To appear in Proc. APSIPA ASC 2025

LAPS-Diff: A Diffusion-Based Framework for Hindi Singing Voice Synthesis With Language Aware Prosody-Style Guided Learning

Sandipan Dhar*, Mayank Gupta*, Preeti Rao
Department of Electrical Engineering, Indian Institute of Technology Bombay, India.
Email: sandipandhartsk03@gmail.com, guptamayank8899@gmail.com, prao@ee.iitb.ac.in

***Abstract*—The field of Singing Voice Synthesis (SVS) has seen significant advancements in recent years due to the rapid progress of diffusion-based approaches. However, capturing vocal style, genre-specific pitch inflections, and language-dependent characteristics remains challenging, particularly in low-resource scenarios. To address this, we propose LAPS-Diff, a diffusion model integrated with language-aware embeddings and a vocal-style guided learning mechanism, specifically designed for Bollywood Hindi singing style. We curate a Hindi SVS dataset and leverage pre-trained language models to extract word and phone-level embeddings for an enriched lyrics representation. Additionally, we incorporated a style encoder and a pitch extraction model to compute style and pitch losses, capturing features essential to the naturalness and expressiveness of the synthesized singing, particularly in terms of vocal style and pitch variations. Furthermore, we utilize MERT and IndicWav2Vec models to extract musical and contextual embeddings, serving as conditional priors to refine the acoustic feature generation process further. Based on objective and subjective evaluations, we demonstrate that LAPS-Diff significantly improves the quality of the generated samples compared to the considered baseline and state-of-the-art (SOTA) model for our constrained dataset that is typical of the low resource scenario.**

***Index Terms*—Singing Voice Synthesis, Diffusion Model, Bollywood Hindi Singing Style**

I. Introduction

Singing Voice Synthesis (SVS) is the process of generating natural-sounding singing voices using statistical or Artificial Intelligence (AI)-based algorithms, guided by musical scores [1]. Unlike Text-to-Speech (TTS) synthesis systems [2], the primary objective of SVS models is to produce a singing voice that is accurately synchronized with the musical composition as given by a music score.

In the recent past, improved variants of diffusion models have exhibited significant advancements in SVS [3], [4]. In particular, the DiffSinger model [1] is widely recognized as a pioneering approach in diffusion-based SVS research and has demonstrated promising results. It includes an auxiliary decoder that generates the target mel-spectrogram from music score embeddings. Specifically, the shallow diffusion mechanism introduced in DiffSinger accelerates the inference by providing the denoiser a boundary predicted intermediate mel-spectrogram rather than the noise input. Learn2Sing 2.0 [2] is also a well-known diffusion-based approach in which the authors adapted the GradTTS framework [5] to transform speech data into singing audio. While diffusion-based approaches have shown impressive results, models like VISinger2 [6] have also exhibited notable performance in SVS by utilizing a conditional variational autoencoder (CVAE)-based framework.

Although the discussed models have shown significant advancements, most of these existing SVS models, including DiffSinger, have been trained and evaluated on singing audio datasets [7] that typically contain more than 5 hours of audio data (pure single voice) in languages such as Mandarin, Korean, Japanese, and English [8]. Moreover, each of these language-specific singing datasets belongs to a distinct musical genre with its own peculiar characteristics for melody (pitch variation), rhythm, and pronunciation. These aspects are also strongly influenced by the individual singer's unique vocal characteristics. However, due to the scarcity of labeled singing data in most languages and genres, it is highly challenging for an SVS model to effectively capture linguistic content, style, and pitch related information from low-resource singing data, highlighting a significant research gap.

To address this challenge of capturing content information, vocal style, and pitch variations from low-resource SVS data, we introduce LAPS-Diff, a diffusion model that incorporates language-aware embeddings and prosody-style guided learning mechanism. The proposed model is built on the DiffSinger framework [1]. As part of this work, we curate a dataset of about one hour duration of Hindi Bollywood-style songs by a single male singer, in particular, audio data obtained by vocal separation from original music productions and process the audio to obtain the music scores. In this work, our primary goal is to adapt an existing singing voice synthesis model to a new language by leveraging pre-trained language models, along with models that capture prosody and musical features, to achieve compatibility with a new singing style. This work is based on the Hindi Bollywood singing style, using singing data recorded in a male singer's voice. To effectively capture the lyrical content information, we leverage additionally two pre-trained language models, IndicBERT [9] and XPhoneBERT [10], for Hindi word and phone-level embeddings. We combine these embeddings with music score embedding [1] to form an enriched content representation. Further, we integrate a

