Received Date:24th May, 2024
Revision Date:15th June, 2024
Accepted Date:23rd July, 2024

Nepali Text-to-Speech Synthesis Using Tacotron2 and WaveGlow

Ashma Rai[1*], Shikshya Shiwakoti[2], Swostika Basukala[3], Er. Suramya Sharma Dahal[5]

[1]*Dept of Electronics and Computer Engineering, Thapathali Campus, IOE, TU, E-mail:* ashma.rai85@gmail.com
[2]*Dept of Electronics and Computer Engineering, Thapathali Campus, IOE, TU, E-mail:* shikshiwakoti@gmail.com
[3]*Dept of Electronics and Computer Engineering, Thapathali Campus, IOE, TU, E-mail:* swostikabasukala1@gmail.com
[4]*Associate Professor, Dept of Electronics, Communication & Information Engineering, Kathmandu Engineering College, E-mail:* suramya.sharma@kecktm.edu.np

Abstract— **This research paper presents the development of a Nepali Text-to-Speech (TTS) system under low-resource conditions by adapting pre-trained English Tacotron2 and WaveGlow models. Tacotron2 has been utilized for spectrogram generation, and WaveGlow has been employed for vocoding, with recognition of the pivotal role played by these components in determining the efficacy of a Text-to-Speech (TTS) system. Our approach entails the adaptation of a pre-trained English Tacotron2 model and WaveGlow architecture to Nepali, leveraging limited data resources to craft a Nepali TTS system capable of producing natural-sounding output under low-resource conditions. Through fine-tuning with a Nepali text corpus aligned with its corresponding audio dataset, the pre-trained Tacotron2 model is optimized for spectrogram generation. Subsequently, WaveGlow, our chosen audio synthesis model, is utilized to convert the spectrogram representations into audible waveforms. It is worth noting that our model exhibits limitations in synthesizing audio for a restricted subset of Nepali texts, attributed to challenges stemming from text cleaning and normalization inadequacies.**

Keywords— Fine-tuning, Text-to-Speech, Synthesis, Tacotron2, WaveGlow

Introduction

Speech synthesis, the process of generating human speech through technology, is facilitated by speech synthesizers, which can be integrated into both hardware and software systems. These systems, commonly referred to as text-to-speech (TTS) systems, convert ordinary Nepali text into spoken language, with the capability to handle numerals, dates, and abbreviations within the input text. The social significance of speech synthesis is evident, particularly in aiding individuals with visual impairments or those who may struggle with traditional reading and writing methods. TTS systems facilitate automated conversion of text into speech, mimicking the natural cadence of a native speaker. These systems typically fall into two categories: limited domain TTS, designed for specific applications with a restricted vocabulary, and generic TTS, capable of reading any text from a document. Essential attributes of synthesized speech include naturalness and intelligibility, aiming to produce speech that is both comprehensible and reminiscent of human speech patterns.

Despite the advancements in speech synthesis technology, research on Nepali speech synthesis remains relatively underdeveloped. The Nepali language is characterized by 11 independent distinct vowels and 33 consonants, and it is primarily written using the Brahmi-descended Devanagari script. Notably, the Devanagari script lacks specific forms for capital letters and is written from left to right, presenting unique challenges and opportunities for speech synthesis research in the Nepali context.

Democratizing AI is a core aspiration in the AI community, yet biases persist in available resources. Tacotron 2 [1] by Google shows near-human performance for English and Chinese, but Nepali TTS remains underdeveloped, yielding subpar results due to limited linguistic resources. Our project aims to bridge this gap by leveraging pre-trained Tacotron with WaveGlow [2] to develop a high-quality Nepali TTS model. Key objectives include training a model for spectrogram-based audio synthesis and converting spectrograms into waveforms. Despite challenges with properly formatting Nepali text, the model holds promise for applications such as assistive technology, customer service, language learning, and more, underscoring its pivotal role in advancing Nepali TTS technology.

Related works

In previous research, various approaches have been explored in Text-to-Speech (TTS) synthesis methods. ATR Interpreting Telecommunications Research Lab published a research paper by A. J. Hunt and A. W. Black in 1996 titled "Unit Selection in a Concatenative Speech Synthesis System using a Large Speech Database" [3] that discussed how it is based on searching and gathering samples of speech components from the voice database. They produce speech synthesized with a more natural voice, however a large quantity of database is needed.

