade-off between alignment stability and naturalness. However, it relies on an explicit duration predictor, which reduces the benefits of alignment-free TTS by reintroducing duration modeling.

Most people in India are bilingual and frequently switch between English and their regional languages during everyday conversations. Consequently, cross-lingual TTS between English and

*equal contribution

[1] The dataset and code recipe are available at: https://github.com/Cross-Lingual-TTS-for-Indian-Languages/Cross-Lingual-TTS-Indic-Test-Set/



*Proceedings of the 1st Workshop on Multi-modality and Multi-lingual Language Learning*, pages 47–55
November 9, 2026 ©2026 Association for Computational Linguistics

Indian languages has become increasingly important. These languages exhibit substantial differences in orthographic characteristics and character-to-speech duration, making accurate duration estimation essential for stable cross-lingual speech synthesis. Despite its practical importance, systematic evaluation of TTS performance for Indian languages in cross-lingual settings remains underexplored due to the lack of standardized evaluation benchmarks and limited investigation of duration adaptation strategies for alignment-free synthesis.

To address these limitations, we propose a cross-lingual evaluation benchmark covering nine Indian languages and English and introduce language-aware inference-time duration estimation strategies for alignment-free TTS. Our approach incorporates language-specific duration-per character statistics directly during inference without requiring retraining or additional model components. Specifically, we design heuristic formulations that adjust the estimated duration according to language-specific duration-per-character statistics, enabling effective cross-lingual duration adaptation.

This work makes the following contributions:

- We construct a cross-lingual evaluation benchmark covering nine Indian languages and English, evaluating two settings: **(i) English Reference and Indic Generation**, where English reference speech is used to generate Indian language speech and **(ii) Indic Reference and English Generation**, where Indian language reference speech is used to generate English speech, to systematically analyze TTS performance in cross-lingual scenarios.

- We train bilingual and multilingual A-DMA models for Indian languages from scratch to investigate their performance in cross-lingual settings.[2]

- We propose inference-time language-aware duration estimation strategies for alignment-free TTS in cross-lingual settings, addressing duration mismatch by incorporating language-specific duration-per character information into the generation process.

[2]**Demo Samples**: https://cross-lingual-tts-for-indian-languages.github.io/Cross-Lingual-TTS-Demo/

| Language Indic+Eng | Duration (Hrs) | No. of Indic Utt. | No. of Eng Utt. | Total Utt. |
|---|---|---|---|---|
| Kannada | 2.00 | 540 | 588 | 1128 |
| Telugu | 2.00 | 531 | 588 | 1119 |
| Tamil | 2.00 | 591 | 588 | 1179 |
| Hindi | 1.35 | 171 | 588 | 759 |
| Malayalam | 2.00 | 554 | 588 | 1142 |
| Punjabi | 2.00 | 520 | 588 | 1108 |
| Bengali | 2.00 | 543 | 588 | 1131 |
| Odia | 2.00 | 532 | 588 | 1120 |
| Marathi | 2.00 | 531 | 588 | 1119 |

Table 1: Cross-lingual evaluation set statistics.

## 2 Methodology

### 2.1 Baselines

We adopt A-DMA (Choi et al., 2025) framework for our experiments. It is based on the F5-TTS (Chen et al., 2025) backbone and introduces a training objective that combines conditional flow matching with dual-modality text and speech representation alignment losses to accelerate convergence. We utilize self-supervised representations (SSL) from the 21st layer of data2vec-AQC (Lodagala et al., 2023) to compute the speech-level alignment loss. The model is trained using the conditional flow matching loss from F5-TTS as the primary generative objective. A Connectionist Temporal Classification (CTC) based text alignment loss (Graves et al., 2006) is applied at an intermediate layer to for early character-level alignment, and a speech-level alignment loss is used at a later layer to enforce alignment in the speech representation space.

$$\mathcal{L} = \mathcal{L}_{\text{CFM}} + \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{speech}}\mathcal{L}_{\text{speech}},$$

where $\lambda_{\text{text}} = 0.1$ and $\lambda_{\text{speech}} = 1.0$. This combined objective improves convergence and speech synthesis quality.

### 2.2 Cross-lingual Evaluation Dataset

To test the cross-lingual generalization capability of the A-DMA (Choi et al., 2025) model, we constructed a dedicated cross-lingual evaluation benchmark covering nine Indian languages and English. In this setting, the reference speech and the text being synthesized belong to different languages.

For each Indian-language and English pair in the cross-lingual evaluation setting, we constructed approximately two hours of evaluation data, allocating one hour to (i) English Reference and Indic Generation and one hour to (ii) Indic Reference and

English Generation. The Indian-language speech was selected from the Rasa (Srinivasa Varadhan et al., 2024) and IndicVoices-R (Sankar et al., 2024) test splits, while the English speech was sampled from the IndicTTS (Consortium et al., 2023) training split. [3]

We considered two cross-lingual evaluation scenarios: (i) Indian-language reference speech with English generation text, and (ii) English reference speech with Indian-language generation text. In both scenarios, reference and target utterances were randomly sampled subject to duration constraints of 4–7 seconds and 4–10 seconds, respectively. Detailed statistics of the constructed benchmark, including the total number of utterances for each language pair, are provided in Table 1.

### 2.3 Model Setup

We trained the original A-DMA (Choi et al., 2025) model for Indian languages in two settings: bilingual and multilingual. For the bilingual setup, each model was trained on one Indian language combined with a 30-hour subset of Indian English data (Consortium et al., 2023), while the multilingual model was trained on 24 languages as described in Section 3.1.

