E-TTS: Expressive Text-to-Speech Synthesis for Hindi Using Data Augmentation

Ishika Gupta(✉) and Hema A. Murthy

Department of Computer Science and Engineering, IIT Madras, Chennai, India
{ishika,hema}@cse.iitm.ac.in

Abstract. Current state-of-the-art text-to-speech (TTS) systems trained on read-speech have reduced issues with repetition or skipping of words and can produce natural-sounding speech. However, E2E systems have difficulty producing conversational speech, especially in generating out-of-domain terms, and lack appropriate prosody.

This paper proposes a novel data augmentation approach to build intelligible and expressive speech using FastSpeech2 (FS2). Two different studies are performed. Conversational-style phrases/short interrogative sentences are synthesized using a baseline FS2 system and a hidden Markov model-based speech synthesis (HTS) system. Both systems are trained on 8.5 h of read speech in Hindi. This results in DS1 (FS2) and DH1 (HTS) synthetic datasets, respectively. Using DS1 and DH1, we train FS2 models, namely S1 and H1. While S1 sounds natural, H1 is more intelligible on OOD words. An attempt is made to further adapt these systems with as little as 11 min of original prosodically-rich story data from the same speaker to produce systems S2 and H2, respectively. We evaluate three FS2-based models: Baseline FS2 (vanilla FS2), the proposed models, S2, and H2. The subjective evaluation shows that systems S2 and H2 outperform the baseline FS2 system with an average MOS of 4.16 and 4.42, respectively. Further, we observe that H2 is better than S2 in terms of both MOS and intelligibility. We also do the objective evaluation tests and analyze the synthesized speech based on prosodic attributes to support our claim.

Keywords: Text-to-speech · Data-augmentation · Conversational · Expressive speech · HMM models · Prosody transfer · FastSpeech2

1 Introduction

End-to-End (E2E) text-to-speech (TTS) systems have become very popular owing to their naturalness. Nevertheless, they still possess a limitation in requiring a considerable amount of context-rich training data to enhance their intelligibility on out-of-domain (OOD) words. Next, prosodic variation is an essential component in the recognition of spoken communication [9]. These systems lack

A. Karpov et al. (Eds.): SPECOM 2023, LNAI 14339, pp. 243–257, 2023.
https://doi.org/10.1007/978-3-031-48312-7_20

the ability to accurately render the expressiveness in the synthesized voice, which can be critical for some tasks. The motivation for our work is as follows. A scenario where an error-free conversational TTS system is essential is in the task of dubbing educational classroom lectures from English [23] into Indian languages [19]. The word order in Indian languages is distinctly different from that of English; therefore, conversational prosodic transfer from English to an Indian language is non-trivial. The source videos are technical lectures/extempore, which use a conversational style of speech. So, expressiveness is essential in TTS systems when it comes to lip-syncing lectures in other Indian languages, isochronously. Conversational speech is characterized by an unplanned set of words and abrupt sentence endings. The state-of-the-art (SOTA) E2E systems like FastSpeech2 [30] that are trained on clean read speech do not scale for conversational speech synthesis. Moreover, building monolingual conversational speech synthesis systems using conversational corpora is quite challenging, owing to disfluencies, significant syllable rate variation, and diverse prosody patterns. These factors contribute to "buzziness" in the synthesized speech, making the task more complex [22].

We address two issues in this paper: 1) the synthesized speech does not have the prosodic style required for spontaneous speech. This may be because, unlike the traditional parametric synthesis models, the current neural TTS systems do not explicitly model the prosodic attributes.

2) Besides this, it has also been noted that current TTS models do not generalize well to unseen contexts, such as situations where they are trained on read speech and applied to generate conversational speech. This is possibly due to the fact that existing E2E systems are trained on limited domain-specific training data, so they may be unable to pronounce words from an unseen domain correctly. Figure 1 shows examples of Hindi text which is translated from a technical text in English. The highlighted words in the table are mispronounced by the vanilla FastSpeech2 (FS2) trained on the Hindi read-speech corpus.

To overcome these issues, we present a novel methodology based on data augmentation, which results in an intelligible and expressive TTS model. In this work, we have used FastSpeech2 [30] as a baseline model for comparison.

Conversational Text containing out-of-domain words
When you already have a sorted list then insertion sort will be fast because insert step will be instantaneous. जब आपके पास पहले से ही एक सॉर्टेड लिस्ट है, तब इंसर्शन सॉर्ट काफी तेज होगा क्योंकि इन्सर्ट स्टेप इंस्टैंटेनियस है.
The meaning of percentile is if we divide the whole strength of students into different groups, then in which group you belong to. पर्सेंटाइल का अर्थ है यदि छात्रों की पूरी श्रृंखला को सौ समूहों में विभाजित किया जाता है, तो आप किस समूह से संबंधित हैं.

Fig. 1. Different examples of conversational text structure.

