StyleDubber: Towards Multi-Scale Style Learning for Movie Dubbing

Gaoxiang Cong[1] **Yuankai Qi**[2*] **Liang Li**[1*] **Amin Beheshti**[2] **Zhedong Zhang**[3]
Anton van den Hengel[4] **Ming-Hsuan Yang**[5] **Chenggang Yan**[3] **Qingming Huang**[1]
[1]Institute of Computing Technology, CAS [2]Macquarie University
[3]Hangzhou Dianzi University [4]University of Adelaide [5]University of California
conggaoxiang@foxmail.com, yuankai.qi@mq.edu.au, liang.li@ict.ac.cn

Abstract

Given a script, the challenge in Movie Dubbing (Visual Voice Cloning, V2C) is to generate speech that aligns well with the video in both time and emotion, based on the tone of a reference audio track. Existing state-of-the-art V2C models break the phonemes in the script according to the divisions between video frames, which solves the temporal alignment problem but leads to incomplete phoneme pronunciation and poor identity stability. To address this problem, we propose StyleDubber, which switches dubbing learning from the frame level to phoneme level. It contains three main components: (1) A multi-modal style adaptor operating at the phoneme level to learn pronunciation style from the reference audio, and generate intermediate representations informed by the facial emotion presented in the video; (2) An utterance-level style learning module, which guides both the mel-spectrogram decoding and the refining processes from the intermediate embeddings to improve the overall style expression; And (3) a phoneme-guided lip aligner to maintain lip sync. Extensive experiments on two of the primary benchmarks, V2C and Grid, demonstrate the favorable performance of the proposed method as compared to the current state-of-the-art. The code will be made available at https://github.com/GalaxyCong/StyleDubber.

1 Introduction

Movie Dubbing (Chen et al., 2022a), also known as Visual Voice Cloning (V2C), aims to convert a script into speech with the voice characteristics specified by the reference audio, while maintaining lip-sync with a video clip, and reflecting the character's emotions depicted therein (see Figure 1 (a)). V2C is more challenging than conventional text-to-speech (TTS) (Shen et al., 2018a; Ren et al., 2021), and has obvious applications in the film industry and audio AIGC, including broadening the audience for existing video.

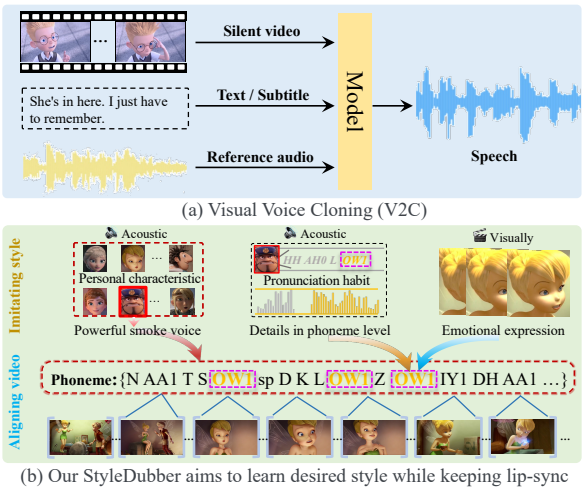

Figure 1: (a) Illustration of the V2C task. (b) Our StyleDubber learns speech styles on two levels: phoneme-level focuses on pronunciation details, while utterance-level emphasizes the overall consistency like timbre.

Existing methods broadly fall into two groups. One group of methods focus primarily on achieving audio-visual sync. For example, a duration aligner is introduced in (Hu et al., 2021; Cong et al., 2023) to explicitly control the speed and pause-duration of speaker content by mapping textual phonemes to video frames. Then, an upsampling process is used to expand the video frame sequence to the length of mel-spectrogram frame sequence by multiplying by a fixed coefficient. However, the frame level alignment makes it hard to learn complete phoneme pronunciations, and often leads to seemingly mumbled pronunciations. The other family of methods focuses on maintaining identity consistency between the generated speech and the reference audio. To enable the model to handle a multi-speaker environment, a speaker encoder is used to extract identity embeddings through averaging and nor-

*Corresponding authors

Findings of the Association for Computational Linguistics: ACL 2024, pages 6767–6779

malizing per speaker embeddings (Hassid et al., 2022). In contrast, (Lee et al., 2023) and (Hu et al., 2021) try to learn desired speaker voices based on facial appearances. Although humans' faces can reflect some vocal attributes (*e.g.*, age and identity) to some extent, they rarely encode speech styles, such as pronunciation habits or accents.

According to (Zhou et al., 2022; Li et al., 2022b), human speech can be perceived as a compound of multi-acoustic factors: (1) unique characteristics, such as timbre, which can be reflected on utterance level (see left panel of Figure 1 (b)); (2) pronunciation habits, such as the rhythm and regional accent, which are usually reflected at the phoneme level (see pink rectangles in Figure 1 (b)). We also note that one's voice can be affected by emotions. For example, the voice can be significantly different when one gets angry. Based on these observations, we propose to learn phoneme level representations from the speaker's pronunciation habits reflected in the reference audio, and take both facial expressions and overall timbre characteristics at the utterance level of the reference audio into consideration when generating speech.

In light of the above, we propose StyleDubber, which learns a desired style at the phoneme and utterance levels instead of the conventional video frame level. Specifically, a multimodal phoneme adaptor (MPA) is proposed to capture the pronunciation styles at the phoneme level. By leveraging the cross-attention relevance between textual phonemes of the script and the reference audio as well as visual emotions, MPA learns the reference style and then generates intermediate speech representations with consideration of the required emotion. Our model also introduce an utterance-level style learning (USL) module to strengthen personal characteristics during both the mel-spectrogram decoding and refining processes from the above intermediate representations. For the temporal alignment between the resulting speech and the video, we propose a Phoneme-guided Lip Aligner (PLA) to synchronize lip-motion and phoneme embeddings. At last, HiFiGAN (Kong et al., 2020) is used as a vocoder to convert the predicted mel-spectrogram to the time-domain waves of dubbing.

The main contributions are summarized as:

- We propose StyleDubber, a style-adaptative dubbing model, which imitates a desired personal style in phoneme and utterance levels. It enhances speech generation in terms of speech clarity and its temporal alignment with video.
- At the phoneme level, we design a multimodal style adaptor, which learns styled pronunciation of textual phonemes and considers facial expressions when generating intermediate speech representations. At the utterance level, our model learns to impose timbre into resulting mel-spectrograms.
- Extensive experimental results show that our model performs favorably compared to current state-of-the-art methods.

2 Related Work

Text to Speech is a longstanding problem, but recent models represent a dramatic improvement (Liu et al., 2024; Tan et al., 2024; Casanova et al., 2022; Wang et al., 2023a; Huang et al., 2023a,b; Ju et al., 2024). FastSpeech2 (Ren et al., 2021), for example, alleviates the one-to-many text-to-speech mapping problem by explicitly modeling variation information. Min et al. (2021), in contrast, improves generalization through episodic meta-learning and generative adversarial networks. Recently Le et al. (2023) proposed a non-autoregressive flow-matching model for mono or cross-lingual zero-shot text-to-speech synthesis. Despite the impressive speech they generate, these methods cannot be applied to the V2C task as they lack the required emotion modelling and lip sync.

