GANDAKI UNIVERSITY
BIT PROGRAM
Gyankunja-32, Pokhara

A

Project Proposal on

Non-Autoregressive Transformer Text-To-Speech (TTS) Model For Low-Resource Nepali: Balancing Efficiency and Prosody

Submitted By:
Nirajan Dhakal

Submitted To:
Suresh Raj Dhakal
BIT Program
2024

Non-Autoregressive Transformer Text-To-Speech (TTS) Model For Low-Resource Nepali: Balancing Efficiency and Prosody

Abstract

This research proposes the development of an efficient Text-to-Speech (TTS) system for Nepali, addressing challenges of limited data, dialect diversity, and tonal complexity. Unlike resource-intensive autoregressive models common in high-resource languages, this project will employ non-autoregressive transformer architectures (e.g., FastSpeech, VITS) for computational efficiency. Cross-lingual transfer learning from Hindi, a linguistically related language, will enhance performance. Hierarchical prosody modeling will address tonal variations in Nepali phonetics. To overcome data scarcity, community-driven data curation will gather authentic text, ensuring cultural relevance. The primary objective is to develop a scalable, open-source TTS tool, bridging the digital divide and enhancing technology access in sectors such as education and accessibility. This research aims to achieve a balance between computational efficiency and prosodic richness, contributing to low-resource language TTS. The outcome will be a functional system deployable across various platforms, fostering inclusivity and linguistic diversity. This project innovates technically and through community engagement, using neural architectures to create a high-quality, efficient state-of-the-art TTS model, empowering Nepali speakers by making technology more relevant to their needs.

Introduction

Nepali, spoken by over 17 million people globally, stands at a critical juncture in terms of accessible speech technology. Despite its rich linguistic heritage and cultural significance, there is a notable gap in the availability of robust text-to-speech (TTS) systems tailored to this low-resource language. This gap can largely be attributed to three primary, intertwined challenges, with the linguistic and sociocultural context making Nepali a compelling candidate for TTS research. Traditional TTS systems for high-resource languages, like English, predominantly rely on autoregressive models, which, while achieving impressive naturalness, are computationally expensive and resource-intensive. These systems prove less viable for low-resource languages like Nepali, requiring alternative approaches for both efficiency and high-quality output.

As a tonal language, subtle changes in pitch can alter word meanings entirely. For example, the distinction between "kān" (ear) and "kăn" (mine) hinges on pitch differentiation, necessitating precise prosody modeling. Additionally, the language exhibits significant dialect diversity, with Eastern, Western, and Central dialects differing in pronunciation, vocabulary, and intonation. These variations complicate the creation of a unified TTS system.

Beyond linguistic complexity, digital inequity exacerbates the problem: limited access to Natural Language Processing (NLP) tools hinders educational, healthcare, and economic opportunities for Nepali speakers. A robust TTS system could democratize access to technology and preserve linguistic heritage in the face of globalization.

First, low-resource constraints severely limit the availability of high-quality text-to-speech datasets necessary for training state-of-the-art models. These constraints are compounded by the fact that Nepali is a low-resource language with limited annotated data available for speech synthesis tasks. Consequently, existing TTS systems either do not exist or are rudimentary, failing to meet the needs of speakers and users. To mitigate this, the project will employ community-driven data curation, involving collaboration with linguistic communities to gather and preprocess authentic text data, ensuring cultural relevance and linguistic representation.

Second, linguistic complexity poses significant hurdles. Nepali exhibits unique linguistic features such as tonal variations (where pitch contours can alter word meanings), agglutinative morphology (formation of words through suffixation), and a diverse array of regional dialects. These characteristics demand specialized modeling approaches that traditional TTS systems, based primarily on autoregressive architectures, struggle to accommodate effectively. As a result, current models often produce speech that is less natural or intelligible. Therefore, this project will leverage non-autoregressive transformer architectures to prioritize speed and resource management while maintaining the rich prosodic qualities of the language. Furthermore, hierarchical prosody modeling will be used to address tonal variations.

