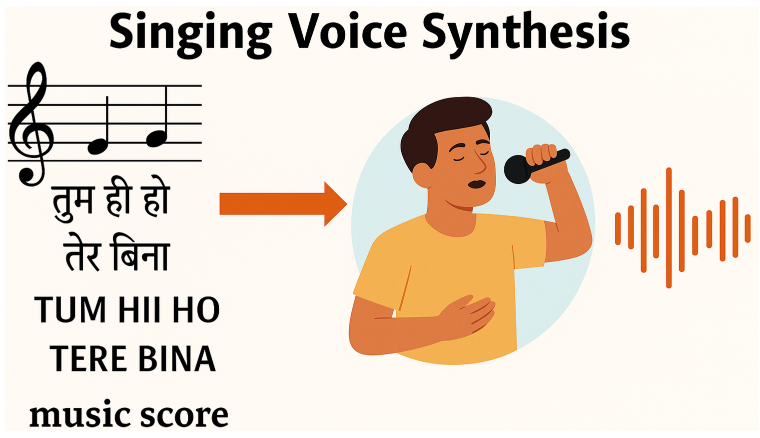
LAPS-Diff: A Diffusion-Based Framework for Hindi Singing Voice Synthesis With Language Aware Prosody-Style Guided Learning

Sandipan Dhar*, Mayank Gupta*, Preeti Rao
Indian Institute of Technology, Bombay

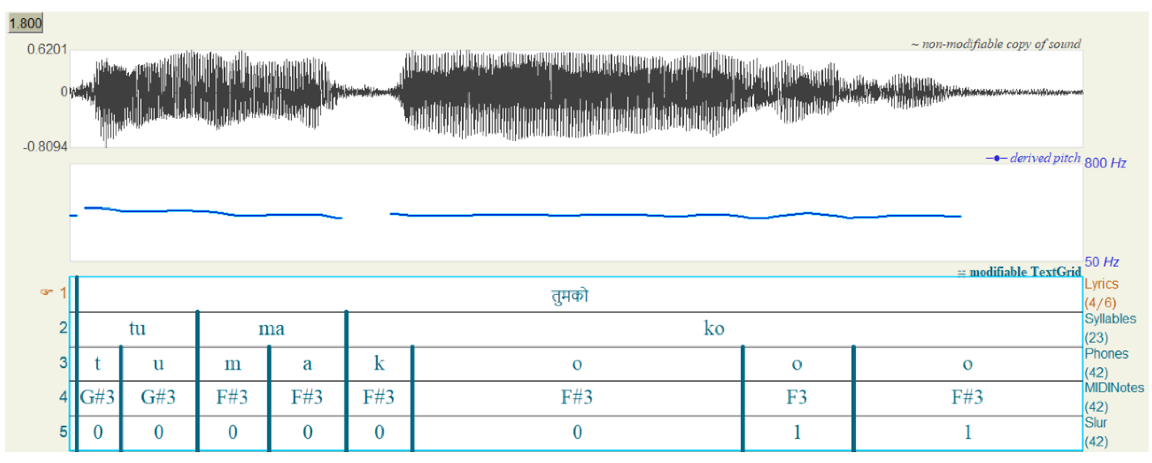
Motivation



- Despite the popularity of Indian music, singing voice synthesis (SVS) for Indian languages remains underexplored due to the lack of suitable datasets.
- Limited labeled singing data poses a challenge for accurately modeling linguistic content, style, and pitch information.
- Cross-lingual finetuning of pretrained SVS models like DiffSinger[1] for Hindi results in accented pronunciation, pitch errors, vowel distortion, and artifacts in slurred regions.

Contributions

Curation of Hindi Bollywood SVS dataset:



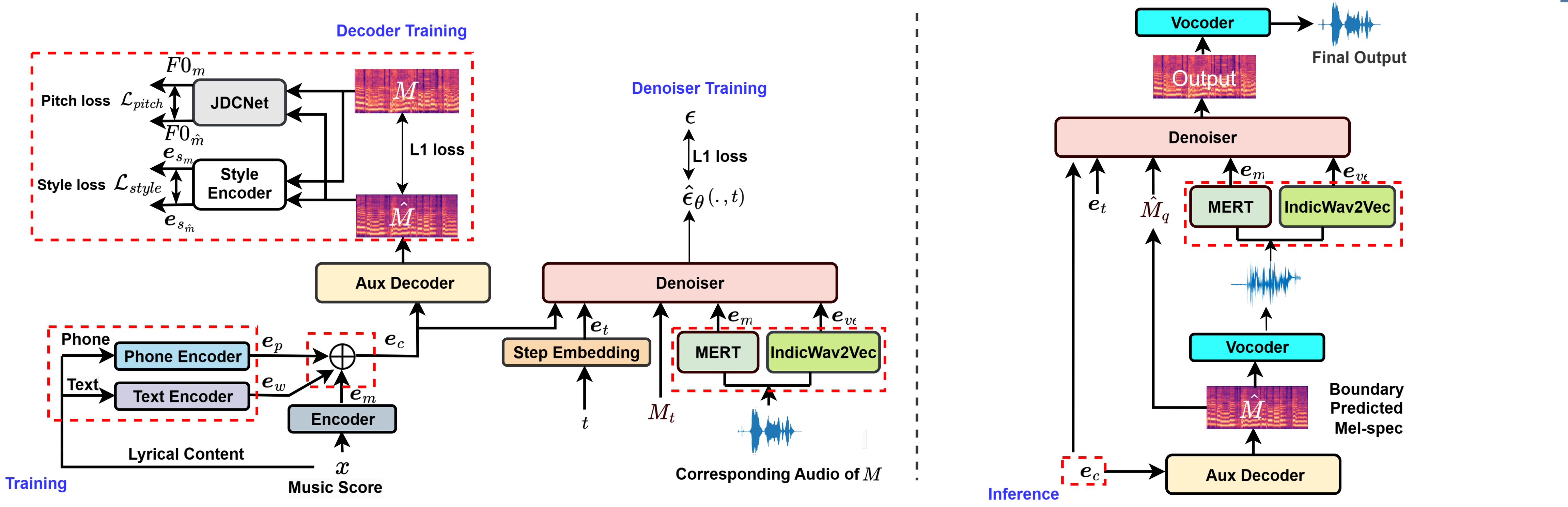
- We adopt the same music score format as the Opencpop dataset for our curated Hindi Bollywood SVS dataset.
- The score includes lyrics, syllables, phonemes, phoneme durations, musical notes (pitch), note durations, and slur indicators.
- Sung lyrics differ from speech, notably in vowel duration and pitch variation.
- Hindi song recordings, featuring a single male singer in Bollywood style, were collected from public sources and processed for vocal separation using Gaudio.
- We perform phone-level audio-text alignment using a hybrid ASR trained on 180 hours of adult Hindi speech from 400 speakers.
- After automatic alignment, we manually correct any misaligned phoneme/notes to achieve precise fine alignment.
- We curated 65 minutes of labeled singing dataset.

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

Adaptation to Hindi singing style with language aware, prosody-style and musical feature guided modeling:

- We extract Hindi word-level and phone-level embeddings, $\mathbf{e}_w$ and $\mathbf{e}_p$ respectively, from two pretrained language models: **IndicBERT** and **XPhoneBERT**. These content embeddings are combined with the music score embedding $\mathbf{e}_m$.
- We integrate a **style encoder** and a pre-trained **JDCNet** pitch extraction model to capture style and pitch (melody) information through corresponding losses.
- The pitch loss incorporates the Concordance Correlation Coefficient (CCC) to mitigate misalignment-induced errors in training.
- We employ the pretrained **MERT** model to extract musical feature embeddings $\mathbf{e}_{mert}$ and **IndicWav2Vec** for contextual embeddings $\mathbf{e}_{vec}$ as conditional priors to the denoiser.

Proposed Work



DiffSinger model[1] with proposed enhancements (in red dashed boxes)

- **Feature fusion** to obtain the fused embedding $\mathbf{e}_c$.

$$\mathbf{e}_c = \mathbf{e}_w + \mathbf{e}_p + \mathbf{e}_m$$

- The denoising process in reverse diffusion, adding MERT and IndicWav2vec embeddings:

$$M_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(M_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}_\theta(M_t, \mathbf{e}_c, \mathbf{e}_t, \mathbf{e}_{mert}, \mathbf{e}_{vec})\right) + \sigma_t z$$

- **Style loss** $\mathcal{L}_{\text{style}}$:

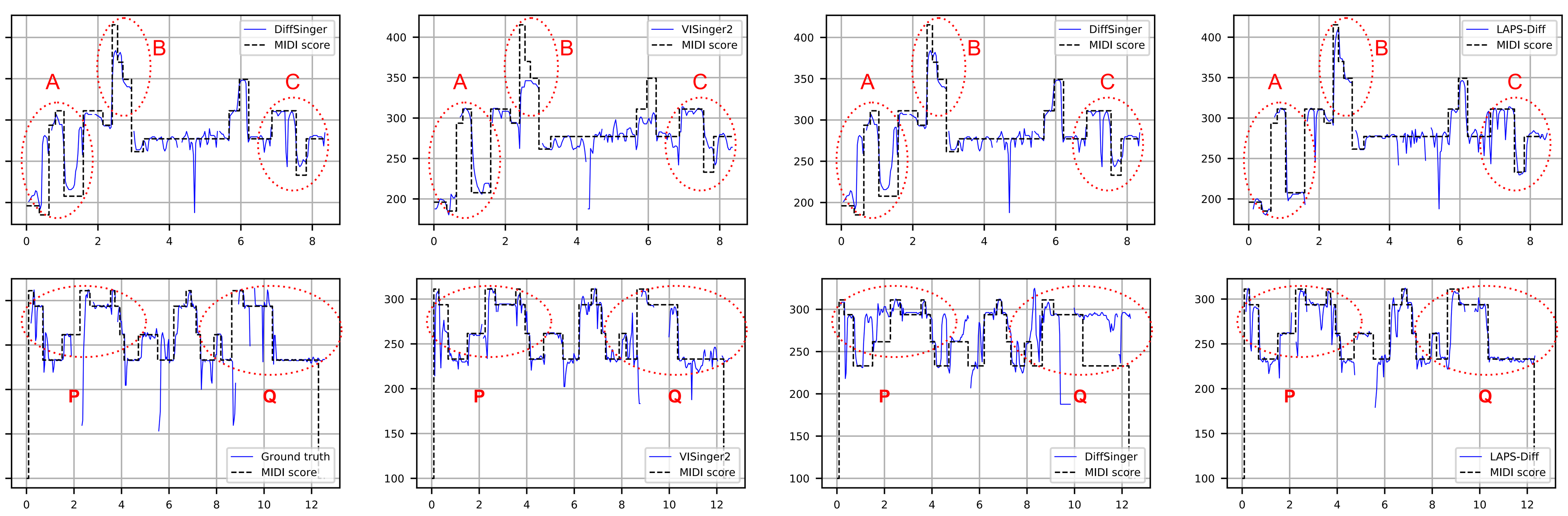
$$\mathcal{L}_{style} = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{e}_{s_{m_i}} - \mathbf{e}_{s_{\hat{m}_i}}\right\|^2,$$

- **CCC-based pitch loss** $\mathcal{L}_{\text{pitch}}$:

$$\text{Pitch Loss} = (1 - \text{CCC}) \times \left(\frac{1}{K}\sum_{i=1}^{K}\|F0_{m_i} - F0_{\hat{m}_i}\|^2\right)$$

$$\text{CCC} = \frac{2\,\rho_{F0_m\,F0_{\hat{m}}}\,\sigma_{F0_m}\sigma_{F0_{\hat{m}}}}{\sigma^2_{F0_m} + \sigma^2_{F0_{\hat{m}}} + (\mu_{F0_m} - \mu_{F0_{\hat{m}}})^2}$$

Results



F0 contour of ground truth and synthesised outputs, with reference to the MIDI score. Vertical axis shows frequency (Hz), and Horizontal axis shows time (s). Top and bottom row contains a sample with faster and slower singing rate respectively.

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 (Language emb)	0.973	0.171	0.890	0.159	7.983	4.200	6.499
Ablation 2 (Music emb)	0.978	0.185	0.869	0.151	9.445	4.151	6.408
Ablation 3 (Pitch loss)	0.978	0.171	0.898	0.118	7.928	3.460	6.355
Ablation 4 (Style loss)	0.986	0.169	0.880	0.145	7.883	3.869	6.511

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

LAPS-Diff outperforms across most metrics, achieving:

- Highest average cosine similarity
- Lowest MAE
- Highest V/UV accuracy
- Second lowest log-F0 RMSE and MCD.

- LAPS-Diff captures speaker characteristics more accurately and ensures closer alignment of spectral features (including pitch) with the ground truth, enhancing the overall quality.
- 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.

References

[1] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11020–11028, Jun. 2022.

[2] Bagus Tris Atmaja and Masato Akagi. Evaluation of error- and correlation-based loss functions for multitask learning dimensional speech emotion recognition. *Journal of Physics: Conference Series*, 1896, 2020.

[3] Yongmao Zhang, Heyang Xue, Hanzhao Li, Lei Xie, Tingwei Guo, Ruixiong Zhang, and Caixia Gong. Visinger2: High-fidelity end-to-end singing voice synthesis enhanced by digital signal processing synthesizer. In *Interspeech 2023*, pages 4444–4448, 2023.

[4] Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *ArXiv*, abs/2502.05139, 2025.