

ALIGN2SPEAK: IMPROVING TTS FOR LOW RESOURCE LANGUAGES VIA ASR-GUIDED ONLINE PREFERENCE OPTIMIZATION

*Shehzeen Hussain, Paarth Neekhara, Xuesong Yang, Edresson Casanova, Subhankar Ghosh
Roy Fejgin, Ryan Langman, Mikyas Desta, Leili Tavabi, Jason Li*

NVIDIA Corporation, USA

ABSTRACT

Developing high-quality text-to-speech (TTS) systems for low-resource languages is challenging due to the scarcity of paired text and speech data. In contrast, automatic speech recognition (ASR) models for such languages are often more accessible, owing to large-scale multilingual pre-training efforts. We propose a framework based on Group Relative Policy Optimization (GRPO) to adapt an autoregressive, multilingual TTS model to new languages. Our method first establishes a language-agnostic foundation for TTS synthesis by training a multilingual baseline with International Phonetic Alphabet (IPA) tokens. Next, we fine-tune this model on limited paired data of the new languages to capture the target language’s prosodic features. Finally, we apply GRPO to optimize the model using only unpaired text and speaker prompts, guided by a multi-objective reward from pretrained ASR, speaker verification, and audio quality estimation models. Experiments demonstrate that this pipeline produces intelligible and speaker-consistent speech in low-resource languages, substantially outperforming finetuning alone. Furthermore, our GRPO-based framework also improves TTS performance in high-resource languages, surpassing offline alignment methods such as Direct Preference Optimization (DPO) yielding superior intelligibility, speaker similarity, and audio quality.¹

Index Terms— Text-to-Speech, Preference Optimization, Speech Synthesis, Low-resource language, Multilingual

1. INTRODUCTION

Recent advances in neural text-to-speech (TTS) synthesis have led to highly natural and intelligible systems, enabled by large-scale datasets and powerful generative models such as autoregressive transformers and diffusion models. Particularly, large language model (LLM) based TTS approaches have become increasingly common, leveraging the strong representation learning capabilities of LLMs to improve prosody, context adherence and naturalness of generated speech [1, 2, 3, 4, 5, 6]. However, the development of TTS for

low-resource languages remains a pressing challenge. Collecting high-quality paired text–speech corpora is expensive and often infeasible for many underrepresented languages. This data scarcity creates a substantial performance gap compared to high-resource languages [7, 8].

In contrast, automatic speech recognition (ASR) systems for low-resource languages have benefited significantly from large-scale multilingual pretraining and transfer learning. Past work such as Whisper [9] and Common Voice [10] have made robust ASR models widely available, even for languages with limited training resources. This asymmetry suggests an opportunity: while curating extensive TTS data is difficult, one may still obtain a reliable ASR model for a given low-resource language, which can serve as a feedback signal for improving TTS models.

A parallel line of research in natural language processing has shown that reinforcement learning (RL) can effectively align generative models with task-specific objectives. Reinforcement learning from human feedback (RLHF) has emerged as a central paradigm for large language models (LLMs), where an external “judge” (human or automated) provides preference signals that guide the model towards producing more useful outputs. Recently, similar ideas have been explored in speech generation. Offline preference optimization methods such as Direct Preference Optimization (DPO) [11] have been applied to TTS synthesis [6, 12], but these approaches lack the dynamic feedback loop provided by online RL methods.

In this work, we investigate whether *ASR models can be leveraged as a reward function for learning TTS in low-resource settings*. Specifically, we introduce a framework based on *Group Relative Policy Optimization (GRPO)* [13], where an autoregressive TTS model trained across multiple languages is adapted to new languages using reinforcement learning. Our method proceeds in three stages: 1) We train a baseline autoregressive TTS model on a set of source languages, using the International Phonetic Alphabet (IPA) as the text representation. 2) The baseline model is adapted to a new target language using a small amount of available paired data, capturing its phonetic and prosodic patterns. 3) We generate text and speaker prompts in the target language