Concatenative synthesis, exemplified by the TD-PSOLA [4] method, has been a prominent technique. This method involves recombining pre-recorded speech segments to

[*] *Corresponding Author*

generate synthetic speech. While TD-PSOLA has shown promise in producing natural-sounding speech, challenges such as limited coverage in the speech dictionary have been observed, leading to unclear articulation for specific phonemes and errors due to data scarcity.

Parametric synthesis techniques, such as those utilizing Hidden Markov Models (HMMs), have also been investigated. These methods aim to model speech using generative models, thereby eliminating the need for a large speech database. However, the quality of the synthesized speech may not always match that of concatenative approaches.

Neural synthesis methods, including WaveNet [5] and Tacotron, have emerged as promising alternatives. WaveNet, as proposed by H.Zen and et al., operates directly on raw audio waveforms and has demonstrated the ability to generate highly realistic speech samples. Similarly, Tacotron [1], employs a sequence-to-sequence model with an attention mechanism to synthesize speech directly from text inputs.

While these methods have advanced the field of TTS synthesis, the TD-PSOLA approach remains particularly relevant for Nepali TTS systems. Further research and development in this area are crucial to address its limitations and enhance the naturalness and accuracy of synthesized Nepali speech.

Char2Wav [6] uses a bidirectional RNN to generate waveforms directly from text, eliminating the need for intermediate features or expert linguistic knowledge. It employs SampleRNN for spectrogram reconstruction and a connectionist temporal classification (CTC) loss function for training. This approach allows direct mapping from character sequences to waveforms. A CNN extracts features from text, and an RNN generates the waveforms. The model is trained end-to-end with backpropagation, using "teacher forcing" to enhance stability by providing the correct waveform at each training step.

Another recently-developed neural model, Deep Voice [7], replaces each component of a traditional TTS pipeline with neural networks, using five DNNs for text analysis, phoneme duration prediction, F0 prediction, spectral envelope generation, and waveform generation. This results in fewer parameters and faster performance compared to WaveNet. However, since each component is trained independently, end-to-end training is challenging.

Another notable paper on text-to-speech synthesis presents a model based on generative adversarial networks (GANs) [8]. This model uses a generator to create raw audio from text and a discriminator to distinguish real from synthesized audio, providing feedback to improve the generator. The GAN-TTS model excels in producing high-quality speech with diverse voices and styles and can be fine-tuned for specific traits like emotional tone or accent. However, it requires a large amount of training data and may produce unnatural speech if the data is imbalanced or the training is inadequate.

Tacotron 2 stands out as a favourable option for text-to-speech (TTS) generation compared to other TTS models due to its ability to produce natural and human-like speech. It utilizes a deep neural network architecture with attention mechanisms and a transformer decoder, enabling it to generate speech with accurate pronunciation, intonation, and rhythm. Additionally, Tacotron 2's capacity to model longer context enhances coherence in the generated speech. These advancements have made Tacotron 2 a popular choice for TTS applications, setting a new standard in the field. Therefore, we have selected Tacotron 2 as our primary mechanism for research to harness its capabilities in improving Nepali speech synthesis

Methodology

Our proposed methodology involves fine-tuning the Tacotron2 model on a Nepali dataset to adapt its parameters to the specific characteristics of the data. Concurrently, hyperparameter tuning optimizes various model settings, including learning rate and batch size, to enhance overall performance. The process begins with spectrogram generation, where input text from the Nepali dataset is transformed into spectrograms, capturing the sound characteristics of the spoken signal. These spectrograms are then used to train the Tacotron2 model, refining its ability to synthesize Nepali speech. Subsequently, the WaveGlow model synthesizes the spectrograms to generate synthetic speech. Hyperparameter tuning follows, aiming to optimize the model's settings for improved performance.

