Parallel and High-Fidelity Text-to-Lip Generation

Jinglin Liu*[1], Zhiying Zhu*[1], Yi Ren*[1],
Wencan Huang[1], Baoxing Huai[2], Nicholas Yuan[2], Zhou Zhao†[1]
[1]Zhejiang University, China
[2]Huawei Cloud
{jinglinliu,zhyingzh,rayeren,huangwencan,zhaozhou}@zju.edu.cn,
{huaibaoxing,nicholas.yuan}@huawei.com

Abstract

As a key component of talking face generation, lip movements generation determines the naturalness and coherence of the generated talking face video. Prior literature mainly focuses on speech-to-lip generation while there is a paucity in text-to-lip (T2L) generation. T2L is a challenging task and existing end-to-end works depend on the attention mechanism and autoregressive (AR) decoding manner. However, the AR decoding manner generates current lip frame conditioned on frames generated previously, which inherently hinders the inference speed, and also has a detrimental effect on the quality of generated lip frames due to error propagation. This encourages the research of parallel T2L generation. In this work, we propose a parallel decoding model for fast and high-fidelity text-to-lip generation (ParaLip). Specifically, we predict the duration of the encoded linguistic features and model the target lip frames conditioned on the encoded linguistic features with their duration in a non-autoregressive manner. Furthermore, we incorporate the structural similarity index loss and adversarial learning to improve perceptual quality of generated lip frames and alleviate the blurry prediction problem. Extensive experiments conducted on GRID and TCD-TIMIT datasets demonstrate the superiority of proposed methods.

1 Introduction

In the modern service industries, talking face generation has broad application prospects such as avatar, virtual assistant, movie animation, teleconferencing, etc. (Zhu et al. 2020). As a key component of talking face generation, lip movements generation (a.k.a. lip generation) determines the naturalness and coherence of the generated talking face video. Lip generation aims to synthesize accurate mouth movements video corresponding to the linguistic content information carried in speech or pure text.

Mainstream literature focuses on speech-to-lip (S2L) generation while there is a paucity in text-to-lip (T2L) generation. Even so, T2L generation is very crucial and has considerable merits compared to S2L since 1) text data can be obtained or edited more easily than speech, which makes T2L generation more convenient; and 2) T2L extremely preserves privacy especially in the society where the deep learning techniques are so developed that a single sentence speech could expose an unimaginable amount of personal information.

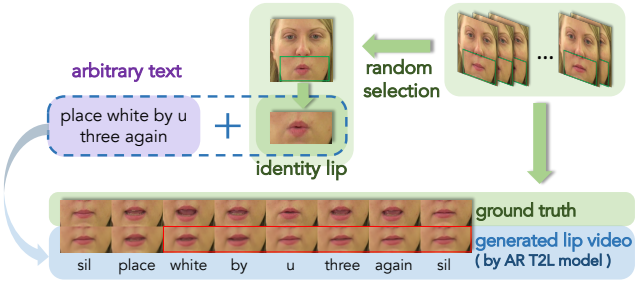


Figure 1: The task description of T2L generation. The model takes in an arbitrary source text sequence and a single identity lip image to synthesize the target lip movements video. And in this figure we can see that the generated video loses the linguistic information gradually and finally becomes fuzzy and motionless (lip frames in the red box), which is the intractable problem existing in AR T2L models due to error propagation.

However, end-to-end T2L task (shown in Figure 1) is challenging. Unlike S2L task where the mapping relationship between the sequence length of source speech and target video is certain (according to audio sample rate and fps), there is an uncertain sequence length discrepancy between source and target in T2L task. The traditional temporal convolutional networks become impractical. Hence, existing works view T2L as a sequence-to-sequence task and tackle it by leveraging the attention mechanism and autoregressive (AR) decoding manner. The AR decoding manner brings two drawbacks: 1) it inherently hinders the inference speed since its decoder generates target lips one by one autoregressively with the causal structure. Consequently, generating a single sentence of short video consumes about 0.5-1.5 seconds even on GPU, which is not acceptable for industrial applications such as real-time interactions with avatar or virtual assistant, real-time teleconferencing and document-level audio-visual speech synthesis, etc. 2) It has a detrimental effect on the quality of generated lips due to error prop-

*Equal contribution.
†Corresponding author

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

agation[1], which is frequently discussed in neural machine translation and image caption field (Bengio et al. 2015; Wu et al. 2018). Worse still, error propagation is more obvious in AR lip generation than in other tasks, because the mistakes could take place at more dimensions (every pixel with three channels in generated image) and there is information loss during the down-sampling when sending the last generated lip frame to predict current one. Although prior works alleviate the error propagation by incorporating the technique of location-sensitive attention, it still has an unsatisfying performance on long-sequence datasets due to accumulated prediction error.

To address such limitations, we turn to non-autoregressive (NAR) approaches. NAR decoding manner generates all the target tokens in parallel, which has already pervaded multiple research fields such as neural machine translation (Gu et al. 2018; Lee, Mansimov, and Cho 2018; Ghazvininejad et al. 2019; Ma et al. 2019), speech recognition (Chen et al. 2019; Higuchi et al. 2020), speech synthesis (Ren et al. 2019; Peng et al. 2020; Miao et al. 2020), image captioning (Deng et al. 2020) and lip reading (Liu et al. 2020). These works utilize the NAR decoding in sequence-to-sequence tasks to reduce the inference latency or generate length-controllable sequence.

In this work, we propose an NAR model for parallel and high-fidelity T2L generation (ParaLip). ParaLip predicts the duration of the encoded linguistic features and models the target lip frames conditioned on the encoded linguistic features with their duration in a non-autoregressive manner. Furthermore, we leverage structural similarity index (SSIM) loss to supervise ParaLip generating lips with better perceptual quality. Finally, using only reconstruction loss and SSIM loss is insufficient to generate distinct lip images with more realistic texture and local details (*e.g.*wrinkles, beard and teeth), and therefore we adopt adversarial learning to mitigate this problem.

