

Supplementary Materials of Submission 1863

Anonymous Authors

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 comparative experiments and 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.

A THE CHALLENGE OF V2C-ANIMATION BENCHMARK

As shown in Table 1, the V2C-Animation benchmark [1] differs significantly from traditional TTS benchmarks in multiple aspects, and its challenges are much greater than them. 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 1, the V2C-Animation benchmark contains only 10,217 samples. Although it is comparable in quantity to LJSpeech [4], the average length of each sample is only about one-third of LJSpeech. The GRID benchmark roughly tripled the data volume with a slightly smaller average length compared to V2C-Animation, LibriTTS [11] 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 of each dataset using a deep learning-based approach [6], and the results are shown in Table 1. As shown in the table, the other three datasets exhibit relatively high signal-to-noise ratios because they are recorded in studio environments, which can provide high-quality speech knowledge for models. However, the V2C-Animation benchmark is excerpted from real movies, which contain background noise and environmental sounds. It poses challenges for models to learn pronunciation accurately. (3) The V2C-Animation benchmark exhibits greater pitch variation. We compute the mean and variance of pitch across different benchmarks and list in Table 1. This further enhances 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 quality of speech, as well as the complexity of the visual scene.

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

Dataset	Sample Number	Avg. Length (s)	SNR (dB)	Pitch (Hz)
LJSpeech [4]	13,100	6.57	26.59	1921.75 \pm 1249.77
LibriTTS [11]	149,736	6.34	26.72	2025.21 \pm 1221.06
GRID [3]	33,000	1.83	23.77	1473.71 \pm 1195.36
V2C-Animation [1]	10,217	2.46	10.15	1955.81 \pm 1301.60

B IMPLEMENTATION DETAILS

B.1 Detail of each module

Our proposed model comprises many modules, including pre-trained and non-pre-trained modules. To facilitate readers in better understanding the various modules in the paper, we provide an overview indicating whether each module is pre-trained as shown in Table 2.

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

Module	Whether pretrained
Phoneme Encoder	Yes
S ³ FD	Yes
EmoFAN	Yes
Emotion Encoder	No
Mel-Timbre Encoder	Yes
Energy/Pitch Predictor	No
Lip Motion Encoder	No
Mel-Decoder	No
Length Regulator	-

B.2 Training Loss in Second Stage

The total loss function of second training stage is:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{mel} + \lambda_2 \mathcal{L}_{pitch} + \lambda_3 \mathcal{L}_{energy} + \lambda_4 \mathcal{L}_{align}, \quad (1)$$

$$\mathcal{L}_{mel} = \frac{1}{L_{mel}} \sum_{t=0}^{L_{mel}-1} \left| \tilde{M}_{Dub}^t - M_{Dub}^t \right|, \quad (2)$$

$$\mathcal{L}_{pitch} = \frac{1}{L_p} \sum_{i=0}^{L_p-1} \left| \tilde{p}_{pho}^i - p_{pho}^i \right|, \quad (3)$$

$$\mathcal{L}_{energy} = \frac{1}{L_p} \sum_{i=0}^{L_p-1} \left| \tilde{E}_{pho}^i - E_{pho}^i \right|, \quad (4)$$

$$\mathcal{L}_{align} = \sum \tilde{A} - \tilde{A} \odot A^*, \quad (5)$$

where the \mathcal{L}_{mel} , \mathcal{L}_{pitch} , and \mathcal{L}_{energy} are the L1 loss between the ground-truth and predicted mel-spectrogram, phoneme-level pitch, and energy, respectively. The \tilde{A} is the ground-truth alignment between the phoneme-level acoustics feature and the video frame-level lip motion feature calculated by the ground-truth duration

Table 3: Supplementary results on GRID benchmark with the same dub setting as the V2C-Animation benchmark.