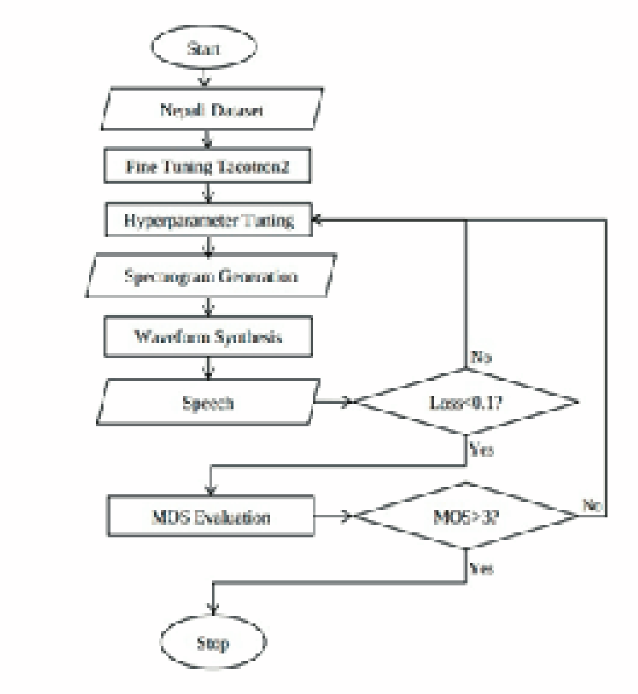

Fig. 1 Proposed Methodology Flowchart for Nepali TTS Synthesis

The system evaluates the effectiveness of the generated speech using the Mean Opinion Score (MOS) method, where human evaluators rate speech quality on a scale of 1 to 5. If the MOS falls below 3, indicating suboptimal performance, the process returns to hyperparameter tuning until the MOS surpasses 3. This comprehensive approach ensures the Tacotron2 and WaveGlow models are effectively trained and evaluated on a Nepali dataset, ultimately leading to high-quality speech synthesis.

A. Data Collection and Preparation

Our dataset preparation approach played a pivotal role in laying a robust foundation for our Nepali Text-to-Speech (TTS) system. We initiated the process by meticulously sourcing high-quality transcribed audio data, comprising WAV and TSV files. Our focus was on obtaining diverse recordings, spanning various speakers and linguistic variations, to ensure comprehensive coverage of Nepali speech patterns.

Following dataset collection, we embarked on a meticulous quality assurance process to refine the dataset further. Our objective was to eliminate any inaccuracies or imperfections within the dataset. Each file underwent thorough scrutiny, with erroneous or defective lines being removed, and audio clips were enhanced to ensure they were free from noise or distortion.

For dataset collection, we relied on OpenSLR's "High quality TTS data for Nepali," which provided a sizable dataset of top-notch audio recordings. This dataset, comprising tsv-formatted transcription texts and wav-format audio files, was meticulously curated to meet the rigorous standards of our TTS system. Each audio waveform underwent a sample rate adjustment to 22050 Hz to ensure compatibility with our system.

In the dataset preparation phase, we conducted a comprehensive review of the transcribed text, ensuring the absence of wrong or broken lines. Additionally, we evaluated the Signal to Noise ratio (SNR) of the voice clips, with an average SNR of 29.19 being considered excellent for TTS applications. The best-performing audios exhibited an SNR of 100, while the lowest recorded SNR was 11.

Moving forward, we leveraged transfer learning to fine-tune the pre-trained Tacotron 2 model on our Nepali dataset. This process involved adapting the model's parameters to better fit the characteristics of the Nepali language, enhancing its performance and efficacy for Nepali speech synthesis. Through meticulous dataset preparation and implementation details, we ensured the optimal functioning of our TTS system for the Nepali language.

B. Transfer Learning using Tacotron2

Initially, the weights of the pre-trained model were fine-tuned using Nepali data, accompanied by adjustments to the hyperparameters to better suit the characteristics of the language. This process involved training the new model for 50 epochs, during which we closely monitored the generated audios and analysed the attention plot to assess performance.

Through iterative refinement, we optimized key hyperparameters including batch size, evaluation batch size, batch group size, epochs, and save checkpoint step. A batch size of 8 was chosen to balance training efficiency and stability. With a larger batch size, the model may converge faster but may also suffer from overfitting, which is why we have lowered the batch size to 8. Eval batch size of 8 was chosen which indicates the number of examples processed at once during evaluation. We chose the evaluation batch size as the same as the training batch size, because it is a good practice for consistency and stability. Every 4 batches were grouped together and processed on the GPU simultaneously in order to improve training efficiency.

In our case, since the dataset is not as extravagant as the original dataset Tacotron2 was trained upon, 50 epochs was considered to be sufficient enough to fine-tune the Tacotron 2 model on our Nepali dataset. The save checkpoint step of 200 was taken to ensure the saving the model's progress during training after every 200 training steps, in case the training process is interrupted. With these adjustments, Tacotron 2 seamlessly adapted to Nepali, resulting in enhanced synthesis quality. Subsequently, the generated mel-spectrogram seamlessly transitioned to the WaveGlow model for audio waveform generation.