Visual Voice Cloning is proposed to address the problem of film dubbing (Chen et al., 2022a) and has attracted a lot of attention in cross-modality alignment field (Tu et al., 2022, 2023, 2024; Li et al., 2022a; Liu et al., 2023; Wang et al., 2023b; Xiao et al., 2023, 2022). Then, Cong et al. (2023) proposed a hierarchical prosody dubbing model by associating with lip, face, and scene and focus on frame-level prosody learning (Hu et al., 2021). To handle multi-speaker scenes, Hassid et al. (2022) matches identities by normalizing each speaker to the unit norm, and Lu et al. (2022) adopts a lookup table to match the d-vector. Recently, Face-TTS (Lee et al., 2023) used biometric information extracted directly from the face image as style to improve identity modelling using a score-based diffusion model. Unlike the above methods, StyleDubber address the challenge of insufficient identity information by introducing adaptive utterance-level embedding and detailed pronunciation variations based on the reference audio and video.

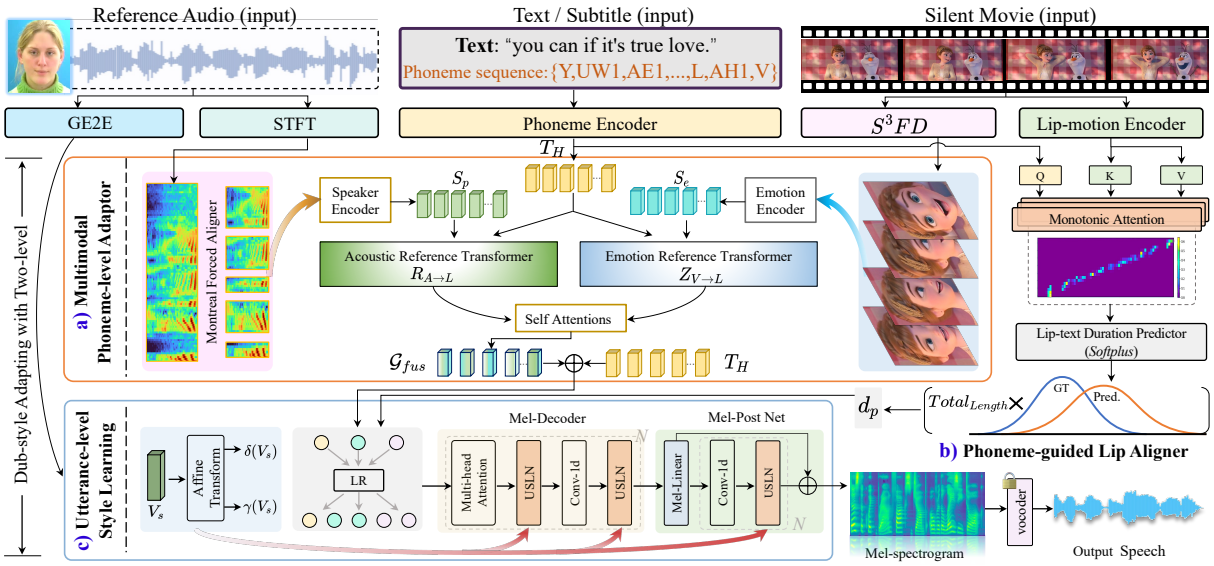

Figure 2: The main architecture of the proposed StyleDubber. It consists of a) Multimodal Phoneme-level Adaptor (MPA) (Sec. 3.2), b) Phoneme-guided Lip Aligner (PLA) (Sec. 3.3), and c) Utterance-level Style Learning (USL) (Sec. 3.4). Note that $\oplus$ is intended to denote vector addition.

Human Pronunciation Modeling aims to learn individual pronunciation variations, which is crucial to generate comprehensible, natural, and acceptable speech (Miller, 1998). Compared with fixed speaker representations, phoneme-dependent methods (Li et al., 2022b; Fu et al., 2019) can better control speech and describe more pronunciation features, as phonemes are the basic sound units in a language (Lubis et al., 2023). Recently, Zhou et al. (2022) analysed the correlation between local pronunciation content and speaker embeddings at the quasi-phoneme level by reference attention. Here, in contrast, we propose a multimodal style adaptor to capture the fine-grained pronunciation variation, which not only imitates the reference style acoustically, but also conveys emotional expression by reference transformer.

3 Proposed Method

3.1 Overview

Our StyleDubber aims to generate a desired dubbing speech $\hat{Y}$, given a reference audio R_a, a phoneme sequence T_p converted from the given script, and a video frame sequence V_l:

$$\hat{Y} = \text{StyleDubber}(R_a, T_p, V_l). \quad (1)$$

The main architecture of the model is shown in Figure 2. Unlike existing prosody dubbing methods, our model learns speech style from phoneme level and utterance level, inspired by human tonal phonetics. First, the textual phoneme sequence is converted from raw text. A phoneme encoder is then used to extract phoneme embeddings. These embeddings are fed into our Multimodal Phoneme-level Adaptor (MPA), which learns to capture and apply phoneme-level pronunciation styles to generate intermediate speech representations, meanwhile taking facial expressions into consideration. Next, our Phoneme-guided Lip Aligner (PLA) predicts the duration for each phoneme by associating lip motion sequence. The duration and intermediate dubbing representation are fed to our Utterance-level Style Learning (USL) module, which learns overall style at the utterance level and applys it during mel-spectrograms decoding and refining processes. We detail each module below.

3.2 Multimodal Phoneme-level Adaptor

Our Multimodal Phoneme-level Adaptor (MPA) contains three steps: (1) learn acoustic style from reference audio; (2) perceive visual emotion from silent movies; (3) generate intermediate speech representations for textual phonemes of the input script, with reference to the captured acoustic styles and emotions in the last two steps.

Learn acoustic style. We extract reference mel-spectrogram R_{mel} from reference audio R_a by Short-time Fourier transform (STFT), and the montreal forced aligner (McAuliffe et al., 2017) is used to clip phoneme. Then, we capture the style feature S_p via a mel-style encoder $\text{E}^{\text{spk}}_{\text{down}}(\cdot)$:

$$S_p = \text{E}^{\text{spk}}_{\text{down}}(R_{mel}), \quad (2)$$

where $\mathrm{E}^{\mathrm{spk}}_{\mathrm{down}}(\cdot)$ comprises a mel-style encoder (Min et al., 2021) and four 1D convolutional downsample layers (Zhou et al., 2022). On the other hand, embeddings of textual phoneme sequence $T_H \in \mathbb{R}^{N_p \times D_m}$ are extracted by a phoneme encoder $T_H = E_{pho}(T_p)$ (Cong et al., 2023), where N_p denotes the length of phoneme sequence. Next, we propose the acoustic reference transformer $R_{A\to L}$ (Acoustic to Language) to calculate the relevance between a textual phoneme embedding and each style feature by crossmodal transformer:

$$
\begin{aligned}
&R^{[0]}_{A\to L} = R^{[0]}_{L},\\
&\hat{R}^{[i]}_{A\to L} = \mathrm{CM}^{[i],\mathrm{mul}}_{A\to L}(\mathrm{LN}(R^{[i-1]}_{A\to L}), \mathrm{LN}(R^{[0]}_{A})) + \mathrm{LN}(R^{[i-1]}_{A\to L}),\\
&R^{[i]}_{A\to L} = f_{\theta^{[i]}_{A\to L}}(\mathrm{LN}(\hat{R}^{[i]}_{A\to L}) + \mathrm{LN}(\hat{R}^{[i]}_{A\to L}),
\end{aligned}
\tag{3}
$$

where $i = \{1, ..., D\}$ denotes the number of feed-forwardly layers, $\mathrm{LN}(\cdot)$ denotes the layer normalization, and f_θ is a positionwise feed-forward sublayer parametrized by θ. $\mathrm{CM}^{[i],\mathrm{mul}}_{A\to L}(\cdot)$ is a multihead attention between S_p and T_H, as follows:

$$
\mathrm{CM}^{[i],\mathrm{mul}}_{A\to L} = \mathrm{softmax}(\frac{T_H {S_p}^\top}{\sqrt{d_{S_p}}})S_p, \tag{4}
$$

where the textual phoneme embedding T_H is used as query and the style feature S_p is used as key and value. Unlike crossmodal transformer in (Tsai et al., 2019), our acoustic reference transformer removes the repeatedly reinforcing and MFCCs frame-level operation, and only focuses on interaction between quasi-phoneme scale of reference audio and script phoneme, which is more conducive to human pronunciation habits.

