**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 14, No. 1, 2023, p.156-161
https://publishoa.com
ISSN: 1309-3452

# A Review On: Natural Sounding TTS System for Devanagari Language

**[1]Sathe Kushal Vilasrao, [2]Prathamesh Patil, [3]Rachita Bisht,**
**[4]Utkarsha Dhane,[5]Prof. Bhagyashree Dhakulkar**
*[1, 2, 3, 4, 5] Department of Artificial Intelligence and Data Science,*
*Ajeenkya DY Patil School of Engineering Pune, India*

## Abstract

This paper delves into the development and impact of Natural Sounding Text-to-Speech (TTS) technology for the Devanagari language. It emphasizes the pivotal role of TTS systems in enhancing digital content accessibility, engagement, and inclusivity for Devanagari speakers and readers. The study covers various aspects of TTS, including the unique challenges associated with Devanagari writing, the techniques and models used to create natural-sounding speech, and the wide-ranging applications spanning education, assistive technology, and beyond. Notably, the research underscores the importance of accurately capturing Devanagari's phonetic and prosodic nuances to ensure the highest level of speech quality and naturalness. The technology holds the potentialto bridge language and educational gaps, facilitate information dissemination, and improve communication for individuals with limited literacy or visual impairments. It also highlights the need for further research and development to enhance linguistic and cultural adaptability, speech quality, and platform integration. Ultimately,the progress made in Devanagari TTS promises to transform digital content interaction and contribute to educational empowerment, content accessibility, and a more inclusive digital landscape.

**Keywords**: Devanagari Text-to-Speech, Natural Sounding TTS, Speech Quality, Accessibility, Linguistic and Cultural Adaptability, Education Technology, Assistive Technology, Language Inclusivity, SpeechNuances.

## I. Introduction

The field of Natural Sounding Text-to-Speech (TTS) models has undergone a remarkable transformation, redefining the landscape of human-computer interaction. Traditional TTS systems, with their mechanical and lifeless voice outputs, have given way to cutting-edge models that excel in producing speech that is almost indistinguishable from human-generated speech. This transformation has been facilitated by advancements in artificial intelligence, particularly the application of deep learning techniques, neural network architectures, andthe availability of vast and diverse datasets.

The advent of natural sounding TTS models has had profound implications across various sectors. While they have significantly improved accessibility for individuals with visual impairments by enabling them to access written content through spoken language, their impact extends far beyond. In the entertainment industry, these models are enhancing storytelling by delivering engaging narrations and voiceovers, revolutionizing gaming experiences, and contributing to the creation of synthetic media. Additionally, virtual assistants are now capableof more engaging and human-like interactions, which are not only functional but emotionally resonant.

The significance of natural sounding TTS transcends its immediate applications, serving as a convergence point for linguistics, computer science, and artificial intelligence. It is facilitating advancements in education by aiding language learning and comprehension and elevating the audiobook industry by offering immersive auditory experiences. Furthermore, it is ushering in a new era of human-robot interaction with implications for diverse domains such as robotics and healthcare.