Results and Analysis

A. Fine-tuned Tacotron2 Results

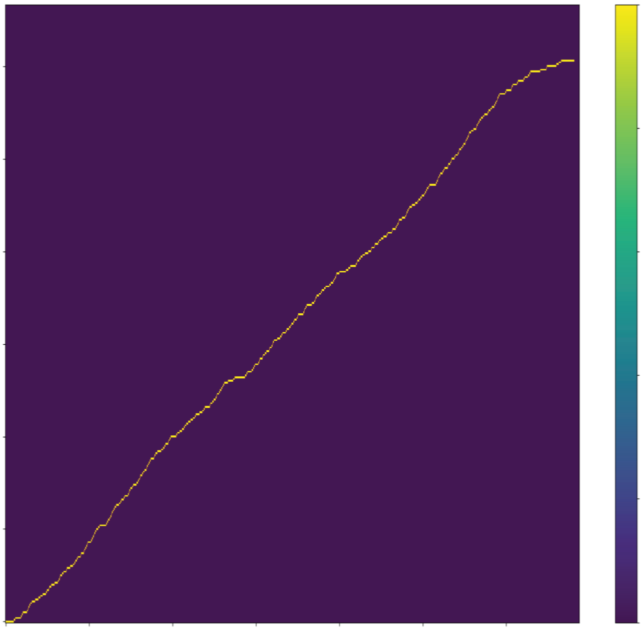

Fig. 2 Alignment Matrix

The graph of encoder timestep versus decoder timestep, often called the "alignment matrix" or "attention matrix," visually represents the alignment between input text and the generated speech waveform. Each element in the matrix corresponds to the attention weight assigned by the decoder to a specific input text feature when producing the output speech waveform at a given timestep. Typically displayed as a heatmap, darker colors indicate higher attention weights, while lighter colors denote lower attention weights. Analyzing this matrix provided insights into how the TTS model utilizes the input text to generate the speech waveform and how the attention mechanism aligns the two sequences.

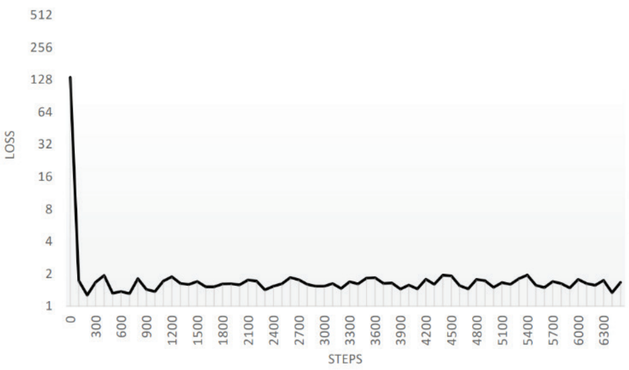

Fig. 3 Training Loss

The training loss graph provides valuable insights into the model's performance trajectory. Initially, a high loss of 134.669 indicated suboptimal performance, gradually decreasing over time. A significant drop to 1.734 after 100 steps marked substantial progress, but fluctuations persisted over the next 6400 steps. Overall, the graph showed improvement over time, albeit with ongoing challenges and fluctuations, indicating areas for optimization.

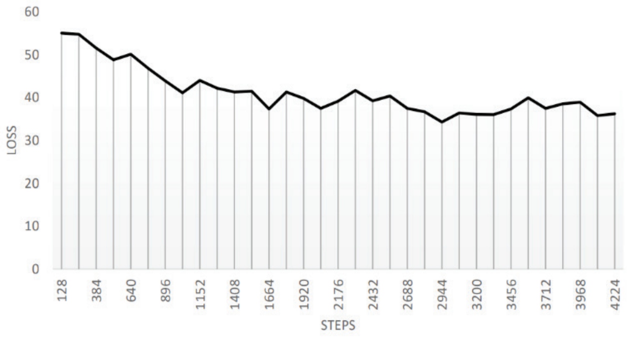

Fig. 4 Validation Loss

Analyzing the validation loss graph showed an initial decrease followed by fluctuations, likely due to dataset complexity or model learning challenges. The lowest validation loss occurred at step 2944 (34.3124), signifying optimal performance at that point. However, fluctuations suggest areas for improvement. Overall, the decreasing trend indicates progress, but further refinement may be needed.