¹Project Webpage: grpotts.github.io

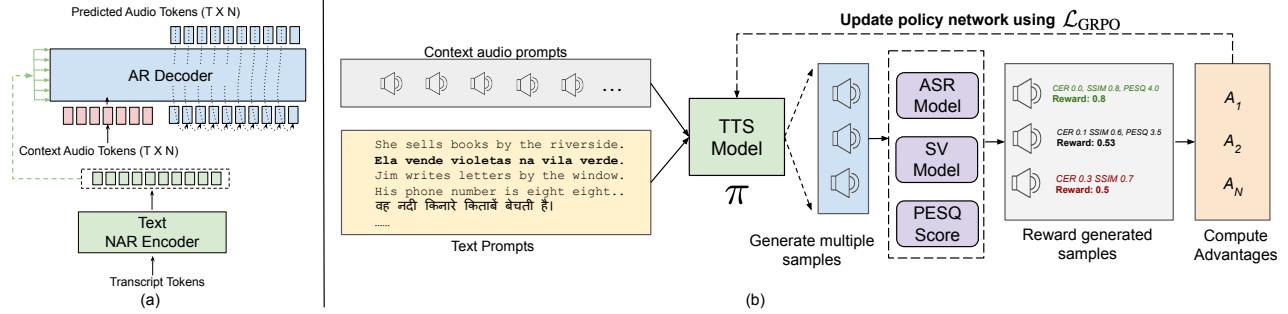


Fig. 1. (a) Encoder-Decoder TTS Model Architecture, (b) Online Preference Alignment for TTS: Model generates multiple outputs for a text and context audio prompt, which are then rewarded using ASR, SV models and PESQ Scores and optimized using GRPO.

and optimize the TTS model across multiple objectives using GRPO, guided by reward metrics from ASR, speaker verification, and PESQ [14] estimation models to improve intelligibility, speaker consistency, and perceptual quality.

Our experiments show that this pipeline of fine-tuning followed by GRPO with automatic reward signals, yields intelligible and speaker-consistent TTS for new low-resource languages, significantly outperforming fine-tuning alone. Furthermore, we show that GRPO provides strong preference alignment even in *high-resource (seen) languages*, outperforming offline alignment methods such as DPO. To the best of our knowledge, this is the first work to extend GRPO to token-based TTS synthesis with multi-objective rewards. These findings suggest that online reinforcement learning offers a general and effective mechanism for aligning TTS models with perceptual and linguistic objectives.

2. METHODOLOGY

This section outlines our proposed framework for adapting a multilingual text-to-speech (TTS) model to a low-resource language using the three-step framework described above.

2.1. Baseline TTS Model

Our synthesis framework is based on Koel-TTS [6], an autoregressive encoder-decoder TTS model that operates on the low-frame rate (21.5 FPS) audio codec representation given by NanoCodec [15]. Koel-TTS comprises an autoregressive (AR) transformer decoder conditioned on text encodings from a non-autoregressive (NAR) transformer encoder using cross-attention (Figure 1a). Specifically, given an input text sequence and acoustic codes of a reference speaker audio, the TTS model autoregressively generates the acoustic codes corresponding to the input text and speaker. To build a foundation model capable of generalizing to new languages, we first pretrain the Koel-TTS model on a multilingual dataset comprising six languages — English, Dutch, Italian, Spanish, French, and German. A key aspect of our pretraining strategy is the use of the International Phonetic Alphabet (IPA) for text representation [16]. IPA provides a universal, standardized representation of speech sounds. The IPA text is tokenized using a byte-level tokenizer, which naturally handles the full range of IPA symbols with just 256 tokens. This phonetic-based approach enables the model to learn a mapping from

universal acoustic units to speech, which is crucial for zero-shot or few-shot adaptation to a new language. The baseline model is trained on triplets of *context audio*, *IPA transcript*, and *target audio* and optimized using only the next-frame prediction loss with a parallel head [2, 6].

2.2. Fine-Tuning on Low-Resource Languages

After pretraining, we adapt the baseline model to a target low-resource language through supervised fine-tuning. This stage uses a limited amount of paired data (i.e., a few hours of speech and corresponding transcripts) from the target language. To prevent catastrophic forgetting of the knowledge acquired during pretraining, we employ a mixed-data training strategy. The fine-tuning dataset is a combination of the original pretraining data and the new low-resource data. We upsample the low-resource data, ensuring that each training batch contains a high proportion of samples from the target language. This approach allows the model to learn the specific phonetic and prosodic characteristics of the new language while retaining its general synthesis capabilities for languages seen during baseline model training. The resulting fine-tuned model serves as the reference policy for the subsequent reinforcement learning stage.