Unlike using cross-entropy loss as style classifier (Zhou et al., 2022), we constrain $\mathrm{E}^{\mathrm{spk}}_{\mathrm{down}}$ via a style consistency loss:

$$
\mathcal{L}_{spk} = \frac{1}{n} \cdot \sum_{j}^{n}(1 - \mathrm{cos\_sim}(\phi(T_j), A(S_p)_j)), \tag{5}
$$

where $\phi(\cdot)$ is a function outputting the embedding by the pre-trained GE2E model (Wan et al., 2018), $A(S_p)$ outputs a style vector via average pooling, and $\mathrm{cos\_sim}(\cdot)$ is the cosine similarity function. T represents the ground truth audio, n is batch size.

Perceive visual emotion. We first use the S^3FD model (Zhang et al., 2017) to detect facial region from each frame of video, and then an emotion face-alignment network (EmoFAN) (Toisoul et al., 2021) is used to extract emotion features $F_p \in \mathbb{R}^{N_v \times D_m}$ from face regions. The N_v represents the number of video frames. Similar to style extraction, emotional feature S_e is obtained by a downsampling equipped encoder:

$$
S_e = \mathrm{E}^{\mathrm{emo}}_{\mathrm{down}}(F_p), \tag{6}
$$

where $S_e \in \mathbb{R}^{N_{dv} \times D_m}$ and N_{dv} is length after down-sample. The difference from $\mathrm{E}^{\mathrm{spk}}_{\mathrm{down}}(\cdot)$ is that $\mathrm{E}^{\mathrm{emo}}_{\mathrm{down}}(\cdot)$ has two 1D convolutional downsample layers. Next, an emotion reference transformer $Z_{V\to L}$ (Visual to Language) is proposed to analyze the correlations between the emotional feature and textual phoneme. The $Z_{V\to L}$ has same architecture with $R_{A\to L}$. The $\mathrm{CM}^{[i],\mathrm{mul}}_{V\to L}(\cdot)$ is multihead attention to calculate correlation between S_e and T_H:

$$
\mathrm{CM}^{[i],\mathrm{mul}}_{V\to L} = \mathrm{softmax}(\frac{T_H {S_e}^\top}{\sqrt{d_{S_e}}})S_e, \tag{7}
$$

where key and value are emotional features S_e to assist script phoneme in selecting related visual emotion expression. Finally, we regard the output $Z^D_{V\to L}$ and $R^D_{A\to L}$ of the last layers of emotion reference transformer and acoustic reference transformer as context visual emotion and acoustic style, respectively. The cross-entropy emotional classification loss $\mathcal{L}_{emo}$ is used to constrain $\mathrm{E}^{\mathrm{emo}}_{\mathrm{down}}(\cdot)$.

Generate intermediate speech representations. We first concatenate the phoneme-level context visual emotion and acoustic style in channel dimension, and then feed it into self-attention blocks $\mathrm{SA}(\cdot)$ to fuse these embeddings:

$$
\mathcal{G}_{fus} = \mathrm{SA}([Z^D_{V\to L}, R^D_{A\to L}]), \tag{8}
$$

where $\mathcal{G}_{fus} \in \mathbb{R}^{N_p \times D_m}$ is the fused multimodal context embedding which is with the same length as the textual phonemes. Finally, we combine textual phoneme embedding and multimodal context embedding $\mathcal{O}_{pho} = \mathcal{G}_{fus} \oplus T_H$, which is viewed as the intermediate dubbing representations.

3.3 Phoneme-guided Lip Aligner

The Phoneme-guided Lip Aligner (PLA) consists of two steps: 1) Monotonic attention is used to learn the contextual aligning feature between lip motion and textual phoneme embedding; 2) Lip-text duration predictor aims to output the duration of each phoneme based on the contextual aligning feature.

Monotonic Attention. The lip-movement hidden representation $L_H = E_{lip}(V_l) \in \mathbb{R}^{N_v \times D_m}$ is obtained using the same lip-motion encoder in (Cong et al., 2023). Then, we encourage PLA to use

textual phoneme embedding to capture related lip motion by multi-head attention with monotonic constraint:

$$C_{lip} = \text{softmax}(\frac{T_H {L_H}^\top}{\sqrt{d_{L_H}}})L_H, \quad (9)$$

where the textual phoneme embedding T_H serves as query, and the lip motion embedding L_H serves as key and value. $C_{lip} \in \mathbb{R}^{N_p \times D_m}$ captures the dependency between lip and textual phoneme. Monotonic Alignment Loss (MAL) (Chen et al., 2020) is used to ensure proper alignment over the time:

$$\mathcal{L}_m = \log(-\frac{\sum_{l=kp-\beta}^{kp+\beta} \sum_{p=0}^{P-1} M_{p,l}}{\sum_{l=0}^{L-1} \sum_{p=0}^{P-1} M_{p,l}}), \quad (10)$$

where β is a hyper-parameter to control bandwidth, k is the slop for length of phoneme P and corresponding of lip length L, and $M_{p,l}$ is the masked attention weight matrix with p-th row and l-th column. To this end, $\mathcal{L}_{mon}$ aims to constrain attention weights to diagonal area to satisfy monotonicity.

Duration predictor. Since the total dubbing times $Total_{Length}$ can known by multiplying time coefficient with video frames N_v in advance (Hu et al., 2021), we transform the alignment problem into inferring the relative time of a phoneme over its total duration. We first use the duration predictor (Ren et al., 2021) to learn the duration from lip-phoneme context C_{lip} and re-scale it into relative duration by using $Total_{Length}$ divide predicted sum:

$$d_p = Total_{Length} \cdot \frac{\text{E}_{\text{Softplus}}(C_{lip}^k)}{\sum_{k=0}^{N_p-1} \text{E}_{\text{Softplus}}(C_{lip})}, \quad (11)$$

where $d_p \in \mathbb{R}^{N_p \times 1}$ represents the relative duration for each phoneme unit. $\text{E}_{\text{Softplus}}(\cdot)$ represents the duration predictor, which consist of 2-layer 1D convolutional with softplus activate function (Song et al., 2021). In this case, we obtain how many mel-frames correspond to lip-phoneme context C_{lip} to ensure the boundary of phoneme unit will not be broken, while syncing with the whole video.

Loss function. The duration loss is optimized with MSE loss, following (Ren et al., 2021):

$$\mathcal{L}_d = \text{MSE}(d_p, \log(g_d)), \quad (12)$$

where $\log(g_d)$ represents the ground-truth duration in the log domain.

3.4 Utterance-level Style Learning

We also consider the utterance-level information of reference audio to enhance global style characteristics. Specifically, we use the GE2E model (Wan et al., 2018) to extract the timbre vector V_s as utterance-level condition, which aggregates global style information to guide the decoding and refinement of mel-spectrograms from intermediate speech representations by affine transform.