The bilingual model used the F5-TTS Small architecture with 18 DiT layers (Peebles and Xie, 2023), 12 attention heads, and 768-dimensional embeddings, together with a 4-layer ConvNeXtV2 (Woo et al., 2023) module with 512-dimensional embeddings, totaling 159M parameters. It was trained for 1.2M steps with 20k warmup updates and an effective batch size of 0.75 hours.

The multilingual model used F5-TTS Base with 22 DiT layers (Peebles and Xie, 2023), 16 attention heads, and 1024-dimensional embeddings, along with a 4-layer ConvNeXtV2 (Woo et al., 2023) module with 512-dimensional embeddings, totaling 335.8M parameters. It was trained for 300k updates with 45k warmup updates and an effective batch size of 2.09 hours. For both settings, 100-channel mel spectrograms were extracted at a 24 kHz sampling rate with a hop size of 256 samples. Waveform synthesis was performed using the pre-trained Vocos vocoder (Siuzdak, 2024), and optimization was carried out using AdamW (Loshchilov and Hutter, 2019) with a peak learning rate of $7.5 \times 10^{-5}$.

[3] Indian English is available only in IndicTTS-v2, so we take 500 utterances from its 250-hour training set as a held-out set. Each language has only two speakers, and all speakers also recorded Indian English.

| Language | Indic Voices-R | Indic TTS-v2 | Indic TTS-v3 | Rasa | Total (Hrs) |
|---|---|---|---|---|---|
| Assamese | 173.77 | 25.73 | 20.33 | 58.32 | 278.15 |
| Bengali | 110.20 | 20.09 | 20.63 | 56.76 | 207.68 |
| Bodo | 170.48 | 0 | 14.47 | 52.31 | 237.26 |
| Gujarati | 4.50 | 26.49 | 18.34 | 25.39 | 74.71 |
| Hindi | 72.75 | 26.23 | 22.51 | 0 | 121.49 |
| Kannada | 43.22 | 22.32 | 11.54 | 54.62 | 131.70 |
| Kashmiri | 63.07 | 0 | 0 | 0 | 63.07 |
| Konkani | 49.33 | 0 | 11.00 | 26.33 | 86.66 |
| Maithili | 80.32 | 0 | 9.73 | 29.34 | 119.39 |
| Malayalam | 80.64 | 20.91 | 25.02 | 51.63 | 178.20 |
| Manipuri | 21.00 | 17.65 | 16.97 | 0 | 55.62 |
| Marathi | 48.73 | 22.37 | 22.78 | 55.36 | 149.24 |
| Nepali | 104.31 | 0 | 7.24 | 55.09 | 166.64 |
| Odia | 69.32 | 19.19 | 22.41 | 47.91 | 158.84 |
| Punjabi | 71.62 | 11.21 | 22.63 | 55.84 | 161.30 |
| Sanskrit | 33.25 | 0 | 22.31 | 51.47 | 107.03 |
| Santali | 72.76 | 0 | 0 | 0 | 72.76 |
| Sindhi | 9.03 | 0 | 14.53 | 0 | 23.56 |
| Tamil | 98.03 | 26.59 | 23.14 | 29.94 | 177.70 |
| Telugu | 134.86 | 36.71 | 23.67 | 52.27 | 247.50 |
| Urdu | 76.36 | 0 | 0 | 0 | 76.36 |
| Indian Eng. | 0 | 251.09 | 0 | 0 | 251.09 |
| Dogri | 0 | 0 | 13.19 | 48.04 | 61.23 |
| Rajasthani | 0 | 20.39 | 20.92 | 0 | 41.31 |
| **Total** | **1587.55** | **546.97** | **363.36** | **750.62** | **3248.50** |

Table 2: Training dataset statistics across multilingual speech corpora in hours.

### 2.4 Language-Aware Heuristic Duration Estimation

#### 2.4.1 Duration Estimation in Original F5-TTS

Alignment-free TTS models require predicting target speech duration prior to synthesis. A-DMA uses the F5-TTS duration estimation formulation defined as:

$$D_{F5} = D_{ref} + \left\lfloor \frac{D_{ref}}{L_{ref}} \cdot \frac{L_{gen}}{s} \right\rfloor \quad (1)$$

where $D_{ref}$ denotes the reference speech duration, $L_{ref}$ and $L_{gen}$ represent the number of characters in reference and generation text, which are converted into UTF-8 byte encoding for duration estimation, and $s$ denotes the local speed factor, which we set to 1.

While the standard F5-TTS formulation performs well when reference and generated speech share the same language, it computes text length using UTF-8 byte-level encoding, which is not suitable for multilingual settings. In cross-lingual generation, this becomes problematic because Indian languages typically require multiple bytes per character, whereas English characters generally require a single byte. Consequently, byte-level length estimation leads to inconsistent scaling in duration prediction across languages. Previous works have

addressed this limitation by incorporating explicit speaking-rate predictors (Liu et al., 2026) or duration predictors (Pankov et al., 2026), which require modifications to the original F5-TTS architecture and model training.

To overcome this issue, we reformulate Original F5-TTS in a character-level space and report results using this variant, referred to as ***Modified F5-TTS***. Additionally, we propose two language-aware duration modeling strategies, *Heuristic-1* and *Heuristic-2*, which incorporate duration-per character information into the F5-TTS formulation through fusion and additive interaction mechanisms, respectively.

