Impacts of Vocoder Selection on Tacotron-based Nepali Text-To-Speech Synthesis

Ganesh Bdr. Dhakal Chhetri[1], Kiran Chandra Dahal[1], and Prakash Poudyal[*2]

[1] *Informatics and Intelligent System Engineering,*
IOE, Thapathali Campus, Kathmandu, Nepal

[2] *Information and Language Processing Research Lab (ILPRL)*
Department of Computer Science and Engineering
Kathmandu University, Nepal
ganesh.078msiise007@tcioe.edu.np, dahalkc@ioe.edu.np, prakash@ku.edu.np

Abstract

Text-to-speech (TTS) technology enhances human-computer interaction and accessibility. While vocoders like WaveNet and MelGAN have been extensively studied for English TTS, their application to Nepali TTS remains underexplored. This research addresses this gap by evaluating the performance of WaveNet and MelGAN vocoders for Nepali text-to-speech synthesis using mel-sepctrograms generated by the Tacotron2 model.

The analysis is based on two datasets: Nepali OpenSLR and News male voice recordings. Performance was measured using Mean Opinion Score (MOS) and Mel-Ceptral Distortion (MCD). The findings reveal that Tacotron2 with MelGAN achieved better naturalness and accuracy compared to Tacotron2 with WaveNet. On the Nepali OpenSLR dataset, Tacotron2 + MelGAN achieved an average MOS score of 4.245, while Tacotron2 + WaveNet scored 3.65. Similarly, on the male voice dataset, Tacotron2 + MelGAN achieved an MOS of 2.885, compared to 2.31 for Tacotron2 + WaveNet.

1 Introduction

Text-To-Speech technology allows machines to turn text into human-like speech. This technology is widely used in applications such as virtual assistants, educational tools, and accessibility solutions (Tan et al., 2024). Recent advancements in deep learning have greatly improved the naturalness and quality of TTS systems (Shen et al., 2018). Tacotron2 is one such model, generating clear and natural speech by converting text into spectrograms (Shen et al., 2018). However, turning these spectrograms into actual audio depends on vocoders, which impact both the quality and speed of the final speech (Van den Oord et al., 2016; Kumar et al., 2019).

TTS technology has evolved significantly, transitioning from rule-based systems to deep learning-driven approaches that produce more natural and intelligible speech (Tan et al., 2024). Among these advancements, Tacotron2 has emerged as a leading model for TTS, using a sequence-to-sequence architecture to convert text into spectrograms, which are then transformed into audio waveforms by vocoders (Shen et al., 2018). Vocoder selection is particularly important for optimizing TTS systems, with different vocoders offering unique strengths.

This paper focuses on using Tacotron2 for Nepali TTS, a language that has seen little development in this area. It compares two vocoders, MelGAN and WaveNet, to evaluate their performance in speech quality and processing speed, aiming to improve TTS for Nepali and similar low-resource languages.

2 Literature Review

Text-to-Speech technology plays crucial role in converting written text into spoken language. It is widely used in applications such as voice assistants, educational aids (Klein et al., 2020), and assistive technologies for individuals with visual impairments (Manirajee et al., 2024).

The concatenation-based approach on the Nepali Text-to-Speech(TTS) has been studied by Ghimire and Bal (2017). The author used existing TTS and enchanced the quality by adding the pre and post processing units. The transformer based Nepali TTS has been published recently by Dongol and Bal (2023). They have achieved a Mean Opinion Score (MOS) of 3.70. Dhakal Chhetri (2023)

*Corresponding Author: prakash@ku.edu.np

Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), pages 185–192
January 19, 2025.

explores the capabilities of deep learning techniques for synthesizing Nepali Text-to-Speech using Tacotron. The goal is to develop a system that produces speech that sounds real. The researchers achieved successful voice output by creating a new Nepali speech dataset and building a model based on Tacotron1. However, the research is deficient in terms of comparisons with alternative vocoders and comprehensive data analysis. This initial research establishes the groundwork for future studies to enhance the model, explore alternative vocoders, and ultimately create more resilient Nepali text-to-speech systems.

