MUNTTS: A Text-to-Speech System for Mundari

Varun Gumma[♠] **Rishav Hada**[♠] **Aditya Yadavalli**[♢]
Pamir Gogoi[♡*] **Ishani Mondal**[♣*] **Vivek Seshadri**[♠♢] **Kalika Bali**[♠]
[♠]Microsoft Corporation [♢]Karya Inc. [♡]Project VANI [♣]University of Maryland
varun230999@gmail.com, kalikab@microsoft.com

Abstract

We present MUNTTS, an end-to-end text-to-speech (TTS) system specifically for Mundari, a low-resource Indian language of the Austo-Asiatic family. Our work addresses the gap in linguistic technology for underrepresented languages by collecting and processing data to build a speech synthesis system. We begin our study by gathering a substantial dataset of Mundari text and speech and train end-to-end speech models. We also delve into the methods used for training our models, ensuring they are efficient and effective despite the data constraints. We evaluate our system with native speakers and objective metrics, demonstrating its potential as a tool for preserving and promoting the Mundari language in the digital age[1].

1 Introduction

India is home to approximately 1652 languages and 22 official languages written in different scripts (Prakash and Murthy, 2023; Gala et al., 2023). Many of these languages are classified as low-resource, as native speakers are moving towards dominant languages that are supported by modern-day technologies (Bali et al., 2019).

According to a UNESCO report, India ranks fourth in the list of critically endangered languages with 41 languages (Kwan, 2022). If a language becomes extinct we lose out a large part of the culture. Motivated by these factors, many communities have self-initiated data collection and partnered with tech organizations to build technologies for their language (Diddee et al., 2022).

Text-to-speech (TTS) systems have been gaining a lot of importance as a vital language technology due to their applications in education, navigation, accessibility, voice assistants etc. (Kumar et al., 2023). However, the development of TTS systems for low-resource languages has several challenges (Pine et al., 2022). Firstly, training current-day TTS systems requires many hours of audio recordings and corresponding text transcriptions which is resource-intensive. Secondly, carefully curating the training data such that it covers the phonetic complexity of the given language requires expert input. This becomes a problem especially when the available data is already scarce. Thirdly, the availability of native speakers who are familiar with technology and can do audio recordings in a high-quality studio setup. Fourthly, availability of enough native speakers, who could systematically evaluate these systems for subjective metrics. Lastly, getting high-quality audio recordings can be very expensive. Overcoming these challenges requires not only technical expertise to extract the most out of limited resources but also significant on-field operational efficiency to collect the right quality and quantity of data. No such data collection is possible without the active participation of the community and other stakeholders.

In this work, we build a TTS system for Mundari. Mundari is an Austro-Asiatic language spoken by Munda tribes in the eastern Indian states of Jharkhand, Odisha, and West Bengal. According to the 2011 India census, there are ≈1M native speakers of this language (2011, Archived from the original on 6 March 2021). Mundari is mainly written in the dominant script of the region where it is spoken, viz. Devanagari, Odia, and Bangla. In this study, we worked with the Mundari spoken in Jharkhand and written in the Devanagari script.

We collected audio recordings for 15,656 unique sentences in Mundari. Our Mundari speech corpus consists of high-quality 26,868 audio recordings in male and female voices consisting of 27.51 hours. The average duration of the recordings in our dataset is 3.7 seconds, and the average sentence length is 8.4 words. Using this data we train three TTS systems: Variational Inference with adversarial learning for end-to-end (E2E) Text-to-Speech

[*]Work done when the author was at Microsoft
[1]Artifacts available at https://aka.ms/MUN-TTS

Proceedings of the The Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 76–82

(VITS) with 22KHz sampling rate (VITS-22K), VITS with 44KHz sampling rate (VITS-44K) and fine-tune XTTS v2 as well. We have audio samples from each of these systems evaluated by native speakers for subjective metrics. VITS-44K gives the best overall performance with MOS = 3.69 $\pm$ 1.18.