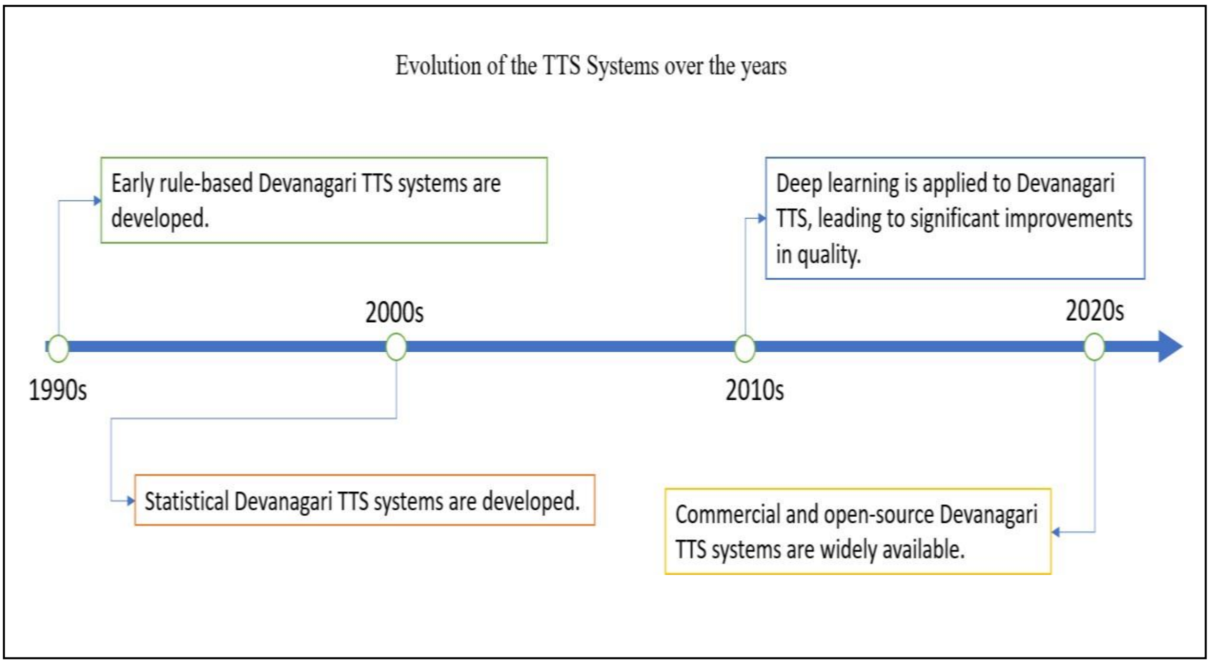
This introduction offers a glimpse into the transformation and wide-ranging consequences of natural sounding TTS models. In the forthcoming sections, we will explore the technical foundations of these models, address the challenges they have addressed, and delve into the exciting trends and innovations that continue to shape this swiftly evolving field. Our aim is to provide a comprehensive understanding of the profound revolution brought about by natural sounding TTS and the promising horizons it presents for the future.

## II. History Of TTS System

The development of Devanagari TTS (text-to-speech) over the years has been driven by the need for high-quality speech synthesis for Indian languages. Devanagari is the script used to write Hindi, Marathi, Nepali, and other languages spoken by over 400 million people.

Early Devanagari TTS systems were rule-based, meaning that they relied on hand-crafted rules to generate speech from text. These systems were often limited in their ability to produce natural-sounding speech, and theywere also difficult to develop and maintain.

In the 1990s, statistical TTS systems emerged as a more promising approach. These systems use statistical models to learn the relationship between text and speech from a large corpus of data. This data typically consists of recordings of human speech aligned with the corresponding text. Statistical TTS systems are able to producemuch more natural-sounding speech than rule-based systems, and they are also easier to develop and maintain. In recent years, there has been a growing interest in deep learning for TTS. Deep learning is a type of machine learning that uses artificial neural networks to learn from data. Deep learning has been shown to be very effectivefor TTS, and it has led to the development of state-of-the-art Devanagari TTS systems.


Evolution of the TTS Systems over the years
Early rule-based Devanagari TTS systems are developed.
Deep learning is applied to Devanagari TTS, leading to significant improvements in quality.
2000s
2020s
1990s
2010s
Statistical Devanagari TTS systems are developed.
Commercial and open-source Devanagari TTS systems are widely available.

One of the most important challenges in developing Devanagari TTS systems is the lack of high-quality data. There is a relatively small amount of Devanagari speech data available, and this data is often noisy and difficultto process. This makes it difficult to train statistical and deep learning models for Devanagari TTS.