In the research, Tan et al. (2024) investigate a new method for Text-to-Speech synthesis by employing a Variational Autoencoder (VAE) in order to produce a voice that is comparable to human quality. Their approach utilizes pre-trained phoneme sequences and a duration predictor to build speech. Unlike conventional autoregressive models such as Tacotron or FastSpeech, this VAE-based system utilizes a non-autoregressive structure, resulting in considerably faster speech creation. The model is trained on the LJSpeech dataset, which undergoes meticulous pre-processing to transform text into phonemes and generate Mel-Spectrograms. The objective of this research is to narrow the underlying gap between existing text-to-speech systems and authentic human speech in terms of quality. This work presents systematic research and the development of NaturalSpeech, with the goal of improving the performance and quality of text-to-waveform production in TTS technology.

Basnet (2021) study utilizes deep learning techniques to develop a Nepali Text-to-Speech synthesis system. The method employs a two-stage technique to transform written Nepali text into speech that sounds natural. Initially, convolutional neural networks examine the text and make predictions about spectrograms. Afterwards, recurrent neural networks equipped with attention mechanisms utilize these spectrograms to produce audio waveforms, prioritizing essential segments of the text to ensure precision. The study investigates the use of convolutional neural networks to predict spectrograms, recurrent neural networks to generate voice, attention mechanisms to focus, and advanced signal processing techniques to refine the results. The model, trained on a dataset of Nepali speech and validated using both subjective and objective approaches, showcases the efficacy of this deep learning technique for Nepali Text-to-Speech. Existing Nepali Text-to-Speech systems face difficulties in generating speech, as they are unable to accurately capture the intricate details of the language (Subedi, 2015). In order to address this disparity, this study suggests an end-to-end deep learning approach. Their network comprises several essential elements: text normalization for consistent processing, an encoder-decoder architecture for fundamental text-to-speech conversion, attention mechanisms to concentrate on significant elements in the text, and a WaveNet vocoder to convert the generated representation into audio. However, the research acknowledges a constraint: the lack of Nepali speech samples restricts further progress. They emphasize the importance of having additional Nepali datasets that are easily accessible.

Additionally, vocoders are integral to the success of Text-to-Speech systems, as they enable the production of high-quality audio. However, many TTS studies typically concentrate on a single vocoder. This research seeks to bridge the gap by evaluating the effectiveness of various vocoders in a Tacotron-based model, with the objective of achieving optimal Nepali audio synthesis.

3 Experimental Workflow

3.1 System Block Diagram

The block diagram in Figure 1 illustrates the overall system flow.

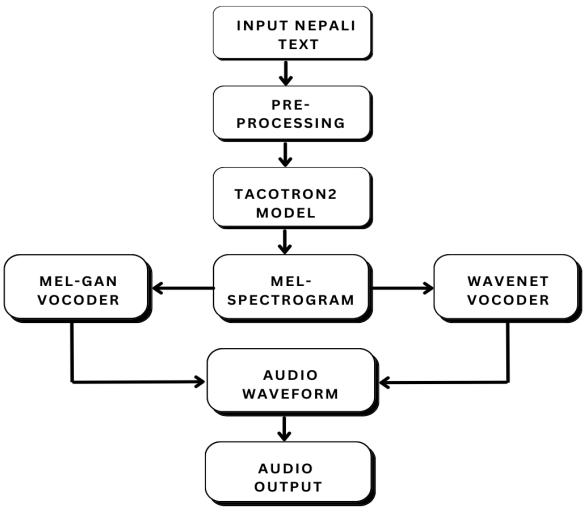

Figure 1: Block diagram for end-to-end TTS synthesis

Pre-Processing: Before inputting the text into the Tacotron2 model, preprocessing is carried out. During the preprocessing stage of this research,

the input text is transformed into Unicode. The 'unidecode' function is employed to transform Nepali text into Unicode format. Here's an example of how to convert Nepali text into Unicode before inputting it into the Tacotron2 model:

Original Nepali Text: प्रमुख अन्तर्राष्ट्रिय खेल घटना

Converted Unicode Text: prmukh antrraassttriy khel ghttnaa

Tacotron2 Model: Character embedding is applied to the input data of a Tacotron2 model before it is used for training. It is a crucial preliminary phase in Tacotron2. The character embedding layer is enhanced with pre-trained vectors. The system converts individual Nepali characters into numerical representations that accurately capture the meaning and context of the language. To create a numerical representation of the term, Character embedding examines the arrangement of characters using a one-dimensional Convolutional Neural Network. For example, the word 'paani' can be decomposed into individual characters: pa, aa, ni, i. An embedding vector can represent each character. pa: 0.7, 0.2, 0.4, aa: 0.1, 0.9, 0.2, ni: 0.2, 0.5, 0.1, i: 0.4, 0.2, 0.6