2 Related Works

- **Neural Speech Synthesis:** The field of neural speech synthesis has experienced a series of transformative developments. WaveNet (van den Oord et al., 2016), utilized a convolutional neural network for raw audio waveform generation, marking a shift from previous heuristic synthesis methods. Tacotron (Wang et al., 2017) further advanced the field with its end-to-end text-to-speech synthesis, streamlining the synthesis process. Tacotron 2 (Shen et al., 2018) built upon this by incorporating WaveNet as a vocoder, enhancing speech quality. Parallel WaveGAN (Yamamoto et al., 2020) introduced a generative adversarial network approach for faster waveform generation. FastSpeech (Ren et al., 2019) and FastSpeech 2 (Ren et al., 2021) utilized feed-forward networks for faster speech generation and enhanced control over speech attributes. VITS (Kim et al., 2021) combined variational autoencoders (Kingma and Welling, 2019) with GANs (Goodfellow et al., 2014) in an end-to-end structure, enabling more expressive speech synthesis. Lastly, XTTS (Coqui, 2023) provided cross-lingual text-to-speech capabilities, representing a notable advancement towards adaptable speech synthesis systems. We refer the readers to Tan et al. (2021) for a comprehensive survey of neural text-to-speech.

- **Relevant TTS systems:** In recent times, there has been a shift towards TTS models for low-resource languages, especially for Indian languages. EkStep Foundation (2021) put forth the first open-source monolingual neural systems for 9 Indic languages using a Glow-TTS + HiFi-GAN combination. Prakash and Murthy (2020) advance it by releasing multilingual TTS models within the same family using a multilingual character map (Prakash et al., 2019) and common label set (Prakash et al., 2019) for Tacotron2 + WaveGlow. Kumar et al. (2023) extend the language coverage to 13 by including 3 low-resource languages, Rajasthani, Bodo, and Manipuri. They also conduct a thorough analysis of different Non-Autoregressive (NAR), flow-based, and end-to-end models in a multi-speaker and multilingual setting and find that single-language models are preferable. Globally, Pratap et al. (2023) expand the text-to-speech coverage to 1017 languages by training individual end-to-end VITS models for each language. However, none of them have developed a dedicated, high-quality, multi-speaker TTS for an extremely low-resource language. In this paper, we present our experiences in developing a high-quality, multi-speaker TTS model for such a language – Mundari.

3 Data

Multiple steps were involved in the data collection process. First, the text data was obtained by translating a Hindi corpus of 100,000 sentences obtained from the Karya database. Karya[2] is a data services organization that takes requests from clients and breaks down these complex requests into simple microtasks that users with little to no digital literacy can perform.

We randomly selected 20,000 of these sentences and manually translated them to Mundari. The translated Mundari sentences were expressed using the Devanagari script. The translators were instructed to prefer fluency of the sentences over faithfulness of the translations wherever they had to make a choice. The translated sentences were then validated for appropriateness by native speakers. This text corpus was then used as the final dataset for recording one male and one female speaker. The male and female speaker was selected from a pool of 12 speakers (6 male and 6 female), who were asked to complete a reading task online. After they submitted the speech samples, the speakers were then evaluated by native speakers and given a score for their reading efficiency and pronunciation. Based on these scores, 3 male and 3 female speakers were shortlisted, and finally, from these 6 speakers, 1 male and 1 female speaker was selected after analyzing some voice quality features. The speakers, i.e., the voice artists, were instructed to record the sentences shown to them without any false starts, filled pauses, hiccups, or any other mistakes. All the recordings are done in a

[2] https://karya.in/

Data	Train	Val	Test
Avg. Sentence Length	8.48	8.57	8.44
Total Duration (in hours)	24.76	1.379	1.375
Male			
Num of Recordings	6302	350	350
Avg. Duration (in seconds)	3.85	3.80	3.88
Total Duration (in hours)	6.74	0.37	0.38
Female			
Num of Recordings	17,879	993	994
Avg. Duration (in seconds)	3.62	3.67	3.62
Total Duration (in hours)	18.02	1.01	0.998

Table 1: Dataset Metrics for the Mundari speech dataset.

studio-quality room with a microphone connected to the Karya crowdsourcing application for the convenience of collecting the data. The recordings' sampling rate is 44.1 KHz with 32 bits per sample.