Mel-Decoder. We use transformer-based mel-decoder (Cong et al., 2023) to decode intermediate speech representations $\mathcal{O}_{pho}$ into a spectrogram hidden sequence:

$$\hat{R} = \text{Decoder}_{\text{USLN}}(\text{LR}(\mathcal{O}_{pho}, d_p), V_s), \quad (13)$$

where $\text{LR}(\cdot)$ is the length regulator (Ren et al., 2021) to expand $\mathcal{O}_{pho}$ to mel-length based on predicted duration d_p. $\hat{R} \in \mathbb{R}^{N_{lr} \times 256}$ denotes a spectrogram hidden sequence, N_{lr} is the predicted total mel-legnth. During decoding, we replace the original layer norm (Ba et al., 2016) in each Feed-Forward Transformer (FFT) block with our utterance-level style learning normalization (USLN):

$$\text{USLN}(h, V_s) = \gamma(V_s) \cdot h_n + \delta(V_s), \quad (14)$$

where $h_n = (h - \mu)/\sigma$ is normalized features by the mean μ and variance σ of input feature h. The $\gamma(V_s)$ and $\delta(V_s)$ represent the learnable gain and bias of the overall style vector by affine transform, respectively, which can adaptively perform scaling and shifting to improve style expression.

Refine mel-spectrogram. We introduce the aforementioned USLN to MelPostNet (Shen et al., 2018b) to inject the style information from timbre vector V_s during refining the final mel-spectrograms stage:

$$\hat{M} = \text{POST}_{\text{USLN}}(\hat{R}, V_s), \quad (15)$$

where $\hat{M} \in \mathbb{R}^{N_{lr} \times 80}$ denotes the predicted mel-spectrograms with 80 channels.

3.5 Training

Our model is trained in an end-to-end fashion via optimizing the sum of all losses. The total loss $\mathcal{L}$ can be formulated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{spk} + \lambda_2 L_{emo} + \lambda_3 \mathcal{L}_r + \lambda_4 \mathcal{L}_m + \lambda_5 \mathcal{L}_d, \quad (16)$$

where $\mathcal{L}_r$ is the reconstruction loss to calculate L1 differences between the predicted and ground-truth mel-spectrograms.

Dataset		V2C-Animation					GRID				
Methods	Visual	SPK-SIM (%) ↑	WER (%) ↓	EMO-ACC (%) ↑	MCD-DTW ↓	MCD-DTW-SL ↓	SPK-SIM (%) ↑	WER (%) ↓	EMO-ACC (%) ↑	MCD-DTW ↓	MCD-DTW-SL ↓
GT	-	100.00	22.55	99.96	0.0	0.0	100.00	22.41	-	0.0	0.0
GT Mel + Vocoder	-	96.96	24.58	97.09	3.77	3.80	97.57	21.41	-	4.10	4.15
Fastspeech2 (Ren et al., 2021)	X	24.87	34.48	42.21	11.20	14.48	47.41	19.05	-	7.67	8.43
StyleSpeech (Min et al., 2021)	X	54.99	106.73	44.12	11.50	15.10	91.06	24.83	-	5.87	5.98
Zero-shot TTS (Zhou et al., 2022)	X	48.98	68.81	42.75	9.98	12.51	86.54	19.13	-	5.71	5.99
StyleSpeech* (Min et al., 2021)	✓	42.53	108.00	42.53	11.62	14.23	90.04	22.62	-	5.74	5.88
Fastspeech2* (Ren et al., 2021)	✓	25.47	33.53	42.39	11.35	14.73	59.58	19.61	-	7.24	7.95
Zero-shot TTS* (Zhou et al., 2022)	✓	48.93	68.05	43.97	10.03	12.01	85.93	20.05	-	5.75	6.40
V2C-Net (Chen et al., 2022a)	✓	40.61	73.08	43.08	14.12	18.49	80.98	47.82	-	6.79	7.23
HPMDubbing (Cong et al., 2023)	✓	53.76	164.16	**46.61**	11.12	11.22	85.11	45.11	-	6.49	6.78
Face-TTS (Lee et al., 2023)	✓	52.81	201.13	44.04	13.44	26.94	82.97	44.37	-	7.44	8.16
Ours	✓	**82.26**	**31.49**	45.62	**9.37**	**9.46**	**93.79**	**18.88**	-	**5.61**	**5.69**

Table 1: Results under the Dub 1.0 setting (Chen et al., 2022a), which uses ground-truth audio as reference audio. The method with "*" refers to a variant taking video embedding as an additional input as in (Chen et al., 2022a). The metric EMO-ACC is not applicable to GRID as it does not have emotional labels.

Methods	Visual	SPK-SIM (%) ↑	WER (%) ↓	EMO-ACC (%) ↑	MCD-DTW ↓	MCD-DTW-SL ↓	MOS-similarity ↑	MOS-naturalness ↑
GT	-	100.00	22.76	99.96	0.0	0.0	4.69 ± 0.12	4.76 ± 0.09
GT Mel + Vocoder	-	96.93	24.83	96.95	3.77	3.80	4.65 ± 0.07	4.63 ± 0.09
Fastspeech2 (Ren et al., 2021)	X	24.17	35.08	42.21	11.20	14.48	2.13 ± 0.09	3.75 ± 0.12
StyleSpeech (Min et al., 2021)	X	75.66	76.58	41.55	11.56	15.10	3.35 ± 0.07	3.24 ± 0.08
Zero-shot TTS (Zhou et al., 2022)	X	47.79	58.82	39.11	10.68	13.52	3.58 ± 0.11	3.72 ± 0.15
StyleSpeech* (Min et al., 2021)	✓	75.67	82.48	42.57	11.58	15.23	3.46 ± 0.16	3.83 ± 0.15
Fastspeech2* (Ren et al., 2021)	✓	25.47	34.08	42.39	11.35	14.73	2.46 ± 0.06	3.77 ± 0.08
Zero-shot TTS* (Zhou et al., 2022)	✓	47.55	58.81	39.30	10.76	13.66	3.68 ± 0.14	3.69 ± 0.09
V2C-Net (Chen et al., 2022a)	✓	34.07	61.61	41.01	14.58	18.73	3.04 ± 0.15	2.78 ± 0.06
HPMDubbing (Cong et al., 2023)	✓	31.42	171.03	**43.97**	11.88	11.98	3.19 ± 0.10	3.06 ± 0.22
Face-TTS (Lee et al., 2023)	✓	51.98	200.18	43.56	13.78	28.03	3.13 ± 0.12	3.09 ± 0.06
Ours	✓	**81.27**	**31.70**	41.35	**10.59**	**10.68**	**3.92 ± 0.11**	**3.86 ± 0.09**

Table 2: V2C-Animation results under Dub 2.0 setting, which uses non-ground truth audio of the desired character as reference audio.

Finally, the generated mel-spectrograms $\hat{M}$ are converted to time-domain wave $\hat{Y}$ via the widely used vocoder HiFiGAN.

4 Experiments

We evaluate our method on two primary V2C datasets, V2C-Animation and GRID. Below, we first provide implementation details. Then, we briefly introduce the datasets and evaluation metrics, followed by quantitative and qualitative results. Ablation studies are also conducted to thoroughly evaluate our model.

