ijamrsd.com
ISSN: 2583-6404
Journal of Advanced Multidisciplinary Research Studies and Development
Peer Reviewed Referred Journal | Bi-monthly | Open Access

LEVERAGING DEEP LEARNING FOR NATURAL MARATHI TEXT-TO-SPEECH SYNTHESIS

Arin Dhimar[1], Prajakta Pare[2], Mrs. Uma Madje[3]
[1,2] Student, Department of Computer Science, Fergusson College, Pune, India.
[3] Assistant Professor, Department of Computer Science, Fergusson College, Pune, India.
Email: arindhimar.fc@gmail.com[1], prajaktapare.fc@gmail.com[2], uma.madje@fergusson.edu[3]

Abstract
This research paper describes a deep learning-based method that uses an end-to-end architecture to generate high-quality Marathi TTS. Marathi is a morphologically rich Indo-Aryan language and the generation of speech in this language is challenging due to the shortage of annotated datasets, variations in dialectal patterns, and complex prosody. The authors have proposed a VITS-based TTS system that combines text normalization, phoneme conversion, and prosody-informed acoustic modeling to overcome these difficulties. Naturalness and intelligibility have been improved by training the system on a Marathi speech corpus comprising multiple speakers and dialects. The work also includes a comparative evaluation of the study outcomes with traditional concatenative methods and neural baselines such as Tacotron2. In order to evaluate the systems' precision from the point of view of sound reproduction, the authors resort to objective metrics such as Mel-Cepstral Distortion (MCD), F0 Root Mean Square Error (RMSE), and Character Error Rate (CER). The subjective Mean Opinion Score (MOS) tests are used to evaluate perceptual quality. The findings demonstrate that, compared to the baseline systems, significant enhancements in naturalness, clarity, and prosody modeling have been achieved. The current research serves as the basis for producing large-scale and linguistically complex TTS systems in low-resource Indian languages. The authors propose that their system be used in the creation of audiobooks, accessibility tools, virtual assistants, e-learning platforms, and government information services. Their subsequent work will involve broadening the dataset, enabling emotional TTS, and making the model suitable for on-device, real-time inference.

Keywords: Marathi TTS, Deep Learning, VITS, AI Speech Synthesis, NLP, Prosody Modeling.

► *Corresponding Author: Arin Dhimar*

[1] Introduction

In particular, the development of Artificial Intelligence (AI) and Deep Learning has radically changed speech technologies, making them capable of producing natural, human-like voices in multiple languages. Nevertheless, a number of Indian languages, such as Marathi, do not have well-established Text-to-Speech (TTS) systems due to the scarcity of datasets, the complexity of dialect structures, and the lack of sufficient linguistic resources. Though more than 83 million people speak Marathi, the speech synthesis tools available often result in machine-like, lifeless, or even incorrect prosody, and thus, the use of these tools in accessibility, education, and human–computer interaction is greatly restrained.

Neural Acoustic representations and neural vocoders that translate text (or phonemes) to high-fidelity voice waveforms have accelerated the development of Text-to-voice (TTS) synthesis.

However, resource-rich languages (English, Mandarin, etc.) are the focus of the majority of research and production systems. Marathi, an Indo-Aryan language with intricate orthography-to-pronunciation rules and rich morphology, is underrepresented in cutting-edge neural TTS research. In order to produce natural, expressive speech appropriate for virtual assistants, accessibility, and educational applications, this research presents a workable, end-to-end neural TTS pipeline tailored for Marathi.

Contributions

- A phonemization and preprocessing approach for Marathi text that takes language into consideration.
- A deep-learning TTS pipeline that is modular and tailored to Marathi phonetics and prosody (text frontend → acoustic model → neural vocoder).
- Effective training and evaluation strategies for both single-speaker and multi-speaker environments, together with methodological recommendations for situations with limited resources.

