ENHANCING ACCESIBILITY AND COMMUNICATION THROUGH TEXT TO SPEECH CONVERSION

J. Kavitha[1], P. V. Chinmayee[2], Nandyala Varun Reddy[3], M. Ramu[4], L. Sharmila[5]
[1,2,3,4]*Department of CSD, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, India*
[1] drkavithaj@veltech.edu.in, [2] vtu22650@veltech.edu.in, [3] vtu23623@veltech.edu.in, [4] vtu22365@veltech.edu.in
[5]*Department of CSE,Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, India*
[5] shar.hariharan@gmail.com

2024 International BIT Conference (BITCON) | 979-8-3315-1839-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/BITCON63716.2024.10985405

***Abstract*—Text-to-speech (TTS) converters have revolutionized accessibility and communication, particularly for individuals with visual impairments or reading disabilities. By converting written text into spoken language, these tools empower individuals to access information and engage in meaningful interactions. This abstract explores the significance of TTS converters in bridging communication gaps, improving educational outcomes, and enhancing overall quality of life. TTS converters offer a versatile solution for a wide range of applications. In educational settings, they enable students with dyslexia or learning disabilities to access textbooks and assignments independently, fostering a more inclusive learning environment. For individuals with visual impairments, TTS converters provide a valuable tool for reading documents, navigating websites, and interacting with digital content. Additionally, these technologies can be used to create audio versions of books, magazines, and other materials, making them accessible to a broader audience.The advancements in TTS technology have led to significant improvements in the naturalness and clarity of synthesized speech. Modern TTS engines utilize sophisticated algorithms and deep learning techniques to produce more human-like voices, enhancing the listening experience. This progress has expanded the potential applications of TTS converters, including their use in virtual assistants, language learning tools, and entertainment systems. TTS converters have emerged as indispensable tools for enhancing accessibility and communication. By breaking down barriers and providing equal opportunities for individuals with disabilities, these technologies contribute to a more inclusive and equitable society. As TTS technology continues to evolve, It'll come with greater advancements in natural language processing and speech synthesis, further expanding the possibilities for their application in various domains.**

***Index Terms*—Text-to-Speech(TTS), Accessibility, Tacotron2, Waveglow, mel-spectrogram.**

I. INTRODUCTION

The text-to-speech system's converter operates on extremely advanced technology of TTS that directly translates the written text into natural and expressive speech through complex algorithms and machine learning techniques. Based on its strength in Natural Language Processing, it analyzes the syntax, grammar, and context of the input text to create a phonetic representation that can capture pronunciation, intonation, and prosody. There have been tremendous advancements in deep learning that have enabled better performance by the system and the ability of the system to generate speech that has the qualities of human-like expressiveness. These have emerged from the models such as those developed by NVIDIA, specifically Tacotron 2 and WaveGlow. Tacotron 2 is an end-to-end deep learning model that translates the input text directly to a mel-spectrogram; this is one way of representing sound frequencies over time. This model pairs up a sequence-to-sequence architecture with attention mechanisms to produce super-natural speech. The next step in this pipeline involves WaveGlow, another deep-learning-based model that produces high quality waveform-level speech audio given the mel-spectrogram generated by the Tacotron2. WaveGlow is a flow-based generative model, which combines normalizing flows with a probabilistic framework to produce coherent and realistic speech output.

It therefore has transformed the face of TTS and brought it closer to a natural voice with qualities, emotions, and expressiveness. Such TTS systems have now opened avenues in assistive technologies for the visually impaired, virtual assistants, e-learning sites, and customer services. TTS systems are now more accessible and user-friendly, thus adding the adaptability needed for different languages and contexts across users worldwide.

TTS converters are numerous in their benefits. They facilitate the filling of communication gaps, improvement of educational outcomes, and better quality of life for disabled persons. Its uses range from academics to entertainment as synthesis methods continually improve them to sound more natural and efficient. Currently, machine learning and NLP have become the front-runner for development in TTS, which enables the transformation of text into speech that is not only correct but full of emotions and contextual. Such systems opened up barriers to a place where "the playing field seemed leveled," further proving TTS in the modern world.