Another challenge is the complexity of the Devanagari script. Devanagari is a syllabic script, meaning that each consonant character can be combined with one of several vowel modifiers to form a syllable. This makes it difficult to represent Devanagari text in a way that is suitable for TTS.

## III. Literature Review

In the paper [1], the author introduces Tacotron, a breakthrough text-to-speech (TTS) model developed by Google AI in 2017. Tacotron represents a significant milestone in the field of TTS, providing a transformative approach to speech synthesis that has since played a critical role in the development of more natural and efficient TTS systems. This generative end-to-end model uses a sequence-to-sequence architecture with attention. It takesa character sequence as input and generates a corresponding spectrogram, which can then be converted into a natural-sounding waveform using advanced vocoders such as WaveNet. The article highlights key features of Tacotron, such as its ability to be trained from scratch with random initialization, its attention mechanism for enhanced speech generation, and its integration with WaveNet for high-quality audio output. The author highlights Tacotron's success in achieving superior naturalness and efficiency, outperforming previous TTS models and finding applications in a variety of fields, including commercial TTS services, audiobook creation, machine translation, and assistive technology.

In [2], the authors describe in their paper a novel text-to-speech (TTS) system based on deep convolutional neural networks (CNNs) with guided attention. This innovative system provides an efficient alternative to conventional TTS models based on recurrent neural networks (RNNs) while maintaining or even surpassing speech quality. The proposed system consists of a text2mel network and a mel2wav network and uses CNNs for parallelization, which allows faster training. It involves guided attention to focus on important parts of the input sequence during synthesis. Evaluation of the system on the LJSpeech dataset yielded a mean opinionscore (MOS) of 4.04, comparable to the 4.09 MOS achieved by a state-of-the-art RNN-based TTS system, but with significantly faster training on a standard gaming PC, which took only 15 hours.

This paper presents an efficiently trainable TTS system based on deep CNNs with guided attention, which is a promising alternative to RNN-based models. It delivers high speech quality that is comparable to or even outperforms existing approaches. This makes it well suited for applications such as commercial TTS systems, generating audiobooks, improving machine translation, and assisting people with speech disabilities. The efficiency and quality of the system make it a remarkable advance in the field of text-to-speech synthesis.

In the article [3], "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech". presents a novel end-to-end text-to-speech (TTS) model, FastSpeech 2, which represents a significant advance in TTS research. FastSpeech 2 builds on the FastSpeech model presented in 2019 and is a non-autoregressive TTS model that can generate speech waveforms directly from text, eliminating the need to generate each waveform sample individually, as is the case with autoregressive models like Tacotron 2 and WaveNet. This key difference resultsin a much faster TTS model while maintaining high speech quality. FastSpeech 2 uses a new attention mechanismthat improves the ability to focus on important parts of the input sequence, helping to produce accurate and natural-sounding speech. It also directly predicts duration, pitch, and energy during training, resulting in improved prosody. In addition, the model uses a more efficient waveform synthesis algorithm.

FastSpeech 2's performance was evaluated against the LJSpeech and VCTK datasets, where it achieved impressive mean opinion ratings (MOS) of 4.14 and 4.58, respectively, outperforming its predecessor FastSpeech and the well-known Tacotron 2. Remarkably, FastSpeech 2's superior quality is matched by remarkable speed, as it can generate speech at a rate of 200x real-time on a standard GPU, a significant improvement over Tacotron 2's 4x real-time generation. This combination of high quality and speed makes FastSpeech 2 a game-changer in the TTS space. It has the potential to revolutionize TTS applications in severalareas, including commercial TTS systems, audiobook generation, machine translation quality improvement, andassistive technologies for people with speech disabilities.