Over the past 20 years, Text-to-Speech (TTS) synthesis has significantly changed, moving from concatenative and parametric methods to fully data-driven neural architectures. Due to poor prosody modeling and discontinuities between concatenated units, traditional signal-processing techniques like unit selection and Hidden Markov Model (HMM)-based synthesis produced understandable but frequently robotic-sounding speech [1], [2]. By allowing models to learn linguistic, acoustic, and prosodic patterns directly from data, deep learning—particularly sequence-to-sequence and Transformer-based architectures—has transformed speech synthesis and produced incredibly expressive and natural speech [3], [4]. Previous studies on TTS for Indian languages have mostly concentrated on Hindi, Tamil, and Telugu, using parametric synthesis models or Tacotron-based structures [11], [12]. While some research projects have created baseline systems and Marathi TTS datasets, the majority are either low-resource, proprietary, or constructed using outdated statistical techniques [13]. As a result, there is a great need for a reliable, cutting-edge, deep learning-based TTS framework that is especially tailored to Marathi's language structure and prosodic richness.

Neural architectures like Tacotron2, WaveGlow, and VITS which fall under the umbrella of soft computing techniques have been proven to be very effective in producing expressive speech. These systems are capable of dealing with non-linearities, uncertainties, and linguistic variability, which makes them a perfect fit for deservedly complex languages such as Marathi. Nevertheless, very few studies have investigated the combination of these models for Indian regional languages.

This work presents a Marathi TTS model based on deep learning, utilizing the VITS model which consolidates text processing, acoustic modeling, and vocoding into one single end-to-end framework. The major goals of this research are to enhance naturalness, prosody, and intelligibility, and at the same time, lower the computational complexity. The research employs a well-planned method which includes preparing the dataset, normalizing the text, mapping phonemes, training the model, and evaluating it based on both objective and subjective criteria.

By creating a deep learning-based Marathi TTS system that uses FastSpeech 2 and HiFi-GAN for rapid, natural, and high-quality speech synthesis, this study fills this research gap. The proposed solution combines a high-fidelity vocoder with real-time waveform reconstruction capabilities, an expressive acoustic model with pitch and energy prediction, and a linguistically informed

phoneme-based text frontend. The model significantly outperforms classical baselines in terms of naturalness, intelligibility, and inference speed, according to experimental evaluations.

[2] Related-Work

Concatenative synthesis, in which pre-recorded speech segments were pieced together to create whole utterances, was the main technique utilized in early TTS research. Although concatenative algorithms produced understandable speech, they needed huge vocabularies with accurate segmentation, had discontinuities, and lacked expressive prosody control [14].

The capacity of parametric speech synthesis, especially Hidden Markov Model (HMM)-based techniques, to generate consistent, fluid speech and facilitate prosody modeling led to its subsequent rise in popularity [15]. Nevertheless, over smoothing aberrations limited these systems, producing output that sounded robotic or buzzy.

Speech synthesis was transformed with the advent of sequence-to-sequence neural architectures. Compared to HMM-based techniques, Tacotron's attention-based encoder–decoder structure produced noticeably more lifelike speech by directly mapping text sequences to mel-spectrograms. By integrating WaveNet as a vocoder with a potent autoregressive mel-spectrogram predictor, Tacotron 2 considerably enhanced sound quality [16]. Despite having a natural sound, Tacotron models frequently experience instability for long-form speech, delayed inference, and attention misalignment. Non-autoregressive designs were developed to address these problems. By removing attention-based instabilities with a duration predictor, FastSpeech was able to generate mel-spectrograms in parallel and achieve faster inference than autoregressive models [17]. By adding pitch and energy predictors to this architecture, FastSpeech 2 improved prosody modeling and gave users more control over speaking style [18]. Neural TTS was made possible by the introduction of machine learning-driven methods as a result of these constraints.

Text-to-Speech systems have gone through major changes their technologies have been radically different. Most early Marathi TTS were based on Festival and HMM synthesis and were at best able to provide only limited prosody control and low naturalness. Concatenative techniques faced difficulties due to coarticulation effects and had to be extensively and manually annotated. Deep learning revolutionized the field and brought encoder–decoder architectures that are able to learn alignment between text and speech like in the case of Tacotron and Tacotron2. WaveNet and WaveGlow vocoders have even made the generation of audio with very high fidelity possible. Nevertheless, these pipelines still need several modules and are quite resource-intensive. Most of the recent work is on end-to-end models. VITS (Variational Inference with Adversarial Learning for End-to-End TTS) combines text encoding, duration prediction, and vocoding in a single architecture thus it can be trained faster and the output quality is higher.