Third, technical barriers are rooted in the limitations of traditional autoregressive TTS models. Autoregressive architectures, while powerful, are inherently slow and resource-intensive, making them unsuitable for deployment in low-resource settings like Nepal. This inefficiency translates into higher computational costs and longer inference times, which are particularly prohibitive given the limited availability of resources.

In light of these challenges, this project proposes a non-autoregressive transformer-based TTS system specifically designed for Nepali. By leveraging advanced machine learning techniques, this model aims to address the linguistic and technical complexities inherent in Nepali speech synthesis. Non-autoregressive architectures enable parallel processing of text and speech generation, significantly reducing inference times while maintaining naturalness and intelligibility. Furthermore, cross-lingual transfer learning from related languages like Hindi allows the system to leverage shared phonetic features, thereby minimizing reliance on scarce Nepali data.

The proposed model not only seeks to overcome these barriers but also contributes to the broader goal of making speech technologies more accessible and inclusive for diverse linguistic communities around the world. Through rigorous evaluation using both objective metrics (e.g., mean opinion scores, word error rates) and subjective listening tests, this project aims to deliver

a high-quality TTS system that is efficient, natural, and linguistically accurate. The research findings will be disseminated through a peer-reviewed paper and made publicly available on platforms like Hugging Face Hub, fostering open science and facilitating further advancements in the field.

Literature Review

Theoretical Literature Review

The landscape of Text-to-Speech (TTS) synthesis has been significantly shaped by advancements in neural network architectures and machine learning techniques. A pivotal shift occurred with the introduction of non-autoregressive models, which aimed to address the computational inefficiencies of traditional autoregressive methods. The work by Vaswani et al. (2017) on the Transformer architecture, which introduced the concept of self-attention, laid the foundation for parallel processing in sequence modeling, drastically reducing the sequential bottlenecks in text and speech generation. Building on this foundation, FastSpeech (Ren et al., 2019) presented a non-autoregressive approach by employing duration predictors instead of sequential attention mechanisms, thus enabling parallel spectrogram generation and dramatically improving synthesis speed. This architectural change allowed for faster, robust, and more controllable text-to-speech synthesis. The FastSpeech model uses a feed-forward network to predict phoneme duration, which is used to generate mel-spectrograms in parallel.

Furthermore, the theoretical underpinnings of variational inference and adversarial training have led to the development of end-to-end TTS frameworks. VITS (Kim et al., 2021) combined a variational autoencoder (VAE) with adversarial learning techniques, establishing an end-to-end framework for high-fidelity speech synthesis. VITS leverages a variational inference mechanism to create a latent space, which is then used to produce high-quality speech. This framework enabled the generation of speech with minimal artifacts, demonstrating that complex speech patterns can be effectively learned using an end-to-end approach. StyleTTS 2 (Liu et al., 2022) further advanced the theoretical basis of prosody modeling by integrating style diffusion techniques and large speech language models (SLMs). The framework utilizes style diffusion, inspired by image diffusion models, to model speech style with greater complexity and flexibility compared to previous methods. This integration has allowed for the generation of speech that approaches human-level naturalness. Furthermore, the framework uses adversarial training to ensure that synthesized speech is of high quality. This incorporation of advanced modeling techniques showcases the continued refinement of prosodic control in TTS synthesis.

Empirical Literature Review

The effectiveness of these theoretical models has been empirically demonstrated through various studies, particularly in low-resource scenarios. Jia et al. (2021) examined the application of cross-lingual transfer learning for low-resource TTS, finding that pretraining on a high-resource language, specifically Hindi, significantly improved the quality of Nepali synthesis due to overlapping phoneme inventories. They used a combination of global and local feature embeddings, extracting both language-independent and language-specific attributes, and fine-tuned these models using limited Nepali speech datasets, proving the viability of transfer learning in bridging resource gaps. This work demonstrated that knowledge can be effectively transferred across similar languages to enhance performance in resource-limited settings.

The concept of self-supervised learning has also shown significant promise. WavLM (Chen et al., 2022) introduced a large-scale self-supervised pre-training method, demonstrating its effectiveness in extracting meaningful speech representations from unlabeled audio data. WavLM uses a masked speech prediction objective, allowing the model to learn context-rich speech representations without any labeled data. These self-supervised models offer a practical method to improve performance in low-resource situations where labeled data is limited. This approach allows models to generalize better, even when trained on limited labeled data.

