Flow-Based Unconstrained Lip to Speech Generation

Jinzheng He[1], Zhou Zhao*[1], Yi Ren[1]
Jinglin Liu[1], Baoxing Huai[2], Nicholas Yuan[2],
[1]Zhejiang University, China
[2]Huawei Cloud
{3170106086,zhaozhou,rayeren,jinglinliu}@zju.edu.cn,
{huaibaoxing,nicholas.yuan}@huawei.com

Abstract

Unconstrained lip-to-speech aims to generate corresponding speeches based on silent facial videos with no restriction to head pose or vocabulary. It is desirable to generate intelligible and natural speech with a fast speed in unconstrained settings. Currently, to handle the more complicated scenarios, most existing methods adopt the autoregressive architecture, which is optimized with the MSE loss. Although these methods have achieved promising performance, they are prone to bring issues including high inference latency and mel-spectrogram over-smoothness. To tackle these problems, we propose a novel flow-based non-autoregressive lip-to-speech model (GlowLTS) to break autoregressive constraints and achieve faster inference. Concretely, we adopt a flow-based decoder which is optimized by maximizing the likelihood of the training data and is capable of more natural and fast speech generation. Moreover, we devise a condition module to improve the intelligibility of generated speech. We demonstrate the superiority of our proposed method through objective and subjective evaluation on Lip2Wav-Chemistry-Lectures and Lip2Wav-Chess-Analysis datasets. Our demo video can be found at https://glowlts.github.io/.

Introduction

Given a silent facial video, lip-to-speech aims to generate corresponding speech. This technology can be widely used in a variety of applications such as video conferencing in silent or noisy environments (Vougioukas et al. 2019), long-range listening for surveillance (Ephrat, Halperin, and Peleg 2017) and artificial voice aid for people suffering from aphonia (Mira et al. 2021).

Currently, most existing lip-to-speech methods (Ephrat and Peleg 2017; Kumar et al. 2019b; Akbari et al. 2018; Vougioukas et al. 2019) are proposed to explore constrained lip-to-speech, where videos are collected in an artificially constrained environment with nearly no head motion and speeches contain a narrow vocabulary, for example, only 56 tokens in GRID corpus (Cooke et al. 2006). Unconstrained lip-to-speech, however, uses real lecture videos, which contain observable head motion (see Figure 1) and nearly 100x larger vocabulary, which puts forward higher requirements for the modeling power of the lip-to-speech model: previous work (Prajwal et al. 2020) investigates the problem using the autoregressive sequence-to-sequence architecture, and demonstrates that such architecture can generate more intelligible speech compared with other previous lip-to-speech methods.

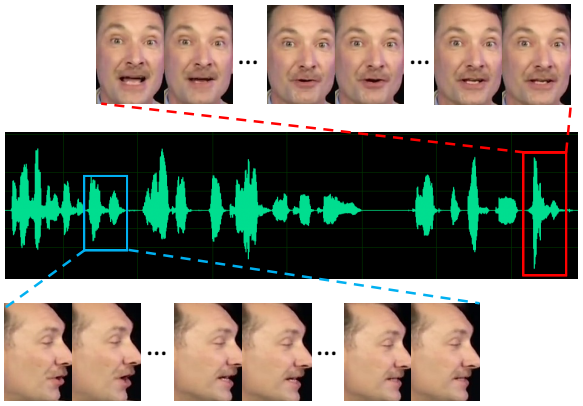


Figure 1: An example of video and audio in unconstrained datasets. In the blue box, the audio pronounces "process", while in the red box, the audio pronounces "hybrid". Large head movements exist in unconstrained settings.

In unconstrained lip-to-speech, naturalness and intelligibility are the crucial parts of speech quality. Moreover, the inference speed is also worthy of consideration (e.g., in some applications like online video conferencing, the speech generation latency can greatly affect the user experience). Current method (Prajwal et al. 2020) designed for unconstrained lip-to-speech faces several challenges. 1) Big gap in naturalness between generated speech and realistic speech: the existing method in unconstrained lip-to-speech adopts the MSE criterion in predicting each spectrogram frame. Such design can not capture the correlation among frequency bins in a frame, which leads to over-smoothness in spectrogram (Sheng and Pavlovskiy 2019). 2) High inference latency: the existing method utilizes the autoregressive architecture, generating current frames conditioned on previous ones. Such architecture suffers from a low inference speed (Gu et al. 2017; Chen et al. 2019).