The proposed techniques described in this paper can be applied to the JETS (Jointly Training FastSpeech2 and HiFi-GAN for End to End Text to Speech) [17] and VITS (Variational Inference with adversarial learning for end-to-end TTS) [12] frameworks as well. We have also tried training the VITS architecture on our Hindi speech corpus, but we couldn't see much improvement over the FastSpeech2 synthesis output. Also, this model cannot disentangle prosody information from speech [18], so we have not considered VITS in the evaluation. The FastSpeech2 model proves to be a strong baseline system, with its ability to predict pitch and energy information in addition to the mel-spectrogram from the text transcription. This added capability allows the model to capture a wide range of speech variations and nuances. We propose two approaches, each of which involves training on read-speech data and usage of conversational style multi-domain text as new training data, followed by fine-tuning. The training overview is shown in Fig. 2. The use of a multi-domain text gives us the added advantage of not only producing labeled context-rich training data but also broadening the domain of the TTS system. The text comprises short conversational phrases and interrogative sentences. Both FS2 and Hidden Markov model (HMM) based speech synthesis (HTS) systems have the ability to learn prosody from the statistics of the training data by utilizing the pitch and energy information from the ground truth. The findings of this work also indicate that the inclusion of short conversational phrases in the synthetic training data enhances the prosody modeling capability of TTS systems.

The novelty of the work is in the use of classical HMM-based [36] synthesis to direct an E2E system. HTS can handle OOD words to a large extent due to tree-based clustering. These systems explicitly model pitch variations using a multi-space probability density function. Studies have also shown that HTS systems, being statistical models, can be intelligibly trained on smaller amounts of data [29].

Main Contributions: 1) We are able to build expressive TTS with only conversational-style synthetic data and give the model an additive improvement in prosody with a small amount of prosodic-rich data.
2) a scalable language and architecture-agnostic approach for building an expressive TTS system.

The rest of the paper is organized as follows. Section 2 reviews the related literature. Section 3 presents the proposed system overview and approach. Section 4 discusses the analysis done on synthesized speech outputs using qualitative and quantitative measures and presents the results of objective and subjective listening tests performed on the synthesized speech output of the trained systems. Finally, Sect. 5 concludes the paper.

2 Related Work

Several studies have been conducted owing to the growing research interest in expressive speech synthesis. Research in conversational speech synthesis has also

led to various approaches towards either recording conversational speech corpora or selectively utilizing existing conversational data.

In TTS applications, [11] generates a large number of synthetic utterances using an auto-regressive (AR) model to improve the quality of non-AR TTS models. [7] uses embedding computed from trained global style tokens (GST) to build multi-style TTS with a good prosodic variation. [24] investigates whether a data mixing strategy can improve conversational prosody for a target voice based on monologue data from audiobooks by adding real conversational data from podcasts. [10] uses synthetic data in the desired speaking style on top of the available recordings generated by applying voice conversion. Further fine-tuning the model on a small amount of expressive samples for the target speaker helps improve naturalness and style inadequacy. [6] focuses on increasing the naturalness of TTS systems on specific domains by adding domain-specific speech to their database. [31] propose an ensemble approach that combines multiple prosody predictors to achieve more accurate and natural-sounding speech. Trained on a diverse dataset of expressive speech samples, the ensemble produces more expressive prosodic patterns by combining predictions from various models.

Parallel efforts to use hybrid paradigms for TTS have also been explored. [20] uses a hybrid approach combining HTS with the neural-network-based Waveglow vocoder [27] using histogram equalization (HEQ) in a low resource setting for improving the quality of speech output. Our work differs in its focus on achieving the challenge of bringing both expressivity and intelligibility in the synthesized speech by combining classical parametric and neural speech synthesis methods.

3 Proposed Approach

We investigate the data augmentation approach using multi-domain text. The overview of our proposed system framework is depicted in Fig. 2. In this study, we propose two TTS systems, namely: FS2 augmented (S1) and HTS augmented TTS system (H1), to build conversational speech synthesis systems. Both the proposed approaches consist of 3 stages: (i) Source model training, (ii) target model training (iii) Fine-tuning the model with as minimal as 11 min of expressive speech data to build proposed systems S2 and H2. The basic modules involved in training the proposed TTS system are described as follows:

3.1 FS2-Based Data Augmentation

Figure 2(a) shows an overview of FS2-based data augmentation. Here, our source model is the FS2 system, which we use to generate synthetic data. Earlier work shows that prosodic phrase breaks can be identified for Indian languages quite accurately if the training data is accurately marked with prosodic phrase marks [26]. During the source model training on read-speech data, the text is manually annotated with phrase breaks to facilitate the learning of the system to predict the pitch for a complete phrase. A fundamental difference between read speech and conversational speech lies in the fact that conversational speech is typically