Mel-Spectrogram: Tacotron2 uses multiple layers, such as character embedding and encoder-decoder networks, to process input text. It generates an important representation called the Mel-Spectrogram. The Mel-Spectrogram captures the intensity of varoius frequencies in speech over time. The frequency axis is represented using Mel scale, which reflects the non-linear nature of human auditory perceptionn. During training, the model learns from paired samples of text and corresponding Mel-Spectrograms. These spectrograms are derived from real voice recordings. Using its encoder-decoder structure, Tacotron2 maps the textual input to the Mel-Spectrogram representation. This allows it to generate natural and intelligible speech.

Vocoders for Speech Waveform Generation: Once Tacotron2 has produced the Mel-Spectrogram, a distinct model known as a vocoder is employed to construct the ultimate voice waveform. In this research, the vocoder models WaveNet and MelGAN are employed in conjunction with Tacotron2.

WaveNet Vocoder: WaveNet is a predictive model that uses the previously generated audio samples and Mel-Spectrogram information to forecast the next audio sample in the speech waveform. It utilizes a sequential prediction approach to effectively capture the complex temporal relationships seen in speech. This paper employs WaveNet to synthesize speech audio waveforms by employing the Mel-Spectrogram produced by the Tacotron2 model.

MelGAN Vocoder: MelGAN is a vocoder model that utilizes Generative Adversarial Networks (GANs) as its basis. The system consists of two networks: a generator and a discriminator. The generator's objective is to produce authentic speech waveforms using the Mel-Spectrogram, whereas the discriminator's goal is to differentiate between genuine and created speech. The adversarial training procedure facilitates the acquisition of the ability for MelGAN to generate speech of high quality. In this research, the MelGAN model, trained separately, can generate audio waveforms from the Mel-Spectrogram generated by the Tacotron2 model, just like WaveNet.

3.2 Dataset

To develop a natural-sounding Nepali speech synthesis system, high-quality audio data with minimal background noise and clear pronunciation was essential. The research collected Nepali audio recordings paired with corresponding transcriptions. Audio files were encoded in .wav format and segmented into fragments of 1 to 12 seconds using Audacity software (Zen et al., 2013). This segmentation aligned with the natural rhythms, making it easier for the model to capture authentic patterns and reducing computational resource requirements (Van den Oord et al., 2016).

The dataset included transcriptions for each segment and recordings from multiple male news presenters to enhance diversity. Both the OpenSLR dataset (Sodimana et al., 2018) and the male voice data collected from Nepal Television were used separately for this research. Details about the male voice dataset from news recording are provided in Table 1.

3.3 Model Prepared

This research used the Tacotron2 model along with the MelGAN and WaveNet vocoders. The

Table 1: News Male Voice Data

Feature	Description
Speaker	Multiple male speakers
Source	Nepal Television broadcasts
Format	Varies, resampled to 22.05 kHz
Size	Approximately 2100 samples
Data Split	80% training, 20% validation/testing

OpenSLR dataset and a male voice dataset were used separately for analysis. The training process followed the parameters setup described by (Kumar et al., 2019; Shen et al., 2018; Van den Oord et al., 2016).

3.3.1 Tacotron2 Model

The Tacotron2 model's effective training depends on key audio parameters that determine its interpretation of raw audio input. For the training process, a learning rate of 0.001 and a batch size of 8 were employed.

3.3.2 MelGAN Model

The MelGAN model was trained by adjusting multiple training parameters, as detailed in Table 2.

Table 2: MelGAN Training Parameters

Parameter	Value
Batch Size	16
Learning Rate	0.0001
Optimizer	Adam
Beta-1	0.5
Beta-2	0.9

3.3.3 WaveNet Model

Similarly, the WaveNet model was trained with various hyperparameters, as detailed in Table 3.

Table 3: WaveNet Training Parameters

Parameter	**Value**
Learning Rate	0.001
Batch Size	8
Optimizer	Adam

3.4 Verification and Validation

3.4.1 Qualitative Approach

A qualitative methodology was employed to compare the audio waveforms generated by the Tacotron2 + MelGAN model and the Tacotron2 + WaveNet model with the original test data audio waveforms. This comparison was done individually for both OpenSLR and male voice data.