*Corresponding author

To tackle these problems, we develop a flow-based non-autoregressive lip-to-speech generation model for unconstrained settings. We adopt generative flow (Glow) (Kingma and Dhariwal 2018) as the main part of the decoder. As a powerful non-autoregressive generative architecture, Glow has been proved to be efficient in modeling images and waveforms (Prenger, Valle, and Catanzaro 2019) by simply maximizing the likelihood of the training data. During the training of our flow-based decoder, the frequency channels are closely connected through affine coupling layers, which contributes to modeling correlation among frequency bins and results in an improvement in speech naturalness. However, we find that the normalizing flow-based decoder sometimes focuses more on local details in mel-spectrogram (e.g., high-frequency part reconstruction) but less on global semantic information (e.g., word-level pronunciation), which is consistent with the findings described in Kirichenko, Izmailov, and Wilson (2020) that normalizing flow tends to focus on pixel-level local correlations. This issue can be more severe in the unconstrained lip-to-speech task since the visual inputs can not provide enough certain information about the pronunciation, leading to a drop in speech intelligibility.

To solve this problem, we propose a condition module to generate coarse but more intelligible speech based on aligned visual features. Then the flow-based decoder is utilized to generate more realistic speech conditioned on the coarse speech. Such design improves the intelligibility of generated speech. Through extensive experiments, we observe an improvement in both naturalness and intelligibility. Our key contributions are as follow:

- We propose GlowLTS, the first flow-based non-autoregressive lip-to-speech method for the challenging unconstrained settings, to generate natural speech with low inference latency.
- We propose a condition module to generate coarse but intelligible speech as the condition of our proposed flow-based decoder, which observably improves the intelligibility of generated speech.
- GlowLTS can generate more intelligible and much more natural speech with a speed of 5.377x faster than the current state-of-the-art model.

The rest of the paper is organized as follows: First, we survey recent progress in constrained lip-to-speech, unconstrained lip-to-speech, and flow-based methods. Then we introduce our proposed GlowLTS in detail. Next, we introduce our experiment settings, report detailed results and conduct analyses. Finally, we conclude this paper.

Related Work

Constrained Lip-to-speech

Constrained lip-to-speech generates speech based on given facial videos from small datasets (Cooke et al. 2006) with narrow vocabulary speech in artificially constrained environments. Early approaches (Ephrat and Peleg 2017; Kumar et al. 2019b) adopt an end-to-end CNN-based method and utilize low-dimensional LPC (Linear Predictive Coding) features as regression targets. However, the low-dimensional LPC features contain insufficient speech information to be converted back to natural speech waveforms. More recently, methods proposed in Ephrat, Halperin, and Peleg (2017); Akbari et al. (2018) adopt the high-dimensional spectrograms as the training targets and utilize vocoders to convert spectrograms back to waveforms. Other methods proposed in Vougioukas et al. (2019); Mira et al. (2021) introduce GANs (Goodfellow et al. 2014) into lip-to-speech and struggle to generate the waveform directly. Our work focuses on unconstrained lip-to-speech with no limitation on head movements and vocabulary.

Unconstrained Lip-to-speech

Unconstrained lip-to-speech is a more challenging task as it aims to generate large vocabulary speech based on real-world videos. To the best of our knowledge, only one prominent work, named Lip2Wav (Prajwal et al. 2020), exists in the current literature. Lip2Wav uses a modified version of Tacotron2 (Shen et al. 2018) model designed for lip-to-speech. Like Tacotron2, Lip2Wav adopts an autoregressive structure and takes mel-spectrogram as the generation target. While it has achieved decent intelligibility, Lip2Wav adopts the MSE loss in predicting each mel-spectrogram frame, which ignores the frequency correlation in each frame and leads to mel-spectrogram over-smoothness. In addition, the inference speed of Lip2Wav is also slow due to the adoption of autoregressive architecture. Our method adopts the flow-based non-autoregressive architecture, which can model the frequency correlation effectively and achieve parallel generation.