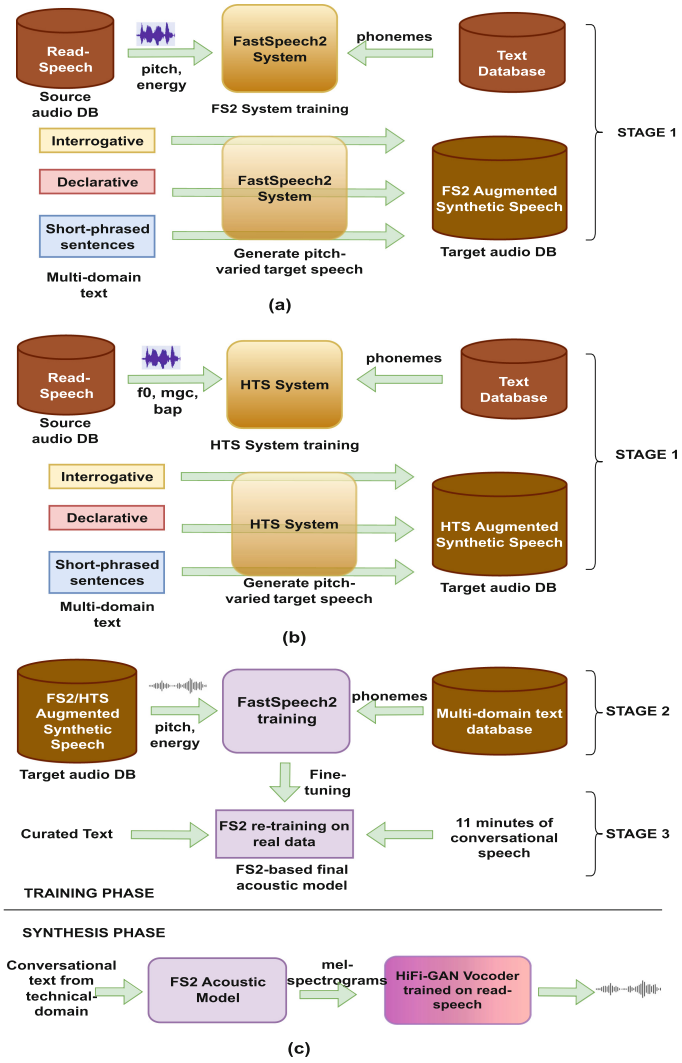

Fig. 2. Overview of the proposed method for building expressive TTS. We have trained the FS2 model using data augmentation: (a) FS2-based data augmentation, (b) HMM-based data augmentation, (c) Shows the training stages 2 and 3 and synthesis phases of both the proposed TTS systems.

composed of short phrases or sentences that are concatenated without punctuation and may have abrupt endings too. A baseline FS2 system attempts to predict the prosody based on the characteristics learned from read speech, which is inadequate for conversational speech owing to its grammatical incorrectness. This is especially evident in conversational speech in Indian languages, where utterances can contain many phrases. Generally, in E2E and HTS systems, prosody at the

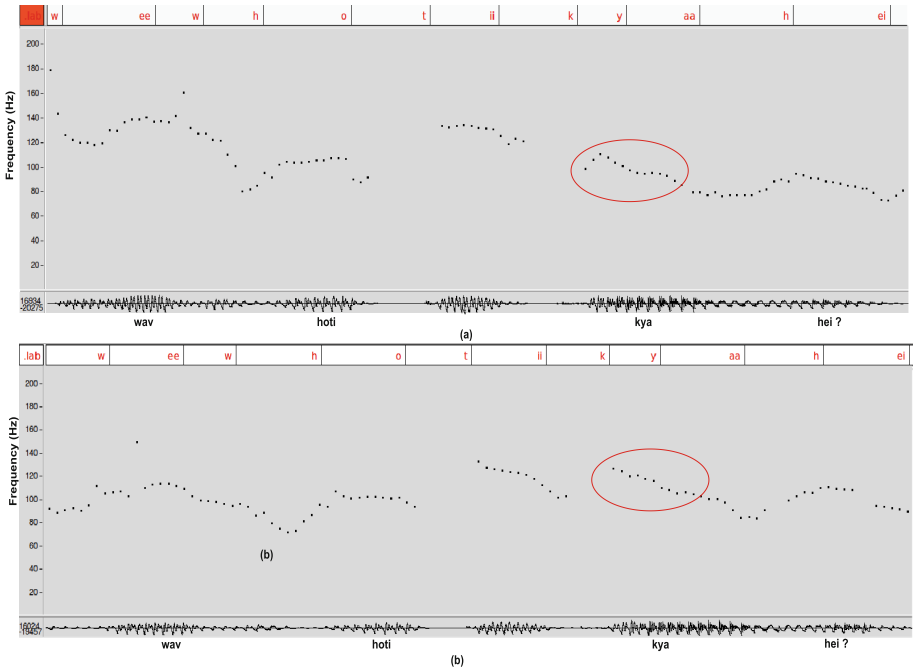

Fig. 3. Pitch contour plotted for a test phrase when (a) it is synthesized as an isolated sentence (b) it occurs in a long sentence.