Our main contributions can be summarized as follows: 1) We point out and analyze the unacceptable inference latency and intractable error propagation existing in AR T2L generation. 2) To circumvent these problems, we propose ParaLip to generate high-quality lips with low inference latency. And as a byproduct of ParaLip, the duration predictor in ParaLip could be leveraged in an NAR text-to-speech model, which naturally enables the synchronization in audio-visual speech synthesis task. 3) We explore the source-target alignment method when the audio is absent even in the training set. Extensive experiments demonstrate that ParaLip generates the competitive lip movements quality compared with state-of-the-art AR T2L model and exceeds the baseline AR model TransformerT2L by a notable margin. In the meanwhile, ParaLip exhibits distinct superiority in inference speed, which truly provides the possibility to bring T2L generation from laboratory to industrial applications. Video samples are available via https://paralip.github.io.

[1]Error propagation means if a token is mistakenly predicted at inference stage, the error will be propagated and the future tokens conditioned on this one will be influenced (Bengio et al. 2015; Wu et al. 2018).

2 Related Work

2.1 Talking Face Generation

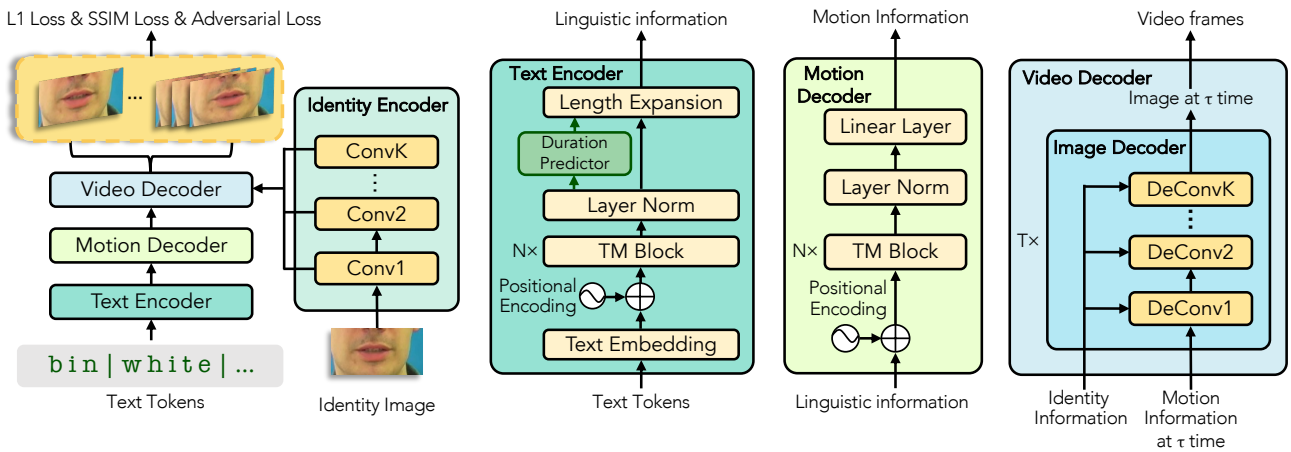
Talking face generation aims to generate realistic talking face video and covers many applications such as avatar and movie animation. There is a branch of works in computer graphics (CG) field exploring it (Wang et al. 2011; Fan et al. 2015; Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017; Yu et al. 2019; Abdelaziz et al. 2020) through hidden Markov models or deep neural network. These works synthesize the whole face by generating the intermediate parameters, which can then be used to deform a 3D face. Thanks to the evolved convolutional neural network (CNN) and high-performance computing resources, end-to-end systems which synthesize 2D talking face images by CNN rather than rendering methods in CG, have been presented recently in the computer vision (CV) field (Kumar et al. 2017; Chung, Jamaludin, and Zisserman 2017; Chen et al. 2018; Vougioukas, Petridis, and Pantic 2019; Zhou et al. 2019; Zhu et al. 2020; Zheng et al. 2020; Prajwal et al. 2020). Most of them focus on synthesizing lip movements images and then transforming them to faces. We mainly take these works in CV field into consideration and broadly divide them into two streams as the following paragraphs.

Speech-to-Lip Generation Previous speech-driven works *e.g.*Chung, Jamaludin, and Zisserman (2017) simply generate the talking face images conditioned on the encoded speech and the encoded face image carrying the identity information. To synthesize more accurate and distinct lip movements, Chen et al. (2018) introduce the task of speech-to-lip generation using lip image as the identity information. Further, Song et al. (2019) add a lip-reading discriminator to focus on the mouth region, and Zhu et al. (2020) add the dynamic attention on lip area to synthesize talking face while keeping the lip movements realistic. Prajwal et al. (2020) propose a pre-trained lip-syncing discriminator to synthesize talking face with speech-consistent lip movements.

Text-to-Lip Generation The literature of direct text-to-lip generation is rare. Some text-driven approaches either cascade the text-to-speech and speech-to-lip generation model(KR et al. 2019; Kumar et al. 2017), or combine the text feature with speech feature together to synthesize lip movements (Yu, Yu, and Ling 2019). Fried et al. (2019) edit a given video based on pure speech-aligned text sequence. Unlike the scenario where source speech or video is given, the sequence length of target lip frames is uncertain with only text input. Existing work (Chen et al. 2020) depends on the attention mechanism and AR decoding method to generate the target lip frames until the stop token is predicted.