#### 2.4.2 Heuristic-1: Language-Aware Duration Model

The heuristic (H1) extends the original F5 duration estimation by incorporating language-specific duration-per-character statistics.

$$D_{H1} = D_{ref} + \alpha \left\lfloor \frac{D_{ref}}{L_{ref}} \cdot \frac{L_{gen}}{s} \right\rfloor + \beta(D_{ind} + D_{eng}) \quad (2)$$

where $\alpha = 0.7$ and $\beta = 0.3$ in all experiments. The effect of varying $\alpha$ and $\beta$ is analyzed in Section 5.2. $D_{F5}$ denotes the duration estimated by F5-TTS. Duration estimations for Indian languages and English are computed as follows to capture language-dependent duration variations:

$$D_{ind} = \frac{d^c_{ind} \times C_{ind}}{s} \qquad D_{eng} = \frac{d^c_{eng} \times C_{eng}}{s} \quad (3)$$

where $D_{ref}$ denotes the reference speech duration, $d^c_{ind}, d^c_{eng}$ denote the duration-per character statistics for Indic languages and English, respectively, $C_{eng}$ and $C_{ind}$ denote the number of English and Indic characters in the generated text, and $s$ denotes the local speed factor, which we set to 1.

The language-specific duration-per character are estimated from multilingual training corpora. For each utterance, the duration-per character is computed as:

$$d^c = \frac{L_{audio}}{C_{text}} = \frac{N_{samples}/H}{C_{text}} \quad (4)$$

where $N_{samples}$ denotes the waveform length, $H$ denotes the hop length, and $C_{text}$ denotes the text length in characters.

To obtain robust language-specific statistics, duration-per-character values are aggregated across all speakers within each language using the corresponding training corpus (Table 2). For example, the Telugu statistic is computed from the Telugu training corpus, and similarly for the remaining languages. The language-specific average duration-per-character is computed as:

$$d^c_{lang} = \frac{1}{N_{\text{utt}}} \sum_{i=1}^{N_{\text{utt}}} D^c_i \quad (5)$$

where $N_{\text{utt}}$ denotes the number of utterances per language.

Fig. 1 shows the utterance-level distribution of duration-per-character values for all languages in the training corpora. Although each language exhibits natural variability due to differences in speakers and speaking styles, we use the average duration per character, $d^c_{lang}$, for each language in the proposed heuristics. The resulting language-specific average duration-per-character statistics are:

Indian English (9.6), Kannada (8.7), Telugu (8.4), Tamil (7.0), Hindi (8.5), Malayalam (6.8), Punjabi (10.1), Bengali (8.1), Odia (8.7), and Marathi (9.8). This formulation combines the original F5 duration estimate with language-aware duration estimation to better handle cross-lingual synthesis scenarios.

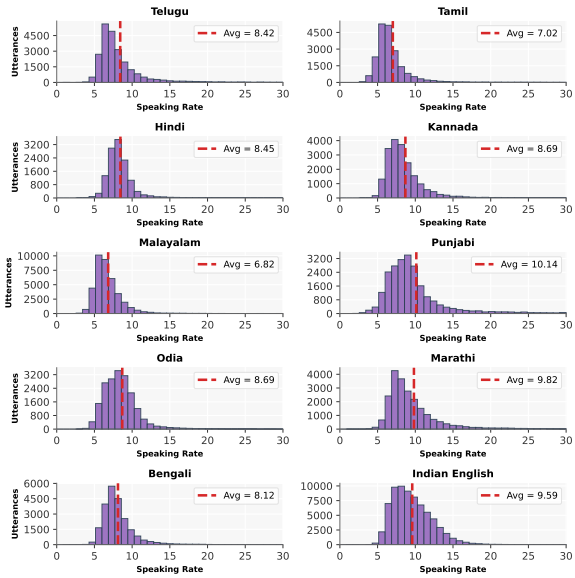



Figure 1: Utterance-level duration-per character distributions for the multilingual training data. The red dashed line indicates the average duration-per character ($d^c_{lang}$) used in the proposed heuristics.

