STORiCo: Storytelling TTS for Hindi with Character Voice Modulation

Pavan Kalyan, Preethi Jyothi, Preeti Rao, Pushpak Bhattacharyya

IIT Bombay
{190020124@, pjyothi@cse, prao@ee, pb@cse}.iitb.ac.in

Abstract

We present a new Hindi text-to-speech (TTS) dataset and demonstrate its utility for the expressive synthesis of children's audio stories. The dataset comprises narration by a single female speaker who modifies her voice to produce different story characters. Annotation for dialogue identification, character labelling, and character attribution are provided, all of which are expected to facilitate the learning of character voice and speaking styles. Experiments are conducted using different versions of the annotated dataset that enable training a multi-speaker TTS model on the single-speaker data. Subjective tests show that the multi-speaker model improves expressiveness and character voice consistency compared to the baseline single-speaker TTS. With the multi-speaker model, objective evaluations show comparable word error rates, better speaker voice consistency, and higher correlations with ground-truth emotion attributes. We release a new 16.8 hours storytelling speech dataset in Hindi and propose effective solutions for expressive TTS with narrator voice modulation and character voice consistency.

1 Introduction

Speech synthesis has considerably evolved over the last few years, going beyond the goal of achieving understandable and natural speech. It now includes aspects such as expressiveness and other notable qualities of the desired speaking style. English has well-established speech datasets for read speech, such as LJSpeech (Ito and Johnson, 2017), M-AILABS (GmbH, 2019), Blizzard 2013 (King and Karaiskos, 2014), and the recently released Storynory (Kalyan et al., 2023). While the Blizzard 2013 and Storynory datasets include expressive storytelling speech for children, there is a lack of such expressive TTS datasets for Hindi. Hindi is the third most spoken language in the world after English and Mandarin. Although there are Hindi TTS datasets available, such as those released by Indic TTS (Baby and Leela, 2016) and Syspin[1], these datasets primarily consist of neutral-toned read speech and lack expressiveness. However, storytelling for children involves more interactive and conversational speech. In storytelling speech, voice modulation by the storyteller for different characters is key to consider. Using appropriate voices for dialogue and maintaining consistency in character voices are crucial in storytelling. Children's stories can have characters of different species, ages, genders, and giving each character a distinct voice helps keep children engaged.

Previous works, like Greene et al. (2012), aim to predict character voice for a text-to-speech (TTS) system in storytelling. However, they only focus on retrieving the correct speaker voice from a given set of audio based on textual descriptions. Xin et al. (2023) explores improving speech synthesis for audiobooks by considering acoustic and textual contexts. The study uses a multi-speaker Japanese audiobook TTS dataset (Takamichi et al., 2022), different from single-speaker storytelling speech. Nakata et al. (2022) explore character acting in Japanese audiobooks by predicting character-appropriate voices using character embeddings derived from the character's name, sentences and surrounding characters. However, the synthesized audio sample lacks expressiveness and does not evaluate the character's voice consistency when attempting to mimic the character's voice. Kato et al. (2020) focuses on synthesizing Rakugo speech, a form of comic storytelling that only includes character dialogues. The authors create a database and annotate character descriptions based on the conversation. However, in storytelling speech, the character descriptions come from the stories themselves. Moreover, storytelling

[1] https://syspin.iisc.ac.in/

Total Duration	16.8 hours
Total utterances	13876
Mean (s.d.) utterance duration	4.4 (2.9) sec
Total unique stories	150
Sampling rate	16 kHz
Avg. num. of characters per story	4
Narrator utterances	9561
Character utterances	4315

Table 1: TTS data statistics

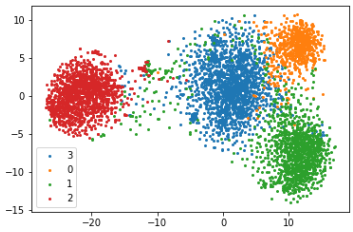

Figure 1: TSNE of speaker embeddings from a speaker encoder. The labels are obtained by applying k-means clustering with $k = 4$.

speech requires controllability in expressiveness, particularly when comparing the narrator's text with that of various characters. Kalyan et al. (2023) present a single-speaker English storytelling TTS dataset that allows shifting the voice from the narrator to the character. In our work, we present an expressive Hindi TTS dataset where the narrator modulates an average of 3-4 character voices apart from the narration.

