Appendix of Prosody-Enhanced Acoustic Pre-training and Acoustic-Disentangled Prosody Adapting for Movie Dubbing

We organize the supplementary materials as follows:

- In Section A, we analyze the challenges of the V2C-Animation benchmark compared to the traditional TTS benchmark and GRID benchmark.
- In Section B, we provide a more detailed description of the model implementation, including the settings of the modules and details of the loss function.
- In Section C, we introduce the baseline models.
- In Section D, we provide additional ablation studies to validate the effectiveness of our method.
- In Section E, we provide additional visualizations of mel-spectrograms to compare with other baseline models.
- In Section F, we discuss the limitations of the proposed method.

A. The Challenge of V2C-Animation Benchmark

As shown in Table A.1, the V2C-Animation benchmark [1] differs significantly from traditional TTS benchmarks in multiple aspects, and it is more challenging. The main reasons for this are as follows: (1) The V2C-Animation benchmark has a smaller data scale and shorter speech duration compared to other datasets. As shown in Table A.1, the V2C-Animation benchmark contains only 10,217 samples. Although it is comparable in quantity to LJSpeech [6], the average length of each sample is only about one-third of LJSpeech. The GRID benchmark roughly triples the data volume with a slightly smaller average length compared to V2C-Animation, LibriTTS [12] far exceeds the V2C-Animation benchmark in both average length and total quantity. (2) The V2C-Animation benchmark exhibits more noticeable background noise compared to other benchmarks. We estimate the signal-to-noise (SNR) ratios and the audio quality of each dataset using deep learning-based approachs [8, 11], and the results are shown in Table A.1. As shown in the table, the other three datasets exhibit relatively high signal-to-noise ratios because they are recorded in studio environments without background sound, which can provide high-quality speech knowledge for models. However, the V2C-Animation benchmark is excerpted from real movies, which contain background and environmental sounds. Its limitations in sound quality are also reflected in the UT-MOS metric, which measures sound quality. Compared to other TTS datasets or the GRID dataset, V2C-Animation exhibits significant differences in speech quality. It poses challenges for models to build dubbing with high acoustic quality and accurate pronunciation. (3) The V2C-Animation benchmark exhibits more significant pitch variation. We compute the mean and variance of pitch across different benchmarks and list in Table A.1. This further increases the challenge of the V2C-Animation benchmark. (4) The V2C-Animation benchmark contains more complex and realistic scenes compared to the GRID benchmark. As a multi-speaker dubbing dataset, all speakers in GRID are recorded using the same fixed perspective and uniform background, while V2C-Animation includes more complex scenes from real movies. Complex scenes and environments increase the difficulty of modeling the prosody and variation information of dubbing from visual information.

Overall, the V2C-Animation benchmark is more challenging than traditional TTS benchmarks or GRID dubbing benchmark, both in terms of the scale and acoustic quality, as well as the complexity of the visual scene.

Table A.1. Difference between V2C-Animation benchmark and other benchmarks.

Dataset	Sample Number	Avg. Length (s)	SNR (dB)	Pitch (Hz)	UT-MOS
LJSpeech [6]	13,100	6.57	26.59	1921.75 ± 1249.77	4.37
LibriTTS [12]	149,736	6.34	26.72	2025.21 ± 1221.06	4.09
GRID [4]	33,000	1.83	23.77	1473.71 ± 1195.36	3.97
V2C-Animation [1]	10,217	2.46	10.15	1955.81 ± 1301.60	2.26

B. Implementation Details

B.1. Dataset Splits

Following VDTTS [5] and HPMDubbing [2], the GRID and V2C-Animation datasets have no individual valid sets, the valid sets are mixed with their train sets. For GRID, we take 100 random videos from each speaker as a test set and use the remainder 900 examples per speaker as training data. For V2C-Animation, the number of training and test data are 6517 and 2779, respectively.

B.2. Detail of Each Module

Our proposed model contains several modules, including an open-source grapheme-to-phoneme (G2P) module[1], pre-trained modules, and those updated only during specific training phases. To facilitate a better understanding of our design, we provide an overview of the training status of each module as shown in Table A.2.

Table A.2. The status of each module in the second stage.