And so, with time advancing in technology, TTS systems will be significantly improved upon in the future. With models like Tacotron 2 and WaveGlow integrated into it, TTS's

future is promising: providing a variety of applications with more natural and audible speech synthesis, thus continuing to give humans all over the world better communication and accessibility.

II. Ease of Use

A. Maintaining the Integrity of the Specifications

Ensuring the integrity of specifications is crucial for the successful implementation of text-to-speech (TTS) conversion systems, particularly when enhancing accessibility and communication for individuals with disabilities. By maintaining accuracy, completeness, and consistency in specifications, organizations can deliver high-quality TTS solutions that meet user needs and contribute to a more inclusive society.

- Prioritize user-centered design throughout the specification development process. Involve individuals with disabilities in the design and testing phases to gather valuable insights and ensure that the TTS system meets their specific needs.
- Adhere to relevant accessibility standards, such as the Web Content Accessibility Guidelines (WCAG) and Section 508, to ensure that the TTS system is compatible with assistive technologies and accessible to a wide range of users.
- Conduct rigorous quality assurance testing to identify and address any issues or inconsistencies in the specifications. This includes testing the TTS system with various text inputs, languages, and accents to ensure that it produces accurate and understandable speech.

By following these strategies,organizations can maintain the integrity of specifications for TTS conversion systems and deliver solutions that truly enhance accessibility and communication for individuals with disabilities.

III. Framework and Working Model

A. Abbreviations and Acronyms

- ACR – Absolute Category Rating
- GPU – Graphical Processing Unit
- MOS – Mean Opinion Score
- NLP – Natural language processing
- TTS – Text to speech
- VRAM – Virtual Random Access Memory
- WCAG – Web Content Accessibility Guidelines

B. Working Procedure

The proposed application Text-to-Speech (TTS) conversion involves transforming written text into spoken language. The working procedure is as follows:

- **Text Preprocessing:** The input text is cleaned and normalized to remove any inconsistencies or errors - for example, punctuation is removed, and it may be tokenized into phonemes or characters, depending on your setup. For example, the text "Hello, this is a text-to-speech conversion using Tacotron2 and WaveGlow." could be tokenized into characters or phonemes.
- **Tacotron 2 Model (Generating Mel-Spectrogram):** The Tacotron 2 model will take the preprocessed text and output a mel-spectrogram-a sequence of 2D arrays, representing the acoustic features of the speech. WaveGlow then makes use of this spectrogram to synthesize the final audio.
- **WaveGlow Model (Generating Waveform):** The mel-spectrogram generated by Tacotron 2 will be passed to the WaveGlow model, which will then convert the spectrogram into an audio waveform. WaveGlow's output is a raw waveform that corresponds to the speech sound. The generated waveform is usually a long array of numbers representing the sound's amplitude over time.
- **Postprocessing & Saving the Result:** The waveform will be saved as a ".wav" file. This file will contain the synthesized speech, which can be played back using any audio player.

C. System Architecture and Diagrams

To provide a comprehensive understanding of the system's structure and functionality, The following diagrams are presented :