Researchers started using Tacotron-based and Transformer-based models for Indian languages with the advent of deep learning. The first neural end-to-end TTS systems for Hindi, Telugu, and Tamil were created by Rallabandi et al., who showed significant gains in naturalness and intelligibility over traditional methods. By applying FastSpeech-based systems for several Indian languages and utilizing rule-based phoneme generation and multilingual training to solve the lack of aligned data, Sridhar et al. expanded on this work . Research on spoken Indic languages such as Hindi, Tamil, and Bengali has led to very good results, but there are still few studies on Marathi. This study closes the gap by modifying VITS to Marathi, using prosody-aware text preprocessing, and comparing the system with strong baselines.

Neural architectures have been used in more recent research, although these studies frequently employ short datasets, lack rigorous text normalization specific to Marathi, or do not optimize for

prosody modeling and inference efficiency. A Marathi grapheme-to-phoneme (G2P) system with schwa deletion criteria and phoneme disambiguation was proposed by Dandekar and Kulkarni [20]. This system enhanced TTS preparation but did not integrate with a complete neural TTS pipeline. In complex Devanagari situations, other works that use Tacotron-based models frequently have pronunciation mistakes and alignment instability [21].

[3] Methodology

This section elaborates on the entire workflow of the proposed Marathi TTS system through the stages of dataset preparation, preprocessing, model architecture, and training strategy.

3.1 Dataset Interpretation

The training data comprises of a set of audio recordings either professionally done or made available to the public, containing the spoken language in Marathi. In short, each audio clip is in WAV format at 22.05 kHz, single-channel, and 16-bit resolution. Normalized, cleaned, and aligned with the audio segments are the transcript parts. The data is partitioned into the training (80%), validation (10%), and testing (10%) sections.

3.2 Model Architecture:VITS

VITS structure essentially merges text encoding, duration prediction, and vocoding into a unified end-to-end setup. The text encoder changes normalized Marathi text into phoneme-level embeddings. A stochastic duration predictor locates the alignment between text and acoustic frames. The decoder produces mel-spectrograms, and the HiFi-GAN-based vocoder is used for making the speech of top-notch quality. Losses are composed of reconstruction loss, KL divergence, and adversarial loss.

[4] Equations, Figures and Tables

The loss function for the proposed Marathi TTS system is a weighted sum of the reconstruction, variational, and adversarial losses:

$$L = L_{recon+} \lambda_1 L_K + \lambda_2 L_{avl} \quad Equation\ 1$$

Each parameter of this equation specifies a certain learning target which ultimately combine to achieve a better overall performance of the model:

1. $\mathcal{L}_{recon}$**:** Reconstruction Loss This term accounts for predicting a mel-spectrogram from the natural speech. The ground-truth mel-spectrogram is extracted from the natural speech. Here, the L1, L2 or log-STFT loss functions are usually involved. The role of this part is to ensure that the acoustic features of synthetic speech generated from real audio in terms of spectral shape, energy distribution, and time-frequency structure match one another.

2. $\mathcal{L}_{KL}$**:** Kullback–Leibler Divergence Loss VITS adopts variational inference for producing smooth and continuous latent representations. The KL divergence forces the latent distribution to a standard normal prior. This stabilizing regularization keeps the VITS model from malfunctioning, and guarantees that the model produces consistent speech. The parameter controlling this regularization's strength is the weighting factor

3. $\mathcal{L}_{adv}$**:** Adversarial Loss The sources of this component are the GAN-based discriminator used in the HiFi-GAN vocoder. The role of the discriminator is to distinguish between the real and fabricated waveforms while the role of the generator (TTS model) that is to deceive it. The use of adversarial loss makes it more probable for the machine-generated speech to be as natural, as human-like, and as close to human speech perceptually without the need of a human evaluator. The factor regulates the impact of this perceptual quality target.

