Proceedings of 14th IOE Graduate Conference
Peer Reviewed
Year: 2023 Month: December Volume: 14
ISSN: 2350-8914 (Online), 2350-8906 (Print)

Nepali Text-to-Speech Using FastSpeech and MelGAN Model

Alina Lamichhane [a], Basanta Joshi [b], Bibha Sthapit [c]

[a, b, c] *Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, Tribhuvan University, Nepal*
[a] alinalamichhane05@gmail.com, [b] basanta@ioe.edu.np, [c] bibha@ioe.edu.np

Abstract
In the realm of deep learning models, a significant amount of attention and interest have been directed towards the field of text-to-speech synthesis particularly in context of Nepali text-to-speech synthesis. Since many years, researchers have attempted to produce speech from targeted Nepali texts, many models have also been developed but obtaining universal application has proven to be difficult. Our initiative is committed to using cutting-edge methods to effectively produce speech synthesis from Nepali text. As, only a limited amount of data is available for Nepali TTS system, we first focus on preparing Nepali News TTS dataset where news data was collected from the Nepal Television (NTV). Further, many data preprocessing has been applied and then the prepared dataset was subjected to the Fastspeech model with the MelGAN, generative adversarial network based neural vocoder. A real-time factor dictated the speech waveform inference time, synthesising 12 s of voice data in 118.069 s where inference time of mel-spectrogram and audio waveform seperately were 0.8549 s and 117.214 s respectively.

Keywords
Synthesis, Fastspeech, Text-to-Speech, MelGAN

1. Introduction

Text-to-Speech (TTS) is a prevalent technology that has witnessed significant advancements, playing a pivotal role in various human-computer interaction applications [1]. TTS systems aims to generate human-like audible voice/audio from digital text serving purposes such as assisting visually impaired individuals, powering chatbots, voice cloning, and more. However, capturing the nuances of natural human speech is very difficult. Despite the difficulty, researchers continuously endeavor to enhance the naturalness and quality of text-to-speech systems. TTS approaches include Concatenative synthesis [2], Parametric synthesis, and Neural synthesis, each with its own strengths and limitations. Among all these, latter approach marks a groundbreaking development from traditional methods by employing deep neural networks. Models like WaveNet, Tacotron, FastSpeech have revolutionized TTS technology by capturing the subtle nuances of human speech, resulting in highly realistic and expressive synthetic voices. These neural synthesis model have extensively explored several languages and have impressive result in terms of Mean Opinion Score(MOS).

FastSpeech model stands out for its efficiency in generating high-quality speech from input text with an impressive Mean Opinion Score (MOS) of 3.84 for English language. However, the main attraction of FastSpeech for text-to-speech technology are it's exceptional attributes of speed, robustness and controllability [3]. Because of its incredible speed, FastSpeech can produce speech quickly, making it ideal for interactive systems that require prompt responses. Moreover, due to it's robustness characteristic, it is reliable in variety of usage scenarios and diverse linguistic contexts. Additionally, this model allows different aspects of speech to be controlled and customized. In a similar vein, various models and methods are applied to conduct experiments in the English language. However, only limited research is done in Nepali, an Indo-Aryan language as there are many challenges in developing TTS system for Nepali language. Furthermore, the findings from these studies are not easily applicable in global context.

Nevertheless, the development of a Nepali TTS system holds a potential to not only facilitate the digital content for Nepali-speaking individuals but also benefit Nepali speaking community in many other aspects. Imagine a situation where a visually impaired person in Nepal is trying to access digital content online. Existing Text-to-Speech systems, typically designed for more commonly spoken languages, may not effectively convey an authentic and culturally meaningful auditory experience. Consider a news article or a piece of literature losing its essence when converted to speech. This is the void we aim to fill with our work which explores the utilization of state-of-the-art deep learning models for high-quality Nepali TTS synthesis. Specially, we propose the integration of FastSpeech and MelGAN for speech synthesis. FastSpeech, an attention-based sequence-to-sequence model excels in generating mel-spectrograms from input text, while MelGAN, a generative adversarial network to convert these mel-spectrograms into high-quality audio waveforms [4]. Our main contributions include:

- We present a novel method that utilizes Fastspeech and MelGAN as the base model to efficiently synthesize the high quality speech.
- The comparative study of systhesized speech with the original human speech and the overall system achieving a total inference time synthesising 12 s of voice data in 118.069 s.