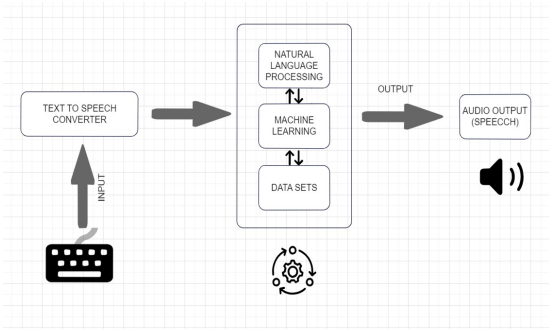

Fig. 1: System Architecture Diagram

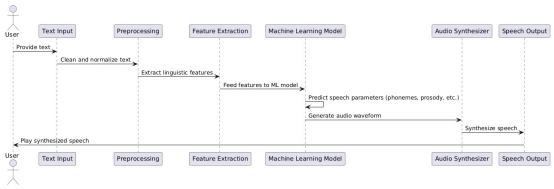

Fig. 2: Sequence Diagram of User Application

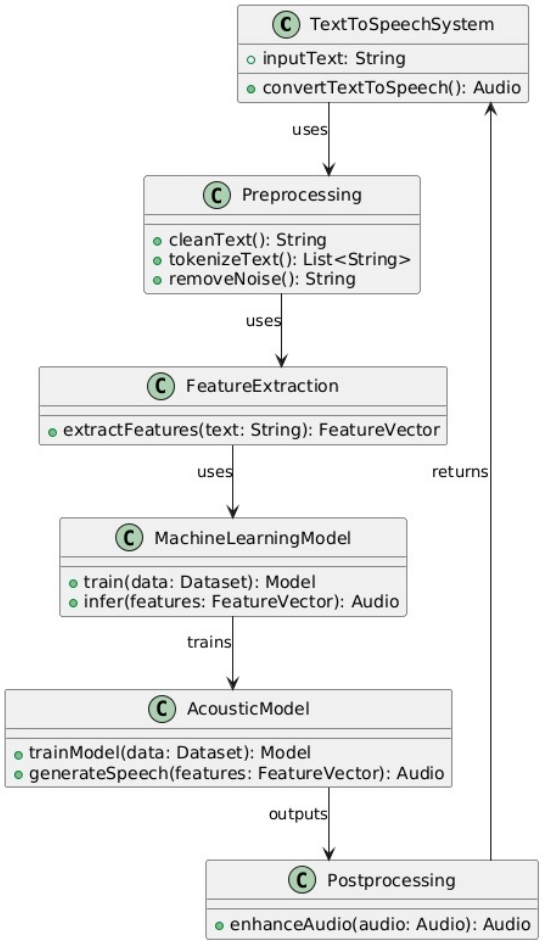

Fig. 3: Class Diagram of System Components

These diagrams illustrate the system's architecture, user interaction sequence, class structure, and data flow, providing a visual representation of the system's design and functionality.

D. *Frequency and Amplitude of Output speech*

- **Frequency of speech :** In voice production and perception, frequency is essential as it affects many facets of communication. Speech perception and production both depend on frequency. It affects how people communicate meaning, emotion, and clarity by altering the pitch and sound quality in various frequency ranges.

  During speech, changes in fundamental frequency can highlight particular points or express various feelings. For instance, a question might be indicated by a rising pitch, while a response might be shown by a dropping pitch.
  The Frequency of the speech can be calculated by:

$$f = 1/T \tag{1}$$

- *f= Frequency in Hertz(HZ)*
- *T= Time period*

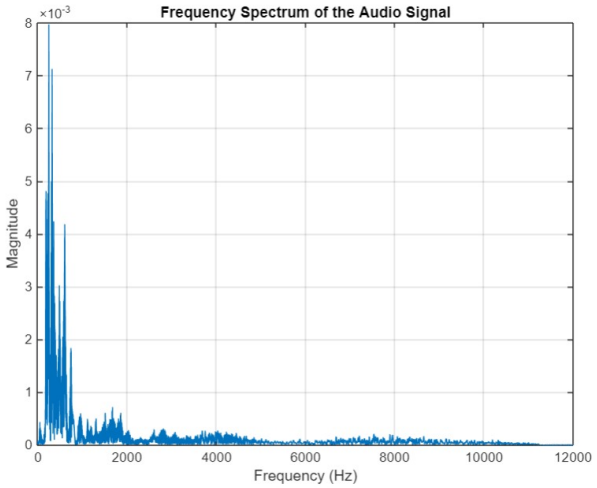

Fig. 4: Frequency graph of the produced speech

- **Amplitude of the produced speech :** A sound's perceived loudness is directly influenced by its amplitude. Louder sounds have a higher amplitude, which is important for speaking since it can emphasize or communicate urgency. For example, the Lombard effect is the tendency for speakers to amplify their voices in busy settings in order to be heard.

  An essential component of speech, amplitude affects vowel quality, loudness, intelligibility, emotional expression, and rhythmic structure. The methods for teaching language, improving communication, and developing assistive technologies for the deaf can all benefit from an understanding of the interactions between amplitude and these components. The amplitude of the sound waves can be calculated by:

$$A = \frac{p_{\max}}{\rho \cdot c \cdot \omega} \tag{2}$$

- *A = Amplitude of the sound wave*
- *Pmax = Maximum pressure variation*
- *c = Speed of sound in the medium, (i.e. air, 343m/s)*
- *ρ = Density of the medium, (i.e. air, 1.225 kg/m^3)*
  ω = Angular frequency of the wave
  ω can be given by $2\pi f$, where f is frequency of the produced speech.