2.2 Non-Autoregressive Sequence Generation

In sequence-to-sequence tasks, an autoregressive (AR) model takes in a source sequence and then generates tokens of the target sentence one by one with the causal structure at inference (Sutskever, Vinyals, and Le 2014; Vaswani et al. 2017). Since the AR decoding manner causes the high inference latency, many non-autoregressive (NAR) models, which generate target tokens conditionally independent of



(a) ParaLip.

(b) Text Encoder with Length Regulator.

(c) Motion Decoder.

(d) Video Decoder with multiple Image Decoders.

Figure 2: The overall architecture for ParaLip. In subfigure (a), Identity Encoder sends out residual information at every convolutional layer. In subfigure (b), Length Regulator expands the text sequence according to ground truth duration in training or predicted duration in inference. In subfigure (c), Motion Decoder models lip movement information sequence from linguistic information sequence. In subfigure (d), there are T Image Decoders placed parallel in Video Decoder. The τ-th Image Decoder takes in motion information at τ time and generates lip image at τ time. T means total number of lip frames.

each other, have been proposed recently. Earliest in the NAR machine translation field, many works use the fertility module or length predictor (Gu et al. 2018; Lee, Mansimov, and Cho 2018; Ghazvininejad et al. 2019; Ma et al. 2019) to predict the length correspondence (fertility) between source and target sequences, and then generate the target sequence depending on the source sequence and predicted fertility. Shortly afterward, researchers bring NAR decoding manner into heterogeneous tasks. In the speech field, NAR-based TTS (Ren et al. 2019; Peng et al. 2020; Miao et al. 2020) synthesize speech from text with high speed and slightly quality drop; NAR-based ASR (Chen et al. 2019; Higuchi et al. 2020) recognize speech to corresponding transcription faster. In the computer vision field, Liu et al. (2020) propose an NAR model for lipreading; Deng et al. (2020) present NAR image caption model not only improving the decoding efficiency but also making the generated captions more controllable and diverse.

3 Method

3.1 Preliminary Knowledge

The text-to-lip generation aims to generate the sequence of lip movement video frames $\mathcal{L} = \{l_1, l_2, ..., l_T\}$, given source text sequence $\mathcal{S} = \{s_1, s_2, ..., s_m\}$ and a single identity lip image l_I as condition. Generally, there is a considerable discrepancy between the sequence length of $\mathcal{L}$ and $\mathcal{S}$ with uncertain mapping relationship. Previous work views this as a sequence-to-sequence problem, utilizing attention mechanism and AR decoding manner, where the conditional probability of $\mathcal{L}$ can be formulated as:

$$P(\mathcal{L}|\mathcal{S}, l_I) = \prod_{\tau=0}^{T} P(l_{\tau+1}|l_{<\tau+1}, \mathcal{S}, l_I; \theta), \tag{1}$$

where θ denotes the parameters of the model.

To remedy the error propagation and high latency problem brought by AR decoding, ParaLip models the target sequence in an NAR manner, where the conditional probability becomes:

$$P(\mathcal{L}|\mathcal{S}, l_I) = \prod_{\tau=1}^{T} P(l_\tau|\mathcal{S}, l_I; \theta). \tag{2}$$

3.2 Model Architecture of ParaLip

The overall model architecture and training losses are shown in Figure 2a. We explain each component in ParaLip in the following paragraphs.

Identity Encoder As shown in the right panel of Figure 2a, identity encoder consists of stacked 2D convolutional layers with batch normalization, which down-samples the identity image multiple times to extract features. The identity image is selected randomly from target lip frames, providing the appearance information of a speaker. It is worth noting that the identity encoder sends out the final encoded hidden feature together with the intermediate hidden feature of convolutional layers at every level, which provides the fine-grained image information.

Text Encoder As shown in Figure 2b, the text encoder consists of a text embedding layer, stacked feed-forward Transformer layers (TM) (Ren et al. 2019), a duration predictor and a length regulator. The TM layer contains self-attention layer and 1D convolutional layer with layer normalization and residual connection (Vaswani et al. 2017; Gehring et al. 2017). The duration predictor contains two 1D convolutional layers with layer normalization and one linear layer, which takes in the hidden text embedding sequence

and predicts duration sequence $\mathcal{D}^* = \{d_1^*, d_2^*, ..., d_m^*\}$, where d_i^* means how many video frames the i-th text token corresponding to. The length regulator expands the hidden text embedding sequence according to ground truth duration $\mathcal{D}$ at training stage or predicted duration $\mathcal{D}^*$ at inference stage. For example, when given source text and duration sequence are $\{s_1, s_2, s_3\}$ and $\{2, 1, 3\}$ respectively, denoting the hidden text embedding as $\{h_1, h_2, h_3\}$, the expanded sequence is $\{h_1, h_1, h_2, h_3, h_3, h_3\}$, which carries the linguistic information corresponding to lip movement video at frame level. Collectively, text encoder encodes the source text sequence S to linguistic information sequence $\widetilde{\mathcal{S}} = \{\widetilde{s}_1, \widetilde{s}_2, ..., \widetilde{s}_{T^*}\}$, where $T^* = \sum_{i=1}^{m} d_i$ at training stage, or $T^* = \sum_{i=1}^{m} d_i^*$ at inference stage.

Motion Decoder Motion decoder (Figure 2c) aims to produce the lip movement information sequence $\widetilde{\mathcal{L}} = \{\widetilde{l}_1, \widetilde{l}_2, ..., \widetilde{l}_{T^*}\}$ from linguistic information sequence $\widetilde{\mathcal{S}}$. It utilizes the positional encoding and self-attention mechanism in stacked TM blocks to enforce the temporal correlation on the hidden sequence. There is a linear layer at the end of this module convert the hidden states to an appropriate dimension.