2.3. Refinement with GRPO

While fine-tuning adapts the model to the target language, its quality is often limited by the small data size. To further enhance intelligibility and speaker similarity, we adapt Group Relative Policy Optimization (GRPO) [13], an online reinforcement learning algorithm designed for policy alignment. This procedure is depicted in Figure 1(b).

Prompt Construction: The GRPO training process requires prompts from which the model generates speech. Each prompt consists of a speaker reference audio and a text transcript to be synthesized. For low-resource scenarios, a significant advantage of this setup is that the content of the speaker audio prompt does not need to match the text transcript. This decouples the need for paired data, allowing us to leverage readily available, non-parallel speech and text data for creating a set of prompts. Specifically, we include 15k text and speaker reference audio pairs as the prompts for each language in our dataset. During GRPO, we construct prompts for all languages, including those seen during baseline model training.

Reward Function: For each generated audio sample, we compute a reward signal R that quantifies its preference. The total reward is a weighted combination of three components: transcription accuracy, measured by Character Error Rate (CER), speaker similarity, measured by a cosine similarity score (SSIM) and perceptual speech quality, measured by a neural PESQ estimator:

$$R = w_{cer} \cdot R_{cer} + w_{ssim} \cdot R_{ssim} + w_{pesq} \cdot R_{pesq} \quad (1)$$

where we set $w_{cer} = w_{ssim} = 0.45$ and $w_{pesq} = 0.1$.

To ensure the reward signal is well-scaled and interpretable, we map the raw CER and SSIM scores to a normalized reward value between 0 and 1 using a piecewise linear function. This function is defined by three key points: the worst possible score maps to 0, the best possible score maps to 1, and the average score observed from the baseline model maps to 0.5. The reward for intermediate values is linearly interpolated. To optimize for PESQ, we normalize the raw PESQ score between 0 and 1 by setting $R_{pesq} = PESQ/4.5$. The maximum CER and minimum PESQ values are clipped at 1 and 0 respectively before normalization. For measuring CER, we use the multilingual Whisper Large V3 model [9]. SSIM is computed between the context audio and the generated audio using the Titanet-large model [17]. PESQ is estimated using the reference-free neural PESQ estimator from torchaudio [14]. The precise reward normalization functions used in our framework are shown in Figure 2.

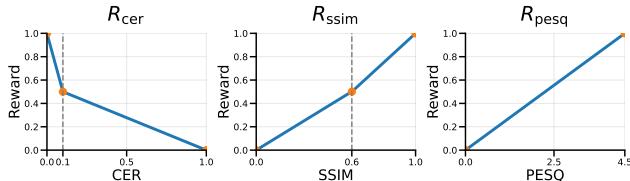


Fig. 2. Reward functions used to map CER, SSIM and PESQ values to normalized rewards between 0 and 1.

GRPO Objective: GRPO optimizes the policy π_θ (our TTS model) to maximize the expected reward from generated samples [13]. For a given text and context audio input $x = (x_{text}, x_{audio})$, the model’s response distribution $\pi(y|x)$ encompasses a range of potential outputs y with varying reward values. For each prompt x_i , we sample K responses $\{y_{i,k}\}_{k=1}^K \sim \pi_{\theta_{old}}$ and obtain rewards $r_{i,k}$. A group baseline is defined as

$$\mu_i = \frac{1}{K} \sum_{k=1}^K r_{i,k}, \quad A_{i,k} \triangleq r_{i,k} - \mu_i. \quad (2)$$

The training objective of GRPO directly maximizes the log-likelihood of responses weighted by these centered advantages $A_{i,k}$:

$$\mathcal{L}_{GRPO}(\theta) = \frac{1}{MK} \sum_{i=1}^M \sum_{k=1}^K A_{i,k} \log \pi_\theta(y_{i,k} | x_i). \quad (3)$$

We omit the KL penalty to the reference model proposed in the original GRPO formulation [13] since we find that relying solely on the group-relative advantages stabilizes learning and speeds up training.