Finally, the curated text used to collect recordings contains 15,656 unique sentences. The average sentence length in the collected text corpus is 8.4 words. Some duplication of sentences across speakers yields a total of 26,868 sentences and recordings in our final dataset. Around 74% of the recordings in our dataset feature a female speaker, while the remaining 26% are attributed to a male speaker. We notice that the female recordings are, on average, slightly shorter than male recordings – females' being 3.62 seconds compared to males' 3.85 seconds. We present more details in Table 1.

4 Experiments

4.1 Pre-Processing

The source sentences were normalized by collapsing repeated punctuations, exclamations, and spaces. Next, all kinds of brackets were removed and newline and tab characters were substituted with spaces. The `indic_nlp_library`[3] was used to further normalize the Devanagari text and appropriately space words with "matras". The dataset was split into train (95%), dev (5%), and test (5%) sets by stratifying on the number of speakers. The exact number of data points per split is available in Table 1.

4.2 Models

We train E2E TTS models using the `coqui-ai`[4] framework. These include a VITS model (Kim et al., 2021) trained from scratch, and a finetuned XTTS v2. Additionally, we also evaluated the zero-short performance of the pretrained XTTS v2 model and MMS-UNR Mundari model[5][6] from Facebook's Massively Multilingual Speech project (Pratap et al., 2023). Since the data curated is of very high-quality and sampled at 44.1KHz, we trained our VITS models with 44.1KHz data and standard 22.05KHz sub-sampled data. The latter was also used for finetuning the XTTS v2 model.

Here, we suggest the usage of single E2E models, as they are found to be significantly faster than two-stage models (Kim et al., 2021) and are optimal for deployment and efficient real-time usage.

4.3 Training Strategies

Both variants of the VITS models were trained with an elevated learning rate of $5e$-4 for the generator and discriminator, batch size of 128, and default ExponentialLR scheduler and AdamW (Loshchilov and Hutter, 2019) optimizer. As for the XTTS v2 finetuning, a significantly lower learning rate of $5e$-6 was used with a batch size of 256, and AdamW with a weight_decay of $1e$-2 was preferred as the optimizer along with a MultiStepLR Scheduler.

All our models were trained, and evaluated on a single A100 80GB GPU and were trained for 2500 epochs and converged within 5 days. A speaker-weighted sampler was also incorporated during the training/finetuning procedure to handle the speaker imbalance on our dataset. The models were checkpointed after every epoch based on the `loss_1` of the dev set and the best model checkpoint was used for evaluation.

5 Results and Discussions

5.1 Post-Processing

We use `ffmpeg` for rudimentary band-pass filtering and noise reduction on synthesized speech. To evaluate the XTTS v2 models, we provide one speaker reference audio from the dev set for conditioning and voice-cloning. Note that, the same reference audio was used for all the test examples for that speaker, and it was manually chosen to be a longer text and speech pair.

[3] https://github.com/VarunGumma/indic_nlp_library

[4] https://github.com/coqui-ai/TTS

[5] https://huggingface.co/facebook/mms-tts-unr

[6] To evaluate the MMS-UNR model, we transliterate our text from Devanagiri to Odia script using `indic_nlp_library`

Model	Full $n = 100$	Male $n = 26$	Female $n = 74$
gt-22k	4.62±0.68	4.59±0.65	4.63±0.69
gt-44k	4.58±0.70	4.47±0.79	4.62±0.66
mms	0.79 ± 1.02	0.79 ± 1.02	–
vits-22k‡	3.04 ± 1.29	2.65 ± 1.34	3.18 ± 1.25
vits-44k†	3.69 ± 1.18	3.39 ± 1.25	3.79 ± 1.13
xtts-finetuned	0.05 ± 0.30	0.13 ± 0.52	0.02 ± 0.16
xtts-pretrained	2.20 ± 1.32	2.10 ± 1.36	2.23 ± 1.31