[*] Sandipan Dhar and Mayank Gupta are co-first authors

style encoder [11] and pre-trained JDCNet pitch extraction model [12] to obtain style and pitch information (melody) to compute the corresponding losses for the training of the auxilary decoder of our proposed LAPS-Diff model. Given the critical importance of preserving expressive pitch inflections in the songs in the genre of interest, we go beyond L1 loss of the fundamental frequency or pitch (F0), and investigate the role of preserving melodic contour shape in the loss function. The resulting style and pitch losses are found to improve the model's ability to effectively capture vocal style and pitch-related information, enabling a closer resemblance to the underlying singing dynamics. Moreover, we utilize pre-trained MERT [13] and IndicWav2Vec [14] models for extracting musical feature embeddings and contextual embeddings, respectively, to use as conditional priors to the denoiser for improving the reverse diffusion process [1] through explicit feature guidance. Finally, we evaluate the performances of the proposed LAPS-Diff model and the considered diffusion-based SVS models using our Bollywood Hindi dataset with objective and subjective measures and discuss the obtained improvements. The proposed model is derived from the DiffSinger model [1] which is therefore considered the baseline.

II. Dataset

In our work, we select the popular genre of Indian music called Bollywood. While considered a pop genre in terms of song structure, Bollywood is known to be influenced by Indian classical traditions, with Hindi as its primary language[1]. We selected a collection of 38 songs performed by the renowned singer Arijit Singh.

The original song audio (as obtained from publicly accessible sites) is processed for vocals separation using a commercial tool [15]. Automatically detected silences greater than 500 ms are used to segment the song into the lyrics lines/phrases that are then manually assigned to each audio segment. Next, we achieve the automatic alignment of the audio and text at the phone level using forced alignment with a hybrid automatic speech recognition system [16] trained on an adult Hindi speech dataset. Sung lyrics exhibit notable differences from speech, chiefly on word pronunciations in terms of vowel duration as well as the nature of pitch variation over this duration. It is also common to find the addition of schwa on the last consonant of words. We account for these differences with a specially constructed lexicon augmented with alternate pronunciations. Syllable boundaries are obtained by merging the corresponding aligned phones. The vocal pitch is extracted at 10 ms intervals using an autocorrelation based method for fundamental frequency and voicing using parselmouth library. Brief uv segments and pauses are linearly interpolated in pitch. Each syllable is assigned a MIDI note corresponding to the quantized mode of the F0 values across the syllable segment. Syllables longer than 200 ms are analysed for pitch fluctuations by computing the MIDI pitch separately for non-overlapping sub-segments and assigning a Slur flag to those segments that register an intra-syllable MIDI note change. This is the first attempt to the best of our knowledge to build an Indian music dataset for an SVS task. We adopt the same music score format as the Opencpop dataset [7].

III. Proposed Method

The proposed model is built upon the DiffSinger framework [1]. The same encoder model $E(.)$ is utilized to extract the music score embedding $\boldsymbol{e}_m$ from the music score x, as depicted in Fig. 1. Details of the working mechanism is discussed in the following subsections.

A. Feature Fusion

To efficiently extract Hindi lyrical content related information we use two pre-trained language models: IndicBERT (as a text encoder $T(.)$) and XPhoneBERT (as a phone encoder $H(.)$) both of which are compatible with the Hindi language. Thereafter, we combine IndicBERT's word level embedding $\boldsymbol{e}_w$ and XPhoneBERT's phone level embedding $\boldsymbol{e}_p$ with $\boldsymbol{e}_m$ to derive the final fused embedding $\boldsymbol{e}_c$ (i.e., fusion of three feature embeddings via summation operation [17]), as described mathematically in Eq.(1) (shown in Fig. 1),

$$\begin{aligned} \mathbf{e}_c &= T(\text{text}) + H(\text{phone}) + E(x), \\ &= \mathbf{e}_w + \mathbf{e}_p + \mathbf{e}_m, \end{aligned} \tag{1}$$

The objective of constructing the fused embedding $\boldsymbol{e}_c$ is to provide a better representation of linguistic content in the embedding space.