- **Peak Amplitude:**
  The peak amplitude is defined as the maximum absolute value of the speech signal at any point in time. Provided

that speech signal can be portrayed as a time-domain waveform, the peak amplitude would be therefore calculated as:

$$Apeak = max(|s(t)|) \tag{3}$$

- A*peak* = *Peak Amplitude*
- *s(t) = speech signal as a function of time.*
- *max(s(t)) = Maximum absolute value of the speech signal over time*
- **Amplitude in Terms of dB (Decibels):**
  The amplitude can be expressed in terms of decibels, or dB, which is a logarithmic scale. The decibel scale is extremely useful for describing audio and speech processing because it duplicates the human perception of loudness. Amplitude in decibels is calculated by:

$$Amplitude\ (dB) = 20log10(A/Aref) \tag{4}$$

A = Amplitude of the speech
Aref= Reference amplitude

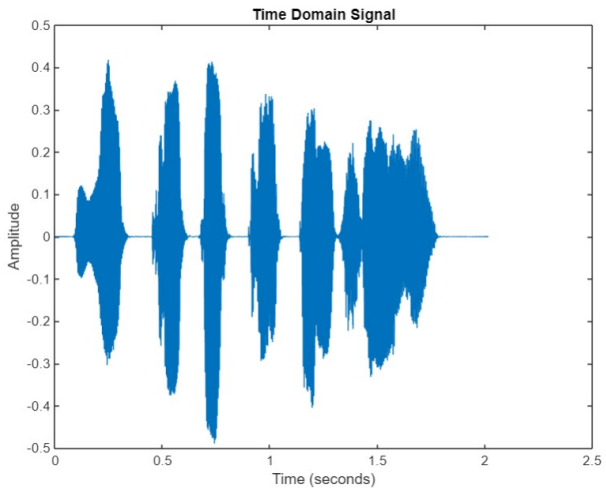

Fig. 5: Amplitude graph of the produced speech

- Figure 4 & 5 depicts about the Frequency and Amplitude graphs of the produced speech, which helps in finding the accuracy, naturalness, rhythm etc.. of the produced speech.

IV. Implementation

By implementing an Text to Speech converter using Tacotron2 and Waveglow, helps an individual with visual impairments, and can be used for languages to hear the correct pronunciation of phrases and practices.

1) Software Components:

- Compiler: Google Colaboratory, Visual Studio Code etc...
- Decision-Making Algorithm: Machine learning models based on deep learning and neural networks like WaveNet, Tacotron, FastSpeech, WaveGlow etc...

2) Software Setup:

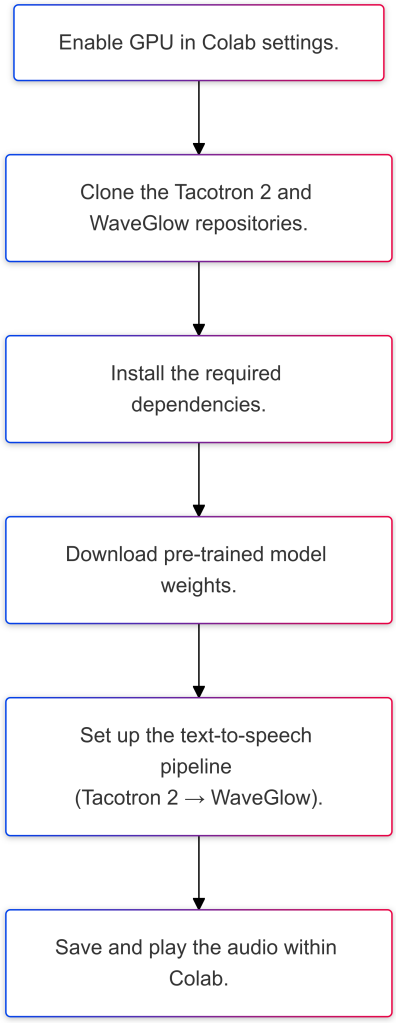

Fig. 6: Implementation of the TTS converter.