Flow-based Generation Architecture

The flow generation model (Dinh, Krueger, and Bengio 2014; Kingma and Dhariwal 2018) is very attractive because of the tractability of the exact log-likelihood. It incorporates a stack of invertible transformations, converting the simple Gaussian distribution to a more complex distribution. The powerful distribution fitting ability of flow models has encouraged many studies in different generation areas. Videoflow (Kumar et al. 2019a) incorporates flow architecture for predicting future video frames. C-flow (Pumarola et al. 2020) brings flow architecture to the field of 3D point clouds. Methods in Miao et al. (2020); Kim et al. (2020) use flow-based architecture in text-to-speech generation. Methods proposed in Prenger, Valle, and Catanzaro (2019); Kim et al. (2018) incorporate flow architecture in the vocoder area and observe an improvement. However, to the best of our knowledge, flow-based methods have not been used in unconstrained lip-to-speech. In this paper, we propose the first flow-based unconstrained lip-to-speech method to model the frequency correlation and generate more natural speech based on lip motions.

Our Method

Problem Definition

In this section, we consider the task of the unconstrained lip-to-speech generation. Given an unconstrained facial video $V = \{v_1, v_2, ..., v_M\}$, where M is the total number of video

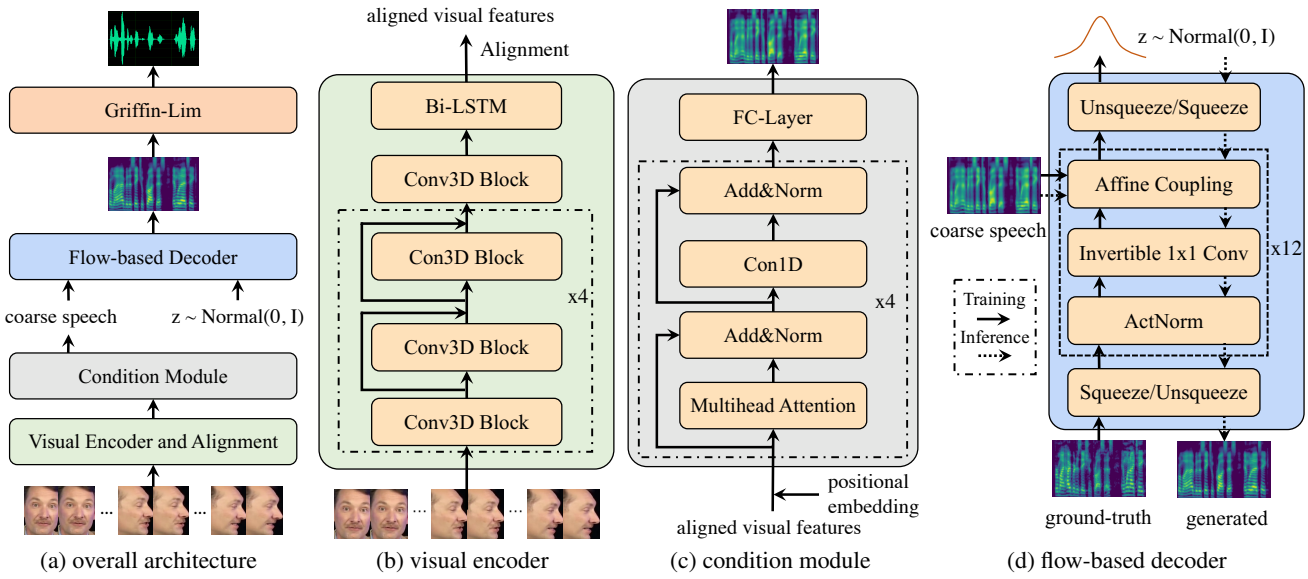


Figure 2: The overall architecture of GlowLTS. In subfigure (b), the visual encoder extracts visual features from the facial video. Visual features are aligned based on the alignment. In subfigure (c), the condition module generate coarse speech. In subfigure (d), the flow-based decoder generate more natural speech conditioned on the coarse speech.

frames, and v_i represents i_{th} video frame in this video. Unconstrained lip-to-speech aims to generate the corresponding mel-spectrogram $A = \{a_1, a_2, ..., a_N\}$ where N is the total number of mel-spectrogram frames, and a_j is the j_{th} frame. In our case, we try to achieve a non-autoregressive unconstrained lip-to-speech synthesis, which means $A = \{a_1, a_2, ..., a_N\}$ is generated in parallel.