beginning and end of phrases is learned accurately. Figure 3 and 4 depicts the change in pitch contour and short-term energy when a short phrase is synthesized in isolation and when synthesized as a part of a long sentence. From Fig. 3(a), we can observe that, there is a rise in the pitch at the word '*kya*' and at the end of the question phrase *"wave hoti kya hai?"*, shown in a red oval, whereas the pitch is not modeled so correctly in Fig. 3(b). Similarly, significant fluctuations are visible in the energy plot in Fig. 4(a) than in Fig. 4(b), leading to more stress on words. We can infer that training the FS2 system at a phrase level can help predict pitch better, consequently improving the prosody of synthesized speech. Therefore, while building system S1 (Table 1) on FS2-generated synthetic speech (DS1), we ensure that corresponding multi-domain training data encompasses short conversational phrases and declarative and interrogative sentences.

3.2 HMM-Based Data Augmentation

Figure 2(b) overviews our HTS-based data augmentation. Here, the source model is the HTS system, which we use to generate synthetic data. During training on read-speech data, the same manually annotated text is passed to help the HMM model learn the prosodic cues. Since HMMs explicitly model prosody, we generate HTS synthetic speech dataset (DH1) corresponding to the multi-

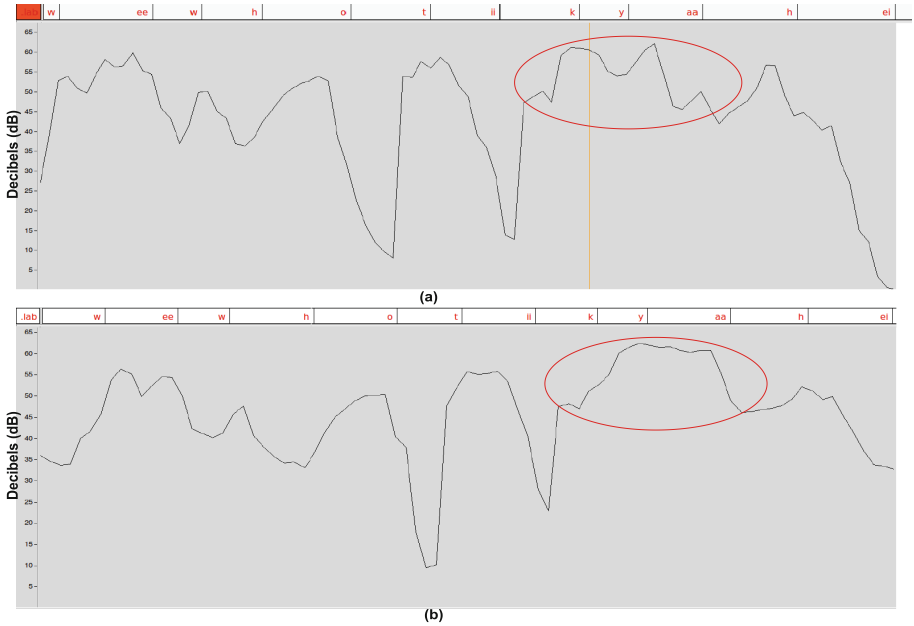

Fig. 4. Short-term energy plotted for a test phrase when (a) it is synthesized as an isolated sentence (b) it occurs in a long sentence.

domain text. While training the FS2 model on this phrasal synthetic speech dataset, the system learns to model the pitch at the start and end of phrases better than what it would have learned with a speech dataset at a sentence level, which is quite depicted in Fig. 3. There have also been some studies that report that prosody modeling is better done at phrase level [26].

3.3 Multi-domain Text Data Collection

We crawled text from Kaggle [2] web source in Hindi for augmenting data. The text data is carefully collected and contains short phrases from various domains ranging from health, lifestyle, science, technology, and ordinary colloquial conversations to interrogative and Yes-No questions. The intent of having such diversity in text data is to train the TTS model with a lot of context information and to enable the TTS model to learn various kinds of phonotactics. Totally we have curated around 14,450 sentences ($\approx$ 15 h) in Hindi. The average number of words per sentence in multi-domain text ranges between 7 to 10.

3.4 HTS Training

To train the HTS model, the text is first transcribed in terms of its constituent phones, which are represented using the common label set (CLS) representation [28]. The speech waveform is aligned at the phone level using a hybrid

HMM-DNN approach. Having a precisely aligned speech database at the syllable/phone level is crucial for accurately mapping linguistic and acoustic units, which is essential for developing high-quality TTS systems [3]. The HTS voice is trained on context-dependent pentaphone units using the aligned speech data. Since the HTS model first segments into phrases and performs embedded reestimation at the syllable level (where syllable boundaries are obtained using signal processing) for training, it produces accurate penta-phone models. Phone HMMs are concatenated to obtain the utterance HMMs [4]. The duration model predicts the number of frames reserved for each phone. The acoustic model predicts the acoustic features for the required number of frames. The HTS engine internally uses a Mel-generalized log spectrum approximation (MLSA) vocoder, which is used to synthesize the utterances. Decision tree-based clustering is performed during HTS training to account for unseen context based on a phonetic yes-no question set, frequently encountered in the technical domain [34]. Thus, HTS can broadly extrapolate to out-of-domain word scenarios.