2. Related Works

Evolving from the early mechanical attempts [5] at simulating human speech to more advance digital methods, a lot of research reported in the literature of speech synthesis and further Text-to-Speech (TTS) with the focus on diverse language. In recent years, end-to-end TTS powered by neural network based has emerged as a promising alternative to the traditional approaches that can directly map text to speech without the need for intermediate representation or use a simplified representation such as Mel-spectrogram. It utilizes the power of deep learning to model complex nonlinear relationships between text and speech, ultimately producing high-quality and natural-sounding speech.

English TTS research has been at the forefront of TTS technology, with numerous pioneering models and methodologies. WaveNet, introduced by A. Oord et al. [6] in 2016, marked a significant milestone by using deep generative networks to synthesize speech, achieving remarkable naturalness. Subsequent research focused on improving efficiency and quality of the synthesized speech with approaches like WaveRNN and Subscale WaveRNN [7] for efficient batch generation, SampleRNN [8] which combined autoregressive multilayer perceptron and stateful recurrent neural networks. Another notable development was MultiSpeech, a multi-speaker TTS system based on the Transformer by M. Chen and Xu Tan [9] addressing the text-to-speech alignment challenge demonstrating state-of-the-art performance on multi-speaker datasets. Additionally, weighted forced attention [10] was introduced to improve the text-to-speech alignment and investigate the effect of removing the duration predictor from FastSpeech.

Furthermore, Hong et al. [11] proposed Dynamic Transformer to overcome the challenges of using Transformer for TTS, improving convergence and computational speed. However, Chen and Rudnicky [12] presents a novel architecture for fine-grained style control in Transformer-based TTS using local style tokens (LST) and cross-attention blocks to capture and transfer speaking styles from reference speech to target text, while preserving linguistic content separation.

While English TTS research has achieved considerable success, Indic languages, including Nepali, have posed unique challenges due to their diverse phonological and linguistic characteristics. In recent years, research in Indic language TTS has gained momentum [13]. Studies focusing on languages such as Hindi, Urdu, Bengali, and Tamil have explored the adaptation of TTS models to accommodate the complexities of Indic phonetics and script. Notable advancements include the integration of prosody, linguistic context, and speaker identity into Indian language TTS models. Urdu TTS traditionally uses Hidden Markov Models (HMMs) [14], which require manual effort and struggle with language nuances. Alternatively, Natural Language Processing (NLP) [15] techniques enhance Urdu TTS but also needed manual rules and face language variation challenges. In 2020, Indian researchers build a generic TTS system and adapt it to new language having similar linguistic behaviour. The findings show that high-quality TTS systems can be created with as little as 7 minutes of adaptation data, maintaining the target speaker's prosody and enabling smooth transitions between speakers and languages [16].

In the context of the Nepali language, Chettri and Bikram Shah [17] introduced ESNOLA method in 2013, which relies on concatenating vocal tract resonance-based elements. Subsequently, Ghimire and Bal [2] improved an existing TTS system by introducing post and preprocessing modules. Later on, Shrestha et al. employed FreeTTS to build a system using a rule-based approach for tasks such as text normalization, syllabification, phonetic transcription, and prosody production. They also used a diphone concatenation-based unit selection approach for voice synthesis. More recently, Ashok [18] developed a comprehensive Nepali speech synthesis system using an encoder-decoder architecture guided by attention mechanisms, with WaveNet as the vocoder. This system underwent training using Nepali speech data from OpenSLR, and its speech synthesis quality received a MOS score of 3.07, which is generally considered acceptable but leaves room for improvement. Possible enhancements could be realized by using a larger dataset for model training, exploring alternative models like FastSpeech, addressing issues related to dataset noise, and overcoming resource constraints that may have impacted the system's performance.

3. Proposed Approach

In the pursuit of synthesizing high-quality, natural-sounding human-like audio waveforms in the context of Nepali Text-to-Speech (TTS), the generation of mel spectrograms is a crucial step. To achieve this, we adopt a two-step approach: first, utilizing the FastSpeech model to convert textual inputs into mel-spectrogram representations, and subsequently, employing the MelGAN vocoder to convert these mel-spectrograms into the desired audio waveforms. This approach ensures the end-to-end synthesis of high-quality Nepali speech with a focus on overcoming the particular phonological and linguistic difficulties of the Nepali language. Figure 1 shows the general block diagram of TTS system.

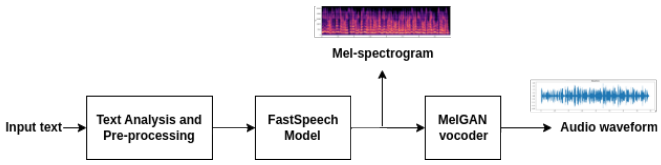

Figure 1: General Block diagram of TTS system

3.1 Text-to-Speech System

3.1.1 Character Embedding

It is a technique that transforms input text is into a sequence of embeddings, which are vector representations of each character in the text which are furthered learned during training process. These embeddings capture the phonetic and semantic information of the text. Additionally, character embedding can handle misspelled words and slang terms.