4.1 Implementation Details

The video frames are sampled at 25 FPS and all audios are resampled to 22.05kHz. The ground-truth of phoneme duration is extracted by montreal forced aligner (McAuliffe et al., 2017). The window length, frame size, and hop length in STFT are 1024, 1024, and 256, respectively. The lip region is resized to 96×96 and pretrained on ResNet-18, following (Martinez et al., 2020; Ma et al., 2021). We use 8 heads for multi-head attention in PLA and the hidden size is 512. The duration predictor consists of 2-layer 1D convolution with kernel size 1. The weights in Eq. 16 are set to $\lambda_1 = 25.0$, $\lambda_2 = 0.1$, $\lambda_3 = 5.0$, $\lambda_4 = 2.0$, $\lambda_5 = 5.0$. For down-sampling encoder in $\mathrm{E}^{\mathrm{spk}}_{\mathrm{down}}$, we use 4 convolutions containing [128, 256, 512, 512] filters with shape 3×1 respectively, each followed by an average pooling layer with kernel size 2. In $\mathrm{E}^{\mathrm{emo}}_{\mathrm{down}}$, 2 convolutions are used to download to quasi phoneme-level, containing [128, 256] filters with shape 3×1. In $R_{A \to L}$ and $Z_{V \to L}$, the dimensionality of all reference attention hidden size is set to 128 implemented by a 1D temporal convolutional layer. We set the batch size to 32 and 64 on V2C-Animation and Grid dataset, respectively. For training, we use Adam (Kingma and Ba, 2015) optimizer with β_1 = 0.9, β_2 = 0.98, ϵ=10^{-9} to optimize our model. The learning rate is set to 0.00625. Both training and inference are implemented with PyTorch on a GeForce RTX 4090 GPU.

4.2 Dataset

V2C-Animation dataset (Chen et al., 2022a) is currently the only publicly available movie dubbing dataset for multi-speaker. Specifically, it contains 153 diverse characters extracted from 26 Disney cartoon movies, specified with speaker identity and emotion annotations. The whole dataset has 10,217 video clips, and the audio samples are sampled at 22,050Hz with 16 bits. In practice, (Chen et al., 2022a) removes video clips less than 1s. In this work, all experiments are conducted on the V2C denoise version. We will publish this version.

GRID dataset (Cooke et al., 2006) is a basic benchmark for multi-speaker dubbing. The whole dataset has 33 speakers, each with 1000 short English sam-

Setting	Explanation	Num.
Dub 1.0 (Original setting (Chen et al., 2022a))	Ground-truth speaker in test set	2,779
Dub 2.0 (Reference speaker setting)	Same-speaker from other movie clips	2,626
Dub 3.0 (Unseen speaker setting)	Unseen speaker	4,851

Table 3: Experimental settings for dub testing in V2C.

ples. All participants are recorded in a noise-free studio with a unified screen background. The train set consists of 32,670 samples, 900 sentences from each speaker. In the test set, there are 100 samples of each speaker.

4.3 Evaluation Metrics

Objective metrics. To measure whether the generated speech carries the desired speaker identity and emotion, speaker identity similarity (SPK-SIM) is calculated by SECS (Casanova et al., 2021), and emotion accuracy (EMO-ACC) is employed by pre-trained speech emotion recognition model (Ye et al., 2023). Besides, we adopt the Mel Cepstral Distortion Dynamic Time Warping (MCD-DTW) to measure the difference between generated speech and real speech. We also adopt the metric MCD-DTW-SL, which is MCD-DTW weighted by duration consistency (Chen et al., 2022a). The Word Error Rate (WER) (Morris et al., 2004) is used to measure pronunciation accuracy by the publicly available whisper (Radford et al., 2023) as the ASR model. Besides, we use the ASV model (WavLM-TDNN (Chen et al., 2022b)) to comprehensively evaluate identity similarity (see Appendix D), following NaturalSpeech 3 (Ju et al., 2024).

Subjective metrics. We also provide subjective evaluation results via conducting a human study using a 5-scale mean opinion score (MOS) in two aspects: naturalness and similarity. Following the settings in (Chen et al., 2022a), all participants are asked to assess the sound quality of 25 randomly selected audio samples from each test set.

4.4 Performance Evaluations

We evaluate our method in three experimental settings as shown in Table 3. The first setting is the same as in (Chen et al., 2022a), which uses target audio as reference audio from test set. However, this is impractical in real-world applications. Thus, we design two new and more reasonable settings: "Dub 2.0" uses non-ground truth audio of the same speaker as reference audio; "Dub 3.0" uses the audio of unseen characters (from another dataset) as reference audio. We compare with six recent related baselines to comprehensively analyze. Furthermore, we will release the detailed configuration for all experiment settings for the GRID and V2C Animation datasets.

Methods	Visual	MOS-S	MOS-N	SPK-SIM	WER
Fastspeech2 (Ren et al., 2021)	X	2.91 ± 0.13	3.02 ± 0.09	21.11	14.05
StyleSpeech (Min et al., 2021)	X	3.17 ± 0.06	3.22 ± 0.15	55.81	77.46
Zero-shot TTS (Zhou et al., 2022)	X	3.53 ± 0.12	3.35 ± 0.07	57.23	19.83
Fastspeech2* (Ren et al., 2021)	✓	2.97 ± 0.12	3.03 ± 0.29	26.79	18.38
StyleSpeech* (Min et al., 2021)	✓	3.31 ± 0.18	3.22 ± 0.10	58.71	89.11
Zero-shot TTS* (Zhou et al., 2022)	✓	3.62 ± 0.09	3.31 ± 0.13	61.12	19.25
V2C-Net (Chen et al., 2022a)	✓	3.05 ± 0.07	2.83 ± 0.09	38.43	112.71
HPMDubbing (Cong et al., 2023)	✓	3.11 ± 0.08	2.92 ± 0.09	49.31	106.81
Face-TTS (Lee et al., 2023)	✓	3.10 ± 0.05	3.17 ± 0.15	33.80	201.98
Ours	✓	**3.94± 0.12**	**3.87± 0.14**	**71.52**	**13.45**

Table 4: The V2C results under Dub 3.0 setting, which use unseen speaker as refernce audio.

Results under Dub 1.0 setting. As shown in Table 1, our method achieves the best performance on almost all metrics on both GRID and V2C-Animation benchmarks. Our method only performs slightly worse in terms of EMO-ACC than the SOTA movie dubbing model HPMDubbing (Cong et al., 2023). Regarding identity accuracy (see SPK-SIM), our method outperforms other models with an absolute margin of 27.27% over the 2nd best method. In terms of MCD-DTW and MCD-DTW-SL, our method achieves 6.11% and 24.38% improvements, respectively. This indicates our method can achieve better speech quality and better duration consistency.

Results under Dub 2.0 setting. We report the V2C results in Table 2. Despite Dub 2.0 is much more challenging than 1.0, our method still outperforms other methods on six metrics. The SPK-SIM and WER significantly improve 7.4% and 6.98% than the fastspeech 2 (TTS-textonly) and meta-stylespeech method, respectively. Additionally, the proposed method (StyleDubber) achieves the lowest MCD-DTW compared to all baselines, which indicates our method achieves minimal acoustic difference even in challenging setting 2.0. Furthermore, the lowest MCD-DTW-SL shows that our method achieves almost the same duration sync as the groud-truth video. Finally, the human subjective evaluation results (see MOS-N and MOS-S) also show that our StyleDubber can generate speeches that are closer to realistic speech in both naturalness and similarity.