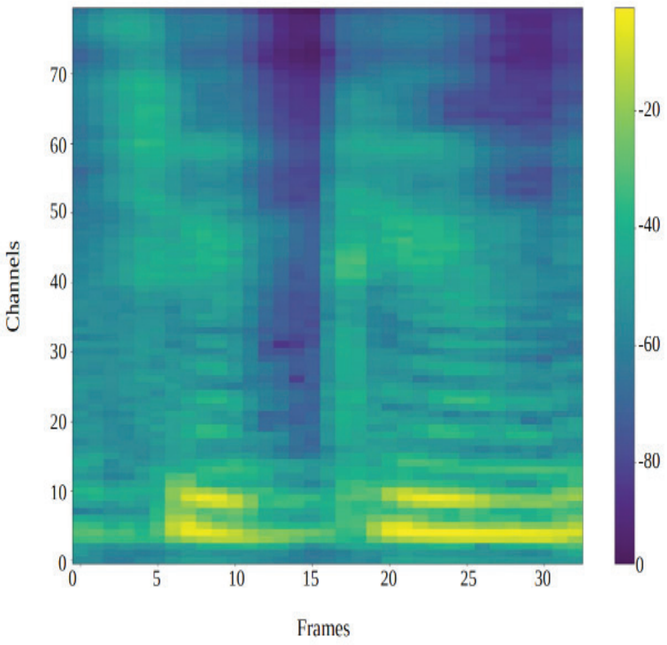

Fig. 5 Predicted Mel Spectrogram

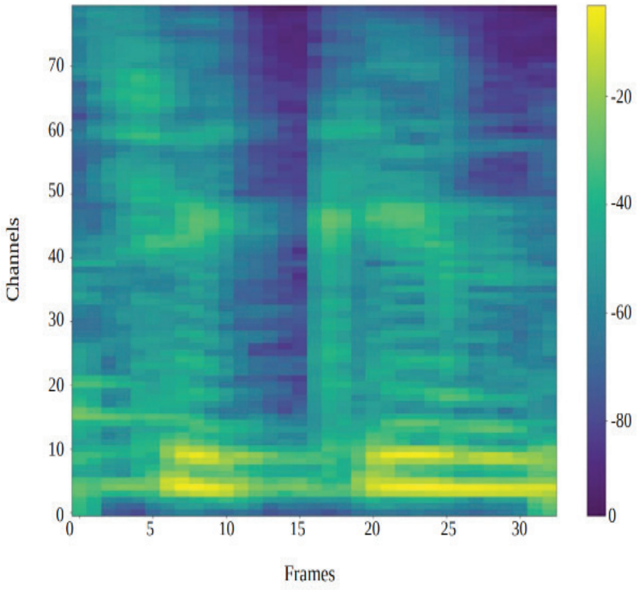

Fig. 6 Target Mel Spectrogram

The provided figure illustrates a training spectrogram generated during training, offering insight into the model's ability to predict spectrograms resembling real data.

Analyzing the Mel spectrogram histogram revealed the frequency content of the audio signal. A skew towards low frequencies suggests higher energy in low-frequency bands, while a skew towards high frequencies indicates higher energy in high-frequency bands. The given figure is of train spectrogram which is a spectrogram image generated during this training process. It could be used to visualize how well the model is predicting the spectrograms that resemble the spectrograms of real data.

B. WaveGlow Results

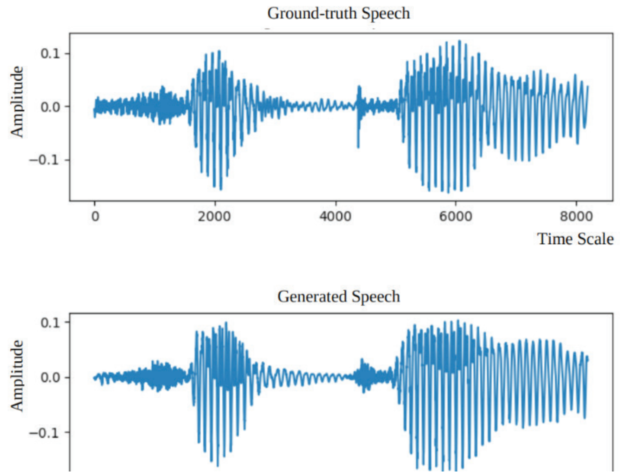

Fig. 7 WaveGlow Result Comparison

Comparing the ground truth and generated speech waveforms, we can see how closely it matches the original speech in terms of timing, intonation, and loudness.