Setting		Dub 1.0				Dub 2.0			
Methods	Visual	SECS (%) ↑	WER (%) ↓	MCD-DTW ↓	MCD-DTW-SL ↓	SECS (%) ↑	WER (%) ↓	MCD-DTW ↓	MCD-DTW-SL ↓
GT	-	100.00	22.41	0.00	0.00	100.00	22.41	0.00	0.00
GT Mel + Vocoder	-	97.57	21.41	4.10	4.15	97.57	21.41	4.10	4.15
Matcha-TTS [7]	X	81.27	18.22	6.38	6.92	67.09	18.22	6.38	6.92
Ours	✓	94.50	17.07	5.34	5.45	85.76	17.42	6.17	6.43

Table 4: Ablation studies about different second-stage strategies on V2C-Animation benchmark. The “p,e” denotes the pitch and energy embedding layer.

Setting		Dub 1.0					Dub 2.0				
Frozen Modules	Visual	SECS (%) ↑	WER (%) ↓	EMO-ACC (%) ↑	MCD-DTW ↓	MCD-DTW-SL ↓	SECS (%) ↑	WER (%) ↓	EMO-ACC (%) ↑	MCD-DTW ↓	MCD-DTW-SL ↓
GT	-	100.00	25.55	99.96	0.00	0.00	100.00	22.55	99.96	0.00	0.00
GT Mel + Vocoder	-	96.96	24.40	97.09	3.77	3.80	96.96	24.40	97.09	3.77	3.80
Phoneme Encoder + Mel-Decoder	✓	81.43	24.74	45.24	9.87	10.05	79.49	23.08	39.71	10.91	11.12
Phoneme Encoder + Mel-Decoder + p,e	✓	81.34	23.50	44.70	9.76	9.95	80.49	23.05	40.33	10.81	11.02
Phoneme Encoder + p,e	✓	81.40	17.60	46.45	9.48	9.66	79.81	17.43	41.20	10.70	10.90
Phoneme Encoder	✓	81.50	17.51	46.80	9.46	9.65	79.86	17.33	43.66	10.64	10.84

Table 5: Ablation studies about pre-train mel-timbre encoder or not on V2C-Animation benchmark.

Setting	Dub 1.0		Dub 2.0	
Frozen Module	SECS (%) ↑	WER (%) ↓	SECS (%) ↑	WER (%) ↓
GT	100.00	25.55	100.00	22.55
GT Mel + Vocoder	96.96	24.40	96.96	24.40
w/o pre-trained Mel-timbre Encoder [8]	78.36	37.58	78.27	44.12
w/ pre-trained Mel-timbre Encoder (Ours)	81.50	17.51	79.86	17.33

of each phoneme. In the \tilde{A} , the values on the ground-truth alignment path are set to 1, while all other elements are set to 0. The \mathcal{L}_{align} aims to optimize the lip motion encoder based on the alignment path of the ground truth, thereby achieving more accurate pronunciation-lip alignment. The selection of loss weights is to adjust all loss items to the same scale.

C BASELINE CHOICE

We compare our model with seven relevant methods for which code is available. 1) **FastSpeech2** [9] is a popular non-autoregressive TTS method that explicitly models duration, pitch, and energy as variation information. 2) **StyleSpeech** [8] is a TTS method based on the FastSpeech2 [9] framework, which utilizes a style encoder and meta-learning to adapt to multi-speaker environments. 3) **Zero-shot TTS** [8] is a content-dependent fine-grained speaker method for zero-shot speaker adaptation. 4) **Matcha-TTS** [7] is a state-of-the-art TTS model based on conditional flow matching. 5) **V2C-Net** [1] is the first visual voice cloning model for movie dubbing. 6) **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. 7) **FaceTTS** [5] is a novel diffusion-based TTS approach attempting to use facial to synthesize voice timbre.

The baseline models include pure TTS and dubbing methods for comprehensive comparative experiments. To ensure fairness in comparison, we provide video embeddings as additional inputs for all pure TTS methods following [1].

D SUPPLEMENTARY EXPERIMENTS

D.1 Supplementary Comparison on GRID benchmark

We supplement the performance of Matcha-TTS [7] on the GRID benchmark. As shown in Table 3, our proposed model still achieves the best performance across all evaluation metrics.