| Model | Original F5-TTS | | | | Modified F5-TTS | | | | Heuristic-1 | | | | Heuristic-2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | WER | SIM | UTMOS | CER | WER | SIM | UTMOS | CER | WER | SIM | UTMOS | CER | WER | SIM | UTMOS | CER |
| **Kannada** | | | | | | | | | | | | | | | | |
| GT | 33.38 | – | 3.461 | 5.88 | 33.38 | – | 3.461 | 5.88 | 33.38 | – | 3.461 | 5.88 | 33.38 | – | 3.461 | 5.88 |
| Bilingual | 109.56 | 0.630 | 3.070 | 57.07 | 26.98 | 0.613 | 3.435 | 4.87 | 26.91 | 0.615 | 3.459 | 4.79 | 27.85 | 0.619 | 3.468 | 5.23 |
| Multilingual | 55.88 | 0.641 | 3.383 | 22.22 | 26.45 | 0.662 | 3.460 | 4.17 | 26.22 | 0.664 | 3.475 | 4.16 | 25.15 | 0.667 | 3.495 | 3.91 |
| **Telugu** | | | | | | | | | | | | | | | | |
| GT | 21.06 | – | 3.119 | 3.07 | 21.06 | – | 3.119 | 3.07 | 21.06 | – | 3.119 | 3.07 | 21.06 | – | 3.119 | 3.07 |
| Bilingual | 82.05 | 0.662 | 3.249 | 48.02 | 20.68 | 0.653 | 3.575 | 3.49 | 21.15 | 0.656 | 3.590 | 3.66 | 23.00 | 0.658 | 3.592 | 5.33 |
| Multilingual | 37.66 | 0.635 | 3.436 | 18.02 | 16.73 | 0.664 | 3.508 | 2.08 | 16.40 | 0.662 | 3.508 | 1.93 | 16.16 | 0.668 | 3.554 | 1.96 |
| **Tamil** | | | | | | | | | | | | | | | | |
| GT | 25.39 | – | 3.357 | 4.87 | 25.39 | – | 3.357 | 4.87 | 25.39 | – | 3.357 | 4.87 | 25.39 | – | 3.357 | 4.87 |
| Bilingual | 105.57 | 0.630 | 2.728 | 52.20 | 25.17 | 0.626 | 3.387 | 5.29 | 23.91 | 0.624 | 3.390 | 4.88 | 26.49 | 0.627 | 3.388 | 6.01 |
| Multilingual | 50.95 | 0.600 | 3.140 | 24.02 | 20.52 | 0.650 | 3.346 | 3.56 | 20.39 | 0.646 | 3.328 | 3.52 | 20.98 | 0.652 | 3.347 | 3.63 |
| **Hindi** | | | | | | | | | | | | | | | | |
| GT | 14.40 | – | 2.573 | 5.21 | 14.40 | – | 2.573 | 5.21 | 14.40 | – | 2.573 | 5.21 | 14.40 | – | 2.573 | 5.21 |
| Bilingual | 62.45 | 0.633 | 3.101 | 47.01 | 12.58 | 0.662 | 3.608 | 4.04 | 13.16 | 0.656 | 3.627 | 4.30 | 12.87 | 0.664 | 3.638 | 4.31 |
| Multilingual | 26.70 | 0.666 | 3.435 | 17.05 | 11.80 | 0.679 | 3.555 | 3.61 | 11.66 | 0.680 | 3.566 | 3.42 | 11.71 | 0.685 | 3.585 | 3.49 |
| **Malayalam** | | | | | | | | | | | | | | | | |
| GT | 25.89 | – | 3.284 | 3.87 | 25.89 | – | 3.284 | 3.87 | 25.89 | – | 3.284 | 3.87 | 25.89 | – | 3.284 | 3.87 |
| Bilingual | 121.85 | 0.579 | 2.301 | 56.82 | 28.91 | 0.622 | 3.359 | 6.22 | 26.49 | 0.622 | 3.403 | 5.03 | 31.42 | 0.624 | 3.377 | 7.73 |
| Multilingual | 53.42 | 0.600 | 3.170 | 23.81 | 22.98 | 0.661 | 3.343 | 3.49 | 22.15 | 0.660 | 3.351 | 3.34 | 22.45 | 0.667 | 3.353 | 3.44 |
| **Punjabi** | | | | | | | | | | | | | | | | |
| GT | 15.75 | – | 3.337 | 4.81 | 15.75 | – | 3.337 | 4.81 | 15.75 | – | 3.337 | 4.81 | 15.75 | – | 3.337 | 4.81 |
| Bilingual | 71.81 | 0.639 | 3.265 | 51.17 | 16.14 | 0.632 | 3.607 | 4.08 | 17.12 | 0.639 | 3.620 | 5.69 | 17.29 | 0.641 | 3.615 | 6.18 |
| Multilingual | 31.65 | 0.674 | 3.475 | 17.62 | 15.53 | 0.681 | 3.592 | 4.44 | 15.49 | 0.687 | 3.623 | 4.35 | 15.38 | 0.688 | 3.574 | 4.45 |
| **Bengali** | | | | | | | | | | | | | | | | |
| GT | 24.78 | – | 3.062 | 6.48 | 24.78 | – | 3.062 | 6.48 | 24.78 | – | 3.062 | 6.48 | 24.78 | – | 3.062 | 6.48 |
| Bilingual | 108.12 | 0.657 | 2.990 | 70.10 | 24.23 | 0.633 | 3.546 | 4.95 | 24.20 | 0.634 | 3.590 | 6.66 | 25.72 | 0.636 | 3.595 | 7.59 |
| Multilingual | 46.78 | 0.668 | 3.494 | 21.57 | 26.54 | 0.692 | 3.578 | 7.21 | 26.60 | 0.688 | 3.427 | 7.07 | 26.92 | 0.699 | 3.584 | 7.13 |
| **Odia** | | | | | | | | | | | | | | | | |
| GT | 23.52 | – | 3.131 | 4.51 | 23.52 | – | 3.131 | 4.51 | 23.52 | – | 3.131 | 4.51 | 23.52 | – | 3.131 | 4.51 |
| Bilingual | 79.53 | 0.676 | 3.399 | 49.29 | 27.51 | 0.646 | 3.513 | 5.83 | 27.83 | 0.649 | 3.531 | 5.83 | 28.29 | 0.653 | 3.542 | 6.16 |
| Multilingual | 43.25 | 0.651 | 3.467 | 18.90 | 26.57 | 0.656 | 3.477 | 5.27 | 26.17 | 0.655 | 3.512 | 5.19 | 26.09 | 0.665 | 3.529 | 5.18 |
| **Marathi** | | | | | | | | | | | | | | | | |
| GT | 10.56 | – | 3.366 | 2.08 | 10.56 | – | 3.366 | 2.08 | 10.56 | – | 3.366 | 2.08 | 10.56 | – | 3.366 | 2.08 |
| Bilingual | 74.51 | 0.596 | 2.900 | 43.51 | 11.46 | 0.605 | 3.475 | 2.47 | 11.72 | 0.611 | 3.496 | 2.47 | 12.04 | 0.605 | 3.480 | 2.76 |
| Multilingual | 30.76 | 0.637 | 3.449 | 15.44 | 10.58 | 0.657 | 3.501 | 2.08 | 10.74 | 0.661 | 3.528 | 2.00 | 10.43 | 0.663 | 3.545 | 1.96 |