Table 2: MOS values for ground truth and various models. The best and second-best scores are represented by † and ‡ respectively. (*gt* = ground truth)

Model	Full $n = 1344$	Male $n = 350$	Female $n = 996$
mms	15.13±4.19	15.13±4.19	–
vits-22k‡	9.45±3.71	10.03±4.05	9.24±3.56
vits-44k†	7.60±3.99	7.27±3.08	7.72±4.25
xtts-finetuned	13.65±5.92	10.73±5.33	14.69±5.77
xtts-pretrained	15.80±7.03	13.89±5.87	16.48±7.27

Table 3: MCD scores. The best and second-best scores are represented by † and ‡ respectively.

5.2 Subjective Metrics

We use the Mean-Opinion-Score (MOS) as the subjective metric for which 100 data points are randomly subsampled from the test set (with speaker stratification). Audio samples generated by various models for this set were sent for human evaluations to native speakers. The task was set up on the Karya platform, and each sample was rated on a scale of 1 to 5 with 0.5 points increments. As discussed earlier, for low-resource languages it is often difficult to find raters for subjective evaluation of the speech samples. In our case, each sample is rated 5 annotators. Using these ratings, we calculate the MOS for the ground truth (both 22.05 KHz and 44.1 KHz) and various models. In total, there were 7 variations for each text sample. Each sample was rated independently, so different variations of a sample were not directly compared. Raters were instructed to use headphones and rate the naturalness of the speech, considering factors such as prosody, intonation, and overall fluency. Detailed instructions are shown in Figure 1. Table 2 shows a comparison of the MOS values for the ground truth and the various models. We can see that the VITS-44K model performs the closest to ground truth. We also noticed a huge gap between the VITS model and the other models we studied. Interestingly, the MOS values for XTTS v2 became much worse on finetuning than using it in a zero-shot setup.

5.3 Objective Metrics

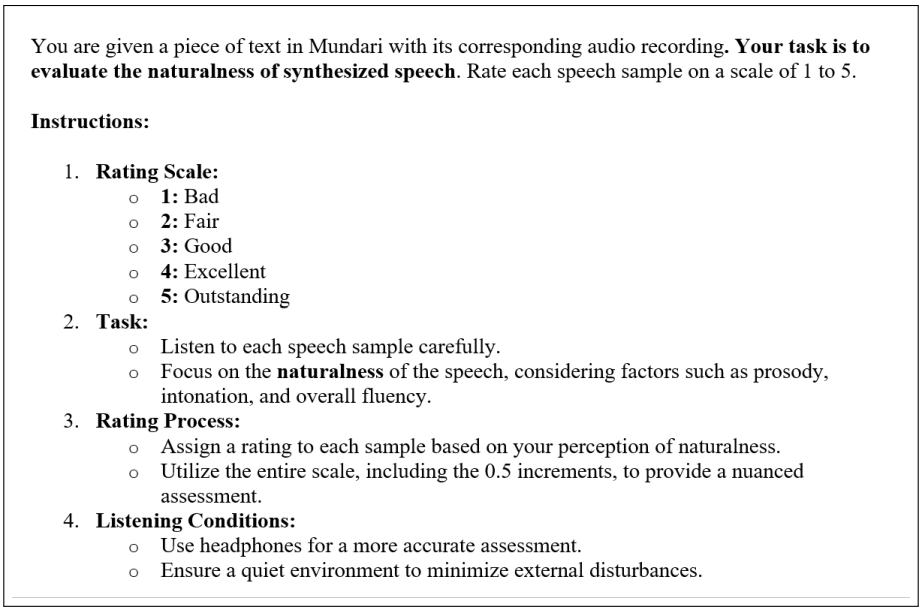
Mel-Cepstral Distortion (MCD) (Kubichek, 1993) is an objective measure used to quantify the difference between two sets of Mel-frequency cepstral coefficients and is useful in evaluating the performance of speech synthesis systems as it provides a numerical indication of how closely the synthesized speech matches the target or reference speech in terms of spectral characteristics.