- A GPU is critical for training and running Tacotron 2 and WaveGlow models efficiently, particularly when dealing with the heavy computational demands of deep learning in TTS applications. GPUs allow the TTS system to scale for larger datasets and more complex models, making it suitable for large-scale applications.A GPU with at least 8 GB of VRAM should be sufficient for most inference tasks, but for training, 12 GB or more VRAM is ideal.
- Cloning the Tacotron2 and WaveGlow repositories refers to downloading the source code of these projects from GitHub to your local system or cloud environment (such as Google Colab). The process involves using the git clone command to fetch the repositories from their respective GitHub URLs.

```
git clone https://github.com/NVIDIA/tacotron2.git
git clone https://github.com/NVIDIA/waveglow.git
```

Listing 1: Cloning Tacotron 2 and Waveglow Repositories from GitHub

- Two of NVIDIA's advanced models, Tacotron2 and WaveGlow, are specifically made for text-to-speech synthesis. Tacotron2 works like any other sequence-to-sequence model that maps input text into mel-spectrograms, which represent the frequency content of speech over time. Using a combination of long short-term memory networks and attention mechanisms, the model learns the mapping from characters or phonemes to the speech features and in this way produces speech sounds extremely close to natural speech. The second is WaveGlow, yet another flow-based generative model, for mapping of mel-spectrograms into the actual waveform of speech-spectrograms to audio signals. The work is therefore based on the Glow architecture with WaveNet-like vocoding for high-quality and efficient speech synthesis. This powerful pipeline, with Tacotron2 and WaveGlow, respectively, is used to synthesize natural speech that almost sounds like a human voice, from text. It gets very wide application in virtual assistants, voice-based applications, and speech synthesis in general.
- Choose several python machine learning models, algorithms, modules and libraries like PyTorch, numpy, scipy, Tacotron 2, WaveGlow, gTTS, etc.. (Any of the above can be used based on the convenience).

```
pip install numpy torch matplotlib scipy gdown
pip install git+https://github.com/NVIDIA/apex.git
```

Listing 2: Installation of dependencies

- Mostly the installation will be done by using pip (eg: !pip install gTTS).
- Give necessary functions and coding.
- Deployment and Testing:
  - Test the TTS engine with various text inputs and languages.
  - Adjust settings as needed to achieve desired results.
  - Each language will be having a separate code (eg: English = 'en', Telugu = 'te', Hindi = 'hi').

V. Results and Impact

Results and Impact of an Enhancing accessibility and communication through text to speech conversion. The use of Tacotron 2 and WaveGlow in the text-to-speech conversion process is highly important, bringing an improvement in naturalness, quality, intelligibility, and speech generation speed. These models have elevated user experiences in applications like virtual assistants, audio books, and accessibility tools to great heights, with a much more fluid interaction and a human touch to it. Future developments are going to make TTS more natural and versatile, hence sophistication expected in emotional expression and speaker diversity.

- **Results:**
  - **Improved Prosody and Intonation:** Tacotron 2 learns the natural rhythm and stress patterns of speech, resulting in more expressive and dynamic output, enhancing conversational tone.
  - **Naturalness of Speech:**Natural Speech Quality: Despite having the ability to generate mel-spectrograms of extremely high quality from text, Tacotron 2 struggles to elicit the natural prosody and intonation in speech. When combined with WaveGlow, it turns these spectrograms into realistic, inaudible audio with minimal noise or distortion.
  - **Faster and Efficient Inference:** WaveGlow enables faster speech synthesis compared to older models like WaveNet, making it ideal for real-time applications.

    Overall, the combination of Tacotron 2 and WaveGlow has significantly enhanced the realism and efficiency of TTS systems, offering better user experiences across multiple domains.
- **Frequency and Amplitude of the result Speech:** The speech outputs from various languages are being collected, and the frequency and amplitude of these speech samples are being calculated and analyzed.