D.2 Ablation Studies on Second Stage Strategy

In the main text, we mention that since speech and dubbing share the same semantic space but differ significantly in tone and prosody, pre-training only the phoneme encoder is the optimal choice. Due to limitations in the length of the main text, we provide experimental evidence here in the appendix. We provide four pre-training strategies, each freezing phoneme encoder, phoneme encoder + mel-decoder, phoneme encoder + mel-decoder + pitch, energy embedding, and phoneme encoder + pitch, energy embedding in the second stage.

As shown in Table 4, the approach freezes only the phoneme encoder during the second stage achieves the best dubbing performance. Two methods that simultaneously freeze the mel-decoder are limited by the prosody differences between speech and dubbing, failing to achieve better MCD-DTW and MCD-DTW-SL performances and decreasing pronunciation clarity (See WER). The approach of freezing the pitch and energy embeddings on top of freezing the phoneme encoder achieves performance closer to the approach of freezing only the phoneme encoder. However, due to the differences in pitch and energy distribution between dubbing and speech, we ultimately chose only to freeze the phoneme encoder as our second-stage training strategy.

D.3 Ablation Studies on Mel-timbre Encoder

We compare the differences between our method and StyleSpeech [8] in terms of timbre extraction on the V2C-Animation benchmark. StyleSpeech [8] trains the style encoder with the model and then integrates style information into the speech. Meanwhile, our model employs a mel-timbre encoder pre-trained with GE2E loss [10] as the timbre feature extraction module. Due to the lack of supervision corresponding to vocal timbre (*i.e.*, style), the approach used in [8]