C. Model Comparison

Following is the comparison of audio features, extracted from the audio synthesized using pretrained Tacotron2 model on the left and the audio synthesized by our finetuned model on the right, for the given Nepali text.

***Text**:* “यहाँबाट हिउँबारीको धेरै सुन्दर दृश्य देखिन्छ ।”

1) Spectrum Based Comparison

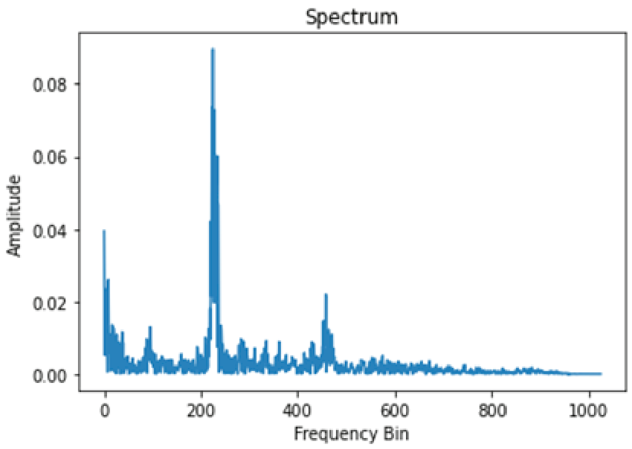

Fig. 8 Original Spectrum

Amplitude: The original audio signal has a lower amplitude peak around 0.08 units at the lowest frequency bin, indicating a much more balanced energy distribution.

Frequency Distribution: The spectrum shows a more natural spread of frequencies with a notable peak around 200 bins, followed by a gradual decline, which suggests the presence of natural harmonics and a richer audio quality.

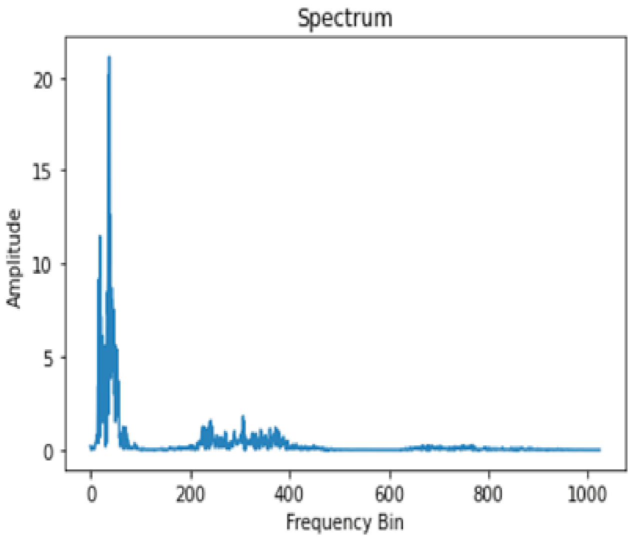

Fig. 9 Pretrained Tacotron2 Spectrum

Amplitude: The pre-trained Tacotron2 model exhibits a high amplitude peak at the lowest frequency bin, exceeding 20 units, which suggests a significant concentration of energy at this frequency.

Frequency Distribution: There are noticeable secondary peaks and a considerable spread of amplitude across various frequency bins up to around 200 bins, after which the energy quickly dissipates.

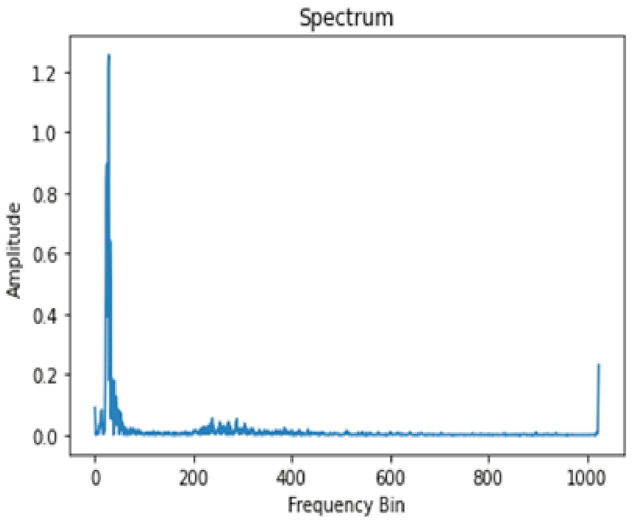

Fig. 10 Fine-tuned Tacotron2 Spectrum

Amplitude: The fine-tuned Tacotron2 model has a much lower amplitude peak at the lowest frequency bin, slightly above 1 unit, indicating a reduction in energy concentration at this frequency.