3. EXPERIMENTS

3.1. Experiment Setup

Baseline Model Training: Our baseline TTS model is trained on English and 5 European languages. For English, the training corpus consists of approximately 18k hours of TTS data from the following datasets: *train-clean-360* and *train-clean-100* subsets of LibriTTS [18], HiFiTTS [19], a 17k-hour subset of the LibriVox MLS dataset [20] and a proprietary, 2-speaker, 63-hour dataset. For European languages, we use 5 languages from the CML dataset [21] that contains 1,562 hours of German, 642 hours of Dutch, 476 hours of Spanish, 283 hours of French, 131 hours of Italian speech data. To enable Classifier Free Guidance (CFG) during inference, we follow the methodology proposed in Koel-TTS [6] to randomly drop out conditioning input during training.

The baseline model is trained on 32 NVIDIA H100 GPUs using a global batch size of 512, optimized using Adam optimizer with an initial learning rate of $1e-4$. The learning rate is annealed every 1k training steps using an exponential decay factor of 0.998. Training for the model converges in around 300k iterations.

Finetuning for low-resource languages: We consider three low-resource languages for our experiments: Polish and Portuguese from the CML dataset [21]; and Hindi from the IndicTTS dataset [22]. For our experiments, we create random subsets comprising 30 minutes, 1 hour and 5 hours of data from each of the three low-resource languages for the fine-tuning stage. For fine-tuning, we mix the data from the new languages with the 21k hours of multilingual data described above. To ensure sufficient exposure to the new languages during fine-tuning, we upsample their data by a factor of 5 compared to their natural proportion in the multilingual corpus. Finetuning is performed at a fixed learning rate of $1e-5$ for a maximum of 30k steps using a global batch size of 512. We choose the checkpoint with the lowest validation loss on a validation set containing only the new languages.

Preference Alignment using GRPO: GRPO is performed for a maximum of 2k mini-batch iterations where each mini-batch contains 64 text and context audio prompts, at a fixed learning rate of $2e-7$. For each prompt, we generate 12 audios using multinomial sampling at temperature 0.7. During GRPO, CFG is used for inference with a probability of 0.5 and CFG scale of 2.5, to improve model alignment both with and without CFG. We validate the model every 50 iterations on a validation set containing the new languages, and choose the checkpoint with the best R_{CER} on this validation set.

Evaluation: We evaluate the generated speech along three

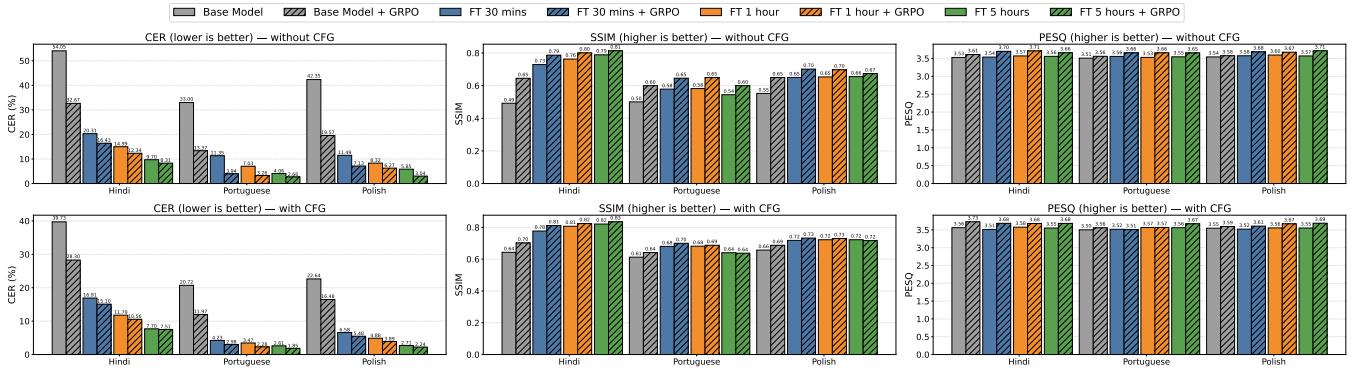


Fig. 3. Intelligibility (CER), Speaker similarity (SSIM) and perceptual quality (PESQ) evaluation of the baseline model and the low-resource adaptations for new languages. Post GRPO, we observe considerable improvements in CER across both baseline and fine-tuned models.