3.4.2 Quantitative Approach

The Mel Cepstral Distortion (MCD) metric is an objective and quantitative measure. The process involves comparing the melspectrograms of generated speech with those of recorded speech. Mathematically it can be computed as in Equation(1).

$$MCD = \sqrt{2 \sum ((mel_1 - mel_2)^2)} \quad (1)$$

where, mel_1 and mel_2 represent the respective melspectrograms.

The MCD values were obtained for both the OpenSLR and male voice datasets using the Tacotron2 + MelGAN and Tacotron2 + WaveNet models. Subsequently, a comparison was conducted between the MCD values acquired for each model and dataset.

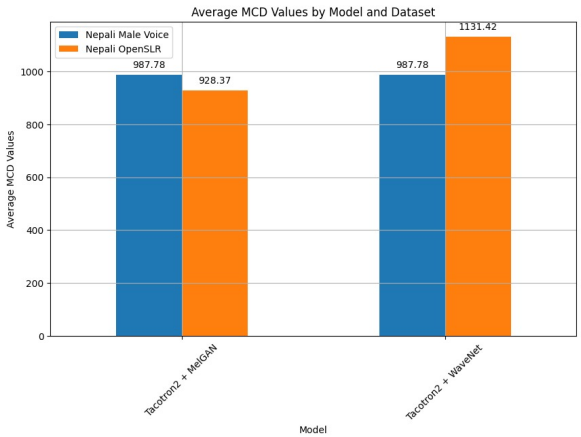

Figure 2: Average MCD values by Model and Dataset

The average MCD values for the Tacotron2 + MelGAN model were 928.37 on the OpenSLR dataset and 987.87 on the News male voice data. On the other hand, the Tacotron2 + Wavenet model yielded mean MCD values of 1131.42 and 987.87, respectively. The findings indicate that Tacotron2 + MelGAN generally outperforms in terms of voice quality on the OpenSLR dataset, however both models demonstrate similar performance on the News male voice data. Figure 2 shows a detailed

analysis.

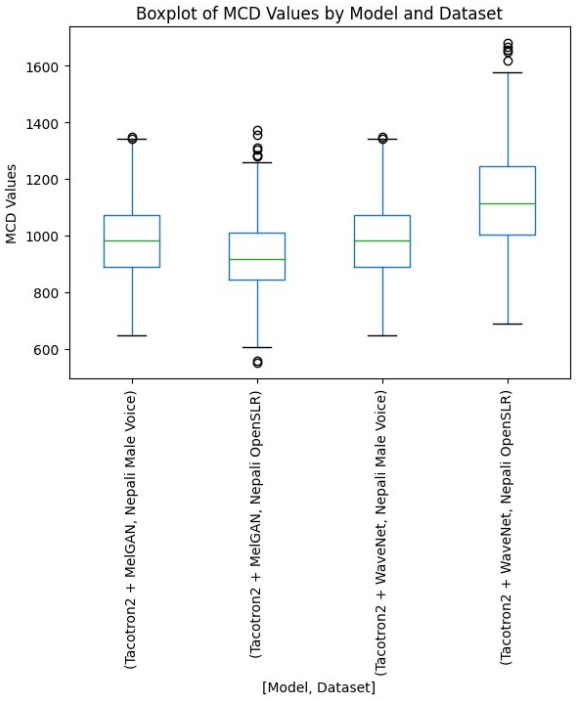

Figure 3: Boxplot of MCD values by Model and Dataset

MelGAN and WaveNet were directly compared to see how well they reproduce target speech. A boxplot in Figure 3 displays their MCD values side by side. The boxplot analysis illustrates that both the Tacotron2 + MelGAN and Tacotron2 + WaveNet models demonstrate almost equal median values for the news male voice dataset. However, the quantity of outliers is significantly reduced in comparison to the OpenSLR dataset. In the OpenSLR dataset, the Tacotron2 + MelGAN model exhibits a relatively lower median value. A lower median value typically indicates superior model performance (Kubichek, 1993). Consequently, drawing from this quantitative analysis of the dataset and model, it can be inferred that the Tacotron2 + MelGAN model attains better results on the OpenSLR dataset.

4 Results

In order to accomplish to assess the efficacy of MelGAN and WaveNet vocoders when used together with the Tacotron2 model for synthesizing Nepali text into speech, a model was first trained and evaluated using the predominantly female-voiced Nepali OpenSLR dataset. Afterwards, the model was retrained using male voice data obtained from Nepal Television.