3.5 Text-to-Speech System Training

Our TTS model consists of two main components: (1) a non-autoregressive FS2-based acoustic model for converting input phoneme sequences into the mel-spectrograms, frame by frame, and (2) a non-autoregressive HiFi-GAN vocoder [13] that is capable of producing high-fidelity speech waveforms from mel-spectro grams. Montreal Forced Aligner (MFA) [21] is used to obtain the alignments between the utterances and the phoneme sequences. Figure 2(c) shows the training process of TTS with the proposed data augmentation. After training the model on the synthetic dataset, we fine-tune the model on 11 min of prosodically-rich expressive speech with the same configurations for 50 more epochs.

4 Analysis of Synthesized Speech

Experiments are conducted in Hindi, which is an Indo-Aryan language, for both male and female data. Studio-recorded Hindi male and female (8.5 h) datasets from the open-source IndicTTS database [5] are considered. We use the HTK [1] [35] and ESPnet toolkits [32] for building voices with default configurations. A HiFi-GAN model is trained separately for Hindi male and female datasets on this 8.5 h of read speech data. Table 1 summarizes various systems trained for each dataset. The text in parentheses indicates the training data used. The synthetic speech produced by baseline FS2, S2, and H2 systems is analyzed on dubbed technical lectures in terms of prosodic parameters.

4.1 Qualitative Analysis

A set of Swayam lectures [25], comprising 347 sentences, is synthesized using the baseline FS2 model (System 2) and our proposed systems S2 and H2. We compare and analyze their pitch and energy distributions.

Table 1. List of systems trained.

Systems	TTS Models
System 1	HTS trained on read-speech (8.5 h)
System 2 (Baseline)	FS2 trained on read speech (8.5 h)
System 3	FS2 trained on FS2 synthesized speech (15 h)**(S1)**
System 4 (Proposed 1)	FS2 trained on FS2 synthesized speech, fine-tuned on conversational speech **(S2)** (8.5 h + 11 min)
System 5	FS2 trained on HTS synthesized speech (15 h) **(H1)**
System 6 (Proposed 2)	FS2 trained on HTS synthesized speech, fine-tuned on conversational speech **(H2)** (8.5 h + 11 min)

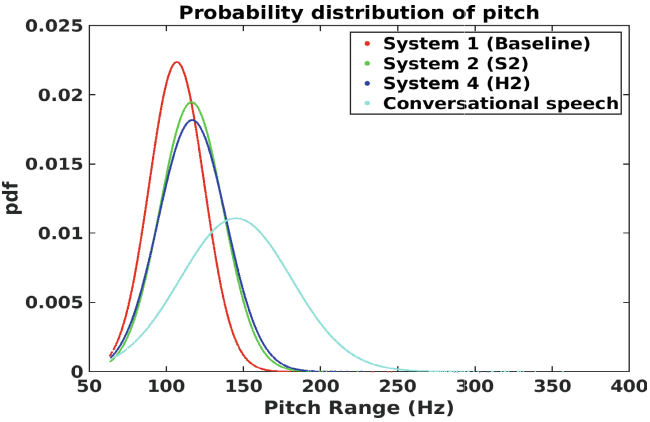

Fig. 5. Pitch plotted for Hindi male synthesized speech on a lecture set and its corresponding English conversational speech.

Pitch. Pitch is a crucial factor in determining the expressiveness of a sentence [31]. From Figure 5, it is observed that there is a wide variation in the pitch trajectory for a conversational speech taken from the lecture. Comparatively, the proposed systems, S2 and H2, are able to capture the finer variations in the pitch better than the FS2 baseline system. This clearly indicates the prosodic richness of S2 and H2 over baseline. A further pitch improvement is noticed in H2 than S2, indicating more closeness toward conversational speech.

Energy. Energy is another critical parameter in speech synthesis quality. The changes in the energy of the speech signal can be used to indicate emphasis on certain words. From Figure 6, we can clearly see a slight increase in the amplitude level for the proposed systems, closer to that of conversational speech in contrast to the baseline. On comparing S2 and H2, further improvement in energy is noticed in system H2.

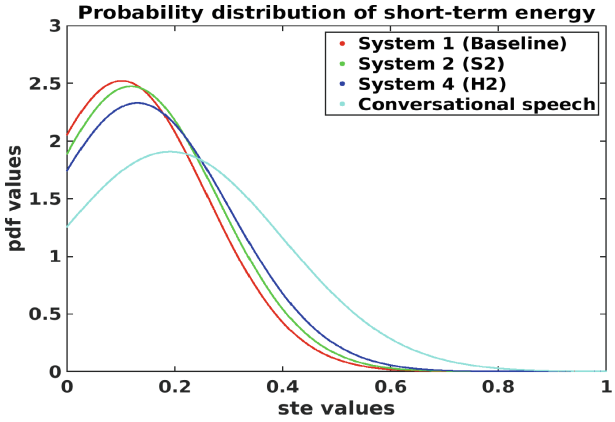

Fig. 6. Energy plotted for Hindi male synthesized speech on a lecture set and its corresponding English conversational speech.