Video Decoder The video decoder generates the target lip movement video $\mathcal{L}^*$ conditioned on the motion information sequence and identity information. As shown in Figure 2d, the video decoder consists of multiple parallel image decoders with all parameters shared, each of which contains stacked 2D deconvolutional layers, and there are skip connections at every level between the identity encoder and each image decoder. The skip connection is implemented by concatenation. Then two extra 2D convolutional layers are added at the end of each decoder for spatial coherence. Finally, the τ-th image decoder takes in lip motion information at τ time $\widetilde{l}_\tau$ and generates lip image l_τ^* at τ time in corresponding shape.

3.3 Training Methods

In this section, we describe the loss function and training strategy to supervise ParaLip. The reconstruction loss and duration prediction loss endow the model with the fundamental ability to generate lip movement video. To generate the lip with better perceptual quality and alleviate the "blurry predictions" (Mathieu, Couprie, and LeCun 2016) problem, the structural similarity index loss and adversarial learning are introduced. We also explore the source-target alignment method when the audio is absent even in the training set, which will be introduced in Section 6.

Reconstruction Loss Basically, we optimize the whole network by adding L_1 reconstruction loss on generated lip sequence $\mathcal{L}^*$:

$$L_{rec} = \sum_{\tau=1}^{T} \|l_\tau - l_\tau^*\|_1. \tag{3}$$

Duration Prediction Loss In the training stage, we add L_1 loss on predicted duration sequence $\mathcal{D}^*$ at token level[2] and sequence level, which supervises the duration predictor to make the precise fine-grained and coarse-grained predictions. Duration prediction loss L_{dur} can be written as:

$$L_{dur} = \sum_{i=1}^{m} \|d_i - d_i^*\|_1 + \|\sum_{i=1}^{m} d_i - \sum_{i=1}^{m} d_i^*\|_1. \tag{4}$$

Structural Similarity Index Loss Structural Similarity Index (SSIM) (Wang 2004) is adopted to measure the perceptual image quality, which takes luminance, contrast and structure into account, and is close to the perception of human beings. The SSIM value for two pixels at position (i, j) in τ-th images l_τ^* and l_τ can be formulated as:

$$SSIM_{i,j,\tau} = \frac{2\mu_{l_\tau^*}\mu_{l_\tau} + C_1}{\mu_{l_\tau^*}^2 + \mu_{l_\tau}^2 + C_1} \cdot \frac{2\sigma_{l_\tau^* l_\tau} + C_2}{\sigma_{l_\tau^*}^2 + \sigma_{l_\tau}^2 + C_2},$$

where $\mu_{l_\tau^*}$ and μ_{l_τ} denotes the mean for regions in image l_τ^* and l_τ within a 2D-window surrounding (i, j). Similar, $\sigma_{l_\tau^*}$ and σ_{l_τ} are standard deviation; $\sigma_{l_\tau^* l_\tau}$ is the covariance; C_1 and C_2 are constant values. To improve the perceptual quality of the generated lip frames, we leverage SSIM loss in ParaLip. Assuming the size of each lip frame to be $(A \times B)$, the SSIM loss between generated $\mathcal{L}^*$ and ground truth $\mathcal{L}$ becomes:

$$L_{ssim} = \frac{1}{T \cdot A \cdot B} \sum_{\tau=1}^{T} \sum_{i}^{A} \sum_{j}^{B} (1 - SSIM_{i,j,\tau})). \tag{5}$$

Adversarial Learning Through experiments, it can be found that only using above losses is insufficient to generate distinct lip images with more realistic texture and local details (*e.g.*wrinkles, beard and teeth). Thus, we adopt adversarial learning to mitigate this problem and train a quality discriminator $Disc$ along with ParaLip. The $Disc$ contains stacked 2D convolutional layers with LeakyReLU activation which down-samples each image to $1 \times 1 \times H$ (H is hidden size), and a 1×1 convolutional layer to project the hidden states to a value of probability for judging real or fake. We use the loss function in LSGAN (Mao et al. 2017) to train ParaLip and $Disc$:

$$L_{adv}^G = \mathbb{E}_{x \sim l^*}(Disc(x) - 1)^2, \tag{6}$$

$$L_{adv}^D = \mathbb{E}_{x \sim l}(Disc(x) - 1)^2 + \mathbb{E}_{x \sim l^*} Disc(x)^2, \tag{7}$$

where l^* means lip images generated by ParaLip and l means ground truth lip images.

To summarize, we optimize the $Disc$ by minimizing Equation (7), and optimize the ParaLip by minimizing L_{total}:

$$L_{total} = \lambda_1 \cdot L_{rec} + \lambda_2 \cdot L_{dur} + \lambda_3 \cdot L_{ssim} + \lambda_4 \cdot L_{adv}^G, \tag{8}$$

where the λ_1, λ_2, λ_3 and λ_4 are hyperparameters to trade off the four losses.

[2]Character level for GRID and phoneme level for TCD-TIMIT following previous works.

4 Experimental Settings

4.1 Datasets

GRID The GRID dataset (Cooke et al. 2006) consists of 33 video-available speakers, and each speaker utters 1,000 phrases. The phrases are in a 6-categories structure following fixed simple grammar: $command^4 + color^4 + preposition^4 + letter^{25} + digit^{10} + adverb^4$ where the number denotes how many choices of each category. Thus, the total vocabulary size is 51, composing 64,000 possible phrases. All the videos last 3 seconds with frame rate 25 fps, which form a total duration of 27.5 hours. It is a typical talking face dataset and there are a considerable of lip-related works (Assael et al. 2016; Chung et al. 2017; Afouras, Chung, and Zisserman 2018; Chen et al. 2018; Zhu et al. 2020; Lin et al. 2021) conducting experiments on it. Following previous works, we select 255 random samples from each speaker to form the test set.