Text	Language	Frequency(in Hz)	Amplitude(in Db)
Anemone	English (en)	150.00 Hz	56.63 dB
Namaste	Telugu (te)	220.63 Hz	52.49 dB
Vanakkam	Tamil (ta)	462.34 Hz	46.76 dB
Kithaab	Hindi (hi)	465.45 Hz	42.59 dB
Hegiddīri	Kannada (kn)	401.23 Hz	51.12 dB
Sukhamaano	Malayalam (ml)	285.13 Hz	46.86 dB
Gamsahamnida	Korean(ko)	146.24 Hz	50.42 dB

TABLE I:
Table for Frequency & Amplitude of the output speeches

- Table I shows the Frequency and Amplitude of the output speech obtained by different words and sentences of various languages. The frequency range of the output speeches are at the range between 100 Hz to 500 Hz and for amplitude the range is between 40 dB to 60 dB. The frequencies and amplitudes of the produced speeches

are ideal for various languages.

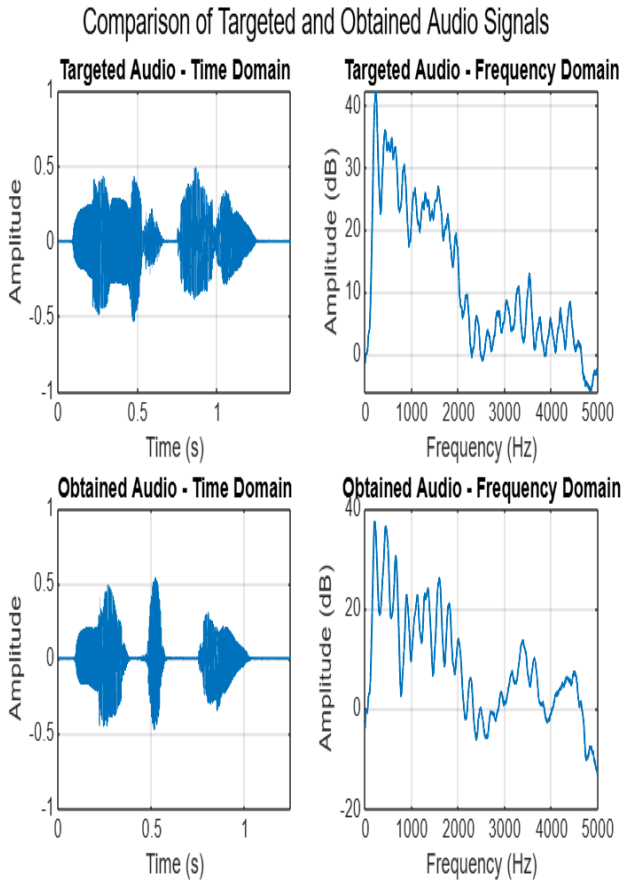

Fig. 7: Graphs of the Targeted speech & the Output speech

- Figure 7 shows the Frequency and Amplitude graphs of the targeted (reference) speech and the produced (synthesized) speech above. These graphs are intended to aid in the testing of how well generated speech can emulate its reference speech in terms of accuracy, naturalness, and rhythm by comparing the frequency and amplitude profiles of the reference speech with that of the generated speech. These factors are examined for the purpose of grading how the synthesized speech compares to natural human speech, in relation to pitch, tone, and expressiveness to produce more natural and more accurate speech synthesis.

- table II & III shows the frequencies and amplitudes of the sample speeches took for reference and output speeches got from the TTS. The values of the sample speech may vary based on the voice tones and loudness of the speech.

Serial no.	Frequency for Sample	Frequency for Output
1	240 Hz	150.0 Hz
2	200 Hz	220.63 Hz
3	367 Hz	462.34 Hz
4	401 Hz	465.45 Hz
5	370.65 Hz	401.23 Hz
6	246.65 Hz	285.13 Hz
7	160.65 Hz	146.24 Hz

TABLE II:
Table of Frequencies for Sample speech & Output speech(in Hertz)

Serial no.	Amplitude for Sample	Amplitude for Output
1	59.65 dB	56.63 dB
2	46.59 dB	52.49 dB
3	56.45 dB	46.76 dB
4	41.66 dB	42.59 dB
5	55.65 dB	51.12 dB
6	52.76 dB	46.86 dB
7	45.98 dB	50.42 dB

TABLE III:
Table of Amplitudes for Sample speech & Output speech(in Decibels)

- **Impact:**
  - Social inclusion: TTS has contributed to a more inclusive society by breaking down barriers and providing equal opportunities.
  - Economic benefits: TTS has enabled individuals with disabilities to participate more fully in the workforce.
  - Quality of life: TTS has significantly improved the quality of life for many individuals by expanding their access to information and communication.
  - Technological advancement: TTS has driven advancements in natural language processing and speech synthesis.