4.2 Subjective Evaluation

Subjective evaluations are performed for the baseline FS2 and the proposed models S2 and H2 for Hindi male and female voices. The comprehension mean opinion score (CMOS) test is performed to assess the system's performance in terms of prosodic variation and intelligibility in the synthetic speech output. Comprehension MOS [33] was proposed to evaluate whether a listener is able to engage with synthetic speech of long durations. Contrast this with sentence-based subjective evaluations. It is a subjective evaluation test that not only assesses the quality of the speech but also checks whether the listener is able to comprehend the content of the audio clearly. It is found in studies that prosodic variations are better realized when long test utterances are presented in context as it helps users to perceive the pitch variations distinctly [8] [15].

We synthesize a coherent paragraph having contextual dependency and then quiz the listener based on the content to check the following: a) expressiveness and b) clarity of the speech. The test includes questions based on the content. 30 native evaluators of Hindi, male, and female, were asked to listen and rate 4 paragraphs presented in random order. MOS is calculated based on the expressiveness of synthetic speech. Listeners were asked to rate the overall quality of synthesized speech on a scale ranging from 1–5, with 1 being poor and 5 being human-like. The intelligibility of the samples is calculated based on the listener's understandability score on a similar scale of 1–5. For the comprehension test, 4 paragraph questions with a mix of translated technical lectures and stories synthesized in Hindi were considered. The comprehension score for each system is obtained by calculating the percentage of correct answers in the evaluation test. CMOS scores are presented in Table 2 and Table 3 for Hindi male and female, respectively. Comparing results of the question: *Rate the audio in terms of expressiveness*, we see that MOS scores are better for H2 than Baseline FS2 and S2, although intelligibility scores are similar for all 3 systems, although H2

Table 2. Comprehensibility MOS scores for Hindi male TTS.

Systems Evaluated	MOS Score	Intelligibility	**Comprehension Score**
Baseline FS2	3.525 ± 0.23	4.425	94.16 %
S2	4.25 ± 0.21	4.575	94.8 %
H2	**4.45 ± 0.18**	**4.65**	**96.4 %**

Table 3. Comprehensibility MOS scores for Hindi female TTS.

Systems Evaluated	MOS Score	Intelligibility	Comprehension Score
Baseline FS2	3.45 ± 0.2	4.325	93.12 %
S2	4.075 ± 0.26	4.495	95.23 %
H2	**4.395 ± 0.19**	**4.557**	**95.71 %**

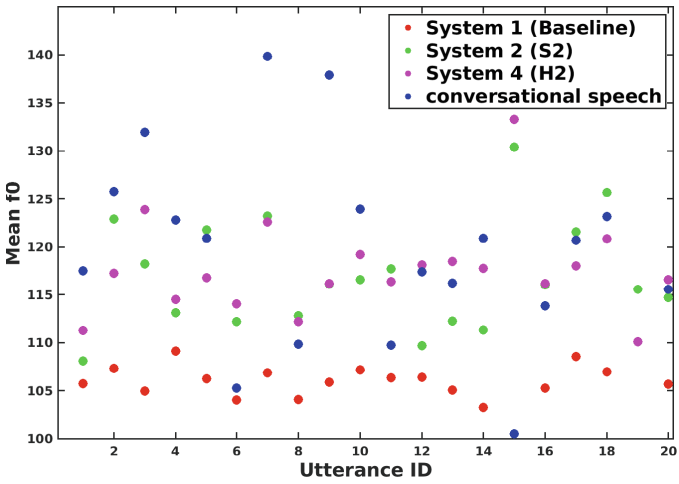

Fig. 7. Mean F0 comparison: baseline system vs. proposed systems.

performs slightly better. A p-test is performed to ensure the results are statistically significant, with both p-values <0.05.
A video lecture demo using the audios synthesized from baseline and the proposed system can be found in this link[1].

4.3 Objective Evaluation

Mean F0: Fundamental frequency (F0) is the most crucial prosody feature [16]. We have used mean F0 as one of the objective evaluation metrics to assess the effectiveness of the proposed systems and how they influence the pitch profile

[1] https://www.iitm.ac.in/donlab/preview/web_demo/index.html.

Table 4. Average increment in the mean and std deviation of F0 when the same text is synthesized using System S2 and H2 vs. baseline TTS.