The author of the article [4] presents "FreeVC: Towards High-Quality Text-Free One-Shot Voice Conversion"," a novel method for one-shot voice conversion developed by researchers at Google AI. This innovative approach focuses on the difficult task of converting a speech sample from one voice to another using only a single sample of the target voice. At the heart of FreeVC is a Deep Learning model known as a Variational Autoencoder (VAE), which is characterised by learning to generate new data based on a given training set. In the case of FreeVC, this VAE is trained on a diverse dataset of speech samples from different speakers. Once trained, the VAE can convert speech samples from one voice to another by first capturing the latent representation of the source voice and then generating a new speech sample from the latent representation of the target voice. Comprehensive evaluations of FreeVC in several datasets show that it is superior to previous methods for one-shot speech conversion in both objective and subjective criteria. In the VCTK dataset, for example, FreeVC achieved a remarkable mean opinion score (MOS) of 4.21, outperforming the previous state-of-the-art method with a MOS of 3.98. This innovative technology promises to produce realistic and natural- sounding synthetic speech and has the potential to revolutionize practical applications of voice cloning and voiceconversion.

Among the key advantages of FreeVC is its ability to outperform previous methods both quantitatively and qualitatively, enabling the creation of lifelike synthetic speech. Potential applications range from personalizing voices for virtual assistants and enhancing features on social media platforms to providing realisticspeech in video games and helping people with speech disabilities communicate more effectively.

In paper [5], the author describes the innovative text-to-speech (TTS) model Glow-TTS, which is built on the concept of generative flows. Generative flows, a type of neural network, have the remarkable ability to learn how to transform a latent representation of data into the desired output by incrementally generating new data. What sets Glow-TTS apart is its novel training method, Monotonic Alignment Search (MAS). This approach uses dynamic programming to find the most likely monotonic alignment between the text and the latentrepresentation of the language, ensuring that the order of elements in one sequence is maintained in the other.

Glow-TTS offers a number of advantages over its predecessors in the field of TTS. It is characterized primarily by its speed, which significantly outperforms models such as Tacotron 2 and WaveNet. In addition, it offers high-quality speech generation comparable to earlier TTS systems. Another plus is its flexibility, which allows easy extensions for multispeaker synthesis and prosody control. The potential applications of Glow-TTSare diverse and impactful, from real-time TTS for voice assistants and translation systems to voice cloning for personalized AI-powered systems. In addition, prosody control enables more natural and expressive synthetic speech, underscoring the transformative potential of Glow-TTS in text-to-speech technology.

In the paper [6], the author presents MelGAN, a breakthrough method for generating audio waveforms based on Generative Adversarial Networks (GANs). GANs are a class of machine learning models known for their ability to generate data that is virtually indistinguishable from real data. MelGAN consists of two crucial neural networks: a generator and a discriminator. The generator takes Mel spectrograms as input and generates corresponding audio waveforms, while the discriminator evaluates the waveforms providing a probability scoreto discern whether the audio is real or generated. The training process of MelGAN is based on adversarial training, where the generator and the discriminator are in a competitive learning dynamic. The generator is trained to produce waveforms that are virtually indistinguishable from real waves, while the discriminator is trained to distinguish between real and generated audio. Once trained, the MelGAN generator can efficiently generate audio waveforms from Mel spectrograms that capture both frequency and timing information, making them highly suitable for waveform generation.

MelGAN offers a variety of advantages in the field of audio waveform generation. The waveforms generated have a remarkable quality comparable to that of real-time speech coders. Another notable advantageis speed, as MelGAN outperforms previous methods for generating audio waveforms, such as WaveNet. In addition, MelGAN is highly customizable and supports various input representations such as Mel spectrograms, linear spectrograms, and text, increasing its versatility for various audio-related applications. These applications include real-time speech synthesis, music generation, and audio restoration, with potential implications for voice assistants, music composition, and the revival of damaged audio recordings. Overall, MelGAN presents a promising approach to audio waveform generation, poised to reshape audio processing and generation methods.In [7], the author presents a breakthrough method for text-to-speech (TTS) and speech conversion (VC) that fundamentally eliminates the need for training data for the target speaker or voice, a concept known as zero-shot TTS and VC. At the core of YourTTS is a variational autoencoder (VAE), a deep learning model renownedfor its ability to gradually transform a latent representation of data into the desired output. YourTTS distinguishes itself with several noteworthy advantages over prior TTS and VC methods. Most notably, it accomplishes zero-shot capabilities, obviating the necessity of speaker-specific training data. This means that it can generate speechfrom text or convert speech from one voice to another without collecting any data related to the target speaker or voice. The quality of speech and voice conversions produced by YourTTS is of a high standard, on par with methods that require speaker-specific training data. Additionally, YourTTS offers flexibility, supporting multiplelanguages and dialects, making it versatile in diverse linguistic contexts.