3.1.2 FastSpeech Block

Embedded text is fed to the FastSpeech block which mainly consists of three components as Phoneme-side block, length

regulator and Mel-side block. Phoneme side FFT block of FastSpeech consists of self attention and 1D convolutional network. The self attention here extracts the cross-position information. The output is further passed to length regulator which have duration predictor to predicts the duration and also length regulator solve the problem of length mismatch between character and mel-spectrogram sequence in FFT block based on the alignments between input text and output mel-spectrogram generated from the teacher model, tacotron. The final linear layer applies the linear transformation to the incoming data. Figure 2 demostrates the block diagram of this block.

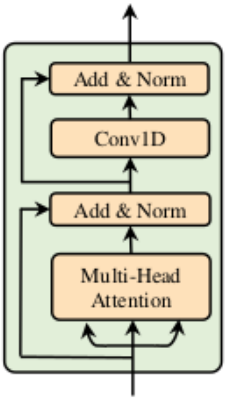

Figure 2: FFT Block

3.1.3 VOCODER

Here the MelGAN vocoder is utilized to synthesize the natural human like speech waveform from the mel-spectrogram produced by FastSpeech module.

3.2 Data Collection and preprocessing

Nepali, an under-resourced language, presents challenges for developing a natural TTS system due to the scarcity of linguistic resources. The resource preparation process is costly and time-consuming, typically requiring linguists or experts. Research in this field primarily focuses on creating TTS systems with limited resources. However, data quality and quantity are crucial for building a reliable model. To address this, the study collects data from two sources: the OpenSLR dataset and news data from Nepal Television (NTV).

- OpenSLR Dataset: The "High-Quality Nepali TTS Dataset" in OpenSLR [19] is a valuable resource for advancing Nepali-language processing and TTS technology. This dataset, thoughtfully compiled by Google in Nepal, features native female Nepali speakers. It includes accurately transcribed Nepali text paired with high-quality audio recordings, emphasizing audio clarity with a 48 kHz, 16-bit, mono format. Comprising literature, news articles, and everyday speech, the dataset offers 2.8 hours of data, 2064 sentences (47114 words), and 1-13 second audio clips from 19 female speakers.
- Nepali News Dataset: The "Nepali News TTS Dataset" is a comprehensive collection of text and audio data sourced exclusively from Nepal Television's 8 PM news broadcasts. It includes news articles, headlines, and transcripts from these broadcasts, meticulously cleaned and standardized for consistency. The audio files, available in WAV format at 48 kHz, may have minimal background noise. This dataset features a diverse range of voices from various news anchors with high-quality audio clips, each ranging from 1 to 12 seconds. Notably, text and audio segments are precisely time-aligned for perfect synchronization.

3.2.1 Data preprocessing

To ensure the quality and integrity of the "Nepali News TTS Dataset" a series of critical steps were meticulously carried out, encompassing both audio and text processing.

Audio Processing : Initially, audio was extracted from Nepal Television's 8 PM broadcasts on YouTube using the YouTube API, preserving the original mp3 format and its inherent quality. However, some background noise was present in the extracted audio so, different noise reduction tools like Audacity and PyTorch packages was used to minimize noise and enhance audio clarity. Subsequently, the audio was converted from mp3 to the more widely recognized WAV format and manually segmented into coherent units by aligning it with sentences. Precise text-audio alignment was executed with meticulous care, as it is crucial for a high-quality TTS system. Finally, all the segmented audio data was systematically organized and stored in a folder structure mirroring the popular English dataset "LJSpeech Dataset".

Text Processing : To preserve consistency and accuracy, the textual data in the "Nepali News TTS Dataset" was subjected for processing. This included tasks including text standardization and cleaning, correcting irregularities, contradictions, and unpleasant textual aspects. The text's authenticity to the original material was carefully preserved, and language conventions and readability standards were followed. Together, these text processing efforts improved the dataset's quality for TTS study and improvement by producing a refined and logically ordered textual component that matches the excellent audio recordings.

4. Result and Discussion

4.1 Training

FastSpeech generates the entire output sequence at once,but it requires advance knowledge of expected output length for a given input. this information is provided through alignments which maps the relationship between the input text and output speech. To acquire the alignments, a tacotron model, functioning as a teacher model, is trained to predict mel-spectrograms from input text. Subsequently, it learns the alignments that establish the correspondence between input text and the resulting mel-spectrogram output.