Results under Dub 3.0 setting. Since there is no target audio at this setting, we only compare SPK-SIM and WER, and make subjective evaluations. As shown in Table 4, our StyleDubber achieves the best generation quality in all four metrics, largely outperforming the baselines. The higher SPK-SIM and MOS-S (mean opinion score of similarity) indi-

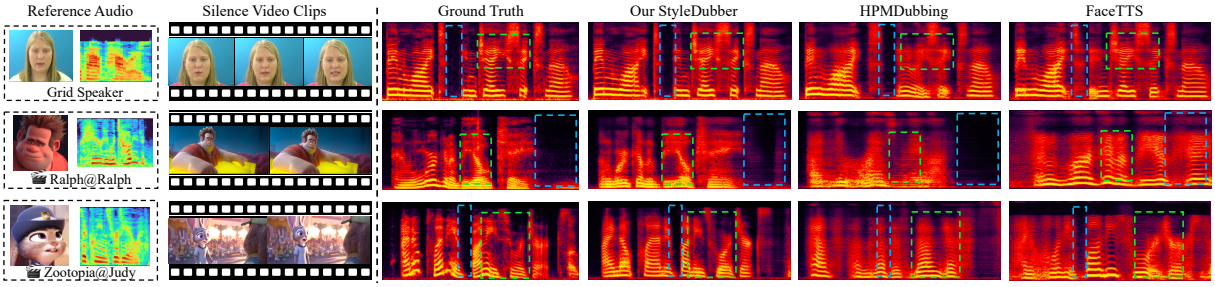

Figure 3: Mel-spectrograms of four synthesized audio samples under the Dub 2.0 setting. The green and blue rectangles highlight key regions that have significant differences in reconstruction details and duration pause.

#	Methods	WER ↓	SPK-SIM ↑	MCD-DTW ↓	MCD-DTW-SL ↓	EMO-ACC ↑
1	w/o MPA	39.74	77.26	9.90	9.98	41.18
2	w/o USL	32.81	47.07	10.23	10.43	42.77
3	w/o PLA	35.90	80.77	9.59	11.47	45.48
4	Quasi-phoneme $v.s.$ frame	33.62	80.76	9.75	9.84	43.83
5	w/o $R_{A\to L}$	38.21	77.79	9.81	9.90	42.61
6	w/o $Z_{V\to L}$	33.87	79.30	9.82	9.91	41.49
7	w/o U-MelDecoder	32.26	47.31	10.31	10.41	42.80
8	w/o U-Post	31.73	80.79	9.52	9.61	44.33
9	Full model	**31.49**	**82.26**	**9.37**	**9.46**	**45.62**

Table 5: Ablation study of the proposed method on the V2C benchmark dataset with 1.0 setting, respectively.

cate the better generalization ability of our methods to learn style adation across unseen speakers. Besides, our method also maintains good pronunciation (see WER). Overall, our StyleDubber achieves impressive results in challenging scenarios.

4.5 Qualitative Results

We visualize the mel-spectrogram of reference audio, ground-truth audio, and synthesized audios by ours and the other two state-of-the-art methods in Figure 3. We highlight the regions in mel-spectrograms where significant differences are observed among these methods in reconstruction details (green boxes), and pause (blue boxes), respectively. We find that our method is more similar to the ground-truth mel-spectrogram, which has clearer and distinct horizontal lines in the spectrum to benefit the fine-grained pronunciation expression of speakers (see green box). By observing the blue boxes, we find that our method can learn natural pauses to achieve better sync by aligning phonemes and lip motion.

4.6 Ablation Studies

To further study the influence of the individual components in StyleDubber, we perform the comprehensive ablation analysis using V2C 1.0 version.

Effectiveness of MPA, USL, and PLA. The results are presented in Row 1-3 of Table 5. It shows that all these three modules contribute significantly to the overall performance, and each module has a different focus. After removing the MPA, the MCD-DTW and WER severely drop. This reflects that the MPA achieves minimal difference in acoustic characteristics from the target speech and better pronunciation by phoneme modeling with other modalities. In contrast, the SPK-SIM is most affected by USL, which indicates decoding mel-spectrograms by introducing global style is more beneficial to identity recognition. Finally, the performance of MCD-DTW-SL drops the most when removing the PLA. This can be attributed to the better alignment between video and phoneme sequences.

Quasi-phoneme $v.s.$ **frame**. To prove the impact of regulating the temporal granularity to quasi-phoneme-scale, we remove the downsample operation and retrain the frame-level information as input of $R_{A\to L}$ and $Z_{V\to L}$. As shown in Row 4 of Table 5, all metrics have some degree of degradation, which means quasi-phoneme level acoustic and emotion representation is more conducive for script phoneme to capture desired information.

Effectiveness of $R_{A\to L}$ **and** $Z_{V\to L}$. To study the effect on each reference transformer in MPA, we remove $Z_{V\to L}$ and $Z_{A\to L}$, respectively. As shown in Row 5-6 of Table 5, $Z_{V\to L}$ has a significant effect on improving emotions, while $Z_{A\to L}$ more focus on local acoustic information to strengthen style and pronunciation.

Effectiveness of U-MelDecoder and U-post. To prove the effect of each module in USL, we remove the utterance-level style learning on mel-decoder and post-net, respectively. In other words, it still keeps an autoregressive manner by transformer-based decoder, and we just cut off the red arrow in Figure 2 (c). As shown in Row 7-8 of Table 5, when removing the U-post, the performance also drops but is not as large as removing the U-MelDecoder. This indicates that U-MelDecoder is critical to the generation of spectrum, while U-post only works

on refining spectrum in 80 channels so that the impact is relatively small.

5 Conclusion

In this work, we propose StyleDubber for movie dubbing, which imitates the speaker's voice at both phoneme and utterance levels while aligning with a reference video. StyleDubber uses a multimodal phoneme-level adaptor to improve pronunciation that captures speech style while considering the visual emotion. Moreover, a phoneme-guided lip aligner is devised to synchronize vision and speech without destroying the phoneme unit. The proposed model sets new state-of-the-art on the V2C and GRID benchmarks under three settings.

6 Limitation

We follow the task definition of Visual Voice Cloning (V2C), which focuses on generating audio only. Truly solving the larger problem would require changing the video to reflect the updated audio. In future, we will add this capability to better support tasks like cross-language video translation.

7 Ethics Statement

The existence of V2C methods lowers the barrier to high-quality and expressive visual voice cloning. In the long term this technology might enable broader consumption of factual and fictional video content. This could have employment implications, not least for current film voice actors. There is also a risk that V2C might be used to generate fake video depicting people apparently saying things they have never said. This is achievable already by an impersonator using entry-level video editing software, so the marginal impact of V2C on this problem is small. The licence for StyleDubber will explicitly prohibit this application, but the efficacy of such bans is limited, not least by the availability of other software that achieves the same purpose.

Acknowledgements

This work was supported in part by National Key R&D Program of China under Grant (2023YFB4502800), National Natural Science Foundation of China: 62322211, 61931008, 62236008, 62336008, U21B2038, 62225207, Fundamental Research Funds for the Central Universities (E2ET1104), "Pionee" and "Leading Goose" R&D Program of Zhejiang Province (2024C01023, 2023C01030). Yuankai Qi, Amin Beheshti, Anton van den Hengel, and Ming-Hsuan Yang are not supported by the aforementioned fundings.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Edresson Casanova, Christopher Shulby, Eren Gölge, Nicolas Michael Müller, Frederico Santos de Oliveira, Arnaldo Candido Jr., Anderson da Silva Soares, Sandra Maria Aluísio, and Moacir Antonelli Ponti. 2021. Sc-glowtts: An efficient zero-shot multi-speaker text-to-speech model. In *Interspeech*, pages 3645–3649.

Edresson Casanova, Julian Weber, Christopher Dane Shulby, Arnaldo Cândido Júnior, Eren Gölge, and Moacir A. Ponti. 2022. Yourtts: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *ICML*, pages 2709–2720.