TCD-TIMIT The TCD-TIMIT dataset (Harte and Gillen 2015) is closer to real cases and more challenging than GRID dataset, since 1) the vocabulary is not limited; 2) the sequence length of videos is not fixed and is longer than that in GRID. We use the 'volunteers' subset of TCD-TIMIT following previous works, which consists of 59 speakers uttering about 98 sentences individually. The frame rate is 29.97 fps and each video lasts 2.5$\sim$8.1 seconds. The total duration is about 7.5 hours. We set 30% of data from each speaker aside for testing following the recommended speaker-dependent train-test splits (Harte and Gillen 2015).

4.2 Data Pre-processing

As for the video pre-processing, we utilize Dlib (King 2009) to detect 68 facial landmarks (including 20 mouth landmarks), and extract the face images from video frames. We resize the face images to 256×256, and further crop each face to a fixed 160×80 size containing the lip-centered region. As for the text pre-processing, we encode the text sequence at the character level for GRID dataset and phoneme level for TCD-TIMIT dataset. And for ground truth duration extraction, we first extract the speech audio from video files, and then utilize "Penn Phonetics Lab Forced Aligner" (P2FA) (Yuan and Liberman 2008) to get speech-to-text alignments, from which we obtain the duration of each text token for training our duration predictor in ParaLip.

5 Results and Analysis

In this section, we present extensive experimental results to evaluate the performance of *ParaLip* in terms of lip movements quality and inference speedup. And then, we conduct ablation experiments to verify the significance of all proposed methods in *ParaLip*.

5.1 Quality Comparison

We compare our model with 1) *DualLip* (Chen et al. 2020), which is the state-of-the-art (SOTA) autoregressive text-to-lip model based on RNN and location-sensitive attention (Shen et al. 2018). And 2) *TransformerT2L*, an autoregressive baseline model based on Transformer (Vaswani et al. 2017) implemented by us, which uses the same model settings with *ParaLip*[3]. The quantitative results on GRID and TCD-TIMIT are listed in Table 1 and Table 2 respectively[4]. Note that we do not add adversarial learning on any model in Table 1 or Table 2, since there is no adversarial learning in *DualLip* (Chen et al. 2020).

Methods	PSNR ↑	SSIM ↑	LMD ↓
AR Benchmarks			
DualLip	**29.13**†	0.872†	1.809 †
TransformerT2L	26.85	0.829	1.980
Our Model			
ParaLip	28.74	**0.875**	**1.675**

Table 1: Comparison with Autoregressive Benchmarks on GRID dataset. † denotes our reproduction under the case w/o GT duration at inference.

Methods	PSNR ↑	SSIM ↑	LMD ↓
AR Benchmarks			
DualLip	27.38†	0.809†	2.351†
TransformerT2L	26.89	0.794	2.763
Our Model			
ParaLip	**27.64**	**0.816**	**2.084**

Table 2: Comparison with Autoregressive Benchmarks on TCD-TIMIT dataset. † denotes our reproduction under the case w/o GT duration at inference.

Quantitative Comparison We can see that: 1) On GRID dataset (Table 1), *ParaLip* outperforms *DualLip* on LMD metric, and keeps the same performance in terms of PSNR and SSIM metrics. However, on TCD-TIMIT dataset (Table 2), *ParaLip* achieves an overall performance surpassing *DualLip* by a notable margin, since autoregressive models perform badly on the long-sequence dataset due to accumulated prediction error; 2) *ParaLip* shows absolute superiority over AR baseline *TransformerT2L* in terms of three quantitative metrics on both datasets; 3) although *DualLip* outperforms AR baseline by incorporating the technique of location-sensitive attention, which could alleviate the error propagation, it is still vulnerable on long-sequence dataset.

Qualitative Comparison We further visualize the qualitative comparison between *DualLip*, *TransformerT2L* and *ParaLip* in Figure 3. It can be seen that the quality of lip frames generated by *DualLip* and *TransformerT2L* become

[3]Most modules and the total number of model parameters in *ParaLip* and *TransformerT2L* are similar.

[4]Note that the reported results are all under the case where the ground truth (GT) duration is not provided at inference (denoted as w/o duration in (Chen et al. 2020)) , since there is no GT duration available in the real case.

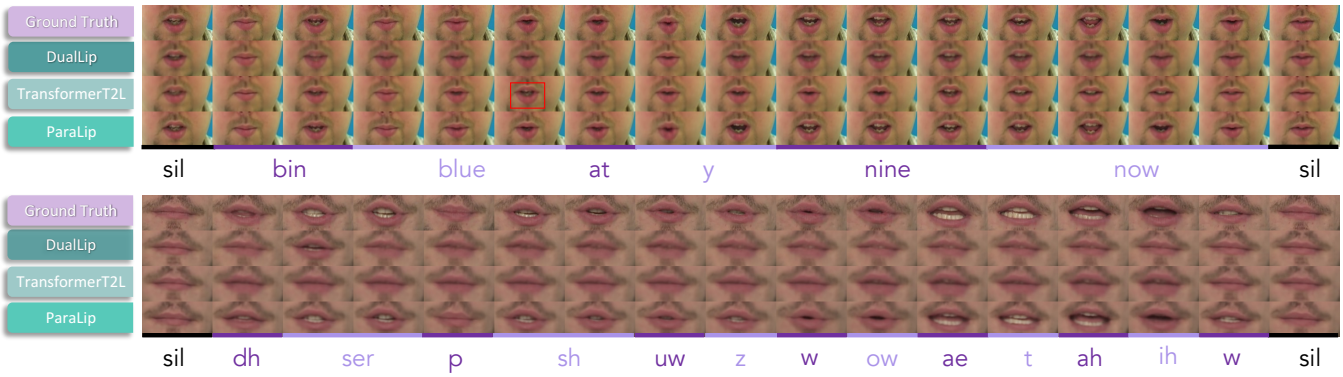


Figure 3: The qualitative comparison among AR SOTA (*DualLip*), AR baseline (*TransformerT2L*) and our NAR method (*ParaLip*). We visualize two cases from GRID dataset and TCD-TIMIT dataset to illustrate the error propagation problem existing in AR generation and verify the robustness of *ParaLip*. In the first case, the lip sequence generated from AR baseline predicts a wrong lip image (the 6-th frame with red box), and as a result, the subsequent lip images conditioned on that image becomes out of synchronization with linguistic information and ends in chaos; *DualLip* alleviates the error propagation to some degree. In the second case, both AR models perform poorly on the long-sequence dataset and generate the frames that look speechless as the time goes further.