To train the FastSpeech model the audio data was resampled to 22050 kHz and the dataset was splitted into two category,

test and train dataset. The prepared train dataset along with the alignments was input as training data for the model to generate Mel-spectrogram and was trained for the iteration of 32k steps in the experimental setup with the 16 core Virtual CPU and total memory of 50 GB. It took 5.8 days (140 hours) to complete the training. Here, the initial learning rate was 0.001 and end learning rate 0.00001, total steps of training was 32000 and batch size was 16. The loss curve for for training set is shown in figure 3 below.

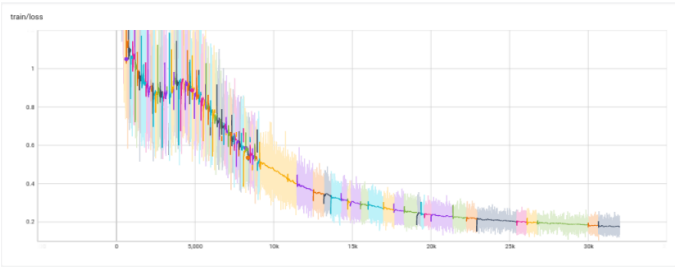

Figure 3: Train loss (Total of mel loss and duration predictor loss)

After training of the text-to-Mel spectrogram model, the pretrained vocoder was used to generate the audio waveform.

4.2 Result

In our experiments, the FastSpeech model underwent two training processes using different datasets, OpenSLR and the self-prepared "Nepali News TTS Dataset." The first process involved 10,000 steps with a batch size of 32 and a learning rate of 0.001, resulting in speech synthesis with background noise and an inference time of 137.050893 seconds for a 24-word sentence. The second process, using the Nepali News TTS Dataset, improved audio quality compared to the first experiment. We increased the training steps to 20,000 and then to 32,000 with a batch size of 16 and an audio sampling rate of 22.05 kHz. The overall synthesis time for given nepali text (Text: निजी क्षेत्रबाट चिकित्सा शिक्षाका कलेज संचालकहरुले नर्सिङ लगाएतका विधामा तत्काल विद्यार्थी भर्ना गर्न पाउनुपर्ने लगाएतका माग राखी सरकार विरुद्द संघर्षको कार्यक्रम घोषणा गरेका छन ।for both 20k and 32k steps are provided in table 1.

Table 1: Detailed time to synthesize audio from the given text in 20k and 32k steps.

Steps	Fastspeech	MelGAN	Total time
20k steps	0.9480	123.303	124.251
32k steps	0.8549	117.214	118.069

4.3 Evaluation

To evaluate the training performance of the text-to-mel spectrogram model, the loss curves of the generated mel-spectrogram on the splitted test set are plotted.

For some set of texts, Mel-spectrogram was generated using the trained models and further converted to audio waveforms.

1. Plot of target text: प्रतिनिधि सभाको सभामुख पदका लागि भएको निर्वाचनमा सत्ता गठबन्धनका उम्मेदवार देवराज घिमिरे विजयी हुनुभएको छ।

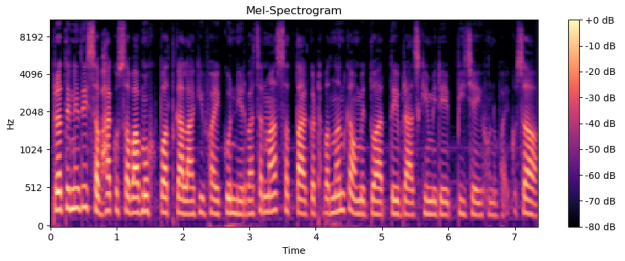

Figure 4: Mel-spectrogram plot of target audio

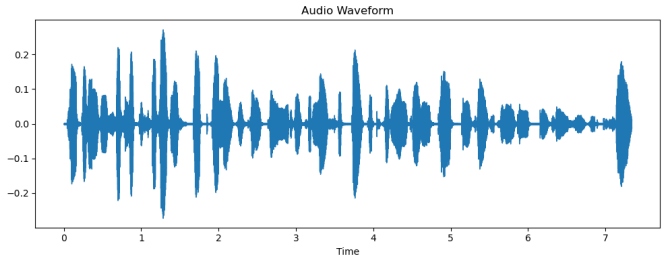

Figure 5: Audio waveform plot of target audio

2. Plot of synthesized text: प्रतिनिधि सभाको सभामुख पदका लागि भएको निर्वाचनमा सत्ता गठबन्धनका उम्मेदवार देवराज घिमिरे विजयी हुनुभएको छ।