Table 3: Cross-lingual TTS evaluation (English Reference and Indic Generation) across Indian languages

#### 2.4.3 Heuristic-2: Additive duration-per character Guided Scaling

The second heuristic extends the original character-based duration estimation by incorporating relative duration-per character differences between languages through an additive correction mechanism. The target duration is computed as:

$$D_{H2} = D_{ref} + \left( \frac{D_{ref}}{C_{ref}} + \frac{d^c_{gen}}{d^c_{ref}} \right) \frac{C_{gen}}{s} \qquad (6)$$

where $D_{ref}$ denotes the reference speech duration, $C_{ref}$ and $C_{gen}$ denote the number of characters in the reference and generated text, and $d^c_{ref}$ and $d^c_{gen}$ denote the language-specific duration-per character statistics corresponding to the reference and target generation languages. $\frac{D_{ref}}{C_{ref}}$ captures the reference duration-per-character, while $\frac{d^c_{gen}}{d^c_{ref}}$ provides a relative duration adjustment that introduces a language-dependent correction to per-character scaling across languages.

## 3 Experimental Setup

### 3.1 Training Dataset

We follow the original A-DMA (Choi et al., 2025) training setup and adapt it for Indian languages in order to evaluate its multilingual performance. The model was trained using four multilingual speech corpora: IndicVoices-R (Sankar et al., 2024) containing 1,587.5 hours of speech from 21 languages; IndicTTS V2 and V3 versions (Consortium et al., 2023) containing 547 hours and 363.3 hours of speech, respectively, covering 14 languages in V2 and 20 languages in V3; and Rasa (Srinivasa Varadhan et al., 2024) containing 750.6 hours of speech across 16 languages. In total, the model was trained on 24 languages, comprising approximately 3,248.5 hours of multilingual speech data. Detailed per-language training hours are provided in Table 2. The trained model are evaluated using the proposed cross-lingual benchmark described in Section 2.2.

### 3.2 Evaluation Metrics

We evaluate synthesized speech using four objective metrics. Word Error Rate (WER) and Character Error Rate (CER) are computed from ASR tran-

scriptions generated using Whisper-large-v3 (Radford et al., 2023) for English and IndicConformer-600M (AI4Bharat, 2024) for Indian languages. Speaker Similarity (SIM) evaluates speaker identity preservation using WavLM-large-based speaker embeddings (Chen et al.). UTMOS (Saeki et al., 2022) measures speech naturalness via an objective MOS prediction model.

## 4 Results

The trained bilingual and multilingual A-DMA models are evaluated using the proposed benchmark under four duration strategies: Original F5-TTS, Modified F5-TTS, Heuristic-1, and Heuristic-2. We evaluate performance in two cross-lingual settings: (i) Indic Reference and English Generation and (ii) English Reference and Indic Generation. Although recent methods such as Cross-Lingual F5-TTS (Liu et al., 2026) and PFluxTTS (Pankov et al., 2026) have proposed alternative duration modeling strategies, their official implementations were not publicly available at the time of our experiments. Consequently, we compare our proposed methods against the publicly available A-DMA baseline.

### 4.1 English Reference and Indic Generation

Table 3 presents cross-lingual results for English reference and Indian-language speech generation. The original F5-TTS exhibits severe degradation in cross-lingual settings, primarily due to byte-level text length estimation, which leads to incorrect duration scaling and high WER/CER across all languages. For instance, in Kannada, the bilingual model yields a WER/CER of 109.56/57.07, while the multilingual model also exhibits degraded performance, with a WER/CER of 55.88/22.22.

The Modified F5-TTS substantially reduces these errors across all languages by correcting text-length estimation, resulting in large relative improvements in both WER and CER. In Kannada, the multilingual model achieves relative reductions of 52.7%/81.2% in WER/CER compared to the Original F5-TTS, indicating better linguistic alignment. Further, Heuristic-1 and Heuristic-2 for the multilingual model, provide relative WER/CER reductions of 0.9%/0.2% and 4.9%/6.2% over the Modified F5-TTS, respectively, while UTMOS increases from 3.460 to 3.475 and 3.495. Similar trends are observed across all languages.

### 4.2 Indic Reference and English Generation

Table 4 reports results for the reverse direction. The original F5-TTS again shows high error rates. For Kannada, the bilingual model yields a WER/CER of 57.01/52.16, while the multilingual model reduces these to 49.66/44.33, respectively.

The Modified F5-TTS significantly improves cross-lingual robustness across all languages. In Kannada, the multilingual and bilingual models achieves relative reductions of 93.2%/97.6% and 91.6%/96.1% in WER/CER over the Original F5-TTS. Further, Heuristic-1 and Heuristic-2 for the multilingual model, provide additional improvements over the Modified F5-TTS, achieving WER/CER reductions of 5.4%/6.6% and 3.0%/4.7%, respectively. SIM remains comparable, while UTMOS shows slight improvement. Similar trends are observed across all languages. Overall, all proposed duration strategies operate within a similar performance range, showing stable improvements across Indian languages.

