JIANTS

The Intelligence of Systems

Nihil: Technical Documentation Text-to-Speech, Speech-to-Text, and Many-to-Many MT5 Toucan Fine-Tuning

The Young JIANTS[1]

[1]JIANTS Research Lab

July 7, 2025

Abstract

This document details the end-to-end design and implementation of Nihil's core services: **Text-to-Speech (TTS)**, **Speech-to-Text (STT)**, and **Many-to-Many Machine Translation** via fine-tuning the MT5-based Toucan model. We cover data collection, preprocessing, model architectures, training protocols, evaluation, and deployment considerations, with rigorous citations to foundational and recent works.

Contents

1 Data Collection and Preparation

1.1 TTS Dataset

We assembled a Basaa speech corpus of ≈15 hours of high-quality recordings from native speakers, sampled at 24 kHz. Text transcripts were normalized (lowercasing, punctuation removal) and phonemized via a custom Basaa grapheme-to-phoneme converter.

1.2 STT Dataset

The STT dataset reuses the TTS audio with manual transcripts, plus few additional spontaneous speech from WhatsApp voice notes. We applied SpecAugment (time warping, frequency masking) to increase robustness [3].

1.3 MT Parallel Corpora

Our machine translation corpora comprise:

- Basaa–French: $\approxeq 80kalignedsentencesharvestedfromscrapedcommunitycontentandvolunteer$

- Basaa–English: $\approxeq 40ksentencesviaback-translationandbilingualvolunteers.$

- French–English: public Europarl subset for transfer learning.

We split each into 90% train, 5% validation, and 5% test.

2 Preprocessing Pipelines

2.1 Acoustic Features

Audio is resampled to 16 kHz and normalized; mel-spectrograms are computed with 80 bins, 50 ms window, 12.5 ms hop. For TTS, mel targets are aligned with text via a monotonic attention layer (Glow-TTS) [7].

2.2 Tokenization and Prefixing

We use the MT5 tokenizer [8]. Every translation input is prepended with a prefix token: “bas: ”, “fra: ”, or “eng: ” to guide decoding direction [9].

3 Model Architectures

3.1 Text-to-Speech

We employ a two-stage TTS model:

1. **FastSpeech2** for mel-spectrogram prediction (encoder, duration predictor, decoder) [6].

2. **HiFi-GAN** vocoder to convert mel-spectrograms into waveform audio [5].

   We also explore VITS (yourTTS) for end-to-end Speech Synthesis.

Training minimizes L1 and adversarial losses:

$$\mathcal{L}_{\text{TTS}} = \|M - \hat{M}\|_1 + \lambda_{\text{GAN}} \sum_k \mathcal{L}_{\text{GAN}}^k. \tag{1}$$

3.2 Speech-to-Text

We fine-tune **wav2vec 2.0 Large** pre-trained on LibriSpeech, with CTC head for Basaa transcription. The CTC objective:

$$\mathcal{L}_{\text{CTC}} = -\log p(\mathbf{l} \mid \mathbf{x}), \tag{2}$$

where $\mathbf{x}$ is the latent audio representation and $\mathbf{l}$ the token sequence [3].

3.3 Many-to-Many Translation: MT5 Toucan

The **Toucan-1.2B** model is a T5-based encoder-decoder with 1.23B parameters, pre-trained on multilingual text. We apply **LoRA** adapters to the query and value projection matrices in each Transformer block, injecting low-rank updates (rank $r = 8$, `lora_alpha`=16) to reduce trainable parameters to <1% of total [10].

4 Training Setup

All experiments ran on NVIDIA T4 GPUs.

4.1 TTS and STT

Hyperparameters:

- Optimizer: AdamW, LR 1e-4 (TTS), 2e-5 (STT)
- Batch size: 16 for FastSpeech2, 8 for HiFi-GAN; 32 sequences for wav2vec2.0
- Epochs: 1000 (TTS), 30 (STT)
- Gradient clipping at 1.0

4.2 MT5 Fine-Tuning

Hyperparameters:

- LoRA rank $r = 8$, dropout 0.05
- Optimizer: paged AdamW 8-bit, LR 5e-5
- Batch size: 32, gradient accumulation 4 (effective 128)
- Epochs: 5, max length 512, warm-up steps 10% total

We use ‘accelerate‘ for distributed offloading and mixed precision.

5 Evaluation Metrics and Results

5.1 TTS

Du to time constraint, we report Mean Opinion Scores (MOS) on 1 listener ratings:

- VITS: mel$_spec = 26.43loss_duration = 2.69loss_disc = 2.64$

5.2 STT

Word Error Rate (WER) on the Basaa test set:

- whisper-tiny: WER = 5.5%

6 Discussion

Adapter-based fine-tuning achieves nearly identical translation quality with <1% of parameters updated, slashing GPU memory usage from 15GiB to 4GiB. TTS pipeline latency is 120ms per utterance; STT throughput is 10x real-time.

7 Future Work

We plan to:

- Extend to additional Cameroonian languages (e.g. Duala, Fulfulde)
- Integrate end-to-end speech translation (speech-to-speech) using cascade and direct models [11]
- Deploy on edge devices via model pruning and quantization

References

[1] A. Elmadany et al., "Toucan: Many-to-Many Translation for 150 African Language Pairs," Findings of ACL 2024.

[2] S. N. Mehdi, "Coqui TTS: Deep Dive Into an Open-Source TTS Framework," Medium 2025.

[3] A. Baevski et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," NeurIPS 2020.

[4] A. Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," arXiv:2212.04356, 2022.

[5] J. Kim et al., "HiFi-GAN: Generative Adversarial Networks for Efficient and High-Fidelity Speech Synthesis," arXiv:2010.05646, 2020.

[6] Y. Ren et al., "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," ICASSP 2021.

[7] J. Kim et al., "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search," arXiv:2005.11129, 2020.

[8] L. Xue et al., "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," arXiv:2010.11934, 2020.

[9] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," JMLR 2020.

[10] E. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," arXiv:2106.09685, 2021.

[11] Y. Jia et al., "Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model," ICASSP 2019.