System	System S2	System H2
Mean F0 change	+12.94	+13.41
Std Dev F0 change	+10.25	+14.45

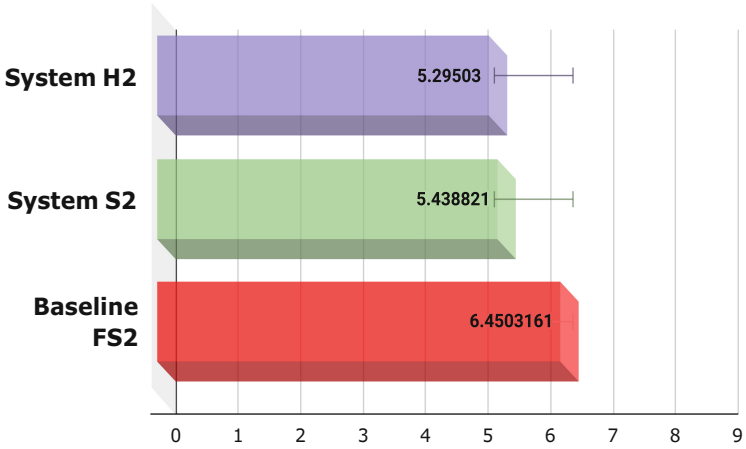

Fig. 8. MCD scores on synthesized utterances across baseline and proposed systems.

of the generated expressive audio with respect to the baseline FS2 system. To calculate the mean F0, 20 synthetic utterances in Hindi male voice were synthesized from the held-out unseen texts using the baseline FS2 and the proposed systems S2 and H2. The mean pitch (F0) is extracted from each system's utterances, including the original conversational speech, and is plotted and compared in Fig. 7.

We observe that mean F0 is higher and varies significantly (like in conversational speech) in both the proposed system's utterances compared to the baseline, even though all utterances were generated in the same speaker's voice. The average difference of mean F0 and standard deviation of F0 between proposed and baseline systems are reported in Table 4.

MCD Score: Mel-cepstral distortion (MCD) [14] serves as a metric to quantify the dissimilarity between two mel cepstral sequences. A smaller MCD score indicates a high degree of similarity between the synthesized and natural speech, indicating that the synthetic speech closely resembles the natural speech. 20 unseen Hindi sentences are synthesized using baseline FS2 and proposed TTS systems S2 and H2 in the male voice. Each system's generated utterances are compared with respect to the corresponding utterances of original conversational speech using dynamic time-warped MCD. The MCD scores for the systems are

depicted in Fig. 8. Clearly, systems S2 and H2 scores seem to have less distortion compared to their baseline counterpart, indicating a closer resemblance to conversational speech.

5 Conclusion and Future Work

This work presents a language and architecture-agnostic approach to building expressive TTS systems. Our proposed method explores FS2-based and HTS-based data augmentation methods to build an intonation-based TTS system. We have demonstrated that the model is able to learn expressiveness from synthetic data and result in prosody transfer when coupled with a few minutes of expressive speech data in Hindi. It means there is no requirement to rely on additional conversational corpora for building conversational TTS systems. This makes our proposed approach vastly scalable to other languages. We plan to extend this work to Indo-Dravidian languages too. It's worth noting that there exist other techniques to introduce prosodic cues into the model, such as the pre-training of the FastSpeech2 variance adapters, whereas our work can obtain a similar effect.

Acknowledgement. We want to thank the Ministry of Electronics and Information Technology (MeitY), Government of India (GoI), for funding the project "Speech Technologies in Indian Languages" (SP/21-22/1960/CSMEIT/003119). Our sincere thanks to Anusha Prakash for her help in training the FastSpeech2 TTS model.

References

1. HMM-based toolkit. https://htk.eng.cam.ac.uk/
2. Kaggle Online Community. https://www.kaggle.com/datasets/rushikeshdarge/hindi-text-corpus?select=hi https://www.kaggle.com/datasets/rushikeshdarge/hindi-text-corpus?select=hi
3. Baby, A., Prakash, J.J., Murthy, H.A.: A hybrid approach to neural networks based speech segmentation. Dimension **40**, 3 (2017)
4. Baby, A., Prakash, J.J., Subramanian, A.S., Murthy, H.A.: Significance of spectral cues in automatic speech segmentation for Indian language speech synthesizers. Speech Commun. **123**, 10–25 (2020)
5. Baby, A., Thomas, A.L., Nishanthi, N.L., Consortium, T.: Resources for Indian languages. In: CBBLR - Community-Based Building of Language Resources, pp. 37–43. Tribun EU, Brno, Czech Republic (2016). https://www.iitm.ac.in/donlab/tts/database.php
6. Chu, M., Li, C., Peng, H., Chang, E.: Domain adaptation for TTS systems. In: ICASSP, vol. 1, pp. I-453. IEEE (2002)
7. Chung, R., Mak, B.: On-the-fly data augmentation for text-to-speech style transfer. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 634–641. IEEE (2021)
8. Clark, R., Silen, H., Kenter, T., Leith, R.: Evaluating long-form text-to-speech: comparing the ratings of sentences and paragraphs. In: Proceedings of 10th ISCA Speech Synthesis Workshop, 2019, pp. 99–104 (2019)