B. Decoder Training

The auxiliary decoder $AD(.)$ used in this work is a mel-spectrogram decoder capable of reconstructing mel-spectrograms $\hat{M}$ from the music score embedding [1]. However, in this work, we provide the combined embedding $\boldsymbol{e}_c$ as the input to the auxiliary decoder to generate mel-spectrograms (i.e., $\hat{M} = AD(\boldsymbol{e}_c)$, where $\hat{M}$ is the generated version that is conditioned on the supplied embedding. To effectively train the auxiliary decoder in capturing relevant information, we incorporate style and pitch losses as optimization objectives in our work.

1) Style Loss: To extract the vocal style related information in terms of style embedding (style vector), we used a residual connection based Convolutional Neural Network (CNN) model as the style encoder $S(.)$ (the architectural framework is same as [11]). The extracted style embedding carries an abstract representation (i.e., latent representation) of the singer's vocal characteristics. To compute the style loss $\mathcal{L}_{style}$, we first extract the style embeddings $\boldsymbol{e}_{s_m}$ and $\boldsymbol{e}_{s_{\hat{m}}}$ from both M and $\hat{M}$ respectively, as illustrated in Fig. 1. We then obtain $\mathcal{L}_{style}$ using the Mean Squared Error (MSE), as defined in the following equation,

$$\mathcal{L}_{style} = \frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{e}_{s_{m_i}} - \boldsymbol{e}_{s_{\hat{m}_i}}\right\|^2, \tag{2}$$

where N is the total length of the embedding. In Eq.(2), $\boldsymbol{e}_{s_m} = S(M)$ and $\boldsymbol{e}_{s_{\hat{m}}} = S(\hat{M})$.

[1] Dataset details, and samples of reference and generated audio, can be accessed https://www.notion.so/LAPS-Diff-Dataset-and-Generated-samples-in-submission-for-APSIPA-2025-1ca8559c5842802b9877f0d5dd4cb2f0

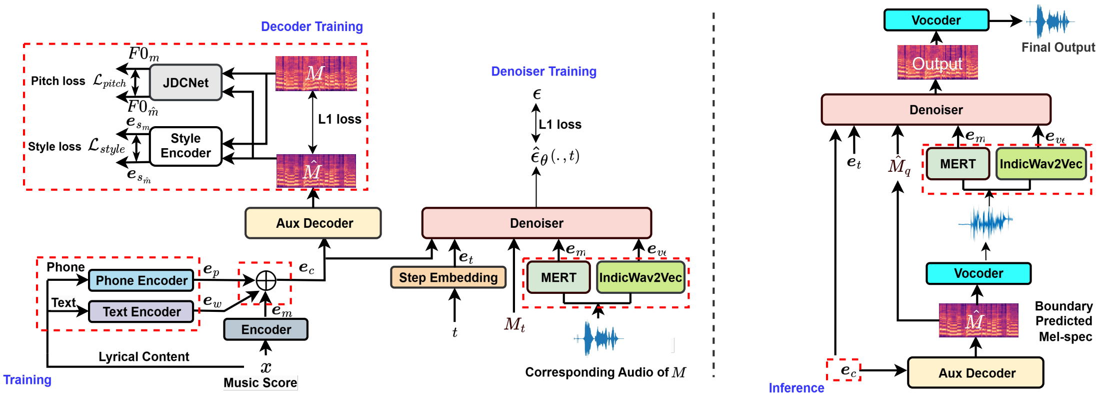

Fig. 1: Schematic overview of the proposed LAPS-Diff model, including training and inference stages. The components enclosed within the red-dashed boxes represent the specific enhancements of this work over the DiffSinger framework.