End-to-end TTS models, such as VAE (Zhang et al., 2019) and GAN-based models (ShuangMa et al., 2019), have demonstrated the ability to generate high-quality speech using phonemes and audio as input. While many TTS models can produce speech comparable to human speech, models utilizing GAN and normalizing flows (Aggarwal et al., 2020) have shown improved expressiveness (Ren et al., 2022). (Kumar et al., 2023) analyses various kinds of neural TTS for Indian languages. Due to its competitive performance for Indian languages, we use VITS TTS (Kim et al., 2021) in a multi-speaker setting.

Our contributions are a) A new, expressive single-speaker Hindi storytelling TTS dataset annotated with character information from the stories. The dataset is unique as the speaker modulates her voice for different characters within a story. b) A Hindi storytelling TTS system with the ability to modulate the voice according to different characters of the story while still maintaining character voice consistency within a story.

2 Dataset

A YouTube channel called Storico[2] offers audio stories in Hindi for children aged 7-12. The stories are collected from the internet, recorded specifically for kids, and narrated by a female speaker in Hindi. The narrator enacts different characters by using multiple voices. The audio recordings include background music and Hindi salutations at the beginning and end of each story. We sought permission from the channel owner to scrape her audio stories from YouTube, with overall 169 stories totaling 19.5 hours. Each story has an average duration of 7.3 minutes and a standard deviation of 3 minutes. The stories were segmented into 10-12 second clips based on silence in the audio. The clips contain background music, animal sounds, etc. but the speaker's speech is clear. The segmented clips were denoised using Meta's Denoiser (Defossez et al., 2020). To ensure quality, 100 random clips were manually verified. All segmented clips were then processed through Nvidia ASR to obtain synthetic transcripts, which were manually corrected and annotated for speaker details.

2.1 Annotation

Two types of annotations were performed on the downloaded data. In the first phase, annotators corrected and added punctuation marks to the transcript of the complete audio story. Four graduates in Hindi literature were hired for this task. In the second phase, four expert Hindi annotators (Appendix B.1) annotated the following information for each corrected transcript of the story:

1. Dialogues: The annotators marked dialogues in the story using quotation marks. They could also identify phrases that described how the dialogue was spoken. Annotators could label the dialogues using 12 emotion labels (if desired).
2. Characters: The annotators identified and labelled all characters in the story. They selected options for each character's gender (Male or Female), age (Adult, Child, or Old), and species (Animal or Human). They also labelled any adjectives or descriptive words for the characters, referred to as keywords.

[2]https://www.youtube.com/@storicokids

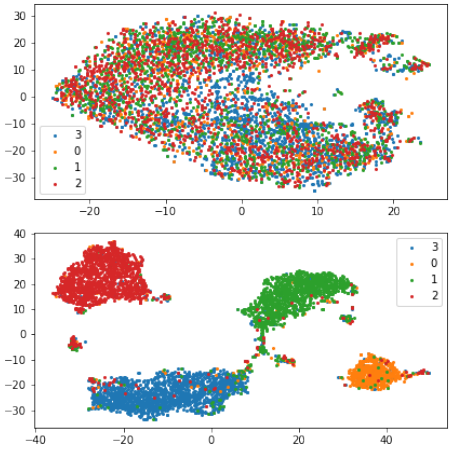

Figure 2: Top: TSNE of the text embeddings obtained from IndicBert (Kakwani et al., 2020) when the speaker description and dialogue are input. Bottom: TSNE of the text embeddings obtained from fine-tuned IndicBert to predict the cluster label obtained from k-means clustering of speaker embeddings

3. Character attribution: After identifying characters and their dialogues, annotators matched each dialogue to the corresponding character directly or through co-reference resolution.

More details about the TTS dataset[3] are described in Appendix A. The final data statistics are provided in Table 1.

3 Methodology

The TTS dataset is unique in two ways. Firstly, it is the first expressive TTS dataset for Hindi that includes speaker information along with the transcript. The speaker information goes beyond just names and includes other details that affect the voice of the characters in the story. Secondly, although the stories were narrated by a single speaker, she modulates her voice for different characters within each story. These characters have different characteristics such as age, gender, and species. However, the speaker can only modulate her voice for a limited number of voice types. Neural speaker identification models treat this modulation as different speakers. We passed the speech of all character instances and randomly sampled 1000 narrator instances, across the stories, through a speaker encoder (Koluguri et al., 2021). The resulting speaker embeddings were visualized using TSNE (Fig. 1), and the plot revealed 4 clusters representing different kinds of voices. We also attempted to identify

[3]Our annotated dataset will be released upon publication. Samples can be found at https://tinyurl.com/4zfxkmxj.