Module	Pretrained	First Stage	Second Stage
Acoustic Text Encoder		✓	
Acoustic Style Encoder		✓	
Audio Decoder		✓	
Prosody Extractor	✓		
Text Aligner	✓		
Prosodic Text Encoder			✓
Prosodic Text BERT Encoder	✓		
Prosodic Style Diffusion			✓
S^3FD	✓		
EmoFAN	✓		
In-Domain Emotion Analysis			✓
Prosody Predictor			✓
Lip Motion Encoder	✓		

B.3. Training Loss in Second Stage

The total loss function of the acoustic-disentangled training stage is:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_p + \lambda_2 \mathcal{L}_e + \lambda_3 \mathcal{L}_d + \lambda_4 \mathcal{L}_{S_p}, \tag{1}$$

$$\mathcal{L}_p = \frac{1}{L_{Mel}} \sum_{i=0}^{L_{Mel}-1} \left\| \widetilde{p}_i - p_i \right\|_1, \tag{2}$$

$$\mathcal{L}_e = \frac{1}{L_{Mel}} \sum_{i=0}^{L_{Mel}-1} \left\| \widetilde{e}_i - e_i \right\|_1, \tag{3}$$

$$\mathcal{L}_d = \frac{1}{L_{pho}} \sum_{i=0}^{L_{pho}-1} \left\| \widetilde{d}_i - d_i \right\|_1, \tag{4}$$

$$\mathcal{L}_{S_p} = \left\| \widetilde{S}_p - S_p \right\|_1, \tag{5}$$

where the $\mathcal{L}_p$, $\mathcal{L}_e$, and $\mathcal{L}_d$ are the L1 loss between the ground-truth and predicted spectrogram frame-level pitch, spectrogram frame-level energy, and phoneme-level duration, respectively. It is noteworthy that the spectrogram frame-level pitch and energy are obtained by upsampling the predicted phoneme-level pitch and energy after alignment with the ground truth. $\mathcal{L}_{S_p}$ is the diffusion L1 Loss at the feature level for prosodic style features.

[1]https://github.com/bootphon/phonemizer

B.4. Ablation Implementation Details

w/o Acoustic Pre-training (AP). We directly train the acoustic system and prosody adaptation modules of the model on the V2C-Animation dataset without employing any prosody enhancement strategies.

w/o Prosody Enhancement (PE). We use the original LibriTTS-460 dataset to pre-train the acoustic system of the model without employing any prosody enhancement strategies.

w/o In-Domain Emotion Analysis (IDEA). We utilize pre-trained modules S^3FD and EmoFAN to extract video frame-level emotion feature sequences from the facial regions of characters in the given movie clips, which are then directly used as Q and V for fusion with prosodic text feature T_p.

w/o Prosodic Style Diffusion (PSD). We remove the acoustic-prosody disentanglement at the style level by not using the prosodic style encoder and prosodic style diffusion modules and directly substitute prosodic style feature S_p with the acoustic style feature S_a.

w/o Prosodic Text BERT Encoder (PTBE). We remove the acoustic-prosody disentanglement at the text modality by removing the prosodic text BERT encoder and directly substituting the prosodic text feature T_p with the acoustic text feature T_a.

C. Baseline Introduction

We compare our model with seven relevant methods for which code is available.

1) StyleSpeech [9] is a TTS method based on the FastSpeech2 [10] framework, which utilizes a style encoder and meta-learning to adapt to multi-speaker environments.

2) Zero-shot TTS [14] is a content-dependent fine-grained speaker method for zero-shot speaker adaptation.

3) V2C-Net [1] is the first visual voice cloning model for movie dubbing. It introduces the visual feature into the modeling of spectrogram frame-level prosody modeling.

4) HPMDubbing [2] is currently the most advanced movie dubbing model. It employs a hierarchical prosody modeling approach to connect the prosody of dubbing with the lip movements, expressions, and scenes in movie clips.

5) FaceTTS [7] is a novel diffusion-based TTS approach attempting to use facial to synthesize voice timbre.

6) StyleDubber [3] is a method that models movie dubbing styles using phoneme-level pronunciation habits and fine-grained character emotions.

7) Speaker2Dubber [13] is a two-stage dubbing method. In the first stage, the model's phoneme encoder is pre-trained using a large-scale speech-text corpus, which significantly enhancing the pronunciation accuracy of the dubbing.

In the experimental tables of the main text, baseline

Table A.3. Results of duration alignment on V2C-Animation benchmark.