9. Cutler, A., Dahan, D., Van Donselaar, W.: Prosody in the comprehension of spoken language: a literature review. Lang. Speech **40**(2), 141–201 (1997)
10. Huybrechts, G., Merritt, T., Comini, G., Perz, B., Shah, R., Lorenzo-Trueba, J.: Low-resource expressive text-to-speech using data augmentation. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6593–6597. IEEE (2021)
11. Hwang, M.J., Yamamoto, R., Song, E., Kim, J.M.: TTS-by-TTS: TTS-driven data augmentation for fast and high-quality speech synthesis. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6598–6602. IEEE (2021)
12. Kim, J., Kong, J., Son, J.: Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: International Conference on Machine Learning, pp. 5530–5540. PMLR (2021)
13. Kong, J., Kim, J., Bae, J.: HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. Adv. Neural. Inf. Process. Syst. **33**, 17022–17033 (2020)
14. Kubichek, R.: Mel-cepstral distance measure for objective speech quality assessment. In: Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, vol. 1, pp. 125–128. IEEE (1993)
15. Latorre, J., Yanagisawa, K., Wan, V., Kolluru, B., Gales, M.J.: Speech intonation for TTS: Study on evaluation methodology. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)
16. Laures, J.S., Bunton, K.: Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions. J. Commun. Disord. **36**(6), 449–464 (2003)
17. Lim, D., Jung, S., Kim, E.: JETS: jointly training fastspeech2 and HiFi-GAN for end to end text to speech. In: Ko, H., Hansen, J.H.L. (eds.) Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18–22 September 2022, pp. 21–25. ISCA (2022)
18. Liu, Z., et al.: Controllable and lossless non-autoregressive end-to-end text-to-speech. arXiv preprint arXiv:2207.06088 (2022)
19. M, M.R.K., Kuriakose, J., S, K.P.D., Murthy, H.A.: Lipsyncing efforts for transcreating lecture videos in Indian languages. In: Proceedings of 11th ISCA Speech Synthesis Workshop (SSW 11), pp. 216–221 (2021)
20. M., M.R.K., Sudhanshu, S., Anusha, P., A, M.H.: A hybrid hmm-waveglow based text-to-speech synthesizer using histogram equalization for low resource Indian languages. In: INTERSPEECH, pp. 2037–2041 (2020)
21. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M.: Montreal forced aligner: trainable text-speech alignment using kaldi. In: Interspeech, pp. 498–502 (2017)
22. Mukherjeee, B., Prakash, A., Murthy, H.A.: Analysis of conversational speech with application to voice adaptation. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 765–772. IEEE (2021)
23. NPTEL: National programme on technology enhanced learning. https://nptel.ac.in/
24. O'Mahony, J., Lai, C., King, S.: Combining conversational speech with read speech to improve prosody in text-to-speech synthesis. In: INTERSPEECH (2022)
25. Paul, P., Bhuimali, A., Kalishankar, T., Aithal, P., Rajesh, R.: Swayam: the platform for modern and enhanced online and flexible education-a knowledge survey. Int. J. Appl. Sci. Eng. **6**(2), 149–155 (2018)

26. Prakash, J.J., Murthy, H.A.: Analysis of inter-pausal units in Indian languages and its application to text-to-speech synthesis. IEEE/ACM Trans. Audio, Speech, Lang. Process. **27**(10), 1616–1628 (2019)
27. Prenger, R., Valle, R., Catanzaro, B.: Waveglow: a flow-based generative network for speech synthesis. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3617–3621. IEEE (2019)
28. Ramani, B., Christina, S.L., Rachel, G.A., Solomi, V.S., Nandwana, S.K., Samudravijaya, K., et al.: A common attribute based unified hts framework for speech synthesis in Indian languages. In: 8th ISCA Workshop on Speech Synthesis (2013)
29. Rao, K.S., Narendra, N.: Source modeling techniques for quality enhancement in statistical parametric speech synthesis. Springer (2019)
30. Ren, Y., et al.: Fastspeech 2: fast and high-quality end-to-end text to speech. In: International Conference on Learning Representations (2020)
31. Teh, T.H., et al.: Ensemble prosody prediction for expressive speech synthesis. In: ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2023)
32. Watanabe, S., et al.: Espnet: end-to-end speech processing toolkit. arXiv preprint arXiv:1804.00015 (2018)
33. Wester, M., Watts, O., Henter, G.E.: Evaluating comprehension of natural and synthetic conversational speech. In: Proceedings of speech prosody. vol. 8, pp. 736–740 (2016)
34. Wu, Z., Watts, O., King, S.: Merlin: an open source neural network speech synthesis system. In: SSW, pp. 202–207 (2016)
35. Young, S.J., Young, S.: The HTK hidden Markov model toolkit: Design and philosophy. Design and philosophy (1993)
36. Zen, H., Nose, T., Yamagishi, J., et al., S.: The HMM-based speech synthesis system (HTS) version 2.0. In: SSW, pp. 294–299. Citeseer (2007)