Frequency Distribution: The distribution of the frequency spectrum is more spread out with smoother transitions and fewer pronounced secondary peaks compared to the pre-trained model. The energy distribution seems to align more closely with the original audio's spectrum.

2) *Spectrogram Based Comparison*

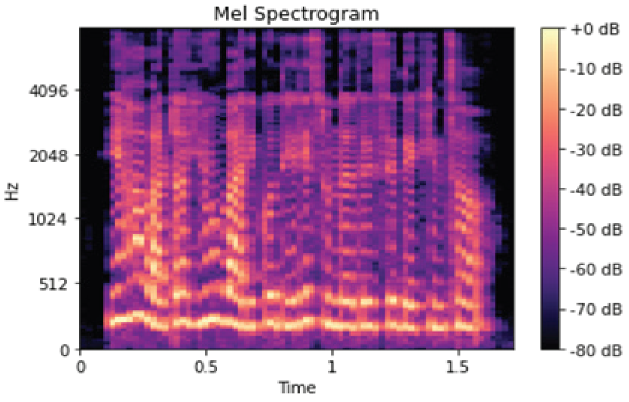

Fig. 11 Original Mel Spectrogram

The original spectrogram exhibits clear and well-defined frequency patterns, reflecting the natural phonetic and prosodic characteristics of Nepali speech. It serves as the reference for evaluating the synthesized outputs.

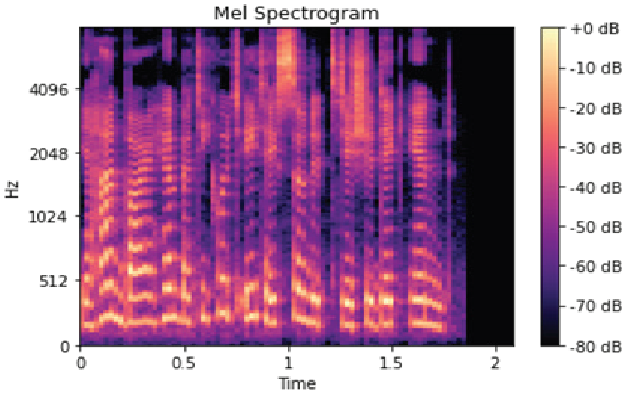

Fig. 12 Pretrained Tacotron2 Mel Spectrogram

The pre-trained Tacotron2 spectrogram, trained on English data, shows noticeable deviations from the original spectrogram. There is a lack of clarity and definition in the frequency patterns, indicating that the model struggles to capture the nuances of Nepali speech. The misalignments and less distinct structures highlight the challenges of using a model trained on a different language without adaptation.

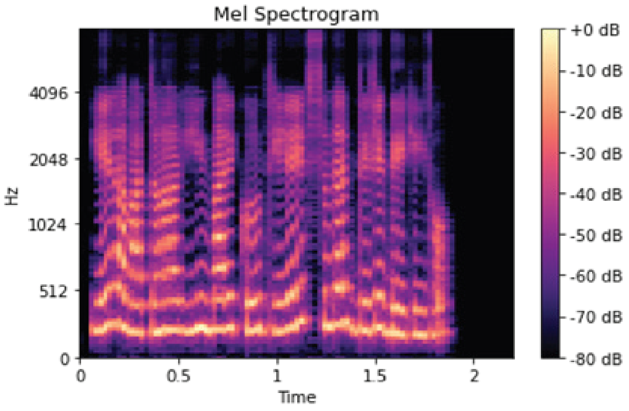

Fig. 13 Fine-tuned Tacotron2 Mel Spectrogram

The fine-tuned Tacotron2 spectrogram shows significant improvements over the pre-trained version. The frequency patterns are more defined and closely resemble those of the original spectrogram. This indicates that the fine-tuning process effectively adapts the model to the phonetic and prosodic characteristics of Nepali, leading to a more accurate and natural-sounding TTS output.