Compared to bilingual models, multilingual models generally achieve lower WER and CER across most languages, indicating improved robustness due to exposure to diverse scripts and duration-per character variations during training.

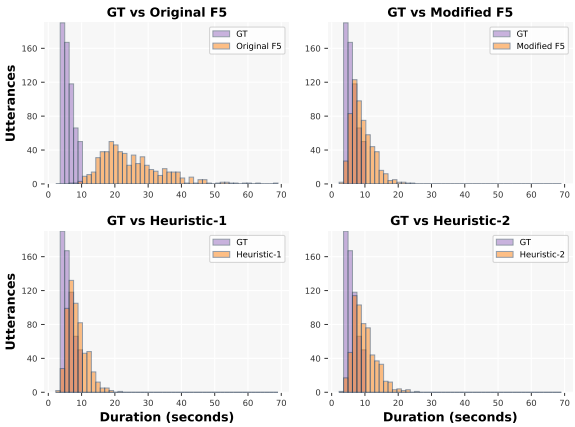



Figure 2: Comparison of ground-truth (GT) and predicted durations for English reference and Tamil speech generation.

## 5 Further Analysis

### 5.1 Qualitative Analysis

Fig. 2 and Fig. 3 show the predicted and ground-truth (GT) duration distributions on our benchmark for two cross-lingual settings: (i) English reference and Tamil speech generation and (ii) Tamil reference and English speech generation. GT durations correspond to natural multi-speaker utter-

| Model | Original F5-TTS | | | | Modified F5-TTS | | | | Heuristic-1 | | | | Heuristic-2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | WER | SIM | UTMOS | CER | WER | SIM | UTMOS | CER | WER | SIM | UTMOS | CER | WER | SIM | UTMOS | CER |
| **Kannada** | | | | | | | | | | | | | | | | |
| GT | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 |
| Bilingual | 57.01 | 0.630 | 3.613 | 52.16 | 4.78 | 0.697 | 3.756 | 2.02 | 4.09 | 0.700 | 3.767 | 1.45 | 4.08 | 0.700 | 3.776 | 1.67 |
| Multilingual | 49.66 | 0.609 | 3.573 | 44.33 | 3.36 | 0.707 | 3.710 | 1.06 | 3.18 | 0.708 | 3.704 | 0.99 | 3.26 | 0.708 | 3.689 | 1.01 |
| **Telugu** | | | | | | | | | | | | | | | | |
| GT | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 |
| Bilingual | 52.09 | 0.626 | 3.181 | 47.81 | 3.86 | 0.722 | 3.414 | 1.53 | 3.60 | 0.720 | 3.374 | 1.24 | 3.74 | 0.724 | 3.408 | 1.45 |
| Multilingual | 45.64 | 0.628 | 3.254 | 41.46 | 3.65 | 0.721 | 3.289 | 1.20 | 3.73 | 0.724 | 3.293 | 1.18 | 3.50 | 0.725 | 3.257 | 1.25 |
| **Tamil** | | | | | | | | | | | | | | | | |
| GT | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 |
| Bilingual | 60.46 | 0.583 | 3.564 | 54.95 | 5.66 | 0.685 | 3.750 | 2.66 | 4.11 | 0.689 | 3.797 | 1.64 | 4.33 | 0.694 | 3.783 | 1.62 |
| Multilingual | 56.75 | 0.602 | 3.440 | 50.59 | 4.11 | 0.701 | 3.636 | 1.51 | 3.49 | 0.708 | 3.635 | 1.14 | 3.35 | 0.706 | 3.635 | 1.14 |
| **Hindi** | | | | | | | | | | | | | | | | |
| GT | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 |
| Bilingual | 50.86 | 0.491 | 3.137 | 45.44 | 7.22 | 0.569 | 3.172 | 3.72 | 5.89 | 0.576 | 3.188 | 2.66 | 6.09 | 0.578 | 3.137 | 2.76 |
| Multilingual | 41.52 | 0.564 | 3.061 | 35.34 | 4.97 | 0.640 | 3.139 | 1.93 | 4.91 | 0.639 | 3.110 | 2.00 | 4.70 | 0.639 | 3.100 | 1.89 |
| **Malayalam** | | | | | | | | | | | | | | | | |
| GT | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 |
| Bilingual | 60.60 | 0.571 | 3.400 | 54.14 | 7.91 | 0.690 | 3.590 | 2.91 | 4.67 | 0.699 | 3.624 | 2.11 | 5.65 | 0.699 | 3.618 | 2.58 |
| Multilingual | 56.55 | 0.589 | 3.436 | 50.85 | 4.84 | 0.704 | 3.617 | 2.23 | 4.04 | 0.708 | 3.633 | 1.56 | 4.68 | 0.708 | 3.628 | 1.97 |
| **Punjabi** | | | | | | | | | | | | | | | | |
| GT | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 |
| Bilingual | 49.27 | 0.599 | 3.400 | 44.57 | 4.57 | 0.690 | 3.438 | 1.80 | 4.32 | 0.692 | 3.440 | 1.83 | 4.76 | 0.692 | 3.426 | 2.15 |
| Multilingual | 39.30 | 0.646 | 3.365 | 33.91 | 3.72 | 0.725 | 3.380 | 1.55 | 3.71 | 0.727 | 3.425 | 1.33 | 3.61 | 0.727 | 3.411 | 1.31 |
| **Bengali** | | | | | | | | | | | | | | | | |
| GT | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 |
| Bilingual | 52.98 | 0.571 | 3.407 | 48.31 | 4.19 | 0.661 | 3.444 | 1.68 | 4.15 | 0.663 | 3.407 | 1.81 | 4.45 | 0.663 | 3.462 | 2.15 |
| Multilingual | 47.22 | 0.595 | 3.324 | 42.60 | 4.14 | 0.687 | 3.325 | 1.54 | 4.19 | 0.687 | 3.315 | 1.51 | 4.06 | 0.687 | 3.282 | 1.42 |
| **Odia** | | | | | | | | | | | | | | | | |
| GT | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 |
| Bilingual | 50.99 | 0.599 | 3.151 | 45.84 | 5.29 | 0.687 | 3.277 | 2.32 | 5.15 | 0.689 | 3.288 | 2.17 | 5.23 | 0.699 | 3.287 | 2.39 |
| Multilingual | 46.57 | 0.606 | 3.243 | 41.51 | 4.17 | 0.691 | 3.317 | 1.50 | 3.93 | 0.708 | 3.634 | 1.44 | 3.92 | 0.695 | 3.314 | 1.44 |
| **Marathi** | | | | | | | | | | | | | | | | |
| GT | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 | 3.54 | – | 4.248 | 1.33 |
| Bilingual | 49.27 | 0.566 | 3.382 | 44.98 | 4.35 | 0.676 | 3.423 | 1.71 | 4.19 | 0.681 | 3.431 | 1.69 | 4.44 | 0.680 | 3.419 | 1.68 |
| Multilingual | 42.28 | 0.603 | 3.324 | 37.96 | 3.16 | 0.699 | 3.397 | 1.10 | 3.04 | 0.700 | 3.415 | 0.92 | 3.20 | 0.701 | 3.404 | 0.94 |