primary dimensions: intelligibility, speaker similarity, and audio quality. Intelligibility is quantified using ASR-derived character error rate (CER) using whisper-large-v3 [9]. Speaker similarity is measured by computing the cosine similarity (SSIM) between embeddings of synthesized utterances and their corresponding context audio. Embeddings are obtained with Titanet-Small [17], distinct from the Titanet-Large model applied during preference alignment. Audio quality is measured using a reference-free neural PESQ estimator [14] and naturalness is evaluated using Squim-MOS [14].

3.2. Offline vs Online Preference Optimization

To provide a direct comparison with the DPO approach introduced in Koel-TTS, we apply GRPO to the *Koel-TTS 380M* English TTS model [6], using the same prompt dataset. For consistency, we exclude the PESQ reward term, ensuring the setup is directly aligned with the conditions reported for DPO. Our results show that GRPO consistently achieves higher intelligibility and speaker similarity scores while maintaining comparable naturalness. This demonstrates that online preference optimization, which leverages continuous feedback and iterative improvement, is more effective than offline alignment techniques such as DPO.

Table 1. Comparison of Offline vs Online Preference Optimization for English TTS Synthesis (with and without CFG). Results reported with 95% confidence intervals averaged across 5 inference runs.

Model	CFG	CER (\downarrow)	SSIM (\uparrow)	Squim-MOS (\uparrow)
Base	No	2.68 ± 1.13	0.637 ± 0.008	4.35 ± 0.02
Base + DPO	No	0.89 ± 0.15	0.667 ± 0.003	4.40 ± 0.01
Base + GRPO	No	0.56 ± 0.24	0.759 ± 0.002	4.39 ± 0.01
Base	Yes	0.57 ± 0.11	0.720 ± 0.004	4.41 ± 0.01
Base + DPO	Yes	0.55 ± 0.10	0.729 ± 0.003	4.41 ± 0.01
Base + GRPO	Yes	0.53 ± 0.16	0.783 ± 0.005	4.40 ± 0.01

3.3. TTS Synthesis for Low Resource Languages

We present the performance of the baseline, fine-tuned, and preference-aligned models on low-resource languages in Figure 3. The baseline model (solid gray), which was not trained on these languages, shows limited generalizability despite

IPA tokenization, evident from its high CER scores. Interestingly, applying our multi-objective GRPO directly on this baseline (dashed gray) substantially reduces CER and boosts SSIM, even without access to paired text–speech data in the target language. This suggests that preference alignment alone can bridge part of the gap in learning TTS for new languages.

Fine-tuning on small amounts of target language data further amplifies performance gains. Models trained with just 30 minutes of supervision already outperform the baseline by a wide margin, with steady improvements as the fine-tuning budget increases to 1 hour and 5 hours. Notably, aligning fine-tuned TTS checkpoints with the proposed GRPO objective consistently lowers CER even further, highlighting the complementary benefits of preference alignment. For example, with only 30 minutes of Portuguese data, CER drops from 33.00% in the baseline to 3.94% after fine-tuning and GRPO, which is a reduction of more than 8×. Similar trends hold across Hindi and Polish, underscoring the robustness of this approach across typologically diverse languages.

Beyond intelligibility, GRPO also enhances SSIM and PESQ scores. GRPO benefits not only unseen languages but also those seen during baseline model training as shown in Table 1 (see project webpage for more details). Overall, combining small-scale fine-tuning with preference alignment enables strong, scalable adaptation to new languages with minimal data.

4. CONCLUSION

In this work, we introduce a GRPO-based framework for adapting multilingual TTS generation models to low-resource languages by leveraging a pretrained ASR model for the new languages. When combining supervised fine-tuning on limited paired data with online preference optimization using multi-objective rewards, our approach delivers consistent gains in intelligibility, speaker similarity, and perceptual quality. Beyond low-resource adaptation, we also demonstrate that GRPO guided by automatic reward functions for TTS, provides stronger preference alignment than offline methods, even for widely used, data-rich languages, highlighting its broad utility for TTS synthesis.