3) *Model Comparison Conclusion*

Energy Concentration: The pre-trained Tacotron2 model shows an excessive concentration of energy at the lowest frequencies, leading to an unnatural audio output. The fine-tuned Tacotron2 model, on the other hand, significantly reduces this concentration, making the audio output more balanced.

Frequency Distribution: The fine-tuned model's frequency distribution more closely matches the original audio, indicating that fine-tuning improves the fidelity and naturalness of the synthesized speech. The pre-trained model has a less accurate frequency distribution with more pronounced artificial peaks.

Effectiveness of Fine-Tuning: Fine-tuning the pre-trained Tacotron2 model on Nepali data substantially improved the quality of the synthesized speech. The fine-tuned model's spectrogram more closely matches the original, demonstrating the importance of adapting pre-trained models to the target language.

Quality of Synthesized Speech: The closer resemblance of the fine-tuned Tacotron2 spectrogram to the original suggests that the adapted model can produce more natural and intelligible Nepali speech.

D. *MOS Comparison*

The MOS analysis highlighted varied perceptions of audio quality, spanning from 2 to 4.5 with an average of 3.1 in terms of naturalness. Notably, the average of MOS on accuracy of the audio seemed to be 3.21, slightly better score than the naturalness. However the subjective nature of MOS scores should be noted down, as it is influenced by individual biases in terms of perceived naturalness and accuracy.

TABLE I
MOS SCORE COMPARISON

Category	MOS
Naturalness	3.1
Accuracy	3.21

Conclusion

This research paper culminates in the successful training of a fine-tuned Tacotron2 model tailored for spectrogram generation in Nepali, complemented by audio synthesis using WaveGlow. While the synthesized audio produced is intelligible, several limitations have been identified, underscoring the imperative for ongoing refinements. These limitations encompass the model's incapacity to synthesize audio for special characters, the suboptimal quality of training data resulting in less natural-sounding speech, and the presence of noise and choppy speech lacking smooth flow.

Additionally, accent issues have led to deviations from the intended Nepali language, thereby compromising the realism of the generated speech for applications like voice assistants or audiobooks. Looking forward, potential enhancements include the deployment of the system as a mobile application, refinement of the audio waveform for specific applications, and addressing clarity and accent issues through advanced techniques such as prosody modeling, articulatory synthesis, and accent-specific training. Furthermore, the exploration of training for multiple speakers and voice cloning could augment the versatility of the system. Despite these challenges, the research represents a notable milestone in the realm of synthesizing human-like speech in Nepali.

Through meticulous fine-tuning of the Tacotron2 model and optimization of hyperparameters, a promising Mean Opinion Score (MOS) of approximately 3.16 was attained in subjective evaluations by 33 participants. This underscores the potential of the approach to mitigate the robotic sound often associated with text-to-speech systems, thus advancing the prospect of more natural and engaging speech synthesis experiences in Nepali.

Acknowledgement

We extend our deepest gratitude to our supervisor, Er. Suramya Sharma Dahal, for his invaluable guidance, expertise, and steadfast support throughout this research project. His insights and encouragement have significantly shaped our work. We also sincerely appreciate the Department of Electronics and Computer Engineering at Thapathali Campus for their continuous support and the resources they provided. We are especially thankful to the faculty and staff for their assistance and for cultivating an environment that promotes research and innovation.

Lastly, we acknowledge the valuable feedback and collaboration from our peers and colleagues, which have been essential throughout this journey.

References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis and Y. Wu, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," 2018.

[2] R. Prenger, R. Valle and B. Catanzaro, "WaveGlow: A Flow-based Generative Network for Speech Synthesis," 2018.

[3] A. J. Hunt and A. W. Black, "Unit Selection in a Concatenative Speech Synthesis System using a large speech database," *ATR Interpreting Telecommunications Research Labs,* no. 1996.

[4] P. Malla, "Nepali Text to Speech using Time Domain Pitch SynchronousOverlap Add Method," *Tribhuvan University,* 2015.

[5] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," 2016.

[6] S. Bengio, O. Vinyals, N. Jaitly and N. Shazeer, "Char2Wav: End-to-End Speech Synthesis," in *International Conference on Learning Representations*, 2017.

[7] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta and M. Shoeybi, "Deep Voice: Real-time Neural Text-to-Speech," *Baidu Silicon Valley Artificial Intelligence Lab,* 7 March 2017.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Networks," 2014.