A. *Impact of the Tacotron2 in Text-to-speech conversion*

Tacotron 2 achieves revolutionary text-to-speech synthesis by producing very natural and human-like speech. The architecture uses a sequence-to-sequence model with an attention mechanism plus a WaveNet-style vocoder, similar to WaveGlow. With such architecture, Tacotron 2 can synthesize expressive prosody and produces better pronunciation and intonation on complex texts in comparison with traditional rule-based and concatenative TTS methods. Thus, learning directly from raw text and speech data enables Tacotron 2 to minimize the requirement for large-scale linguistic preprocessing; it is an advancement toward creating realistic applications in several domains for TTS.

B. *Impact of Waveglow*

It has dramatically changed the face of text-to-speech (TTS), thanks to the strong and efficient vocoder used in the generation of high-quality, natural-sounding audio. Unlike traditional vocoders that normally struggle with the subtleties of human speech WaveGlow uses a generative model based on normalizing flows in generating realistic waveforms from mel-spectrograms. Its upside is that it generates audio in real-time and does so without sacrificing quality, an aspect making it very suitable for low-latency applications. With this architecture, WaveGlow also does away with the need for complex autoregressive decoding and therefore leads to more efficient and scalable TTS systems. It has improved the performance and accessibility of neural TTS solutions.

C. *Collaborative Problem-Solving*

It is essential for developing effective text-to-speech (TTS) conversion systems that meet the needs of diverse users and enhance accessibility and communication. By fostering collaboration among stakeholders, organizations can identify and address challenges, improve the quality of TTS systems, and ensure that they are inclusive and accessible to all.organizations can develop TTS systems that are truly inclusive, accessible, and meet the needs of a diverse user base. This will ultimately enhance communication and improve the quality of life for individuals with disabilities.

D. *Naturalness & Accuracy of the produced speech*

The Mean Opinion Score (MOS) is commonly utilized in the field of telecommunications to evaluate the quality of voice and video sessions. It measures the subjective assessment of quality through human evaluations, usually on a scale ranging from 1 (poor) to 5 (outstanding).

The MOS quantifies the total quality of an audio or video experience using numbers, which indicates user satisfaction.
Rating Scale:
The most common scale used for MOS is the Absolute Category Rating (ACR) scale:
5: Excellent
4: Good
3: Fair
2: Poor
1: Bad

When the score is around 3.5, it suggests that users have encountered below-average quality. Scores exceeding 4.0 are typically deemed satisfactory for high-quality calls.

Participants in controlled settings listen to audio samples or take part in live calls and provide ratings based on clarity, loudness, delay, and background noise. The regular words from various languages are collected and retained for evaluation.

Text	Language	Naturalness Score	Accuracy
Anemone	English (en)	4.0	4.0
Namaste	Telugu (te)	3.8	3.5
Vanakkam	Tamil (ta)	3.6	3.0
Kithaab	Hindi (hi)	3.9	4.0
Hegiddiri	Kannada (kn)	3.4	3.0
Sukhamaano	Malayalam (ml)	4.0	3.9
Gamshamnida	Korean(ko)	3.0	3.1

TABLE IV: MOS table for Uniqueness & Naturalness of speech

E. *Overall Impact*

The overall impact of Enhancing accessibility and communication through text to speech conversion is significant.TTS has enabled individuals with visual impairments to access information and content independently. TTS has helped individuals with dyslexia or other reading difficulties to overcome challenges and improve their comprehension.TTS has facilitated communication between people who speak different languages.TTS has had a transformative impact on society by breaking down barriers, improving communication, and enhancing accessibility. As technology continues to advance.