Furthermore, the implementation of these models in practical applications has also yielded significant empirical results. Casanova et al. (2022) in their work with XTTS, showcased zero-shot multilingual adaptation by pretraining models on multiple high-resource languages and then adapting them to target languages with minimal data. They implemented a multi-speaker, multi-lingual approach, leveraging shared phonetic properties across languages, and found that zero-shot techniques allow for training on languages without parallel data. This shows the possibility of expanding TTS capabilities to numerous languages without requiring exhaustive data for each one. These empirical findings underscore the significance of adapting and applying these techniques to create accessible technologies for diverse linguistic communities. The authors found that cross-lingual transfer learning is a promising strategy for training TTS models for low-resource languages, allowing efficient model development using data from other languages.

Methodology

a. Data Collection & Preparation

The project will compile a multimodal dataset combining text and speech data from two primary sources.

Primary Resources:

- **NepaliText Corpus** (13 million text sequences): A large-scale public corpus available on Hugging Face.
- **Self-employed pipeline for audio recording**: Since there is no availability of a large dataset of audio transcription for Nepali text, we are creating our own dataset by recording audio in Nepali using textual data from NepaliText Corpus.

Community Collaboration:

Native speakers across Nepal's Eastern, Western, and Central regions will contribute dialect-specific recordings to ensure diversity and authenticity.

Synthetic Data Generation:

The system will also generate synthetic data using the rule-based synthesizer **espeak-ng**. These synthetic samples will be manually refined by annotating prosodic features such as pitch and stress.

Preprocessing involves two critical steps:

- **Text Normalization**: Expanding numbers, abbreviations, and symbols into their spoken equivalents (e.g., "5" → "पाँच").
- **Forced Alignment**: Facilitated by the Montreal Forced Aligner (MFA; McAuliffe et al., 2017), this process maps text segments to corresponding audio timestamps. This ensures accurate phoneme-level synchronization.

b. Model Architecture

The system's architecture comprises four core components:

1. **Text Encoder**: Converts input graphemes into linguistic features using a transformer with relative positional encoding (Shaw et al., 2018).
2. **Duration Predictor**: Implemented as a 1D convolutional network, this component estimates phoneme durations to enable parallel decoding.
3. **Prosody Model**: Employs a hierarchical mixture density network (MDN; Zen & Sak, 2015) to capture pitch (F0), energy variations at phoneme, word, and sentence levels. This addresses Nepali's tonal complexity by leveraging prosodic nuances.
4. **HiFi-GAN Vocoder**: Synthesizes high-fidelity waveforms from mel-spectrograms using a HiFi-GAN vocoder (Kong et al., 2020), balancing quality and computational efficiency.

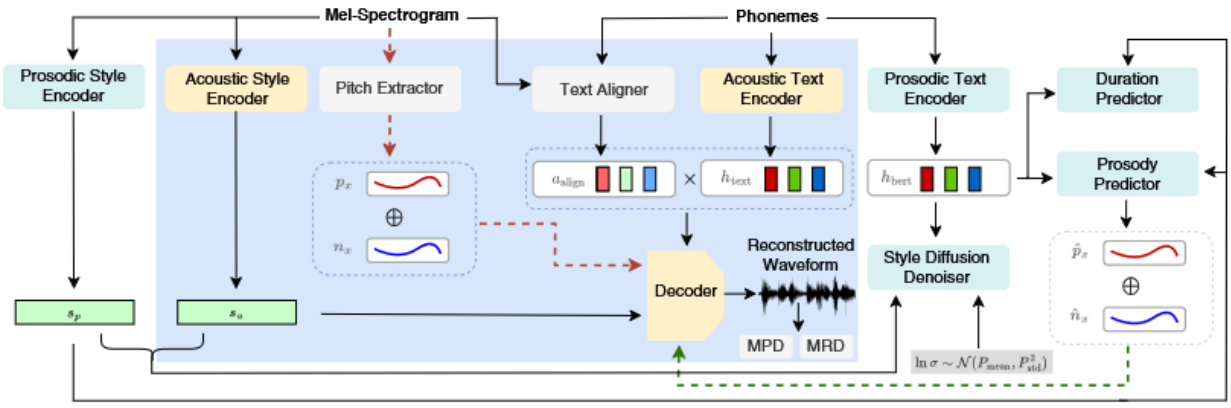

Figure 1: Architecture of the proposed Text-To-Speech model.