Model Overview

In this section, we introduce the overall architecture of our proposed GlowLTS. As shown in Figure 2(a), GlowLTS is built from the following blocks: the visual encoder, the condition module, and the flow-based decoder. Visual features are first extracted through the visual encoder. Due to the inconsistency in length between video and mel-spectrogram frames, we expand these visual features to the length of the corresponding mel-spectrogram according to the temporal synchronization of video and audio. Next, we feed the aligned visual features to the condition module to generate a coarse but intelligible mel-spectrogram. Finally, we provide the coarse mel-spectrogram as a condition to the flow-based decoder to generate a more natural and realistic mel-spectrogram. We convert the mel-spectrogram back to the waveform through the Griffin-Lim algorithm (Griffin and Lim 1984) for a fair comparison with other methods.

Visual Encoder and Alignment

In this section, we introduce the visual encoder and alignment method used in our method. The visual encoder (Figure 2(b)) is adopted to extract semantic information from facal videos. Given a facial video $V = \{v_1, v_2, ..., v_M\}$, where v_i is of size $H \times W \times 3$, we use a stack of 3D-CNN (Ji et al. 2012) blocks with batch normalization and relu activation to downsample each of the video frame v_i to a D-dimensional vector. The 3D-CNN blocks can leverage the neighboring contexts and contribute to reducing the homophenes. The result features $F = \{f_1, f_2, ..., f_M\}$, where $f_i \sim \mathcal{R}^D$, are then fed into a bidirectional LSTM (Hochreiter and Schmidhuber 1997) to leverage the long-range contexts.

On top of the visual encoder, we utilize a simple alignment method to map M video features to N expanded video features through directly repeating. The alignment in lip-to-speech is unique and exact, owing to the temporal synchronization between the video stream and the audio stream. Given M video frames, N mel-spectrogram frames and $M < N$, if N is M-divisible, the alignment is $\{N/M, N/M, ...\}$. If not, we try to ensure the temporal synchronization of audio and video to the maximum extent. For example, if we are given 240 mel-spectrogram frames and 90 video frames, we set the alignment as $\{3, 3, 2, 3, 3, 2, 3, 3, 2, ...\}$.

Condition Module

In this section, we introduce the proposed condition module which adopts the non-autoregressive architecture and aims to generate intelligible speech as a condition for the flow-based decoder. Concretely, as shown in Figure 2(c), we feed the aligned visual features to a stack of feed-forward transformers (Ren et al. 2019) with layer normalization and multi-head attention mechanism (Vaswani et al. 2017). These feed-forward transformers contribute to better utilization of contexts and alleviating ambiguity. A fully connected layer is designed to linearly project the feed-forward transformer output to the same channel number of

mel-spectrogram, namely 80 in our settings. Then we use MSE loss $\mathcal{L}_{mse}$ to constrain the condition module. $\mathcal{L}_{mse}$ is defined as:

$$\mathcal{L}_{mse} = ||cond - y||^2, \tag{1}$$

where $cond$ is the output of the condition module, and y is the ground-truth mel-spectrogram.

The adoption of MSE loss provides more direct supervision for extracting semantic information from facial videos.