VI. Conclusion

Text-to-speech (TTS) conversion has emerged as a powerful tool for enhancing accessibility and communication, particularly for individuals with visual impairments or reading disabilities. By transforming written text into spoken language, TTS technology has broken down barriers, provided equal opportunities, and improved the quality of life for countless people.As TTS technology continues to evolve, The world can expect to see even greater advancements in natural language processing, speech synthesis, and accessibility features. By investing in research, development, and implementation, governments, organizations, and individuals can further harness the potential of TTS to create a more inclusive and equitable society.Text-to-Speech (TTS) conversion has revolutionized accessibility and communication, particularly for individuals with visual impairments or reading disabilities. By breaking down barriers and providing equal opportunities, TTS has empowered countless people to access information, engage in meaningful interactions, and participate fully in society.As TTS technology continues to advance, The world can anticipate even greater improvements in natural language processing, speech synthesis, and accessibility features. This will further expand the potential applications of TTS and enhance its benefits for users. By investing in research, development, and implementation, governments, organizations, and individuals can contribute to a more inclusive and equitable future.

References

[1] Smith and A.Johnson,"Efficient Text-to-Speech Conversion Using Neural Networks", in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 123-128.

[2] R. Gupta and S. Kumar, "Enhancing Text-to-Speech Conversion Through Deep Learning Models," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 932-945, May 2017.
[3] M. Lee and B. Kim, "Robust Text-to-Speech Conversion Using WaveNet Models," in *IEEE Signal Processing Letters*, vol. 28, pp. 112-116, Jan. 2021.
[4] S. Singh et al., "A Comparative Study of Deep Learning Architectures for Textto-Speech Conversion," in *IEEE International Conference on Multimedia and Expo Workshops*, 2023, pp. 345-350.
[5] M. Kimura, J. Martinez,"Enhancing Image Recognition using Convolutional Neural Networks," in *IEEE International Conference on Computer Vision*, 2022, pp. 112-118.
[6] J. Shen, R. Pang, and R. J. Weiss, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779-4783.
[7] Z. Wang, S. Chang, and E. Yang, "Tacotron: Towards End-to-End Speech Synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4779-4783.
[8] S. Karita, S. Watanabe, T. Iwata, M. Delcroix, A. Ogawa, and T. Nakatani, "Semi-supervised end-to-end speech recognition using text-to-speech and autoencoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6166–6170.
[9] Swetha, N., and Anuradha, K. ."Text to speech conversion" in *International Journal of Advanced Trends in Computer Science and Engineering*,2013, 2(6), 269-278.
[10] P. Denisov and N. T. Vu, "Pretrained semantic speech embeddings for endto-end spoken language understanding via cross-modal teacher–student learning," in *Proc. Interspeech*, Oct. 2020, pp. 881–885.
[11] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2021, pp. 244–250
[12] M. Huzaifah and I. Kukanov, "An analysis of semantically-aligned speechtext embeddings," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2023, pp. 747–754
[13] Klatt, Dennis. "The Klattalk text-to-speech conversion system." In *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7, pp. 1589-1592. IEEE, 1982.
[14] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proc. 39th Int. Conf. Mach. Learn.*, vol. 162, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., 2022, pp. 1298–1312.
[15] S. Aryal and R. Gutierrez-Osuna, "Can voice conversion be used to reduce non-native accents", *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 7879-7883, 2014.
[16] D. Felps, C. Geng and R. Gutierrez-Osuna, "Foreign accent conversion through concatenative synthesis in the articulatory domain",in *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 8, pp. 2301-2312, Oct. 2012.
[17] J.-X. Zhang, Z.-H. Ling, Y. Jiang, L.-J. Liu, C. Liang and L.-R. Dai, "Improving sequence-to-sequence voice conversion by adding text-supervision", *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 6785-6789, 2019.
[18] Y. Ren et al., "Fastspeech: Fast robust and controllable text to speech", *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.
[19] Y. Ren et al., "Fastspeech 2: Fast and high-quality end-to-end text to speech", *Proc. Int. Conf. Learn. Representations*, 2020.
[20] . Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, pp. 4171-4186, Jun. 2019.
[21] X. Zhu, Y. Zhang, S. Yang, L. Xue and L. Xie, "Pre-Alignment Guided Attention for Improving Training Efficiency and Model Stability in End-to-End Speech Synthesis," in *IEEE Access*, vol. 7, pp. 65955-65964, 2019.