increasingly worse as the time goes further. Concretely, The lip image becomes fuzzy and out of synchronization with linguistic contents. We attribute this phenomenon to the reason that: error propagation problem is serious in AR T2L since the wrong prediction could take place at more dimensions (every pixel with three channels in generated image) and there is information loss during the down-sampling when sending the last generated lip frame to predict current one. What's worse, on TCD-TIMIT, a long-sequence dataset, *DualLip* and *TransformerT2L* often generate totally unsatisfying results that look like speechless video. By contrast, the lip frames generated by NAR model *ParaLip* maintain high fidelity to ground truth all the while, which demonstrates the effectiveness and robustness of NAR decoding.

5.2 Speed Comparison

In this section, we evaluate and compare the average inference latency of *DualLip*, *TransformerT2L* and *ParaLip* on both datasets. Furthermore, we study the relationship between inference latency and the target video length.

Comparison of Average Inference Latency The average inference latency is the average time consumed to generate one video sample on the test set, which is measured in seconds. Table 3 exhibits the inference latency of all systems. It can be found that, 1) compared with *DualLip*, *ParaLip* speeds up the inference by 13.09× and 19.12× on average on two datasets; 2) *TransformerT2L* has the same structure with *ParaLip*, but runs about 50% slower than *DualLip*, which indicates that it is NAR decoding manner in ParaLip speedups the inference, rather than the modification of model structure; 3) In regard to AR models, the time consumption of a single sentence increases to 0.5-1.5 seconds even on GPU, which is unacceptable for real-world application. By contrast, *ParaLip* addresses the inference latency problem satisfactorily.

Datasets	Methods	Latency (s)	Speedup
GRID	*DualLip*	0.299	1.00 ×
	TransformerT2L	0.689	0.43 ×
	ParaLip	0.022	13.09 ×
TIMIT	*DualLip*	0.650	1.00 ×
	TransformerT2L	1.278	0.51 ×
	ParaLip	0.034	19.12 ×

Table 3: The comparison of inference latency on GRID and TCD-TIMIT dataset. The computations are conducted on a server with 1 NVIDIA 2080Ti GPU.

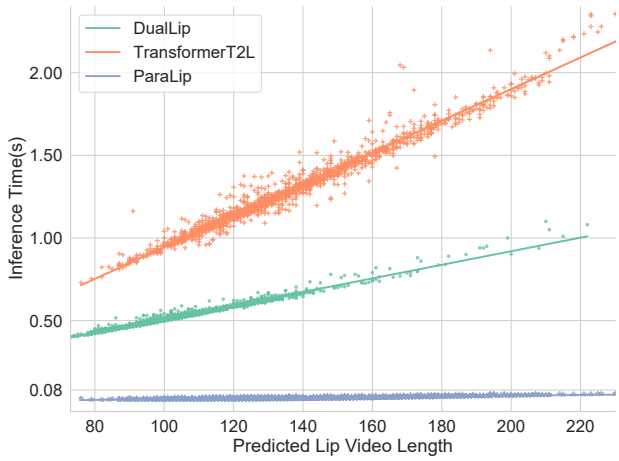


Figure 4: Relationship between inference latency (seconds) and predicted video length for *DualLip*, *TransformerT2L* and *ParaLip*.

Relationship between Inference Latency and Video Length In this section, we study the speedup as the sequence length increases. The experiment is conducted on TCD-TIMIT, since its videos are not in a fixed length. From Figure 4, it can be seen that 1) *ParaLip* model speeds up the inference obviously due to high parallelization compared with AR models; 2) *ParaLip* is insensitive to sequence length and almost holds a constant inference latency, but by contrast, the inference latency of *DualLip* and *TransformerT2L* increase linearly as the sequence length increases. As a result, the speedup of *ParaLip* relative to *DualLip* or *TransformerT2L* also increases linearly as the sequence length increases.

5.3 Ablation Study

Model	PSNR ↑	SSIM ↑	LMD ↓	FID ↓
Base model	30.24	0.896	0.998	56.36
+SSIM	30.51	0.906	0.978	55.05
+ADV	25.70	0.736	2.460	65.88
+SSIM+ADV	28.36	0.873	1.077	39.74

Table 4: The ablation studies on GRID dataset. Base model is trained only with L_1 loss; "+SSIM" means adding structural similarity index loss and "+ADV" means adding adversarial learning to the base model. FID means Fréchet Inception Distance metric. To focus on the frames quality, we provide the GT duration for eliminating the interference caused by the discrepancy of predicted length.

We conduct ablation experiments on GRID dataset to analyze the effectiveness of the proposed methods in our work. All the results are shown in Table 4. Experiments show that:

- Adding only SSIM loss obtains the optimal score on PSNR/SSIM/LMD ("*+SSIM*");
- Adding only adversarial training causes performance drop on PSNR/SSIM/LMD, which is consistent with previous works (Song et al. 2019) ("*+ADV*");
- Adding SSIM to model with adversarial training can greatly alleviate the detriment on PSNR/SSIM/LMD brought by adversarial training; make the GAN-based model more stable; obtain the best FID score, which means the generated lips look more realistic. ("*+SSIM+ADV*").