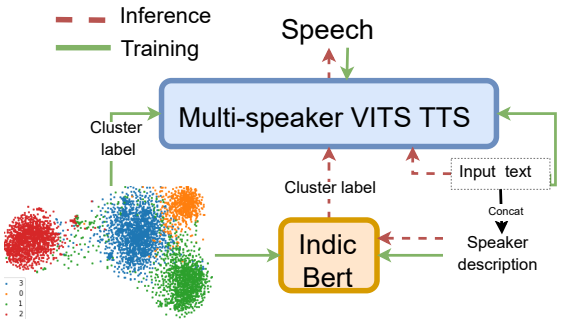

Figure 3: The multi-speaker storytelling TTS pipeline uses character labels predicted by IndicBert that has been fine-tuned on the training speech cluster labels. This allows obtaining labels from text and speaker descriptions during inference.

a clustering based on the speaker descriptions using IndicBert (Kakwani et al., 2020), but no such clustering was observed (Fig. 2).

To address narrator voice modulation and character voice consistency in storytelling speech synthesis, we trained a multi-speaker speech synthesis model on single-speaker data. This approach involved using k-means clustering on speaker embeddings to obtain cluster labels, which were then used as speaker labels in the training data for the text-to-speech (TTS) model. VITS TTS was utilized in a multi-speaker setting, with speaker embedding incorporated for conditioning. We fine-tuned IndicBert to predict one of the 4 cluster labels. The fine-tuned model achieved 75% accuracy on the test set in predicting cluster labels and improved text embedding clustering compared to the pretrained model embeddings (Figure 2). Global conditioning is applied to incorporate the speaker embedding similar to WaveNet (Oord et al., 2016). Since Hindi is a phonetic language, unlike English, we directly use graphemes instead of phoneme sequences as input to the text encoder.

4 Experiments and Results

We conducted the following experiments:

1. VITS SS: VITS on our single-speaker dataset.
2. VITS NC: Multi-speaker VITS with "Narrator" or "Character" as speaker label as in (Kalyan et al., 2023).
3. VITS CL: Cluster labels obtained from k-means clustering used as speaker labels to the multi-speaker VITS TTS model.

The train-test split details and VITS training configuration can be found in Appendix C.

Systems	MOS (CI)	Similarity with ground truth	Character voice consistency
VITS SS	3.25 (0.163)	13.53	22.94
VITS NC	3.18 (0.162)	25.88	7.65
VITS CL	3.62 (0.139)	60.59	69.41
Ground Truth	4.36 (0.105)	-	-

Table 2: Results for subjective tests: 1) MOS for expressiveness (95% confidence intervals) 2) similarity with ground truth 3) character voice consistency. For 2) and 3) values are in percentages, indicating a preference for one system over other systems.

4.1 Subjective Evaluation

We conducted three types of tests for each of the above-mentioned systems. In the first test, listeners rated the expressiveness of the audio on a 1 to 5 scale based on a given sentence and context. A random story from the test set was selected, and 20 sentences containing both narrator and different character sentences were chosen. Each system received ratings for at least five sentences from 39 listeners (Appendix B.2), resulting in 195 judgments per MOS. In the second test, 33 listeners were presented with five sentences and asked to select the audio that is closest to the ground truth in terms of expressiveness and character voice quality. In the third test, listeners selected the pair of audio clips that were closest to each other in the speaker's voice. Table 2 shows the results of all subjective tests. VITS CL outperforms VITS NC and VITS SS, and performs closer to the ground truth. The results demonstrate that multi-speaker training enhances the expressiveness of the generated samples. VITS CL is chosen more than 60% of the time in the last two tests. While the generated audio clips of VITS NC are relatively closer to the ground truth than VITS SS in terms of expressiveness, they are not consistent with the voice of the character.