For all the models, we compute the MCD scores with dynamic-time wrapping and weighted by speech length with respect to the ground truth subsampled to the sampling rate of the generated speech, if required. We present those in Table 3. Similar to the MOS scores, VITS-44K achieves the lowest error, followed by VITS-22K. Despite XTTS v2 employing speaker conditioning, the scores the significantly worse compared to the best model, VITS-44K.

XTTS v2 does not natively support Mundari but is pretrained with Hindi, which shares the same characters and pronunciations. Spot-checking some of the audios of the models revealed that the vanilla pretrained XTTS v2 had long pauses between words and made it sound unnatural. However, it captured the intonation and pronunciation well due to its voice-cloning capabilities. The process of finetuning the model resulted in a notable degradation in performance, leading to the generation of nonsensical outputs despite successful convergence. This might due to the catastrophic forgetting induced by the finetuning. We also observed that XTTS v2, which is based on GPT2 (Radford et al., 2019), generated phantom speech in many cases similar to hallucinations in Large Language Models. This phenomenon manifested as the introduction of random Hindi words and gibberish towards the end of the sentence.

6 Conclusion

In this work, we develop a TTS system for a low-resource language, Mundari, a low-resource language spoken by ≈ 1M people in India. We also analyze existing models for this language and evaluate popular multilingual and multi-speaker models by finetuning them. We show that the VITS-44K model achieves a mean MOS score of 3.69 and is evaluated as the best among the ones compared by native speakers. We release our model publicly and



You are given a piece of text in Mundari with its corresponding audio recording. **Your task is to evaluate the naturalness of synthesized speech**. Rate each speech sample on a scale of 1 to 5.

Instructions:

1. **Rating Scale:**
   - **1:** Bad
   - **2:** Fair
   - **3:** Good
   - **4:** Excellent
   - **5:** Outstanding
2. **Task:**
   - Listen to each speech sample carefully.
   - Focus on the **naturalness** of the speech, considering factors such as prosody, intonation, and overall fluency.
3. **Rating Process:**
   - Assign a rating to each sample based on your perception of naturalness.
   - Utilize the entire scale, including the 0.5 increments, to provide a nuanced assessment.
4. **Listening Conditions:**
   - Use headphones for a more accurate assessment.
   - Ensure a quiet environment to minimize external disturbances.

Figure 1: MOS guidelines provided to the annotators.

hope this research further promotes the development of speech systems for endangered and low-resource languages, aiding in bridging the digital divide in India.

7 Limitations

- Our primary emphasis in this study centers on E2E TTS, deliberately excluding the consideration of combinations involving Acoustic models and Vocoders, as observed in prior works (EkStep Foundation, 2021; Prakash and Murthy, 2020; Kumar et al., 2023). The motivation behind this choice is the intention to construct a simple unified system for speech synthesis, designed for straightforward deployment and ease of use by the general public.

- We explicitly recognize the inherent bias in the speaker distribution employed for our study. The challenge of recruiting native proficient speakers, capable of dedicating extended hours and effort to the recording process, contributed to a noticeable synthesis disparity, particularly evident in the diminished quality of male speech synthesis outputs.

8 Ethical Considerations

We use the framework by Bender and Friedman (2018) to discuss the ethical considerations for our work.

- **Institutional Review:** All aspects of this research were reviewed and approved by Karya.

- **Data:** Our data is collected in multiple steps as described in section 3. We first source the Hindi sentences and manually translate them to Mundari. Specific guidelines for translations were provided. These Mundari sentences were then recorded in a studio by 2 speakers.