Table 4: Cross-lingual TTS evaluation (Indic Reference and English Generation) across Indian languages

ances, while generated speech is conditioned on a reference utterance. Since GT and generated samples are not speaker-aligned, we compare duration distributions instead of absolute differences. The original F5-TTS underestimates durations relative to GT due to byte-level UTF-8 encoding. The modified F5-TTS alleviates this issue by switching to character-level encoding; however, it still underestimates durations, resulting in speech that sounds too fast in setting (ii). Meanwhile, the proposed Heuristic-1 and Heuristic-2 methods produce durations that align better with the GT distribution in both settings. This indicates that incorporating language-specific duration-per character statistics during inference improves cross-lingual duration modeling and yields better alignment with ground-truth.

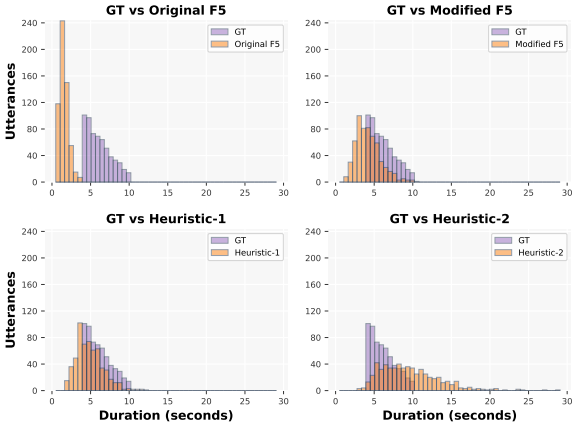



Figure 3: Comparison of ground-truth (GT) and predicted durations for Tamil reference and English speech generation.

### 5.2 Effect of $\alpha : \beta$ ratio in Heuristic-1

To investigate the sensitivity of Heuristic-1 to the interpolation coefficients $\alpha$ and $\beta$, we evaluate three ratio configurations: 6:4, 7:3, and 8:2 using the multilingual model across Telugu, Hindi, and Kannada in both cross-lingual settings. As shown in Table 5, the performance remains similar across ratio configurations, with minor improvements. For Telugu in the (i) English Reference and Indic Generation setting, the 7:3 ratio achieves the best WER/CER of 16.40/1.93, with relative improvements of 2.1%/2.5% over 6:4 and 0.2%/5.9% over 8:2, while SIM remains nearly unchanged.

In the reverse setting, (ii) Indic Reference and

| $\alpha : \beta$ | Ref-English and Gen-Indian | | | | Ref-Indian and Gen-English | | | |
|---|---|---|---|---|---|---|---|---|
| | WER | SIM | UTMOS | CER | WER | SIM | UTMOS | CER |
| Telugu | | | | | | | | |
| GT | 21.06 | - | 3.119 | 3.07 | 3.54 | - | 4.248 | 1.33 |
| 6:4 | 16.75 | 0.661 | 3.243 | 1.98 | 3.21 | 0.725 | 3.379 | 1.03 |
| 7:3 | 16.40 | 0.662 | 3.508 | 1.93 | 3.73 | 0.724 | 3.293 | 1.18 |
| 8:2 | 16.43 | 0.662 | 3.243 | 2.05 | 3.49 | 0.723 | 3.379 | 1.21 |
| Hindi | | | | | | | | |
| GT | 14.40 | - | 2.573 | 5.21 | 3.54 | - | 4.248 | 1.33 |
| 6:4 | 11.72 | 0.678 | 3.608 | 3.48 | 4.41 | 0.642 | 3.172 | 1.71 |
| 7:3 | 11.66 | 0.680 | 3.566 | 3.42 | 4.91 | 0.639 | 3.110 | 2.00 |
| 8:2 | 11.70 | 0.676 | 3.608 | 3.46 | 4.79 | 0.642 | 3.172 | 1.75 |
| Kannada | | | | | | | | |
| GT | 33.38 | - | 3.461 | 5.88 | 3.54 | - | 4.248 | 1.33 |
| 6:4 | 25.70 | 0.660 | 3.435 | 4.00 | 3.20 | 0.708 | 3.756 | 0.97 |
| 7:3 | 26.22 | 0.664 | 3.475 | 4.16 | 3.18 | 0.708 | 3.704 | 0.99 |
| 8:2 | 26.04 | 0.660 | 3.435 | 4.12 | 3.19 | 0.709 | 3.756 | 1.00 |