4.1 OpenSLR Data (Model Training)

The Tacotron2 model completed its training after 112,000 steps, which took around 54 hours. The ultimate training loss reached a convergence point of 0.314. Significantly, the validation loss at step 112,000 was measured to be 0.459. After the completion of training the Tacotron2 model on the OpenSLR dataset, the training of the MelGAN model was initiated. The training procedure of MelGAN involves a dynamic interaction between the generator and discriminator. In this research, the MelGAN training ended after completing 516,444 steps. The generator loss was 89.587, and the discriminator loss was 59.432 at this stage.

This study also trained the WaveNet model to assess the effectiveness of different vocoders. During the training process, the model's performance is evaluated by considering both the training loss and validation loss. The training procedure finished after approximately 1,000 epochs, which is equivalent to 220,000 training steps. The ultimate loss at the end of the training phase was 4.818. The validation loss at this point was 5.436.

4.2 Male Voice Data (Model Training)

The Tacotron2 model underwent additional training after being trained on the OpenSLR dataset, which consisted of male voices data from various speakers on Nepal Television. The model training process involved monitoring the loss for both training and validation data. The number of training steps equaled to 70,000 in total. At this stage, the training loss value achieved was 0.28. The validation loss value achieved during this training stage was 0.65. Figure 4 displays the training and validation graph for the male voice dataset using the Tacotron2 model.

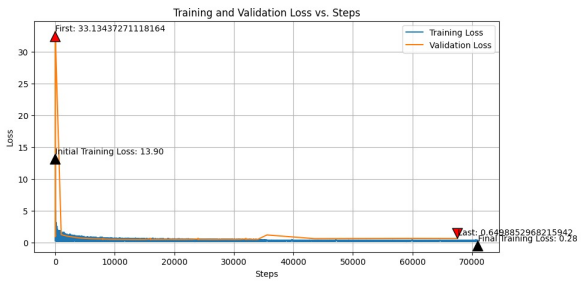

Figure 4: Tacotron2 training & validation loss

The MelGAN model was retrained with Nepali news male voice data, just as it was trained on the OpenSLR dataset. The model underwent a training

process, during which it completed 600,000 steps. After training the models, the combined Tacotron2 and MelGAN models were used to generate the final speech output by running inference on the Nepali news male voice test dataset. An assessment of the loss generated by the generator and discriminator was carried out during the complete training procedure of the MelGAN model. At the $600,000^{th}$ step of training, the generator's training loss had reached 201.84, while the discriminator's training loss was 52.92. The generator and discriminator have validation losses of 319.35 and 208.89 respectively. Figure 5 and 6 illustrate the training and validation losses for the generator and discriminator.

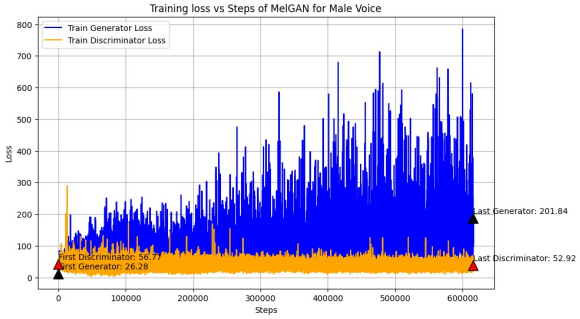

Figure 5: Generator, Discriminator training loss

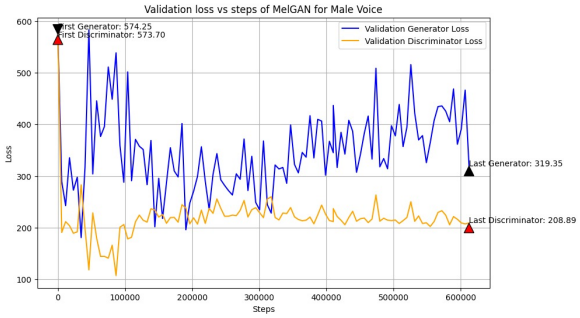

Figure 6: Generator, Discriminator validation loss

The training and validation losses of the WaveNet model are illustrated in Figure 7. The WaveNet model underwent training using the male voice dataset for a duration of up to 250,000 steps. The training loss value was 4.77, while the validation loss value was 5.24 at this stage.

After conducting independent training of the models using the OpenSLR and male voice data, the Tacotron2 + MelGAN model and the Tacotron2 + WaveNet model were utilized to perform inference on the test data from both OpenSLR and male voice datasets. This resulted in predictions for speech synthesis.