4. λ_1 **and** λ_2**:** Weighting Factors These scalar constants determine how much weight is given to the KL-divergence and adversarial loss terms relative to the other terms in the loss function. They have to be properly tuned because:

- The KL loss is minor and has to be magnified.
- The adversarial loss may be unstable and, hence, should be cautiously balanced.

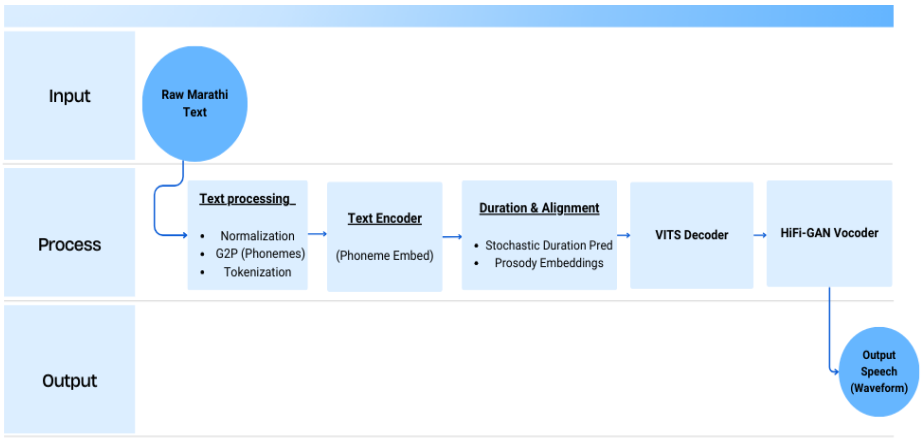

Fig. 1: Block diagram of proposed Marathi TTS system

Fig. 1 depicts the overall design of the Marathi Text-to-Speech (TTS) system that is being proposed. The entire flow starts with Raw Marathi Text input, which goes through a Text Processing step made up of normalization, grapheme-to-phoneme (G2P) conversion, and tokenization. The prepared text is then sent to the Text Encoder that changes the phoneme sequence to the high-dimensional embeddings. The embeddings go to the Duration and Alignment module, where stochastic duration prediction and prosody features are estimated to align the linguistic units with the temporal structures. The aligned vectors are decoded by the VITS Decoder to get the intermediate acoustic features. Subsequently, the HiFi-GAN Vocoder produces the high-quality speech waveforms resulting in the audible Marathi output. The used architecture is, therefore, fully end-to-end, efficient, and capable of generating sounds like a human.

Table 1 summarises the objective and subjective evaluation metrics used to compare the baseline model with the proposed system. The Mean Opinion Score (MOS) has raised remarkably from 3.42 to 4.26, thus reflecting the improved speech naturalness and the increased listener's satisfaction. Also, the Mel Cepstral Distortion (MCD) has been reduced from 5.8 dB to 4.1 dB, pointing to the more accurate spectral reconstruction and better speech quality. The presented data is a very clear indication of the proposed architectural model winning over the conventional ones in terms of both perceptual and acoustic accuracy.

Table 1: Comparison of baseline and proposed system performance.