- **Speaker Demographic:** We recruited 2 speakers to record the audio. Their payment was set after deliberation and contracts were signed. Speakers were paid INR 8 per recording. The average duration of a sample is $\approx$ 3.7 seconds.

- **Annotator Demographics:** Annotators for MOS rating were recruited through an external annotator services company. All annotators were native speakers of the language. The pay was INR 2 per sample, with an average sample length of $\approx$ 3.7 seconds.

- **Annotation Guidelines:** We draw inspiration from the community standards set for similar tasks. These guidelines were created following best practices after careful research. Annotators were asked to rate the speech samples on naturalness. A detailed explanation was given for the task. Annotator's identity was hidden from the authors to limit any bias.

- **Methods:** In this study, using our Mundari speech dataset, we trained 2 models: VITS-22K and VITS-44K, and finetuned the XTTS v2 model. We release the models for the benefit of the Mundari and the research community.

9 Acknowledgements

We sincerely thank the voice artists, Roshan and Meenakshi, for lending us their voices to create the speech dataset. We also extend our gratitude to Prof. Bornini Lahiri and Prof. Dripta Piplai from IIT Kharagpur for their advice on data collection and processing. Finally, we thank Praveen SV (Ph.D. Student, IIT Madras) and Gokul Karthik (MLE, Technology Innovation Institute) for their walk-throughs of IndicTTS[7] and Coqui.

References

India Census 2011. Archived from the original on 6 March 2021. Statement 1: Abstract of speakers' strength of languages and mother tongues – 2011. Office of the Registrar General & Census Commissioner, India.

Kalika Bali, Monojit Choudhury, Sunayana Sitaram, and Vivek Seshadri. 2019. Ellora: Enabling low resource languages with technology. In *Proceedings of the 1st International Conference on Language Technologies for All*, pages 160–163, Paris, France. European Language Resources Association (ELRA).

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Coqui. 2023. XTTS Models - Coqui Documentation. Accessed: 2023-12-14.

Harshita Diddee, Kalika Bali, Monojit Choudhury, and Namrata Mukhija. 2022. The six conundrums of building and deploying language technologies for social good. In *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, COMPASS '22, page 12–19, New York, NY, USA. Association for Computing Machinery.

EkStep Foundation. 2021. Vakyansh: Open source speech recognition. Accessed: 2023-12-14.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5530–5540. PMLR.

Diederik P. Kingma and Max Welling. 2019.

R. Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128 vol.1.

Gokul Karthik Kumar, Praveen S V, Pratyush Kumar, Mitesh M. Khapra, and Karthik Nandakumar. 2023. Towards building text-to-speech systems for the next billion users. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Michael Kwan. 2022. On the edge: Critically endangered languages in top countries. Taken from: UNESCO Atlas of the world's languages in danger 2010.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Aidan Pine, Dan Wells, Nathan Brinklow, Patrick Littell, and Korin Richmond. 2022. Requirements and motivations of low-resource speech synthesis for language revitalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7346–7359, Dublin, Ireland. Association for Computational Linguistics.

Anusha Prakash, Anju Leela Thomas, S. Umesh, and Hema A Murthy. 2019. Building Multilingual End-to-End Speech Synthesisers for Indian Languages. In *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, pages 194–199.

Anusha Prakash and Hema A. Murthy. 2020. Generic Indic Text-to-Speech Synthesisers with Rapid Adaptation in an End-to-End Framework. In *Proc. Interspeech 2020*, pages 2962–2966.

Anusha Prakash and Hema A. Murthy. 2023. Exploring the role of language families for building indic speech synthesisers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:734–747.

[7] https://github.com/AI4Bharat/Indic-TTS

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. 2023. Scaling speech technology to 1,000+ languages. *arXiv*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio.

Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards End-to-End Speech Synthesis. In *Proc. Interspeech 2017*, pages 4006–4010.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. 2020. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203.