

Tokenizer Transplantation: Mitigating Autoregressive Collapse in Edge-Efficient Bengali ASR

Reducing tokenizer fertility by 85.8% restores stable Bengali decoding

Sanjid Hasan¹ Md. Abdur Rahman²

July 3, 2026

¹Khulna University of Engineering & Technology

²Military Institute of Science and Technology

The Goal

Deploy lightweight ASR models (e.g., Moonshine) on edge devices for fast, offline transcription.

The Problem

Optimized architectures fail on morphologically rich, non-Latin languages. English-centric tokenizers fracture words into massive byte chains.

Autoregressive Collapse

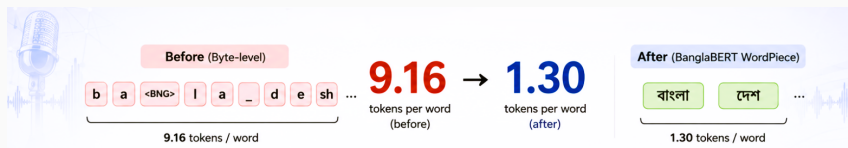
High-fertility token chains exponentially increase inference sequence length, causing the decoder to drift and collapse.

$$\phi = \frac{\text{Total Tokens}}{\text{Total Words}}$$

The Solution: Solving Tokenizer Fertility (Φ)

By replacing the original byte-fallback tokenizer with a native-script BanglaBERT WordPiece tokenizer, we compress the representation space.

Tokenizer Architecture	Fertility (Φ)	Sequence Length Drop
Original Moonshine (Byte-Fallback)	9.16	-
Transplanted (BUET BanglaBERT)	1.30	85.8% Reduction

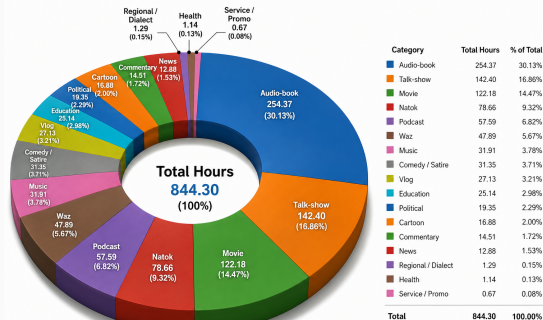


Key Result

Decreasing sequence length by 85.8% entirely mitigates decoding instability.

The Lipi-Ghor Dataset

- **Scale:** 844.3 hours, 15 real-world content categories.
- **Diversity:** Audio-book (30%) and Talk-show (17%) lead, but Movie, Natok, Podcast, and News all contribute meaningfully.
- **Long Tail:** Regional/Dialect, Health, and Service/Promo push the model to generalize beyond high-resource genres.



The Transplantation Pipeline

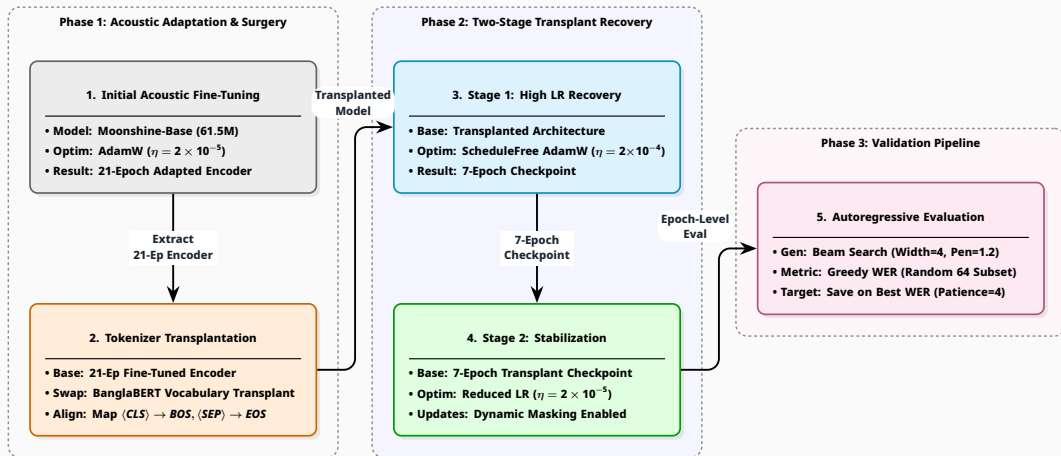


Figure 1: Overview of the Tokenizer Transplantation Methodology.