4.2 Objective Evaluation

We conducted three objective evaluations. First, we tested the synthesized samples using IndicWav2Vec ASR (Javed et al., 2022) to measure their intelligibility. Table 3 shows the Word Error Rate (WER) for all systems. VITS SS has a slightly better WER compared to the other two systems, as transcribing expressive speech can be challenging for the ASR. In the second test, we used Nvidia TitaNet Large (Koluguri et al., 2021) as the speaker verification model to determine if the same speaker spoke the pairs of generated and ground truth audio. Table 3 shows that VITS CL performs better than VITS SS and VITS NC in terms of speaker voice consistency. The slightly lower performance of VITS NC compared to VITS SS aligns with the subjective test results for character voice consistency in Table 2. Finally, we used a 3-dimensional emotion recognition model (Wagner et al., 2022) to extract valence, arousal, and dominance values for each speech segment. Table 3 shows the Pearson correlation values between the ground truth and each system for all three emotion dimensions ($p < 0.001$). Lower correlation values may be due to the emotion recognition model being trained only on real audio files, not synthetic ones. However, VITS CL shows a higher correlation with the ground truth compared to VITS SS and VITS NC for all emotion dimensions.

Systems	WER	Speaker verification	V	A	D
VITS SS	35.76	79.22	0.29	0.28	0.26
VITS NC	38.89	77.73	0.19	0.24	0.25
VITS CL	37.42	83.22	0.33	0.30	0.31

Table 3: Results for objective tests: 1)WER 2)speaker voice verification 3)V,A,D denote Valence, Arousal and Dominance respectively. For test 3) values are Pearson correlation between ground truth and generated samples. For tests 1) and 2) values are given in percentages.

5 Conclusion

This work focuses on Hindi TTS synthesis for expressive storytelling for children. We present a new dataset consisting of expressive narration by a single speaker who modulates her voice for different characters. The annotations provide details about the character voices, including gender, age, species, and keywords from the story text. Despite being a single-speaker dataset, the neural speaker encoder can identify four different speakers arising from the story-telling context. Objective and subjective evaluations demonstrate that training a multi-speaker TTS model on the single-speaker dataset enhances expressiveness and consistency in character voices. As expected, labelling all character voices as one speaker reduces expressiveness and consistency throughout the story. Future work may involve better prediction of the speaker based on the textual speaker descriptions. Additionally, text descriptions can be implicitly used to model the character voice instead of predicting the label.

Limitations

The current approach is a sentence-level text-to-speech system, which is not scalable when dealing with long texts such as lengthy stories or novels. Further research is needed to ensure high-quality synthesis and maintain consistent and expressive character voices throughout extensive narratives. We attempted to identify clustering based on speaker descriptions extracted from the story's text. However, we did not observe any clear clustering, suggesting that relying solely on speaker descriptions and dialogues may not help in reliably determining the speaker's voice types. To enhance accuracy, additional information from the story's content may be needed.

References

Vatsal Aggarwal, Marius Cotescu, Nishant Prateek, Jaime Lorenzo-Trueba, and Roberto Barra-Chicote. 2020. Using Vaes and normalizing flows for one-shot text-to-speech synthesis of expressive speech. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6179–6183.

Arun Baby and Anju Leela. 2016. Resources for indian languages.

Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. 2020. Real time speech enhancement in the waveform domain.

Munich Artificial Intelligence Laboratories GmbH. 2019. The M-AILABS speech dataset. https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/. Accessed: 2023-01-12.

Erica Greene, Taniya Mishra, Patrick Haffner, and Alistair Conkie. 2012. Predicting character-appropriate voices for a tts-based storyteller system. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Keith Ito and Linda Johnson. 2017. The lj speech dataset.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Towards building asr systems for the next billion users. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

T Pavan Kalyan, Preeti Rao, Preethi Jyothi, and Pushpak Bhattacharyya. 2023. Narrator or character: Voice modulation in an expressive multi-speaker tts. *Proc. INTERSPEECH 2023*, pages 4808–4812.

Shuhei Kato, Yusuke Yasuda, Xin Wang, Erica Cooper, Shinji Takaki, and Junichi Yamagishi. 2020. Modeling of rakugo speech and its limitations: Toward speech synthesis that entertains audiences. *IEEE Access*, 8:138149–138161.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Simon King and Vasilis Karaiskos. 2014. The Blizzard Challenge 2013.

Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg. 2021. Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context.