Setting		Dub 1.0						Dub 2.0		
Methods	T-S	SECS (%) ↑	WER (%) ↓	UT-MOS ↑	EMO-ACC (%) ↑	MCD-DTW ↓	MCD-DTW-SL ↓	SECS (%) ↑	WER (%) ↓	UT-MOS ↑
Ours + In-Domain Lip Motion	✓	75.01	8.61	3.06	48.23	9.34	9.37	75.31	11.71	3.05
Ours + Style Duration	✓	74.93	8.39	3.07	47.64	9.38	9.41	74.85	12.05	3.06
Ours	✓	**75.46**	**8.04**	**3.10**	**48.93**	**9.29**	**9.32**	**75.39**	**11.50**	**3.09**

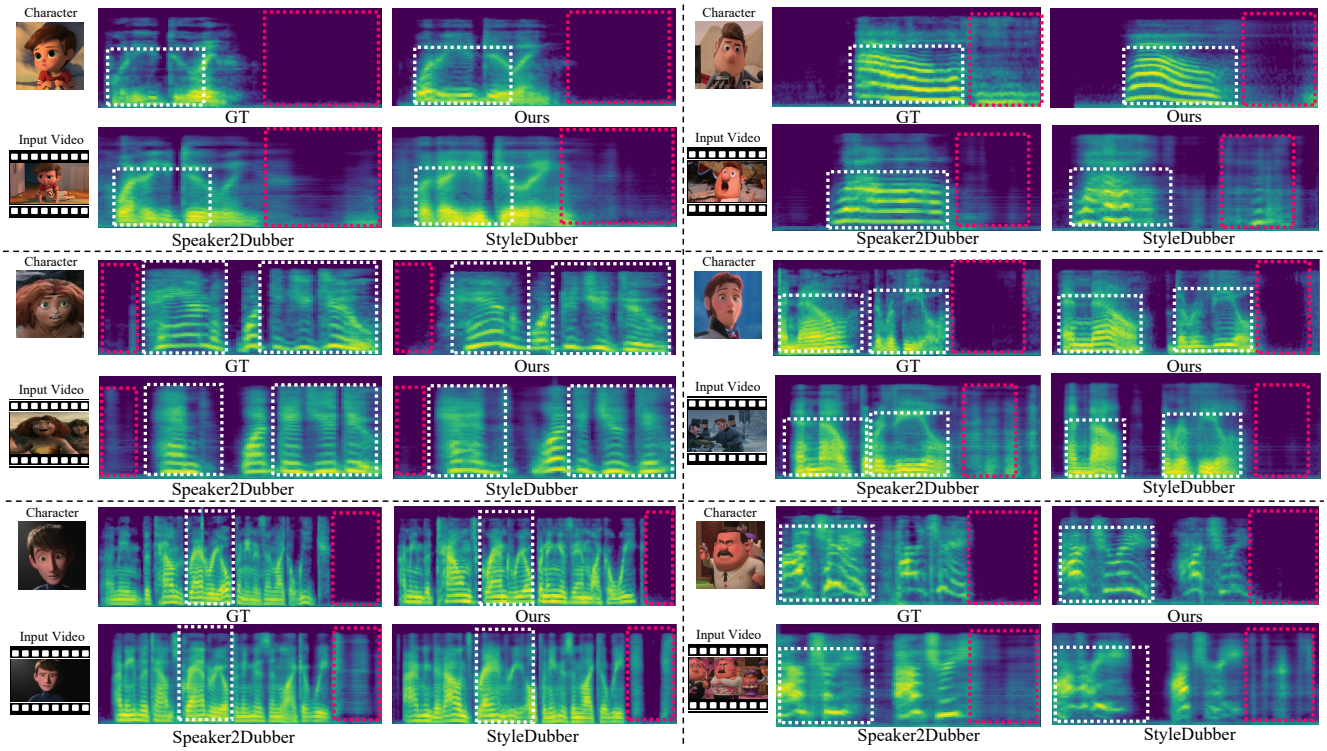

Figure A.1. More visualization of the mel-spectrograms of ground truth and synthesized dubbing of different models. The red and white bounding boxes highlight regions where different models exhibit significant differences in audio quality and pronunciation details.

models marked with "*" indicate TTS models that incorporate additional visual feature inputs to adapt to the movie dubbing task following [1]. "T-S" indicates that the model employs a two-stage training strategy, similar to Speaker2Dubber [13], which involves pre-training the phoneme encoder. The diverse selection of baselines allows for a more comprehensive comparison of the model's performance.