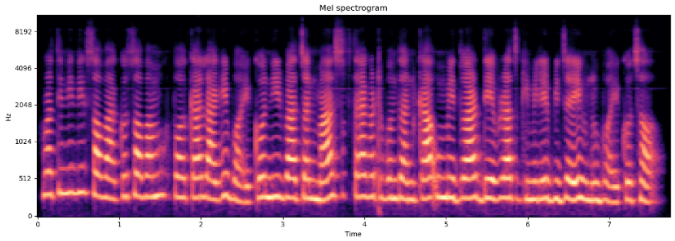

Figure 6: Mel-spectrogram plot of synthesized audio

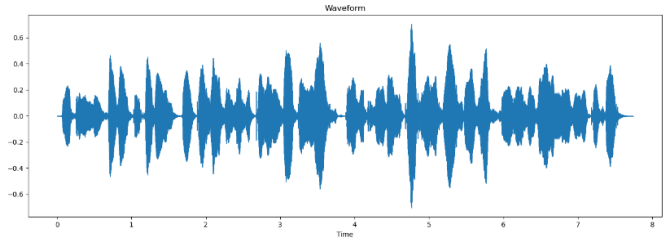

Figure 7: Audio waveform plot of synthesized audio

5. Conclusion

This work has successfully demonstrated the feasibility of constructing an end-to-end Nepali Text-To-Speech synthesis network utilizing deep neural networks. Different sets of hyperparameters were used during the training process. With the use of two stage approach, we succesfully generated the output audio waveform with intermediate Mel-spectrogram result which was somehow similar to the targeted audio waveform and its plot.

Acknowledgments

This work is carried out under the Department of Electronics and Computer Engineering, IOE, Pulchowk Campus. The authors are grateful to the Nepal Television (NTV) for their support in providing the news data for Dataset preparation.

References

[1] J.O. Onaolapo, Francis Idachaba, Joseph Badejo, Tiwalade Odu, and O.I. Adu. A simplified overview of text-to-speech synthesis. *Lecture Notes in Engineering and Computer Science*, 1:582–584, 07 2014.

[2] Roop Bajracharya, Santosh Regmi, Bal Krishna Bal, and Balaram Prasain. Building a natural sounding text-to-speech system for the nepali language - research and development challenges and solutions. pages 152–156, 08 2018.

[3] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech, 2019.

[4] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville. Melgan: Generative adversarial networks for conditional waveform synthesis, 2019.

[5] Brad Story. *History of Speech Synthesis, Chapter 1, Routledge Handbook of Phonetics, W. Katz and P. Assmann, Eds, 2019*, pages 9–32. 01 2019.

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

[7] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis, 2018.

[8] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. Samplernn: An unconditional end-to-end neural audio generation model, 2017.

[9] Mingjian Chen, Xu Tan, Yi Ren, Jin Xu, Hao Sun, Sheng Zhao, Tao Qin, and Tie-Yan Liu. Multispeech: Multi-speaker text to speech with transformer, 2020.

[10] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Ming Liu, and Ming Zhou. Neural speech synthesis with transformer network, 2019.

[11] Zhenhou Hong, Jianzong Wang, Xiaoyang Qu, Jie Liu, Chendong Zhao, and Jing Xiao. Federated learning with dynamic transformer for text to speech, 2021.

[12] Li-Wei Chen and Alexander Rudnicky. Fine-grained style control in transformer-based text-to-speech synthesis, 2022.

[13] Nilesh FalDessai, Gaurav Naik, and Jyoti D. Pawar. Review of syllable based text to speech systems: Strategies for enhancing naturalness for devanagari languages. 2017.

[14] Omer Nawaz and Tania Habib. Hidden markov model (hmm) based speech synthesis for urdu language. 2014.

[15] Humayun Kabir. Natural language processing for urdu tts system. 2002.

[16] Anusha Prakash and Hema A. Murthy. Generic indic text-to-speech synthesisers with rapid adaptation in an end-to-end framework. In *Interspeech 2020*. ISCA, oct 2020.

[17] Bhusan Chettri and Krishna Shah. Nepali text to speech synthesis system using esnola method of concatenation. *International Journal of Computer Applications*, 62:24–28, 01 2013.

[18] Ashok Banset, Basanta Joshi, and Suman Sharma. Deep learning based voice conversion network. 10 2021.

[19] Keshan Sodimana, Knot Pipatsrisawat, Linne Ha, Martin Jansche, Oddur Kjartansson, Pasindu De Silva, and Supheakmungkol Sarin. A Step-by-Step Process for Building TTS Voices Using Open Source Data and Framework for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 66–70, Gurugram, India, August 2018.