5. REFERENCES

- [1] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, et al., “AudioLM: a language modeling approach to audio generation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.
- [2] Paarth Neekhara, Shehzeen Hussain, Subhankar Ghosh, Jason Li, Rafael Valle, Rohan Badlani, and Boris Ginsburg, “Improving robustness of LLM-based speech synthesis by learning monotonic alignment,” *INTERSPEECH*, 2024.
- [3] Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Haohan Guo, Xuankai Chang, et al., “UniAudio: Towards universal audio generation with large language models,” in *ICML*. PMLR, 2024.
- [4] Yakun Song, Zhuo Chen, Xiaofei Wang, Ziyang Ma, and Xie Chen, “ELLA-V: Stable neural codec language modeling with alignment-guided sequence reordering,” in *AAAI*, 2025.
- [5] Xiaofei Wang, Manthan Thakker, Zhuo Chen, Naoyuki Kanda, Sefik Emre Eskimez, Sanyuan Chen, et al., “SpeechX: Neural codec language model as a versatile speech transformer,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [6] Shehzeen Hussain, Paarth Neekhara, Xuesong Yang, Edresson Casanova, Subhankar Ghosh, Mikyas T Desta, Roy Fejgin, Rafael Valle, and Jason Li, “Koel-TTS: Enhancing LLM based speech generation with preference alignment and classifier free guidance,” in *EMNLP*, 2025.
- [7] Yuan-Jui Chen, Tao Tu, Cheng-chieh Yeh, and Hung-Yi Lee, “End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning,” in *INTERSPEECH*, 2019.
- [8] Florian Lux, Julia Koch, and Ngoc Thang Vu, “Low-resource multilingual and zero-shot multispeaker TTS,” *arXiv preprint arXiv:2210.12223*, 2022.
- [9] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” in *ICML*. PMLR, 2023.
- [10] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, “Common Voice: A massively-multilingual speech corpus,” in *Proc. of the Twelfth Language Resources and Evaluation Conference*, 2020.
- [11] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn, “Direct preference optimization: Your language model is secretly a reward model,” *NeurIPS*, 2023.
- [12] Dong Zhang, Zhaowei Li, Shimin Li, Xin Zhang, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu, “SpeechAlign: Aligning speech generation to human preferences,” *NeurIPS*, 2024.
- [13] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al., “Deepseekmath: Pushing the limits of mathematical reasoning in open language models,” *arXiv:2402.03300*, 2024.
- [14] Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, et al., “Torchaudio-Squim: Reference-less speech quality and intelligibility measures in torchaudio,” in *ICASSP*. IEEE, 2023.
- [15] Edresson Casanova, Paarth Neekhara, Ryan Langman, Shehzeen Hussain, Subhankar Ghosh, et al., “NanoCodec: Towards high-quality ultra fast speech LLM inference,” in *INTERSPEECH*, 2025.
- [16] International Phonetic Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, Cambridge University Press, 1999.
- [17] Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg, “Titanet: Neural model for speaker representation with 1D depth-wise separable convolutions and global context,” in *ICASSP*. IEEE, 2022.
- [18] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, “LibriTTS: A corpus derived from LibriSpeech for text-to-speech,” *INTERSPEECH*, 2019.
- [19] Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang, “Hi-fi multi-speaker English TTS dataset,” in *INTERSPEECH*, 2021.
- [20] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert, “MLS: A large-scale multilingual dataset for speech research,” in *INTERSPEECH*, 2020.
- [21] Frederico S Oliveira, Edresson Casanova, Arnaldo Cândido Junior, Anderson S Soares, and Arlindo R Galvão Filho, “CML-TTS: A multilingual dataset for speech synthesis in low-resource languages,” in *Proc. Text, Speech, and Dialogue*. Springer, 2023.
- [22] Arun Baby, Anju Leela Thomas, NL Nishanthi, TTS Consortium, et al., “Resources for Indian languages,” in *Proc. of Text, Speech and Dialogue*, 2016.