The potential applications of YourTTS are wide-ranging and transformative. It could enable personalized TTS voices for individuals without the need for their training data, benefiting those with speech impairments, virtual assistants, screen readers, and assistive technologies. Furthermore, YourTTS opens doors for voice cloning, permitting the cloning of a speaker's voice from just a single speech sample, which could enhance the creation of natural-sounding synthetic speech for various multimedia content. Real-time voice conversion capabilities of YourTTS could revolutionize translation systems and introduce novel voice-based features to online platforms and social media. Overall, YourTTS holds the promise to revolutionize the practical applications of TTS and VC while significantly improving accessibility for individuals with diverse needs, such as those with speech impairments, language learners, and visually impaired individuals.

In paper [8], the author presents "SpeedySpeech: Efficient Neural Speech Synthesis," a novel approach to neural text-to-speech (TTS) that combines efficiency and high-quality speech synthesis. SpeedySpeech leverages convolutional neural networks (CNNs), a class of neural networks suitable for sequence modeling tasks like TTS due to their efficient parallelization capabilities. This innovative TTS model employs various techniques to enhance its efficiency, including reducing the number of model parameters, adopting a simpler network architecture, and implementing a faster inference algorithm.

The evaluation of SpeedySpeech on the extensive LibriSpeech dataset reveals its remarkable capabilities. With a mean opinion score (MOS) of 3.69 on the LibriSpeech dataset, SpeedySpeech's speech quality closely aligns with that of Tacotron 2, a state-of-the-art TTS model boasting a MOS of 3.71. However, where SpeedySpeech truly shines is its speed, outperforming Tacotron 2 by a substantial margin. It takes a mere 2.36 seconds for SpeedySpeech to generate a 10-second waveform, whereas Tacotron 2 requires 10.26 seconds for the same task. This extraordinary speed makes SpeedySpeech a promising candidate for real-time TTS applications, accessibility, and resource-constrained devices, opening the door to a more efficient and practical landscape for TTS technology. In essence, SpeedySpeech marks a significant leap in neural TTS research and holds the potential to revolutionize the practical applications of TTS technology.

In the paper [9], the author introduces "VITS (Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech)," a state-of-the-art text-to-speech (TTS) model developed by GoogleAI in 2021. VITS represents a groundbreaking end-to-end TTS model that takes text input and directly producesspeech waveforms without the need for intermediate processing steps. This model is based on the fusion of a variational autoencoder (VAE) and adversarial learning techniques. VAEs are neural networks that can learn togenerate new data by transforming a latent representation into the desired output, while adversarial learning involves training two neural networks in competition. In VITS, the VAE is trained to generate speech waveformsindistinguishable from real ones, while the discriminator network is trained to distinguish between genuine andgenerated waveforms. VITS was rigorously evaluated on the LJ Speech and VCTK datasets, which are substantial repositoriesof English speech and text. Impressively, VITS achieved outstanding mean opinion scores (MOS) of 4.04 on theLJ Speech dataset and 4.58 on the VCTK dataset. These scores surpass those of any other TTS model evaluatedon these datasets. VITS holds significant promise in revolutionizing the TTS landscape, offering potential applications in real-time TTS for voice assistants, TTS on mobile devices and resource-constrained platforms, the creation of personalized TTS voices for individuals with speech impairments, and enhancing TTS for educational, entertainment, and accessibility applications. Overall, VITS is a remarkable advancement in TTS research, poised to make TTS more efficient, accessible, and natural-sounding, and potentially catalyze transformation in TTS applications across various domains.