2) Pitch Loss: The pitch loss $\mathcal{L}_{pitch}$ is calculated using MSE between the pitch values $F0_m$ and $F0_{\hat{m}}$ extracted from both M and $\hat{M}$ respectively, using the pre-trained JDCNet model [12] denoted as $J(.)$. The pre-trained JDCNet model used in this work is adopted from StyleTTS2 [11]). Due to StyleTTS2's impressive voice cloning abilities in terms of naturalness and expressiveness, we kept the framework of $S(.)$ and $J(.)$ the same as in StyleTTS2. Moreover, JDCNet [12] is extensively trained on singing voice data, making it highly effective in capturing fine pitch details from mel-spectrograms [18]. In a novel approach in this work, we introduce the concordance correlation coefficient (CCC) [19] to exploit the linear correlation between matched pitch contours $F0_m$ and $F0_{\hat{m}}$. That is, the CCC (given in the Eq.(3) below) balances the requirements of pitch height and correlation, making it particularly suitable for evaluating the perceptual similarity of two pitch contours. The CCC value ranges from -1 to 1. To influence the pitch loss, $\mathcal{L}_{pitch}$, we incorporate $(1-\text{CCC})$ as a multiplicative factor such that a highly correlated pair of $F0_m$ and $F0_{\hat{m}}$ results in a lower penalty, whereas a lower correlation increases the penalty. The pitch loss $\mathcal{L}_{pitch}$ is mathematically defined in Eq.(3),

$$\mathcal{L}_{\text{pitch}} = (1 - \text{CCC}) \times \left(\frac{1}{K} \sum_{i=1}^{K} \|F0_{m_i} - F0_{\hat{m}_i}\|^2 \right) \tag{3}$$

where K represents the length of the $F0$ sequence, $F0_m = J(M)$ and $F0_{\hat{m}} = J(\hat{M})$. To prevent the loss from becoming zero for highly correlated pairs, we introduce a minimum multiplicative factor of 0.01.

Apart from the introduced losses, all other losses remain the same as in DiffSinger's Github repository [20].

C. Denoiser Training

In the training phase of the denoiser $D(.)$ in DiffSinger, the model takes a noisy mel-spectrogram M_t from the t-th $(t \in 0, T)$ diffusion step, along with the step embedding $\boldsymbol{e}_t$ for time step t and the music score embedding $\boldsymbol{e}_m$, to estimate the noise $\hat{\epsilon}_\theta(\cdot)$ [1] for obtaining $(t-1)$-th denoised melspectrogram M_{t-1} as given in [1] denoiser training of DiffSinger.

To improve the denoising process in reverse diffusion, particularly for accurate estimation of the denoised mel-spectrogram having precise detail of the musical and linguistic information, we introduce two additional feature embeddings as conditional priors to provide more explicit guidance during denoising. Between the two, one captures musical characteristics through MERT embedding [13] (focusing on tonal and pitch-related attributes), and the other captures linguistic information in terms of IndicWav2Vec [14] embedding to obtain a richer representation in the denoised mel-spectrogram, as shown in Fig. 1. MERT and IndicWav2Vec are two pre-trained models (trained using self-supervised learning approach) that efficiently extract respective embeddings from singing audio data (shown in Fig. 1), as they are trained on singing (music) data and multilingual speech data (including Hindi), respectively.

D. Inference Stage

In the inference stage, the optimal auxiliary decoder $AD^*(\cdot)$ generates the mel-spectrogram $\hat{M}$. The pre-trained HiFi-GAN vocoder [1] is then used to reconstruct the audio waveform from $\hat{M}$, from which $\boldsymbol{e}_{mert}$ and $\boldsymbol{e}_{vec}$ are extracted. These embeddings along with $\boldsymbol{e}_c$, $\boldsymbol{e}_t$, and the intermediate boundary-predicted [1] representation $\hat{M}_q$ are used to condition the optimal denoiser $D^*(\cdot)$, as illustrated in Fig. 1 to generate the final output. Here, $\hat{M}_q$ is derived from $\hat{M}$ through the shallow diffusion mechanism of DiffSinger [1], where q denotes the shallow diffusion step.

IV. Experiments

We present our experimental setup with implementation details.

1) Data Preprocessing: The overall duration of the singing voice segments of our Hindi Bollywood SVS dataset, introduced in Section II, is approximately 65 minutes. With 38 unique songs, segmented into 397 sung phrases each with duration ranging from 5 to 15 seconds. The dataset splits are based on distributing the songs across training, validation, and test sets with details shown in Table I. Each audio file is sampled at 16 kHz with 16-bit quantization. The Hindi text (Devanagari script) is converted into its corresponding phonemes using the IIT-M Hindi phoneset [21]. The size of the phoneme vocabulary considered is 43. We extract mel-spectrograms considering 80 mel-bins using frame size of

512 and hop size of 128. Similar to DiffSinger [1], the mel-spectrograms are normalized to the range $[-1, 1]$.