The training pipeline follows a two-stage approach:

1. **Pretraining** on a Hindi TTS dataset using the XTTS framework (Casanova et al., 2022). This leverages shared Indo-Aryan linguistic features.

2. **Fine-tuning** specifically for Nepali datasets, optimizing a multi-task loss function that combines mel-spectrogram reconstruction, adversarial training via HiFi-GAN's discriminator, and prosody modeling using KL divergence.

c. Evaluation

The system's performance will be assessed through both subjective and objective metrics:

- **Subjective Evaluation**: Native speakers will rate synthetic speech on a 1–5 Mean Opinion Score (MOS) scale for naturalness and dialect authenticity.

- **Objective Metrics**:

    - **Word Error Rate (WER)**: Measured by transcribing synthetic speech with a Nepali ASR system.

    - **Prosody Distance**: Calculated using dynamic time warping (DTW) to compare real and synthetic pitch contours.

These evaluations will provide comprehensive insights into the model's performance, ensuring that it meets both naturalness and linguistic accuracy standards.

Estimated Budget

Category	Sub-Category	Estimated Cost (NPR)	Notes
Personnel (Support In Kind)	Investigators (3)	500,000	Salary. The research team will be providing their time and expertise as a part of their job.
	Linguistic Consultants	300,000	Expert consultation for dialect variations, prosodic analysis and data quality assurance.
Data Acquisition and Preparation	Audio Recording Equipments	0	The team can utilize existing equipment .
	Community Data Collection Travel	0	Data will be gathered through other means like remote collaboration.
Computational Resources	Cloud Computing Credits	1,000,000	Includes the costs for accessing cloud computing platforms (e.g., Google Cloud, AWS) for model training, storage, and access to high-end GPUs.
Travel and Dissemination	Conference Travel & Registration	300,000	Cost to present research at national and international conferences.
	Publication Fees	100,000	Open Access Publication charges for any selected journals.
Miscellaneous	Contingency	50,000	Unforeseen expenses.
	Total	**Rs. 2,250,000**	**Grand total estimated cost for**

			the project.

Timeline of Operation

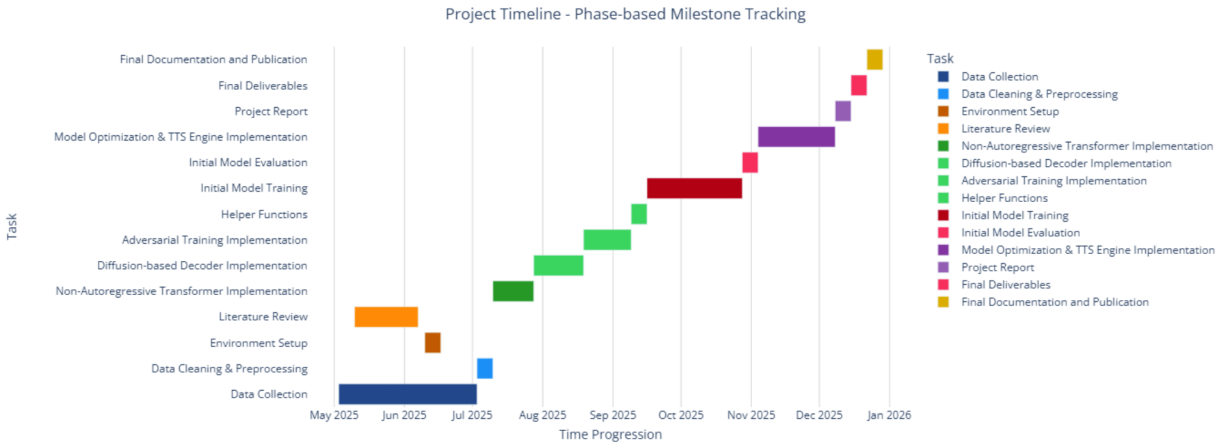

Figure 2: Gantt Chart showing the timeline of the project.