Previous works (Song et al. 2019) claim that 1) PSNR and SSIM cannot well reflect some visual quality; 2) adversarial learning encourages the generated face to pronounce in diverse ways, leading to diverse lip movements and LMD decrease. Although "*+SSIM+ADV*" causes marginally PSNR/SSIM/LMD scores losses, "*+SSIM+ADV*" obtain the best FID score and tends to generate distinct lip images with more realistic texture and local details (e.g. wrinkles, beard and teeth). The qualitative results are shown in Figure 5.

6 Further Discussions

In the foregoing sections, we train the duration predictor using the "GT" duration extracted by P2FA, but it is not applicable to the case where the audio is absent even in the training set. Thus, to obtain the "GT" duration in this case, we tried a lipreading model with monotonic alignment searching (MAS) (Tillmann et al. 1997; Kim et al. 2020) to find the alignment between text and lip frames. Specifically, we 1) first trained a lipreading model by CTC loss on the training set; 2) traveled the training set and for each ($\mathcal{L}$, $\mathcal{S}$) pair, we extracted $O_{ctc} \in R^m$ (the CTC outputs corresponding to the label tokens) from $O_{ctc} \in R^V$ (O_{ctc} is the original CTC outputs; V is the vocabulary size); 3) applied softmax on the extracted O_{ctc} to obtain the probability matrix $P(A(s_i, l_j)) = \frac{e^{O^i_{ctc,lj}}}{\sum_m e^{O_{ctc,lj}}}$; 4) conducted MAS by dynamic programming to find the best alignment solution. This method achieves similar results with P2FA on GRID dataset, but could cause deterioration on TIMIT dataset: PSNR:27.09, SSIM:0.816, LMD:2.313.

Theoretically, obtaining the alignment between lip frames and its transcript directly from themselves has more potential than obtaining this alignment indirectly from audio sample and its transcripts, since there are many cases when the mouth is moving but the sound hasn't come out yet (e.g. the first milliseconds of a speech sentence). This part of video frames should be aligned to some words, but its corresponding audio piece is silent, which will not be aligned to any word, causing contradictions. We think it is valuable to try more and better methods in this direction.

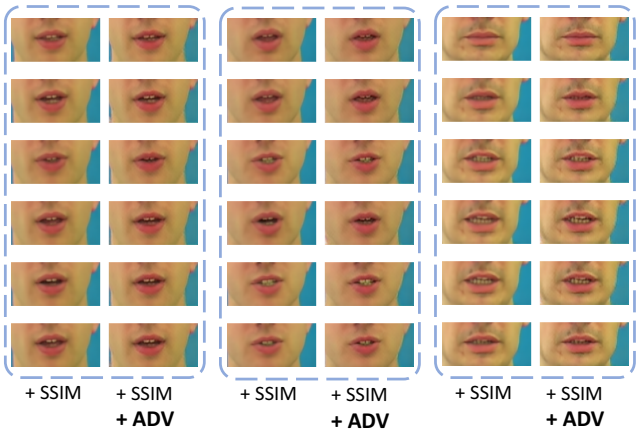


Figure 5: The qualitative evaluation for adversarial learning. "+ADV" tends to generate more realistic lip images.

7 Conclusion

In this work, we point out and analyze the unacceptable inference latency and intractable error propagation existing in AR T2L generation, and propose a parallel decoding model ParaLip to circumvent these problems. Extensive experiments show that ParaLip generates lip movements with competitive quality compared with the state-of-the-art AR T2L model, exceeds the baseline AR model TransformerT2L by a notable margin and exhibits distinct superiority in inference speed, which provides the possibility to bring T2L generation from laboratory to industrial applications.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant No.2020YFC0832505, No.62072397, Zhejiang Natural Science Foundation under Grant LR19F020006.

References

Abdelaziz, A. H.; Kumar, A. P.; Seivwright, C.; Fanelli, G.; Binder, J.; Stylianou, Y.; and Kajarekar, S. 2020. Audio-visual Speech Synthesis using Tacotron2. *arXiv preprint arXiv:2008.00620*.

Afouras, T.; Chung, J. S.; and Zisserman, A. 2018. Deep Lip Reading: A Comparison of Models and an Online Application. In *Proc. Interspeech 2018*, 3514–3518.

Assael, Y. M.; Shillingford, B.; Whiteson, S.; and De Freitas, N. 2016. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*.

Bengio, S.; Vinyals, O.; Jaitly, N.; and Shazeer, N. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, 1171–1179.

Chen, L.; Li, Z.; K Maddox, R.; Duan, Z.; and Xu, C. 2018. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 520–535.

Chen, N.; Watanabe, S.; Villalba, J.; and Dehak, N. 2019. Non-Autoregressive Transformer Automatic Speech Recognition. *arXiv preprint arXiv:1911.04908*.

Chen, W.; Tan, X.; Xia, Y.; Qin, T.; Wang, Y.; and Liu, T.-Y. 2020. DualLip: A System for Joint Lip Reading and Generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 1985–1993. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

Chung, J.; Jamaludin, A.; and Zisserman, A. 2017. You said that? *British Machine Vision Conference 2017, BMVC 2017*.

Chung, J. S.; Senior, A.; Vinyals, O.; and Zisserman, A. 2017. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3444–3453. IEEE.

Cooke, M.; Barker, J.; Cunningham, S.; and Shao, X. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America*, 120(5): 2421.

Deng, C.; Ding, N.; Tan, M.; and Wu, Q. 2020. Length-Controllable Image Captioning. *ECCV*.

Fan, B.; Wang, L.; Soong, F. K.; and Xie, L. 2015. Photo-real talking head with deep bidirectional LSTM. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4884–4888. IEEE.