Mingjian Chen, Xu Tan, Yi Ren, Jin Xu, Hao Sun, Sheng Zhao, and Tao Qin. 2020. Multispeech: Multi-speaker text to speech with transformer. In *Interspeech*, pages 4024–4028.

Qi Chen, Mingkui Tan, Yuankai Qi, Jiaqiu Zhou, Yuanqing Li, and Qi Wu. 2022a. V2C: visual voice cloning. In *CVPR*, pages 21210–21219.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022b. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518.

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 *CVPR*, pages 14687–14697.

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):2421–2424.

Ruibo Fu, Jianhua Tao, Zhengqi Wen, and Yibin Zheng. 2019. Phoneme dependent speaker embedding and model factorization for multi-speaker speech synthesis and adaptation. In *ICASSP*, pages 6930–6934.

Michael Hassid, Michelle Tadmor Ramanovich, Brendan Shillingford, Miaosen Wang, Ye Jia, and Tal Remez. 2022. More than words: In-the-wild visually-driven prosody for text-to-speech. In *CVPR*, pages 10577–10587.

Chenxu Hu, Qiao Tian, Tingle Li, Yuping Wang, Yuxuan Wang, and Hang Zhao. 2021. Neural dubber: Dubbing for videos according to scripts. In *NeurIPS*, pages 16582–16595.

Rongjie Huang, Yi Ren, Ziyue Jiang, Chenye Cui, Jinglin Liu, and Zhou Zhao. 2023a. Fastdiff 2: Revisiting and incorporating gans and diffusion models in high-fidelity speech synthesis. In *ACL*, pages 6994–7009.

Rongjie Huang, Chunlei Zhang, Yi Ren, Zhou Zhao, and Dong Yu. 2023b. Prosody-tts: Improving prosody with masked autoencoder and conditional diffusion model for expressive text-to-speech. In *ACL*, pages 8018–8034.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *NeurIPS*, pages 17022–17033.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *arXiv preprint arXiv:2306.15687*.

Jiyoung Lee, Joon Son Chung, and Soo-Whan Chung. 2023. Imaginary voice: Face-styled diffusion model for text-to-speech. In *ICASSP*, pages 1–5.

Liang Li, Xingyu Gao, Jincan Deng, Yunbin Tu, Zheng-Jun Zha, and Qingming Huang. 2022a. Long short-term relation transformer with global gating for video captioning. *TIP*, 31:2726–2738.

Xiang Li, Changhe Song, Jingbei Li, Zhiyong Wu, Jia Jia, and Helen Meng. 2022b. Towards multi-scale style control for expressive speech synthesis. In *Interspeech*, pages 4673–4677.

Rui Liu, Yifan Hu, Haolin Zuo, Zhaojie Luo, Longbiao Wang, and Guanglai Gao. 2024. Text-to-speech for low-resource agglutinative language with morphology-aware language model pre-training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1075–1087.

Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Zechao Li, Qi Tian, and Qingming Huang. 2023. Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding. *PAMI*, 45(3):3003–3018.

Junchen Lu, Berrak Sisman, Rui Liu, Mingyang Zhang, and Haizhou Li. 2022. Visualtts: TTS with accurate lip-speech synchronization for automatic voice over. In *ICASSP*, pages 8032–8036.

Yani Lubis, Fatimah Azzahra Siregar, and Cut Ria Manisha. 2023. The basic of english phonology: A literature review. *Jurnal Insan Pendidikan dan Sosial Humaniora*, 1(3):126–136.

Pingchuan Ma, Brais Martinez, Stavros Petridis, and Maja Pantic. 2021. Towards practical lipreading with distilled and efficient models. In *ICASSP*, pages 7608–7612.

Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2020. Lipreading using temporal convolutional networks. In *ICASSP*, pages 6319–6323.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, pages 498–502.

Corey Andrew Miller. 1998. *Pronunciation modeling in speech synthesis*. University of Pennsylvania.

Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang. 2021. Meta-stylespeech : Multi-speaker adaptive text-to-speech generation. In *ICML*, pages 7748–7759.

Andrew Cameron Morris, Viktoria Maier, and Phil D. Green. 2004. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Interspeech*, pages 2765–2768.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*, pages 28492–28518.

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 *ICLR*.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ-Skerrv Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018a. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *ICASSP*, pages 4779–4783.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ-Skerrv Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018b. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *ICASSP*, pages 4779–4783.

Wei Song, Xin Yuan, Zhengchen Zhang, Chao Zhang, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. Dian: Duration informed auto-regressive network for voice cloning. In *ICASSP*, pages 8598–8602.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, Sheng Zhao, Tao Qin, Frank Soong, and Tie-Yan Liu. 2024. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *PAMI*, pages 1–12.

Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nat. Mach. Intell.*, 3(1):42–50.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, pages 6558–6569.

Yunbin Tu, Liang Li, Li Su, Shengxiang Gao, Chenggang Yan, Zheng-Jun Zha, Zhengtao Yu, and Qingming Huang. 2022. I 2 transformer: Intra-and inter-relation embedding transformer for tv show captioning. *TIP*, 31:3565–3577.

Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, and Qingming Huang. 2024. Smart: Syntax-calibrated multi-aspect relation transformer for change captioning. *PAMI*.

Yunbin Tu, Chang Zhou, Junjun Guo, Huafeng Li, Shengxiang Gao, and Zhengtao Yu. 2023. Relation-aware attention for video captioning via graph learning. *PR*, 136:109204.

Li Wan, Quan Wang, Alan Papir, and Ignacio López-Moreno. 2018. Generalized end-to-end loss for speaker verification. In *ICASSP*, pages 4879–4883.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Hao Wang, Zheng-Jun Zha, Liang Li, Xuejin Chen, and Jiebo Luo. 2023b. Semantic and relation modulation for audio-visual event localization. *PAMI*, 45(6):7711–7725.

Jiayu Xiao, Liang Li, Henglei Lv, Shuhui Wang, and Qingming Huang. 2023. R&b: Region and boundary aware zero-shot grounded text-to-image generation. *arXiv preprint arXiv:2310.08872*.

Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, and Qingming Huang. 2022. Few shot generative model adaption via relaxed spatial structural alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11204–11213.

Jiaxin Ye, Xin-Cheng Wen, Yujie Wei, Yong Xu, Kunhong Liu, and Hongming Shan. 2023. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In *ICASSP*, pages 1–5.

Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. 2017. S3fd: Single shot scale-invariant face detector. In *CVPR*, pages 192–201.

Yixuan Zhou, Changhe Song, Xiang Li, Luwen Zhang, Zhiyong Wu, Yanyao Bian, Dan Su, and Helen Meng. 2022. Content-dependent fine-grained speaker embedding for zero-shot speaker adaptation in text-to-speech synthesis. In *Interspeech*, pages 2573–2577.

Appendix

We organise the supplementary materials as follows.

- In Section A, we analyze the challenges of the V2C benchmark compared with the traditional TTS benchmark.
- In Section B, we introduced related baseline methods.
- In Section C, we report the WER result by different Whisper versions on the V2C-Animation dataset.
- In Section D, we report the speaker similarity result by the large WavLM-TDNN model on the V2C-Animation dataset and GRID dataset.