TABLE I: Description of dataset splits

Split	Songs	Segments	Duration (Min)
Train	31	344	57.14
Validation	3	25	3.76
Test	4	28	3.76

2) Implementation Details: In our work, we use the IndicBERT model [9], pre-trained on 12 Indian languages, including Hindi. In contrast, the pre-trained XPhoneBERT model [10] is trained on nearly 100 languages, also including Hindi. The embedding dimension for both IndicBERT and XPhoneBERT is 768. Linear projection is applied to reduce the dimensionality to 256, ensuring compatibility with the 256-dimensional music score embedding $\boldsymbol{e}_m$.

The MERT model considered in our work is trained on diverse datasets comprising musical instruments and singing voices [13]. The IndicWav2Vec [14] is trained on nine Indian languages including Hindi. We employ the pre-trained IndicWav2Vec to obtain content embeddings. The size of MERT and IndicWav2Vec embedding is 1024. We consider batch size as 48 and optimize the model using AdamW optimizer [1] with a learning rate of 1×10^{-3}, for a total training iteration of 2×10^5. Validation is performed in every 2×10^3 iterations. We set the value of $T = 100$ and β value increases linearly from from 1×10^{-4} to 6×10^{-2} over the diffusion steps. The LAPS-Diff experiments use the same library versions as DiffSinger. All experiments run on an NVIDIA A100 GPU with 80GB memory. The training process takes approximately 2 GPU days.

V. Results and Discussion

The performance of the proposed LAPS-Diff model is compared with the SOTA DiffSinger and VISinger2 [6] models using both objective and subjective tests, on our Hindi Bollywood SVS dataset. Additionally, an ablation study is carried out to demonstrate the effectiveness of the individual components integrated into LAPS-Diff. For objective evaluation, we have used MCD, logF0 RMSE, MAE, cosine similarity and voiced/unvoiced (V/UV) ratio [22]–[24]. Additionally, we have utilized the metas's Audiobox Aesthetic Score (AES) [25] for perceptual evaluation. Whereas, for subjective evaluation, we have considered Mean Opinion Score (MOS) [22].

Four ablation settings (as summarised in Table II and Table III) are considered in our experiments: (i) training DiffSinger from scratch with text and phone embeddings (Ablation 1), (ii) training DiffSinger from scratch with MERT and IndicWav2Vec features as conditional priors (Ablation 2), (iii) training DiffSinger from scratch with JDCNet-based CCC pitch loss (Ablation 3), and (iv) training DiffSinger from scratch with style loss (Ablation 4). All settings were trained using the same dataset, preprocessing steps, number of training iterations, and overall hyperparameters to ensure consistency.

A. Objective Evaluation

The objective evaluation results in Table II are the averages computed across all the test data. We note the superior performance of the proposed LAPS-Diff model compared to the DiffSinger and VISinger2. LAPS-Diff outperforms across most metrics, achieving the highest average cosine similarity, lowest MAE, highest V/UV accuracy, second lowest logF0 RMSE and MCD. These results suggest that LAPS-Diff captures speaker characteristics more accurately and ensures closer alignment of the spectral features (including pitch) with the ground truth, thereby enhancing the overall quality of the synthesized singing voice. Moreover, the higher V/UV ratio achieved by the proposed model indicates its effectiveness in capturing the details of voiced and unvoiced regions, contributing to greater naturalness in the synthesized singing voice.

Based on the objective scores of the ablation studies reported in Table II, it can be seen that the addition of each component independently leads to an improvement in the corresponding metric. Each component contributes to the expressiveness of the singing audio data, the JDCNet pitch loss improves the logF0 RMSE, the style loss improves the MCD and the IndicBERT, XphoneBERT, MERT and IndicWav2Vec embeddings improve the overall PQ and CE metrics of Meta's Audiobox. The analyses of the ablation study consistently highlight the importance of each component in the proposed model.