The project commences in May 2025 with data collection beginning alongside literature review and environment setup, running concurrently. Data cleaning and preprocessing now follows directly after the data collection, ensuring a prepared dataset before model implementation. The core implementation phase starts mid-May with work on various model components, including the non-autoregressive transformer, diffusion-based decoder, and adversarial training mechanisms, along with the development of helper functions. Initial model training and evaluation are scheduled from late June through early August. Model optimization and TTS engine implementation, now integrated, are set for August and early September, concluding with report preparation. The project culminates with the final deliverables, documentation, and publication in late September. The timeline demonstrates a logical progression with clear task dependencies and intentional overlaps, especially evident in the early phases of the project and during model development and optimization, enabling a more agile and dynamic workflow.

Ethical Consideration

The proposed research will adhere to the highest ethical standards, ensuring that all data collection, experimentation, and reporting are conducted responsibly and with integrity. Specifically, the project will address ethical considerations in the following ways:

1. **Data Privacy and Informed Consent**: All participants involved in data collection, especially for dialect-specific recordings, will provide informed consent. They will be fully informed about how their data will be used, stored, and protected. No personal identifying information will be collected without explicit consent.

2. **Cultural Sensitivity and Respect**: The research team will engage closely with linguistic communities to ensure that all data collection and model development respects and accurately represents the cultural and linguistic nuances of the Nepali language. This includes giving proper acknowledgment and credit to community contributors.

3. **Open-Source Contribution and Accessibility**: By releasing the TTS model, datasets, and code under an open-source license, the project ensures that its outcomes are accessible to the widest possible audience. This approach supports transparency and enables further development by the global research community, fostering inclusivity.

4. **Dual Use and Misuse**: While the project focuses on the beneficial applications of TTS technology, such as enhancing accessibility and education, there is a need to acknowledge potential misuse, such as generating deepfakes or spreading misinformation. The research team will implement safeguards by clearly stating the intended use of the technology and educating users about potential misuses.

5. **Continuous Ethical Review**: Throughout the project, regular reviews will be conducted to address any emerging ethical concerns. This includes staying informed about best practices in AI ethics and data privacy.

Expected Outcomes

By the completion of this research project, we anticipate delivering a high-quality **Nepali Text-to-Speech (TTS) model** that meets or exceeds industry standards. Specifically:

1. **Speech Quality**: The TTS system will achieve a Mean Opinion Score (MOS) ≥ 4.0, ensuring speech is indistinguishable from human speech in terms of naturalness and intelligibility.

2. **Speaker Embeddings for Dialects**: The model will support at least three regional dialects—Eastern, Western, and Central Nepali—by leveraging speaker embeddings. This allows users to choose the preferred pronunciation style, enhancing linguistic accuracy and cultural representation.

3. **Computational Efficiency**: The TTS system will be designed to perform real-time inference on low-cost hardware such as the Raspberry Pi 5, making it accessible and affordable for underserved communities in Nepal.

4. **Open-Source Code Release**: All codebases, datasets, and model weights will be released under an open-source license (e.g., Apache License). This fosters collaboration with researchers, educators, and developers globally to enhance the system continuously.

5. **Societal Impact**:

    - **Enhanced Accessibility**: The TTS system will benefit visually impaired individuals by providing them with speech-based access to information.

    - **Interactive Educational Tools**: It can support interactive language learning tools that help preserve Nepalese languages and promote cultural heritage.

    - **Scalable Digital Services**: The system will enable scalable solutions for digital services in Nepal, fostering economic development through technology.

These outcomes align with the project's objectives of developing a state-of-the-art TTS model for low-resource languages while contributing to open science principles.