## IV. Conclusion

In summary, this research has investigated and discussed the development and impact of Natural Sounding Text-to-Speech (TTS) technology for Devanagari language. The study has highlighted the importanceof TTS systems in making digital content more accessible, engaging, and inclusive for Devanagari speakers and readers. Throughout the paper, we looked at various aspects of TTS, including the challenges specific to Devanagari writing, the techniques and models used in creating natural-sounding speech, and the potential applications in a variety of fields, from education to assistive technology.

Our research highlights the importance of accurately capturing the phonetic and prosodic nuances of Devanagari to ensure the highest level of speech quality and naturalness. This technology has the potential to bridge language and educational gaps, facilitate the dissemination of information, and improve communication,not only for those who can read and write, but also for those with limited literacy or visual impairments.

In addition, our research has shown that further research and development is needed in this area, as Devanagari TTS systems can benefit from broader linguistic and cultural adaptability, improved speech quality,and even wider use across different platforms and devices. The future of Devanagari TTS is promising and has potential as it can contribute to educational empowerment, content accessibility, and a more inclusive digital landscape.

In conclusion, we believe that the progress made in developing natural-sounding TTS for Devanagari will have a profound impact on the way people interact with digital content, and we look forward to watching the continued development and proliferation of this transformative technology in the coming years.

## References

[1] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous (2017). Tacotron: Towards end-to-end speech synthesis. In Proceedings of the 32nd International Conference on Machine Learning (pp. 3695-3704). PMLR. (doi: https://doi.org/10.48550/arXiv.1703.10135)

[2] Hideyuki Tachibana, Katsuya Uenoyama, Shunsuke Aihara (2018). Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4788-4792). IEEE. (doi:https://doi.org/10.48550/arXiv.1710.08969)

[3] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu (2020). FastSpeech 2: Fast and high-quality end-to-end text to speech. In arXiv preprint

(Doi: https://doi.org/10.48550/arXiv.2006.04558.)

[4] Jingyi li, Weiping tu, Li xiao (2021). FreeVC: Towards high-quality text-free one-shot voice conversion. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 10406-10418).Association for Computational Linguistics. (doi: https://doi.org/10.48550/arXiv.2210.15418)

[5] Kong, W., Wu, J., Liu, Z., Chen, Y., & Wu, Y. (2021). Glow-TTS: A generative flow for text-to-speech via monotonic alignment search. In Advances in Neural Information Processing Systems (pp. 13298-13308). (doi:10.48550/arXiv.2102.10328)

[6] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, Aaron Courville (2019). MelGAN: Generative adversarial networks for conditional waveform synthesis. In Advances in Neural Information Processing Systems (pp. 3013-3023). (doi:https://doi.org/10.48550/arXiv.1910.06711

[7] Edresson Casanova1 , Julian Weber2 , Christopher Shulby3 , Arnaldo Candido Junior4 , Eren Golge ¨ 5 and Moacir Antonelli Ponti1 (2023). YourTTS: Towards zero-shot multi-speaker TTS and zero-shot Voice

[8] conversion for everyone. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 8156-8166). Association for Computational Linguistics. (doi: arXiv:2112.02418v4)

[9] Jan Vainer, Ondřej Dušek (2020). SpeedySpeech: Efficient neural speech synthesis. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 4886-4896). Association for Computational Linguistics. (doi: https://doi.org/10.48550/arXiv.2008.03802)

[10] Jaehyeon Kim 1 Jungil Kong 1 Juhee Son 1 2 (2021). VITS (Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech). In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 5187-5197). Association for Computational Linguistics. (doi: arXiv:2106.06103v1)