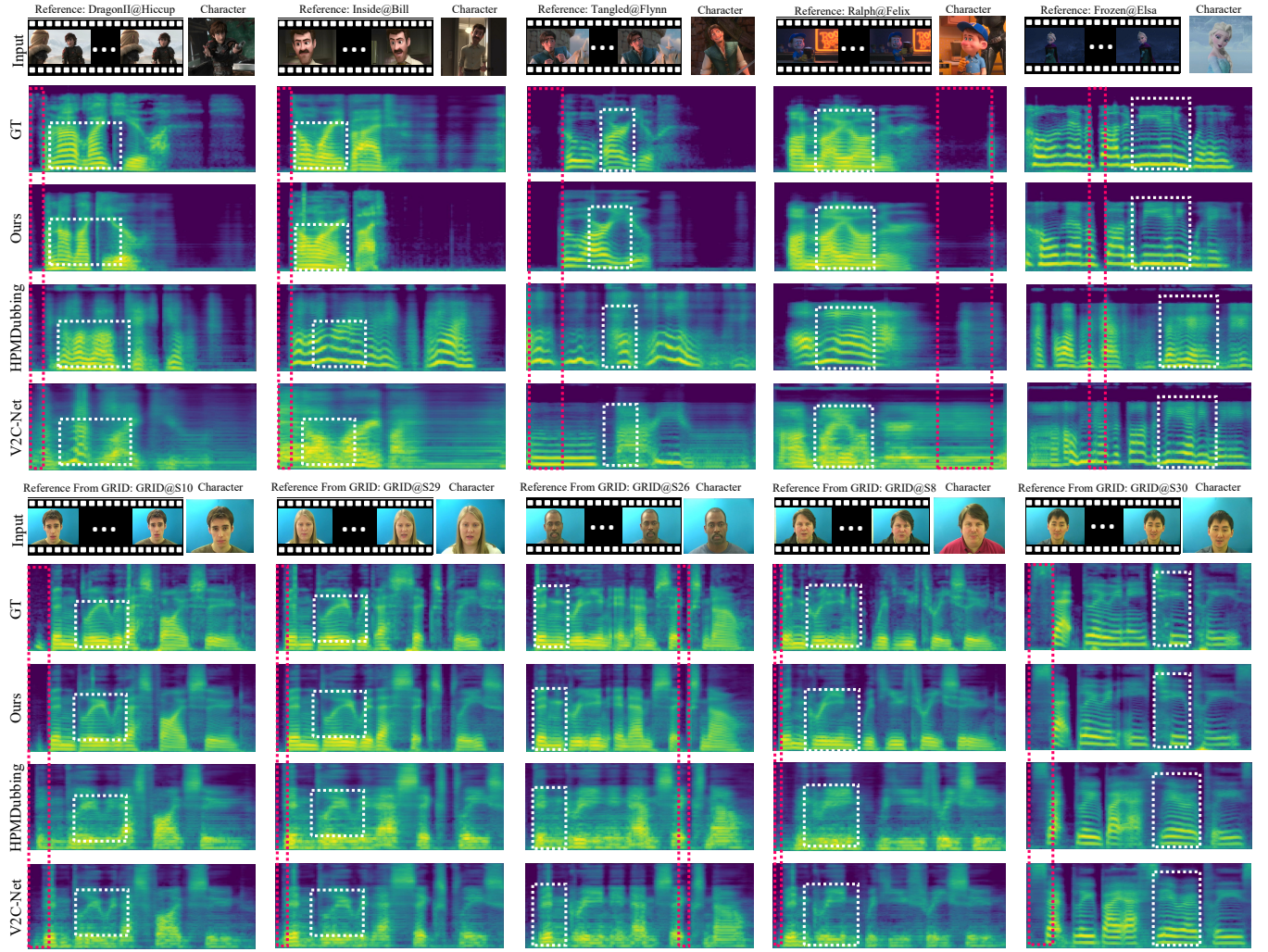


Figure 1: The visualization of the mel-spectrograms from ground truth and synthesized audios by different models. The red and white bounding boxes highlight regions where different models exhibit significant differences in duration pausing and pronunciation details.

cannot better efficiently and accurately extract timbre features in the V2C-Animation benchmark. It also remains a risk of information leakage. Moreover, the nature of environmental sounds in the V2C-Animation dataset is not conducive to the generalization of the style encoder. As shown in Table 5, our method outperforms the approach as [8] on SECS in both dubbing settings. The poor generalization ability also leads to an inferior pronunciation performance.

E QUALITATIVE ANALYSIS

We visualize the mel-spectrograms of ground-truth and synthesized audios by our model and the other two state-of-the-art methods in Figure 1. The red and white bounding boxes represent regions where different models exhibit significant differences in duration consistency and pronunciation details compared to the ground truth. Through the observation of the red bounding box, it is evident that

our model outperforms others in maintaining duration consistency. The duration of pronunciation and pauses in the mel-spectrogram generated by our model is notably closer to the ground truth dubbing. This phenomenon is more pronounced in the V2C-Animation benchmark due to its complex speaking speed variation. Additionally, from the clearer spectrum lines in the white bounding box, it can be observed that the dubbing generated by our model exhibits clearer and more natural pronunciation details.

REFERENCES

- [1] Qi Chen, Minghui Tan, Yuankai Qi, Jiaqi Zhou, Yuanqing Li, and Qi Wu. 2022. V2C: Visual Voice Cloning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. 21210–21219.
- [2] Gaoxiang Cong, Liang Li, Yuankai Qi, Zheng-Jun Zha, Qi Wu, Wenyu Wang, Bin Jiang, Ming-Hsuan Yang, and Qingming Huang. 2023. 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*. 14687–14697.

- [3] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America* 120, 5 (2006), 2421–2424.
- [4] Keith Ito and Linda Johnson. 2017. The Lj Speech Dataset. <https://keithito.com/LJ-Speech-Dataset/>.
- [5] Jiyoung Lee, Joon Son Chung, and Soo-Whan Chung. 2023. 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*. 1–5.
- [6] Hao Li, DeLiang Wang, Xueliang Zhang, and Guanglai Gao. 2020. Frame-Level Signal-to-Noise Ratio Estimation Using Deep Learning.. In *Interspeech*. 4626–4630.
- [7] Shivam Mehta, Ruiho Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. 2024. Matcha-TTS: A fast TTS architecture with conditional flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 11341–11345.
- [8] Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang. 2021. 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*, Vol. 139. 7748–7759.
- [9] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. 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*. <https://openreview.net/forum?id=piLPYqxtWuA>
- [10] Li Wan, Quan Wang, Alan Papir, and Ignacio López-Moreno. 2018. Generalized End-to-End Loss for Speaker Verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*. 4879–4883.
- [11] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. 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*. 1526–1530.