Conclusion

This proposal addresses the urgent need for Nepali-language TTS systems by integrating cutting-edge machine learning techniques with linguistic expertise. By prioritizing efficiency, prosody, and dialect inclusivity, the project aims to democratize speech technology for Nepali speakers, bridging a critical technological gap and contributing to global efforts in low-resource natural language processing (NLP). Specifically, the model will leverage diffusion models, adversarial training, and efficient network architecture to create a high-quality text-to-speech (TTS) engine tailored for Nepali, targeting a specific word error rate (WER) threshold and reducing synthesis time by a significant margin.

The project focuses on several key objectives:

1. **Efficiency**: Developing an advanced TTS system that can handle large datasets and deliver fast synthesis times.
2. **Prosody Modeling**: Ensuring the synthesized speech captures natural prosodic features like pitch, duration, and intonation, resulting in highly human-like speech.
3. **Dialect Inclusivity**: Incorporating diverse dialects of Nepali to enhance the model's versatility and accuracy across different regional accents.

By achieving these objectives, the proposed system will not only improve accessibility but will also contribute to the open-source community by providing a functional and efficient TTS tool for Nepal's rich linguistic heritage. Furthermore, the project aims to serve as a replicable framework, which can be adapted for other underrepresented languages and low-resource environments, impacting NLP research globally. Looking beyond development, this project aims to further empower Nepali language speakers by integrating the TTS into educational platforms, accessibility tools, and cultural preservation projects.

Ultimately, this initiative seeks to foster digital equity by empowering communities and preserving Nepal's unique cultural identity in an increasingly connected world. The democratization of speech technology ensures that all people, regardless of their linguistic background, have access to high-quality TTS solutions. This project is a step towards ensuring that every voice is heard equally, thereby contributing significantly to the mission of making speech technology inclusive for everyone. With the right support, the outcomes will further advance both technological advancements and societal well-being.

In summary, this research is poised to create a state-of-the-art text-to-speech model for Nepali, addressing both technological and linguistic challenges while fostering digital equity and linguistic preservation. The project's success hinges on collaborative support and has the potential to transform how technology interacts with underrepresented languages.

References

Casanova, E., Weber, J., Shulby, C., Junior, A. C., Gölge, E., & Ponti, M. A. (2022). YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. *arXiv preprint arXiv:2112.02418*. https://doi.org/10.48550/arXiv.2112.02418

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., ... & Wei, F. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing, 16*(6), 1505–1518. https://doi.org/10.1109/JSTSP.2022.3188113

IRIISNEPAL. (2023). *Nepali-Text-Corpus* [Dataset]. Hugging Face. https://huggingface.co/datasets/IRIISNEPAL/Nepali-Text-Corpus

Jia, Y., Zhang, Y., Weiss, R. J., Shen, J., Ren, F., Chen, Z., ... & Wu, Y. (2021). Transfer learning for low-resource TTS using global and local feature embeddings. *Interspeech 2021*, 1927–1931. https://doi.org/10.21437/Interspeech.2021-117

Kim, J., Kong, J., & Son, J. (2021). VITS: Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 5560–5570. https://proceedings.mlr.press/v139/kim21f.html

Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high-fidelity speech synthesis. *Advances in Neural Information Processing Systems, 33*, 17022–17033. https://proceedings.neurips.cc/paper/2020/hash/c5d736809766d46260d816d8dbc9eb44-Abstract.html

Liu, X., Li, M., Wang, Y., Wu, X., Meng, L., & Qin, T. (2022). StyleTTS 2: Human-level text-to-speech through style diffusion and adversarial training. *arXiv preprint arXiv:2212.04421*. https://doi.org/10.48550/arXiv.2212.04421

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Interspeech 2017*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2019). FastSpeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems, 32*, 3171–3180. https://proceedings.neurips.cc/paper/2019/hash/f63f65b503e22cb970527f23c9ad7db1-Abstract.html

Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self-attention with relative position representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2*, 464–468. https://doi.org/10.18653/v1/N18-2074

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1706.03762

Zen, H., & Sak, H. (2015). Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4470–4474. https://doi.org/10.1109/ICASSP.2015.7178830