D. Supplementary Experiments

We design two variants for the duration alignment. The first uses a similar approach to in-domain emotion analysis to eliminate the visual domain gap in lip motion features, while the second employs AdaIN to incorporate prosodic style features into lip motion features. The experimental results are shown in the Table A.3.

The experimental results indicate that the lip motion encoder itself already considers the impact of the visual domain gap and focuses more on capturing lip motion and changes. Therefore, in-domain analysis does not enhance the model's performance in duration alignment. Additionally, duration alignment should only be related to lip motion and phoneme pronunciation features. Therefore, incorporating prosodic style features does not improve the model's performance.

E. Qualitative Analysis

We visualize the mel-spectrograms of ground truth and dubbing generated by different models for comparison in Figure A.1. The red bounding boxes represent regions where different models exhibit significant differences in audio quality and the white regions present the difference in pronunciation details. Through the observation of the red bounding box, it is evident that our method exhibits less noise in the non-pronunciation intervals of dubbing compared to other models, indicating better acoustic quality. Additionally, the clearer spectrum lines in the white bounding box and the closer resemblance to the ground truth in terms of prosody (the shape of spectrum lines) show that our method also achieves better pronunciation quality and

prosody alignment performance.

F. Limitations

Although our proposed method significantly improves the acoustic quality of the generated dubbing, there remains a gap compared to speech generated by Text-to-Speech models. An ideal movie dubbing model should function like a dubbing actor, generating dubbing with acoustic quality comparable to speech, while also incorporating exaggerated and diverse prosody that matches character performances, the proposed model has not yet achieved this. Additionally, the prosody-enhanced acoustic pre-training does not fully meet expectations in terms of voice cloning. How to more effectively extract timbre from low-quality audio in movie dubbing datasets remains an unresolved challenge.

References

[1] Qi Chen, Mingkui Tan, Yuankai Qi, Jiaqiu Zhou, Yuanqing Li, and Qi Wu. V2C: visual voice cloning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 21210–21219, 2022. 1, 2, 3

[2] Gaoxiang Cong, Liang Li, Yuankai Qi, Zheng-Jun Zha, Qi Wu, Wenyu Wang, Bin Jiang, Ming-Hsuan Yang, and Qingming Huang. Learning to dub movies via hierarchical prosody models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14687–14697, 2023. 1, 2

[3] Gaoxiang Cong, Yuankai Qi, Liang Li, Amin Beheshti, Zhedong Zhang, Anton van den Hengel, Ming-Hsuan Yang, Chenggang Yan, and Qingming Huang. Styledubber: Towards multi-scale style learning for movie dubbing. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6767–6779, 2024. 2

[4] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006. 1

[5] Michael Hassid, Michelle Tadmor Ramanovich, Brendan Shillingford, Miaosen Wang, Ye Jia, and Tal Remez. More than words: In-the-wild visually-driven prosody for text-to-speech. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10577–10587, 2022. 1

[6] Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017. 1

[7] Jiyoung Lee, Joon Son Chung, and Soo-Whan Chung. Imaginary voice: Face-styled diffusion model for text-to-speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5, 2023. 2

[8] Hao Li, DeLiang Wang, Xueliang Zhang, and Guanglai Gao. Frame-level signal-to-noise ratio estimation using deep learning. In *Interspeech*, pages 4626–4630, 2020. 1

[9] Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang. Meta-stylespeech : Multi-speaker adaptive text-to-speech generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 7748–7759, 2021. 2

[10] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 2

[11] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: utokyo-sarulab system for voicemos challenge 2022. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 4521–4525. ISCA, 2022. 1

[12] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1526–1530, 2019. 1

[13] Zhedong Zhang, Liang Li, Gaoxiang Cong, Haibing Yin, Yuhan Gao, Chenggang Yan, Anton van den Hengel, and Yuankai Qi. From speaker to dubber: Movie dubbing with prosody and duration consistency learning. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 7523–7532, 2024. 2, 3

[14] Yixuan Zhou, Changhe Song, Xiang Li, Luwen Zhang, Zhiyong Wu, Yanyao Bian, Dan Su, and Helen Meng. Content-dependent fine-grained speaker embedding for zero-shot speaker adaptation in text-to-speech synthesis. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 2573–2577, 2022. 2