A The challenges in V2C benchmark

The V2C benchmark significantly differs from traditional TTS benchmark , and it is more challenging in the following aspects: (1) The data scale of V2C dataset is much smaller in terms of either the number of data items or speech length (see Figure 4 (a)-(b)). There are only 9374 video clips in V2C, and most of its audio is shorter than 5s. In contrast, FS2 and Stylespeech are trained on LJspeech and LibriTTS with 13,100 and 149,753 samples, most of which are longer than 5s. Although LJspeech also looks similar in size to V2C, it is a single-speaker dataset, so V2C allocates fewer samples to each speaker. (2) V2C has the largest variance of pitch compared to TTS tasks due to exaggerated expressions of cartoon characters (see Figure 1 (c) and more details in Tab. 2 of V2C-Net). (3) The audio of V2C contains background noise or music, like car whistle and alarm clock sound, *et al.* Signal-to-noise Ratio (SNR) of V2C is the lowest (Figure 4 (d)). In summary, unlike the large-scale clean TTS datasets, V2C is much more challenging, and the well-known TTS methods suffer performance degradation. In this work, all experiments are conducted on the V2C denoise version. We will publish this version.

B Baselines

We compare our method against six closely related methods with available codes. 1) StyleSpeech (Min et al., 2021) is a multi-speaker voice clone method that synthesizes speech in the style of the target speaker via meta-learning; 2) FastSpeech2 (Ren et al., 2021) is a popular multi-speaker TTS method

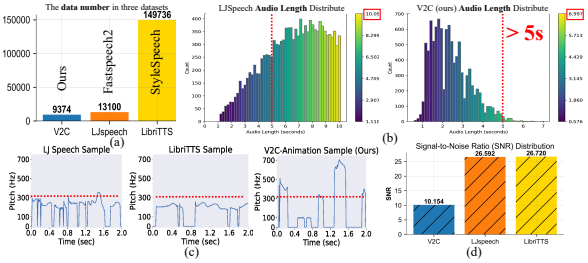

Figure 4: V2C dataset is more challenging than TTS-baseline datasets: (a) fewer samples (only 6567 for training), (b) shorter duration (mostly smaller than 5s), (c) greater variance (pitch), (d) more noise (background sound and music).

#	Whisper's Version	WER on V2C-Animation (%) ↓
1	Whisper (base)	45.58
2	Whisper (large-v1)	23.88
3	Whisper (large-v2)	23.85
4	Whisper (large-v3)	**22.55**

Table 6: The WER test (the ground truth result) for various versions of Whisper on the V2C benchmark dataset.

for explicitly modeling energy and pitch in speech; 3) Zero-shot TTS (Zhou et al., 2022) is a content-dependent fine-grained speaker method for zero-shot speaker adaptation. 4) V2C-Net (Chen et al., 2022a) is the first visual voice cloning model for movie dubbing; 5) HPMDubbing (Cong et al., 2023) is a hierarchical prosody modeling for movie dubbing, which bridges video representations and speech attributes from three levels: lip, facial expression, and scene. 6) Face-TTS (Lee et al., 2023) is a novel face-styled speech synthesis within a diffusion model, which leverages face images to provide a robust characteristic of speakers. In addition, for a fair comparison, for the pure TTS method, we adopt the setting as (Chen et al., 2022a), which takes video embedding as an additional input, before the duration predictor to predict the duration.

C Whisper test on V2C-Animation dataset

In Table 6, the results show that large-v3 achieved the lowest WER, thus we re-selected large-v3 as the measurement tool to get more convincing results on the V2C-Animation dataset. Note that Whisper-Large V3 has not been fine-tuned in the V2C-Animation dataset, there is still some gap (WER in GT is 22.55 %), but it is enough to serve as a fair comparison. All results (Dub1.0, 2,0, and 3.0) still reflect that our StyleDubber is the best (see Table 1,2,4) and is more conducive to

Methods	Visual	Sim-O ↑	Sim-R ↑
Ground Truth	-	0.79	n/a
Fastspeech2 (Ren et al., 2021)	X	0.10	0.19
StyleSpeech (Min et al., 2021)	X	0.14	0.23
Zero-shot TTS (Zhou et al., 2022)	X	0.12	0.21
Fastspeech2* (Ren et al., 2021)	✓	0.10	0.18
StyleSpeech* (Min et al., 2021)	✓	0.14	0.23
Zero-shot TTS* (Zhou et al., 2022)	✓	0.13	0.22
V2C-Net (Chen et al., 2022a)	✓	0.08	0.15
HPMDubbing (Cong et al., 2023)	✓	0.11	0.19
Face-TTS (Lee et al., 2023)	✓	0.09	0.12
Ours	✓	**0.25**	**0.34**

Table 7: The V2C-Animation dataset results under the WavLM-TDNN similarity testing.

Methods	Visual	Sim-O ↑	Sim-R ↑
Ground Truth	-	0.87	n/a
Fastspeech2 (Ren et al., 2021)	X	0.38	0.42
StyleSpeech (Min et al., 2021)	X	0.74	0.79
Zero-shot TTS (Zhou et al., 2022)	X	0.70	0.75
Fastspeech2* (Ren et al., 2021)	✓	0.48	0.52
StyleSpeech* (Min et al., 2021)	✓	0.74	0.79
Zero-shot TTS* (Zhou et al., 2022)	✓	0.69	0.74
V2C-Net (Chen et al., 2022a)	✓	0.43	0.55
HPMDubbing (Cong et al., 2023)	✓	0.46	0.56
Face-TTS (Lee et al., 2023)	✓	0.42	0.51
Ours	✓	**0.75**	**0.80**

Table 8: The GRID dataset results under the WavLM-TDNN similarity testing.

the clarity of movie dubbing. Considering the inference speed and computation memory, the Grid dataset still retains the original "Whisper-base" as the test benchmark. The "Whisper-base" achieves the 22.41 % GT WER on the GRID test as similar to the VDTTS (Hassid et al., 2022) result in Table 2 (GRID evaluation).

D WavLM-TDNN Similarity Testing

We employ the SOTA speaker verification model, WavLM-TDNN, to evaluate the speaker similarity between the prompt (*i.e.*, the reference audio in the V2C task) and synthesized speech, following VALL-E (Wang et al., 2023a), VoiceBox (Le et al., 2023), and NaturalSpeech 3 (Ju et al., 2024). WavLM-TDNN achieved the top rank at the VoxSRC Challenge 2021 and 2022 leaderboards and it is suitable as the SPK-SIM metric for the challenging V2C-nimation dataset (Chen et al., 2022a). The similarity score predicted by WavLM-TDNN is in the range of [-1; 1], where a larger value indicates a higher similarity of input samples. Specifically, two metrics need to be calculated: (1) SIM-R represents the similarity with resynthesized audio, which is not comparable across models using different vocoders; (2) SIM-O is used to measure similarity against the original reference audio. Note that the report results based on the GE2E model (Table 1, 2, 4) are compared with the original waveform.

As shown in Table 7,8, the results on two datasets show that even if the similarity measurement method is replaced, our StyleDubber still achieves the best performance in Sim-O and Sim-R. The result proves the effectiveness of StyleDubber, which proposes multi-scale style learning at phoneme and utterance levels and captures precise pronunciation in acoustic details and visual emotion for dubbing. Besides, we have several findings: (1) Compared with the GT result on the GRID dataset (0.87), the Sim-O result of V2C-Animation is lower (0.79), which may be due to the influence of noise and background music. In contrast, the GRID dataset is recorded in a studio environment. (2) Changing the measurement metric has relatively little impact on the GRID dataset. V2C has a very obvious decline in WavLM-TDNN similarity, which is not captured by the GE2E model (GE2E score can reach more than 80%). In the future, we will investigate more robust timbre extractors and use the denoising diffusion probabilistic models to further improve the generated wave quality.