Meta's AES evaluation metrics [25], adopted in our work, are designed to assess the aesthetic quality of diverse audio types including speech, music and general sounds, using a pre-trained model that closely aligns with human perceptual judgments. We select the measures relevant to synthesis from music score (i.e. content is pre-defined). Production Quality (PQ) capturing the technical clarity and fidelity of the audio, including its dynamics, frequency balance, and spatial attributes. Content Enjoyment (CE) evaluates the emotional and artistic expressiveness of the audio—highlighting aspects like creativity, mood, and subjective appeal. Since we have a single voice, the AES ”production complexity” measure (PC) is irrelevant. Table II clearly shows that the LAPS-Diff model outperforms all other models across both AES dimensions, achieving the highest scores (except for the reference scores, as expected).

Visual Analysis: Next, we present a visual comparison of pitch contours from selected audio segments for LAPS-Diff, DiffSinger, VISinger2, and the reference, corresponding to faster and slower singing rates shown in Fig. 2. For the fast singing rate, the pitch contours in regions A and C of LAPS-Diff closely resemble those of the corresponding ground truth regions, whereas DiffSinger and VISinger2 show noticeable deviations. The region B contains a pitch transition to a much higher value, and Fig. 2 shows LAPS-Diff appropriately synthesizes the high-pitched part, matching the ground truth. In contrast, DiffSinger and VISinger's pitch pattern in region B deviates visibly from the music score MIDI. This indicates LAPS-Diff's ability in handling high-pitched regions and highlights its effectiveness in capturing fine-grained pitch details across diverse frequency ranges. A similar observation holds for region P (Fig. 2) in the slower singing rate scenario,

TABLE II: Objective and perceptual evaluation of proposed LAPS-Diff, baseline DiffSinger, and various ablation settings

Model	Cosine Similarity (↑)	MAE (↓)	V/UV Accuracy (↑)	Log F0 RMSE (↓)	MCD (↓)	Audiobox CE (↑)	Audiobox PQ (↑)
Reference	-	-	-	-	-	**6.206**	**7.637**
LAPS-Diff (Proposed)	**0.987**	**0.165**	**0.907**	0.141	7.897	4.770	6.552
DiffSinger (Baseline)	0.982	0.197	0.890	0.155	8.200	4.004	6.340
VISinger2	0.975	0.207	0.883	0.149	8.741	3.280	6.313
Ablation 1 (Baseline + IndicBERT + XPhoneBERT)	0.973	0.171	0.890	0.159	7.983	4.200	6.499
Ablation 2 (Baseline + MERT + IndicWav2Vec)	0.978	0.185	0.869	0.151	9.445	4.151	6.408
Ablation 3 (Baseline + JDCNet pitch loss)	0.978	0.171	0.898	**0.118**	7.928	3.460	6.355
Ablation 4 (Baseline + style loss)	0.986	0.169	0.880	0.145	**7.883**	3.869	6.511