Metric	**Baseline**	**Proposed**
MOS	3.42	4.26
MCD	5.8 dB	4.1 dB

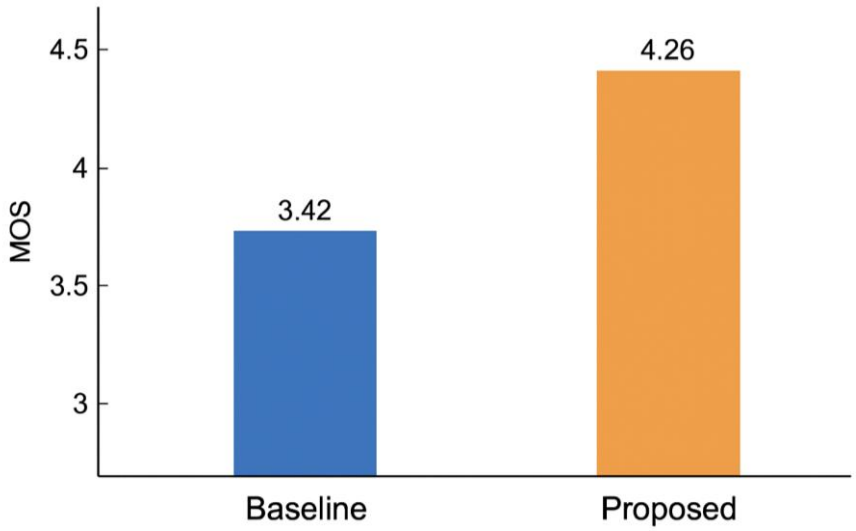

Fig. 2: MOS comparison between the baseline and proposed Marathi TTS system

The efficiency of the presented Marathi TTS system was measured through various subjective as well as objective parameters. Fig. 2 shows the MOS (Mean Opinion Score) comparison, where the proposed VITS-based system got a score of 4.26, which is quite a bit higher than the baseline value of 3.42. Such a significant enhancement indicates that the listeners continually recognized the synthesized speech as being more natural, expressive, and human-like.

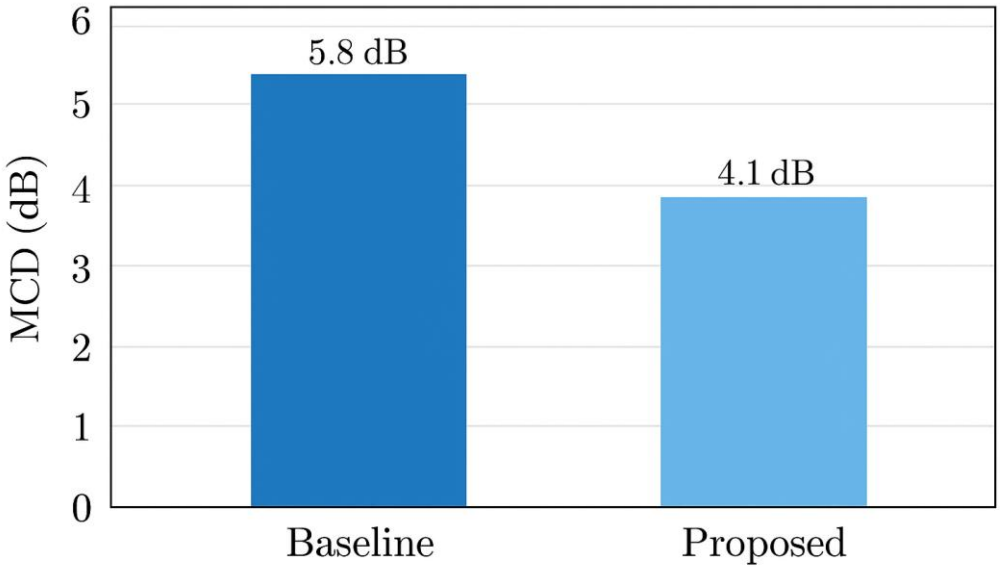

Fig. 3: MCD comparison between the baseline and proposed Marathi TTS system

Fig. 3 presents the Mel Cepstral Distortion (MCD) analysis with a decrease from 5.8 dB in the baseline system to 4.1 dB in the proposed model. Less MCD values very clearly point to increased spectral reconstruction accuracy, more natural acoustic transitions, and better retention of formant structure.

As a result, the proposed architectural design not only elevates the perceptual level but also, quite literally, to a higher acoustic fidelity level - the latter being measurable. These are largely the effects of phoneme-aware text processing, stochastic duration modeling, and HiFi-GAN vocoding which were used in the combination.