Fried, O.; Tewari, A.; Zollhöfer, M.; Finkelstein, A.; Shechtman, E.; Goldman, D. B.; Genova, K.; Jin, Z.; Theobalt, C.; and Agrawala, M. 2019. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)*, 38(4): 1–14.

Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and Dauphin, Y. N. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1243–1252.

Ghazvininejad, M.; Levy, O.; Liu, Y.; and Zettlemoyer, L. 2019. Mask-Predict: Parallel Decoding of Conditional Masked Language Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6114–6123.

Gu, J.; Bradbury, J.; Xiong, C.; Li, V. O.; and Socher, R. 2018. Non-Autoregressive Neural Machine Translation. In *International Conference on Learning Representations*.

Harte, N.; and Gillen, E. 2015. TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech. *IEEE Transactions on Multimedia*, 17(5): 603–615.

Higuchi, Y.; Watanabe, S.; Chen, N.; Ogawa, T.; and Kobayashi, T. 2020. Mask CTC: Non-Autoregressive End-to-End ASR with CTC and Mask Predict. *INTERSPEECH*.

Kim, J.; Kim, S.; Kong, J.; and Yoon, S. 2020. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. *Advances in Neural Information Processing Systems*, 33.

King, D. E. 2009. *Dlib-ml: A Machine Learning Toolkit*. JMLR.org.

KR, P.; Mukhopadhyay, R.; Philip, J.; Jha, A.; Namboodiri, V.; and Jawahar, C. 2019. Towards Automatic Face-to-Face Translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1428–1436.

Kumar, R.; Sotelo, J.; Kumar, K.; de Brébisson, A.; and Bengio, Y. 2017. Obamanet: Photo-realistic lip-sync from text. *arXiv preprint arXiv:1801.01442*.

Lee, J.; Mansimov, E.; and Cho, K. 2018. Deterministic Non-Autoregressive Neural Sequence Modeling by Iterative Refinement. In *EMNLP*, 1173–1182.

Lin, Z.; Zhao, Z.; Li, H.; Liu, J.; Zhang, M.; Zeng, X.; and He, X. 2021. *SimulLR: Simultaneous Lip Reading Transducer with Attention-Guided Adaptive Memory*, 1359–1367. New York, NY, USA: Association for Computing Machinery. ISBN 9781450386517.

Liu, J.; Ren, Y.; Zhao, Z.; Zhang, C.; Huai, B.; and Yuan, J. 2020. FastLR: Non-Autoregressive Lipreading Model with Integrate-and-Fire. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4328–4336.

Ma, X.; Zhou, C.; Li, X.; Neubig, G.; and Hovy, E. 2019. FlowSeq: Non-Autoregressive Conditional Sequence Generation with Generative Flow. In *EMNLP-IJCNLP*, 4273–4283.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2794–2802.

Mathieu, M.; Couprie, C.; and LeCun, Y. 2016. Deep multi-scale video prediction beyond mean square error. In *4th International Conference on Learning Representations, ICLR 2016*.

Miao, C.; Liang, S.; Chen, M.; Ma, J.; Wang, S.; and Xiao, J. 2020. Flow-TTS: A Non-Autoregressive Network for Text to Speech Based on Flow. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7209–7213. IEEE.

Peng, K.; Ping, W.; Song, Z.; and Zhao, K. 2020. Non-Autoregressive Neural Text-to-Speech. ICML.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, 484–492.

Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2019. Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems*, 3165–3174.

Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4779–4783. IEEE.

Song, Y.; Zhu, J.; Li, D.; Wang, A.; and Qi, H. 2019. Talking Face Generation by Conditional Recurrent Adversarial Network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 919–925. International Joint Conferences on Artificial Intelligence Organization.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Suwajanakorn, S.; Seitz, S. M.; and Kemelmacher-Shlizerman, I. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4): 1–13.

Tillmann, C.; Vogel, S.; Ney, H.; and Zubiaga, A. 1997. A DP-based search using monotone alignments in statistical translation. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, 289–296.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Vougioukas, K.; Petridis, S.; and Pantic, M. 2019. End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs. In *CVPR Workshops*, 37–40.

Wang, L.; Han, W.; Soong, F. K.; and Huo, Q. 2011. Text driven 3D photo-realistic talking head. In *Twelfth Annual Conference of the International Speech Communication Association*.

Wang, Z. 2004. Image Quality Assessment : From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*.

Wu, L.; Tan, X.; He, D.; Tian, F.; Qin, T.; Lai, J.; and Liu, T.-Y. 2018. Beyond Error Propagation in Neural Machine Translation: Characteristics of Language Also Matter. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3602–3611. Brussels, Belgium: Association for Computational Linguistics.

Yu, C.; Lu, H.; Hu, N.; Yu, M.; Weng, C.; Xu, K.; Liu, P.; Tuo, D.; Kang, S.; Lei, G.; et al. 2019. Durian: Duration informed attention network for multimodal synthesis. *arXiv preprint arXiv:1909.01700*.

Yu, L.; Yu, J.; and Ling, Q. 2019. Mining audio, text and visual information for talking face generation. In *2019 IEEE International Conference on Data Mining (ICDM)*, 787–795. IEEE.

Yuan, J.; and Liberman, M. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, 123(5): 3878.

Zheng, R.; Zhu, Z.; Song, B.; and Ji, C. 2020. Photorealistic Lip Sync with Adversarial Temporal Convolutional Networks. *arXiv preprint arXiv:2002.08700*.

Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; and Wang, X. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9299–9306.

Zhu, H.; Huang, H.; Li, Y.; Zheng, A.; and He, R. 2020. Arbitrary Talking Face Generation via Attentional Audio-Visual Coherence Learning. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 2362–2368. International Joint Conferences on Artificial Intelligence Organization. Main track.