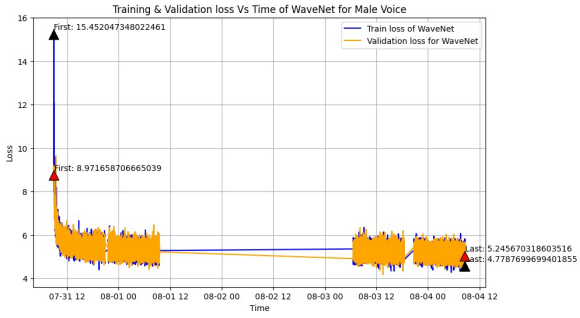

Figure 7: WaveNet training & validation loss

5 Discussion and Analysis

5.1 Inference Time Comparison: MelGAN vs. WaveNet

The speech synthesis on the test data involved the utilization of two models: Tacotron2 + MelGAN and Tacotron2 + WaveNet. Both the Tacotron2 + MelGAN and Tacotron2 + WaveNet models were inferred using an NVIDIA GTX 960M GPU. The average time required for each synthesis process is compared in Table 4.

Table 4: Inference Time Comparison

Model	Inference Time (s)
Tacotron2 + MelGAN	0.142
Tacotron2 + WaveNet	1320

The processing design of MelGAN provides a clear advantage over WaveNet in terms of its speed. Compared to WaveNet, which produces output in a sequential manner, MelGAN is capable of processing the full input at the same time. By using parallel processing, MelGAN achieves a reduction in inference time, making it a highly efficient option.

5.2 MOS Score Comparison: WaveNet vs. MelGAN

The MOS score was calculated using data collected through Google Forms. Each model produced four samples for evaluation. Participants rated the samples on a scale of 1 to 5, assessing naturalness and accuracy. A total of 40 participants provided the ratings used to calculate the final MOS score. The MOS scores for naturalness and accuracy between Tacotron2 + MelGAN and Tacotron2 + WaveNet on the OpenSLR dataset are presented in Table 5. Similarly, the MOS scores for the male voice dataset are shown in Table 6.

Table 5: MOS Score Comparison (OpenSLR Data)

Model	Naturalness	Accuracy
Tacotron2 + MelGAN	4.21	4.28
Tacotron2 + WaveNet	3.56	3.74

Table 6: MOS Score Comparison (Male Voice Data)

Model	Naturalness	Accuracy
Tacotron2 + MelGAN	2.97	2.80
Tacotron2 + WaveNet	2.33	2.29

These findings indicate that the Tacotron 2 + MelGAN model outperforms the Tacotron 2 + WaveNet model in terms of both the naturalness and accuracy of the generated output.

The OpenSLR dataset, which predominantly emphasizes female voices, demonstrates reduced variability in pitch and frequency across different speakers. In contrast, the inclusion of male voice data, which consists of various speakers and different tones even within individual speakers for distinct transcripts, presents more difficult obstacles for models to generalize data. The intrinsic variability of male voice data and the limited amount of data have a substantial impact on MOS scores, resulting in lower overall ratings. Moreover, the complexity of male voice data could result in increased noise production during the process of speech synthesis.

5.3 Comparison with Existing System

The research titled "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Prediction" by (Shen et al., 2018), stated that the WaveNet model achieved a Mean Opinion Score of 4.53 when it was trained using the US English dataset.

In the study "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis" conducted by (Kumar et al., 2019), the MelGAN model demonstrated a Mean Opinion Score of 3.61 ± 0.06 after being trained on the LJ Speech dataset. After being trained on the VCTK dataset, the MelGAN model got a Mean Opinion Score of 3.49 ± 0.09. Similarly, in the research "Attention and WaveNet Vocoder Based Nepali Text-to-Speech Synthesis" by (Basnet, 2021), a MOS of 3.07 was forecasted.

Table 7 presents the average MOS score calculated by taking the mean of the naturalness and accuracy rating in this research.

Table 7: Average MOS Score (OpenSLR & Male Voice Data)

Model	Dataset	Average MOS
Tacotron2 + MelGAN	OpenSLR	4.245
Tacotron2 + MelGAN	Male Voice	2.885
Tacotron2 + WaveNet	OpenSLR	3.65
Tacotron2 + WaveNet	Male Voice	2.31

The Tacotron2 + MelGAN model achieved an average MOS score of 4.245 for the Nepali OpenSLR dataset. This number is greater than the MOS scores presented for both the LJ speech data and VCTK English datasets. Nevertheless, the mean opinion score achieved for male voice data using the Tacotron 2 + MelGAN model is comparatively lower than the MOS scores obtained in previous studies (Kumar et al., 2019) for both LJ Speech data and VCTK English datasets using the MelGAN model.