Gokul Karthik Kumar, SV Praveen, Pratyush Kumar, Mitesh M Khapra, and Karthik Nandakumar. 2023. Towards building text-to-speech systems for the next billion users. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, Yuki Saito, Yusuke Ijima, Ryo Masumura, and Hiroshi Saruwatari. 2022. Predicting vqvae-based character acting style from quotation-annotated text for audiobook speech synthesis. In *Proc. Interspeech*, pages 4551–4555.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Yi Ren, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2022. Revisiting over-smoothness in text to speech. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8197–8213, Dublin, Ireland. Association for Computational Linguistics.

ShuangMa, Daniel McDuff, and Yale Song. 2019. Neural TTS stylization with adversarial and collaborative games. In *International Conference on Learning Representations (ICLR)*.

Shinnosuke Takamichi, Wataru Nakata, Naoko Tanji, and Hiroshi Saruwatari. 2022. J-mac: Japanese multi-speaker audiobook corpus for speech synthesis. *arXiv preprint arXiv:2201.10896*.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Florian Eyben, and Björn Schuller. 2022. Dawn of the transformer era in speech emotion recognition: Closing the valence gap.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 45:10745–10759.

Detai Xin, Sharath Adavanne, Federico Ang, Ashish Kulkarni, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2023. Improving speech prosody of audiobook text-to-speech synthesis with acoustic and textual contexts. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. 2019. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949.

A Data formatting

The standard TTS dataset format requires audio segments of 10-12 seconds, along with the corresponding text and speaker labels. To meet this format, the annotated data was formatted accordingly. Each story text was divided into sentences using end-of-sentence punctuations and quotation marks, and standard text normalization was applied. CTC segmentation, similar to Kalyan et al. (2023), was used to obtain audio segments corresponding to these sentences. Nvidia-Nemo[4] was used for CTC segmentation, and any misaligned segments were removed. To include speaker information, sentences labelled as dialogues had annotated speaker information, while all other sentences were labelled as narrator-spoken sentences. As a result, the final TTS formatted data consists of 16.8 hours of single-speaker expressive audio segments, along with their transcripts and speaker information. The data will be released after publication for research and educational purposes.

B Human annotators

B.1 Annotation

Four graduates in Hindi literature from a reputable university were hired for the task of ASR transcript correction and punctuation. These graduates were paid per story based on the market price for ASR transcription. Four expert annotators working in the organization of authors were employed for the rest of the annotation process.

B.2 Evaluation

The listeners of the subjective test are university students who signed up for a 15-minute task after their lecture hours. As a token of appreciation, light snacks were provided for these listeners. The instructions presented to human listeners are provided on the samples page.[5]

C Training

C.1 Train-test split

Similar to LJSpeech (Ito and Johnson, 2017), we created a training split with 12,206 instances, a test set with 1,275 instances, and a validation set with 395 instances. The split was created by selecting ten stories for the test set, totalling 1.14 hours of speech; five stories for the validation set, totalling 0.53 hours of speech; and the remaining stories for the train set, totalling 15.12 hours of speech. The distribution of different factors such as age, gender, and species is balanced across the train, validation, and test sets to maintain consistency. The training process is similar to the VITS model, as detailed in the section below. For fine-tuning the IndicBert (Kakwani et al., 2020), we used all 4,315 character sentences and randomly sampled 1,000 narrator sentences. A normal 90-10 train-test split was used, resulting in a test accuracy of 75%.

C.2 Training details

Training proceeded similarly to VITS (Kim et al., 2021) utilizing the AdamW optimizer with $\beta 1 = 0.8$, $\beta_2 = 0.99$, and a weight decay of $\lambda = 0.001$. The initial learning rate was set at $2e^{-4}$, and the scheduler reduced it by a factor of $0.999^{1/8}$ after each epoch. Two NVIDIA A100 GPUs were employed for training for all experiments with a batch size of 64 per GPU. Each model underwent training for up to 400k steps.

D Potential risks

Developing and using Text-to-Speech (TTS) models raises some ethical concerns. Due to automation, there is a risk of job displacement in fields like acting and broadcasting. There is potential misuse in voice cloning without consent and the creation of deceptive content for children. The work can be employed to generate speech for immoral stories in the voices of known people, harming the social values of young children.

[4]https://github.com/NVIDIA/NeMo

[5]Link to the sample page: https://tinyurl.com/4zfxkmxj