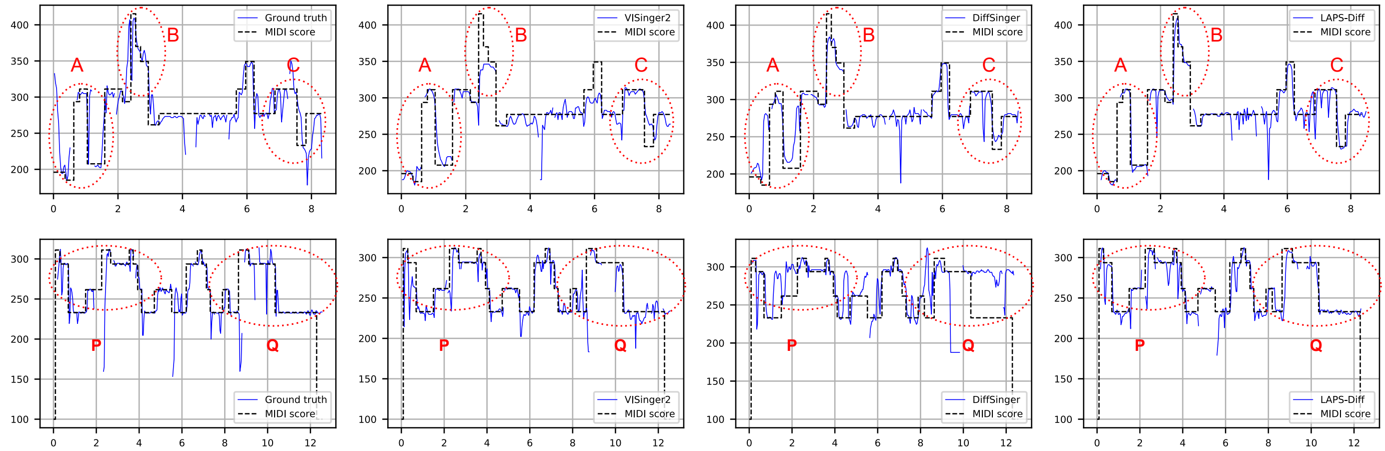

Fig. 2: Visualization of the F0 contour comparing ground truth with synthesized outputs from VISinger2, DiffSinger and LAPS-Diff, all with reference to the MIDI score. The vertical axis shows frequency (Hz), and the horizontal axis represents time (seconds). Top row contains a sample with faster singing rate, and bottom shows a sample with slower singing rate.

where LAPS-Diff better matches the ground truth contours. Moreover, in region Q, which corresponds to a held (and therefore long duration) vowel, DiffSinger fails to follow the within-vowel pitch variation, while LAPS-Diff shows high similarity with the ground truth. These comparisons indicate that LAPS-Diff captures pitch information more accurately and consistently than DiffSinger.

B. Subjective Evaluation

MOS Testing: We conducted MOS subjective evaluation test to assess the naturalness of the generated audio samples and the ground truth. A total of 15 listeners, aged between 20-35 years and without any known hearing impairments, participated. Our test set consists of 4 songs: one each with fast and slow singing rates and two with average singing rates. The evaluation included 6 segments (out of a total of 28) from our test set. We take segments from each of the considered 8 categories, resulting in a total of 48 samples. These were randomly permuted and arranged in 6 test sets of 8 each. The results of the MOS subjective evaluation test are shown in Table III.

As observed from the MOS scores in Table III, the proposed LAPS-Diff model achieves a significantly higher MOS compared to the baseline, confirming the benefit of our proposed enhancements. Based on the ablation study results, it can be seen that IndicWav2vec, MERT and JDCNet leads to a high increase in MOS, indicating the need for a robust metric of pitch, music and language information in the training process.

TABLE III: MOS with 95% confidence intervals

Model	MOS (↑)
Reference	**4.53 ± 0.26**
LAPS-Diff (Proposed)	3.40 ± 0.34
DiffSinger (Baseline)	2.87 ± 0.44
VISinger2	2.47 ± 0.38
Ablation 1 (Baseline + IndicBERT + XPhoneBERT)	2.83 ± 0.58
Ablation 2 (Baseline + MERT + IndicWav2Vec)	3.01 ± 0.18
Ablation 3 (Baseline + JDCNet pitch loss)	3.05 ± 0.38
Ablation 4 (Baseline + style loss)	2.95 ± 0.47

VI. Conclusion

In this work, we introduced LAPS-Diff, a novel SVS model specifically designed for the low-resource Hindi Bollywood singing style. We present a new labeled dataset for Hindi Bollywood music of one-hour duration (audio provided via available public links). By integrating linguistic content embeddings, incorporating style and pitch supervision through loss functions, and leveraging conditional priors, LAPS-Diff achieves notable improvements in the expressiveness of the synthesized singing voice compared to the DiffSinger baseline and VISinger2 model. This work primarily demonstrates how pre-trained language models can be effectively leveraged in SVS systems to make them adaptable to new languages. Future efforts will address computational optimization of training and inference. We will study gains in generated quality that are achievable as the dataset size is expanded, including multi-lingual SVS, with an emphasis on enhancing generalization across a variety of vocal styles.

VII. Acknowledgement

Contributions by Ankit Anand and Gowtham S. in dataset preparation are gratefully acknowledged, as also the funding support by Saregama India Ltd.

References

[1] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "Diffsinger: Singing voice synthesis via shallow diffusion mechanism," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 11 020–11 028, Jun. 2022.