Vocabulary Surgery (Phase 2)

```
# Aligning BanglaBERT to Moonshine
model.resize_token_embeddings(
    len(banglabert_tokenizer)
)

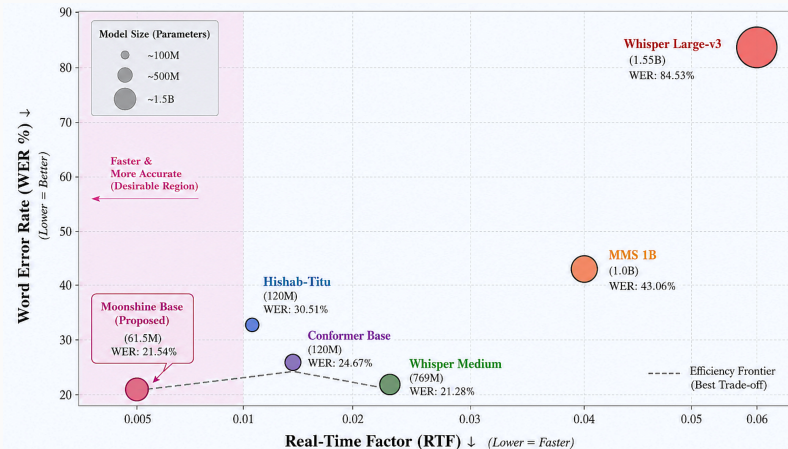
model.config.bos_token_id = \
    banglabert_tokenizer.cls_token_id

model.config.eos_token_id = \
    banglabert_tokenizer.sep_token_id
```

Stabilization Metrics (Phase 3)

- **Training Loss Drop:** 2.5 $\rightarrow$ 0.12
- **WER Stabilization:** Native word-level tokens massively reduce exposure bias compared to teacher-forced training.
- **Metric Extraction:** Greedy WER subset sampling dynamically verifies autoregressive generation.

Efficiency Frontier of Bengali ASR Models



Key Takeaway: The proposed Moonshine model achieves the optimal trade-off between size, speed, and accuracy.

Model Architecture	Params	RTF	WER (%)	CER (%)
Seamless M4T-v2 (Zero-Shot)	~ 2.3B	-	66.71	45.54
OpenAI Whisper large-v3	~ 1.55B	-	84.53	72.92
Faster Whisper Medium	~ 769M	0.0190	21.28	11.18
Conformer Baseline (Fine-Tuned)	~ 120M	0.0120	24.67	15.56
Moonshine Base (Transplanted)	~ 61.5M	0.0053	21.54	10.79

- **Accuracy:** Achieves the lowest Character Error Rate globally.
- **Speed:** Natively operates **3.5× faster** than highly engineered Whisper CTranslate2 pipelines, with a 10× parameter reduction.

Limitations

- Trained on **read speech**; degrades on spontaneous, conversational, and disfluent Bengali
- Limited **code-switching** support for Bengali-English mixed speech
- Performance drops on **phone-recorded audio** due to codec compression artifacts
- RTF benchmarks are hardware-specific;

Future Works

- Collect and train on **spontaneous conversational Bengali** to bridge read-to-spontaneous gap
- Morphologically-aware tokenization respecting Bengali segmentation
- **Codec augmentation** during training for telephony robustness
- Evaluation on **telephony speech** as a voice agent deployment benchmark
- Multi-device RTF benchmarking across low-end hardware (e.g., Raspberry Pi)

- **Vocabulary Surgery is Essential:** Cross-script ASR transfer requires native morphological alignment.
- **Mitigating Exposure Bias:** High tokenizer fertility makes teacher-forcing misleading; sequence compression is required for edge models.
- **A Scalable Blueprint:** Provides a reproducible framework for democratizing lightweight ASR systems in low-resource languages.

Thank You

Project Page



[Sanjidh090/moonshine-base-bn](https://github.com/Sanjidh090/moonshine-base-bn)