Table 5: Heuristic-1 based cross-lingual TTS performance across $\alpha : \beta$ duration ratios.

English Generation, the 6:4 ratio performs slightly better in some cases. Overall, the optimal ratio varies slightly with the language pair and direction, and we **recommend experimenting with different values**. We adopt $\alpha = 0.7$ and $\beta = 0.3$ for all experiments to ensure fair comparisons.

## 6 Conclusion

We propose a cross-lingual evaluation benchmark and language-aware duration strategies for alignment-free TTS. Our approach introduces Modified F5 for improved duration estimation and two language-aware duration prediction methods, Heuristic-1 and Heuristic-2, which incorporate language-specific duration-per character statistics. Extensive experiments on both bilingual and multilingual models, evaluated across nine Indian languages and two cross-lingual directions, show that the proposed heuristic methods consistently improve duration alignment. Qualitative analysis further highlights the importance of language-aware duration modeling for robust multilingual speech synthesis.

## 7 Limitations

Our approach has a few limitations that we plan to address in future work. The proposed duration estimation relies on language-level average duration-per character rates and modifies the inference process without retraining; therefore, it does not explicitly account for speaker- or utterance-level variations in speaking style and prosody. In addition, our evaluation covers nine Indian languages and English, focusing on English to Indic and Indic to English cross-lingual directions. As future work, we plan to extend the evaluation to Indic to Indic language pairs to investigate broader cross-lingual generalization. Finally, the proposed duration adjustments are heuristic rather than learned, and our comparison is primarily with the publicly available alignment-free TTS such as A-DMA/F5-TTS baselines, as implementations of some recent cross-lingual TTS methods were unavailable during our experiments.

## References


AI4Bharat. 2024. IndicConformerASR.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing, 2022*, 16(6):1505–1518.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025. F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching. In *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271, Vienna, Austria. Association for Computational Linguistics.

Jeongsoo Choi, Zhikang Niu, Ji-Hoon Kim, Chunhui Wang, Joon Son Chung, and Xie Chen. 2025. Accelerating Diffusion-based Text-to-Speech Model Trainingwith Dual Modality Alignment. In *Proc. Interspeech*, pages 3459–3463.

Speech Technology Consortium, Hema A Murthy, and S Umesh. 2023. Indic TTS: A Text-to-Speech Database for Indian Languages.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, Yanqing Liu, Sheng Zhao, and Naoyuki Kanda. 2024. E2 TTS:

Embarrassingly Easy Fully Non-Autoregressive Zero-Shot TTS. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 682–689.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist Temporal Classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. International Conference on Machine Learning (ICML)*, page 369–376.

Qingyu Liu, Yushen Chen, Zhikang Niu, Chunhui Wang, Yunting Yang, Bowen Zhang, Jian Zhao, Pengcheng Zhu, Kai Yu, and Xie Chen. 2026. Cross-Lingual F5-TTS: Towards Language-Agnostic Voice Cloning and Speech Synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17362–17366.

Vasista Sai Lodagala, Sreyan Ghosh, and S. Umesh. 2023. Data2vec-Aqc: Search for the Right Teaching Assistant in the Teacher-Student Training Setup. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proc. International Conference on Learning Representations (ICLR)*.

Vikentii Pankov, Artem Gribul, Oktai Tatanov, Vladislav Proskurov, Yuliya Korotkova, Darima Mylzenova, and Dmitrii Vypirailenko. 2026. PFluxTTS: Hybrid Flow-Matching TTS with Robust Cross-Lingual Voice Cloning and Inference-Time Model Fusion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17612–17616.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *J. Mach. Learn. Res.*, 25(1).

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. International Conference on Machine Learning (ICML)*, pages 28492–28518.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge. In *Proc. Interspeech*, pages 4521–4525.

Ashwin Sankar, Srija Anand, Praveen Srinivasa Varadhan, Sherry Thomas, Mehak Singal, Shridhar Kumar, Deovrat Mehendale, Aditi Krishana, Giri Raju, and Mitesh Khapra. 2024. Indicvoices-r: Unlocking a massive multilingual multi-speaker speech corpus for scaling indian tts. *NeurIPS 2024 Datasets and Benchmarks*.

Hubert Siuzdak. 2024. Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis. In *Proc. International Conference on Learning Representations (ICLR)*, pages 25719–25733.

Praveen Srinivasa Varadhan, Ashwin Sankar, Giri Raju, and Mitesh M Khapra. 2024. Rasa: Building Expressive Speech Synthesis Systems for Indian Languages in Low-resource Settings. In *Proc. Interspeech*, pages 1830–1834.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. 2023. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16133–16142.