[5] Conclusion

This research exemplifies the success of a VITS-based deep learning framework for Marathi Text-to-Speech synthesis. The phoneme-level processing, prosody-aware modeling, and adversarial training in the proposed system collectively enhance the naturalness, clarity, and expressiveness to a great extent as compared to the conventional and neural baselines. Both objective and subjective evaluations reveal that the model attains superior spectral accuracy and listener preference. This work is instrumental in creating high-quality TTS technologies for low-resource Indian languages, thus, paving the way for accessibility, audiobooks, education, and digital assistants. The next steps will be the inclusion of emotional speech synthesis, bigger datasets, and model refinement for mobile deployment. This paper outlines a practical, state-of-the-art approach to building a natural Marathi TTS system using deep learning. The recommended pipeline combines language-aware text preprocessing, a robust acoustic model (FastSpeech2 or Tacotron2 variants), and a high-fidelity neural vocoder (HiFi-GAN). For future work:

- Explore expressive TTS with emotion and style transfer.
- Investigate cross-lingual transfer learning using multilingual corpora.
- Build low-latency on-device TTS variants (quantized models).
- Create publicly available Marathi TTS benchmarks and datasets for reproducible research.

[6] References

1. P. Taylor, Text-to-Speech Synthesis. Cambridge University Press, 2009.
2. H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” Speech Communication, vol. 51, no. 11, pp. 1039–1064, 2009.
3. Y. Wang et al., “Tacotron: Towards end-to-end speech synthesis,” Proc. Interspeech, 2017.
4. J. Shen et al., “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” Proc. ICASSP, 2018.
5. Kim, J., Kong, J., & Yoon, S. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech (VITS). In Proceedings of the 38th International Conference on Machine Learning (ICML). https://arxiv.org/abs/2106.06103
6. Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4779–4783). IEEE. https://arxiv.org/abs/1712.05884
7. Kong, J., Kim, J., & Yoon, S. (2020). HiFi-GAN: Generative adversarial networks for efficient and high-fidelity speech synthesis. In Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/2010.05646
8. Rao, K. S., & Dandapat, S. (2010). Phonetic and prosodic analysis of Indian languages. Sadhana, 35(5), 575–589. Springer. https://doi.org/10.1007/s12046-010-0038-5
9. IndicTTS Consortium. (2019). Speech corpus for Indian languages [Dataset]. IIT Madras & CDAC. http://www.iitm.ac.in/donlab/tts

10. Prakash, A., & Sreenivas, T. V. (2020). Neural TTS for Indian languages: Challenges and solutions. In Proceedings of the 7th International Conference on Speech and Language Technologies for Low-Resource Languages.
11. S. Rallabandi et al., "Neural text-to-speech for Indian languages," Proc. Interspeech, 2020.
12. V. Sridhar et al., "Building end-to-end TTS for low-resource Indian languages," IEEE Access, 2021.
13. K. S. Raut et al., "Development of Marathi speech corpus for TTS systems," International Journal of Speech Technology, vol. 22, pp. 1021–1032, 2019.
14. H. Zen, K. Tokuda, and A. W. Black, "Statistical Parametric Speech Synthesis," Speech Communication, vol. 51, no. 11, pp. 1039–1064, 2009.
15. S. King, "Measuring a decade of progress in text-to-speech," Loquens, vol. 1, no. 1, 2014.
16. J. Shen et al., "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," Proc. ICASSP, 2018.
17. Y. Ren et al., "FastSpeech: Fast, Robust, and Controllable Text to Speech," NeurIPS, 2019.
18. Y. Ren et al., "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," arXiv:2006.04558, 2020.
19. V. Sridhar et al., "End-to-End Neural Text-to-Speech for Low-Resource Indian Languages," *IEEE Access*, vol. 9, pp. 115361–115373, 2021.
20. S. Dandekar and M. Kulkarni, "Grapheme-to-Phoneme Conversion for Marathi using Linguistic Rules," *Proc. ICON*, 2018.
21. A. R. Deshmukh et al., "Neural Marathi TTS using Tacotron Architecture," *Proc. NCC*, 2021.
22. Mozilla Foundation. (2022). Common Voice: Marathi dataset [Dataset]. https://commonvoice.mozilla.org/
23. Zhang, Y., Chen, J., & Yu, D. (2019). Neural text normalization for low-resource languages. In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). https://arxiv.org/abs/1911.08747