Table 8: MOS Score Comparison (WaveNet Model)

Model	Average MOS
WaveNet (US English)	4.53
Existing Nepali TTS ((Basnet, 2021))	3.07
Tacotron2 + WaveNet (Nepali OpenSLR)	3.65
Tacotron2 + WaveNet (Male Voice)	2.31

Based on the data presented in Table 8, the WaveNet model achieved the highest Mean Opinion Score when trained on the English dataset. Its performance surpassed that on the Nepali OpenSLR and male voice datasets. It also outperformed results from prior Nepali Text-To-Speech studies.

6 Conclusion and Future work

The research findings clearly demonstrate that the selection of a vocoder has significant impacts on the performance of a Nepali Text-to-Speech system. When combined with Tacotron2, MelGAN consistently outperformed WaveNet in terms of the naturalness, accuracy, and overall quality of speech. The superiority of the Tacotron2 + MelGAN model is apparent by the higher Mean Opinion Scores (MOS) and lower Mel-cepstral Distortion (MCD) it achieves.

Furthermore, MelGAN exhibited a significant benefit in terms of inference time. Compared to WaveNet, MelGAN has a lower computing time need for voice generation, making it a more efficient option for real-time applications. However, the research was hindered by the scarcity of Nepali datasets with multiple-speaker male voices. This impeded the assessment of the model's effectiveness on a broader spectrum of datasets. Moreover, the model encountered difficulties in precisely articulating numbers and symbols, highlighting a constraint that should be tackled in future investigations.

Although there are several restrictions, the Tacotron2 + MelGAN model is considered the best setup for creating Nepali TTS systems of superior quality. Future research can improve efficiency by optimizing single-speaker male voice datasets. Customizing the model to handle numbers and symbols can enhance usability. Exploring context-based approaches and hybrid architectures, like combining Tacotron2 with Transformer-based models, may further advance Nepali TTS systems.

Acknowledgments

The authors would like to extend sincere thanks to the reviewers for their constructive comments and suggestions.

References

Ashok Basnet. 2021. Attention and wavenet vocoder based nepali text-to-speech synthesis. Master's thesis.

Ganesh Bdr. Dhakal Chhetri. 2023. Nepali text to speech using tacotron. Master's thesis, Thapathali Campus, Tribhuvan University.

Ishan Dongol and Bal Krishna Bal. 2023. Transformer-based Nepali text-to-speech. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 651–656, Goa University, Goa, India. NLP Association of India (NLPAI).

Rupak Raj Ghimire and Bal Krishna Bal. 2017. Enhancing the Quality of Nepali Text-to-Speech Systems. In Alla Kravets, Maxim Shcherbakov, Marina Kultsova, and Peter Groumpos, editors, *Creativity in Intelligent Technologies and Data Science*, volume 754, pages 187–197. Springer International Publishing.

Andreas Klein, Andreas Hinderks, Maria Rauschenberger, and Jorg Thomaschewski. 2020. Exploring voice assistant risks and potential with technology-based users.

Robert F. 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. IEEE.

Kundan Kumar, Ritesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C. Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in Neural Information Processing Systems*, 32.

Lalitha Manirajee, Siti Qatrunnada Hanis Shariff, and Syar Meeze Mohd Rashid. 2024. Assistive technology for visually impaired individuals: A systematic literature review (slr). *International Journal of Academic Research in Business and Social Sciences*, 14.

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 *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

Keshan Sodimana, Knot Pipatsrisawat, Linne Ha, Martin Jansche, Oddur Kjartansson, Pasindu De Silva, and Supheakmungkol Sarin. 2018. A step-by-step process for building tts voices using open source data and framework for bangla, javanese, khmer, nepali, sinhala, and sundanese. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 66–70, Gurugram, India. Association for Computational Linguistics.

Kaushal Subedi. 2015. Nepali text-to-speech. Master's thesis.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. 2024. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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.

Heiga Zen, Andrew Senior, and Mike Schuster. 2013. Statistical parametric speech synthesis using deep neural networks. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7962–7966. IEEE.