[2] H. Xue, X. Wang, Y. Zhang, L. Xie, P. Zhu, and M. Bi, "Learn2sing 2.0: Diffusion and mutual information-based target speaker svs by learning from singing teacher," in *Interspeech*, 2022.

[3] S. Kim, M. Jeong, H. Lee, M. Kim, B. J. Choi, and N. S. Kim, "Makesinger: A semi-supervised training method for data-efficient singing voice synthesis via classifier-free diffusion guidance," in *Interspeech 2024*, 2024, pp. 1865–1869.

[4] J.-S. Hwang, S.-H. Lee, and S.-W. Lee, "Hiddensinger: High-quality singing voice synthesis via neural audio codec and latent diffusion models," *Neural networks : the official journal of the International Neural Network Society*, vol. 181, 2023.

[5] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *International Conference on Machine Learning*, 2021.

[6] Y. Zhang, H. Xue, H. Li, *et al.*, "Visinger2: High-fidelity end-to-end singing voice synthesis enhanced by digital signal processing synthesizer," in *Interspeech 2023*, 2023, pp. 4444–4448.

[7] Y. Wang, X. Wang, P. Zhu, *et al.*, "Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis," in *Interspeech 2022*, 2022, pp. 4242–4246.

[8] J. Shi, Y. Lin, X. Bai, *et al.*, "Singing voice data scaling-up: An introduction to ace-opencpop and ace-kising," *Interspeech 2024*, 2024.

[9] S. Doddapaneni, R. Aralikatte, G. Ramesh, *et al.*, "Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages," in *Annual Meeting of the Association for Computational Linguistics*, 2022.

[10] L. T. Nguyen, T.-L.-G. Pham, and D. Q. Nguyen, "Xphonebert: A pre-trained multilingual model for phoneme representations for text-to-speech," in *Interspeech*, 2023.

[11] Y. A. Li, C. Han, V. S. Raghavan, G. Mischler, and N. Mesgarani, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," *Advances in neural information processing systems*, vol. 36, 2023.

[12] S. Kum and J. Nam, "Joint detection and classification of singing voice melody using convolutional recurrent neural networks," *Applied Sciences*, vol. 9, no. 7, 2019, ISSN: 2076-3417.

[13] Y. LI, R. Yuan, G. Zhang, *et al.*, "MERT: Acoustic music understanding model with large-scale self-supervised training," in *The Twelfth International Conference on Learning Representations*, 2024.

[14] T. Javed, S. Doddapaneni, A. V. Raman, *et al.*, "Towards building asr systems for the next billion users," in *AAAI Conference on Artificial Intelligence*, 2021.

[15] G. Lab, *Gaudio studio: Ai-powered audio processing*, 2024. [Online]. Available: https://studio.gaudiolab.io/.

[16] D. Povey, A. Ghoshal, G. Boulianne, *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, IEEE Signal Processing Society, 2011.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019.

[18] Y. A. Li, C. Han, and N. Mesgarani, "Styletts: A style-based generative model for natural and diverse text-to-speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 19, pp. 283–296, 2022.

[19] B. T. Atmaja and M. Akagi, "Evaluation of error- and correlation-based loss functions for multitask learning dimensional speech emotion recognition," *Journal of Physics: Conference Series*, vol. 1896, 2020.

[20] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, *Diffsinger: Singing voice synthesis via shallow diffusion mechanism*, https://github.com/MoonInTheRiver/DiffSinger, 2021.

[21] D. o. C. S. IIT Madras and Engineering, *Indictts common label set*, 2020. [Online]. Available: https://www.iitm.ac.in/donlab/indictts/commonLabelSet.

[22] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.

[23] J. Xue, Y. Deng, Y. Han, Y. Li, J. Sun, and J. Liang, "Ecapa-tdnn for multi-speaker text-to-speech synthesis," *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 230–234, 2022.

[24] Z. Şentürk, Ö. E. Yetgin, and Ö. Salor, "Voiced-unvoiced classification of speech using autocorrelation matrix," in *2014 22nd Signal Processing and Communications Applications Conference (SIU)*, 2014, pp. 1802–1805.

[25] A. Tjandra, Y.-C. Wu, B. Guo, *et al.*, "Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound," *ArXiv*, vol. abs/2502.05139, 2025.