Flow-based Decoder

In this section, we propose the flow-based decoder (Figure 2(d)) which aims to convert the coarse spectrogram $cond$ generated by the condition module to more detailed speech spectrogram y. We parameterize the conditional distribution $P(y|cond)$ by using an invertible neural network f_θ, which is defined as:

$$\begin{aligned} z &= f_\theta(y; cond), \\ y &= f_\theta^{-1}(z; cond), \end{aligned} \tag{2}$$

where $z \sim \mathcal{N}(0, I)$. The probability density $p_{y|cond}$ can be explicitly calculated as:

$$p(y|cond, \theta) = p_z(f_\theta(y; cond)))|det\frac{\partial f_\theta}{\partial y}(y; cond)|. \tag{3}$$

With the explicit probability density, we can train the networks by minimizing the negative log-likelihood which is calculated as:

$$\begin{aligned} \mathcal{L}_{mle} &= -logp(y|cond, \theta) \\ &= -logp_z(f_\theta(y; cond))) - log|det\frac{\partial f_\theta}{\partial y}(y; cond)|. \end{aligned} \tag{4}$$

In the rest of this section, we introduce the key designs of our flow-based decoder. On the whole, we follow the scheme of squeeze, actnorm, invertible 1x1 conv, affine coupling, and unsqueeze. For the brevity of introduction, we give the following definition: x and y represent the input and output of each module during the training process separately, and their size is $t \times c$, where t is the temporal dimension, and c is the frequency channel dimension.

Squeeze and unsqueeze As utilized in Kim et al. (2020), in the squeeze layer, we split 80-channel mel-spectrogram frames into two halves along the temporal dimension and group them into one 160-channel feature map. The unsqueeze layer is the inverse operation of the squeeze layer to restore the mel-spectrogram shape.

Actnorm The actnorm layer is designed as a replacement for batch normalization to train deep models. We use the original design proposed in Kingma and Dhariwal (2018). The Jacobian log-determinant of actnorm layer is calculated as $sum(\log(|s|)) \times t$, where s is the scale.

Invertible 1x1 conv This invertible 1x1 conv layer is firstly proposed in Kingma and Dhariwal (2018) and used as a generalization of the permutation operation. It reweights each channel to achieve channel information fusion. In our model, following Kim et al. (2020), we split input channels into 40 groups. Then the weight W is defined as a $\frac{c}{40} \times \frac{c}{40}$ matrix. The transformation is then defined as:

$$y_i = Wx_i, \tag{5}$$

where W is a $\frac{c}{40} \times \frac{c}{40}$ weigth matrix, and x_i, y_i represent i_{th} group of x, y. The Jacobian log-determinant is calculated as $40 \times \log(|det(W)|) \times t$.

Affine coupling The affine coupling layer (Dinh, Sohl-Dickstein, and Bengio 2016) is utilized to implement an invertible neural network. In our case, we utilize a conditional setting:

$$\begin{aligned} y_b &= x_b, \\ (\log sc, tc) &= func(x_b, cond), \\ y_a &= sc \cdot x_a + tc, \end{aligned} \tag{6}$$

where $x = (x_a, x_b)$ is a partition in the channel dimension. $cond$ is the output of our proposed condition module and is utilized as the guidance for speech generation. Here $func$ can be any transformation. The affine coupling layer preserves invertibility for the overall network, even though $func$ does not need to be invertible. In our case, following Prenger, Valle, and Catanzaro (2019), $func$ uses layers of dilated convolutions with gated-tanh nonlinearities, as well as residual connections and skip connections. The corresponding Jacobian log-determinant is simply computed as $sum(\log(|sc|))$.

From Equation (6), one frequency channel partition y_a is closely connected with the other channel partition x_b during the training of the affine coupling layer, which contributes to learning the correlation among frequency channels. Therefore, the proposed flow-based decoder can greatly improve the naturalness.

Training Details

In our experiment, we adopt a two-stage training methodology. In the first stage, we only train the visual encoder and the condition module via loss $\mathcal{L}_{mse}$. After training, the condition module can output coarse but intelligible mel-spectrogram. In the second stage, with the pretrained condition module and visual encoder, we add $\mathcal{L}_{mle}$ to train the flow-based decoder:

$$\begin{aligned} \mathcal{L}_{mle} = 0.5 \times (\log(2\pi) + \frac{||z||^2}{t \times c}) \\ - \frac{\sum_i^{Act} sum(\log(|s_i|))}{c} - \frac{\sum_k^{Couple} sum(\log(|sc_k|))}{t \times c} \\ - \frac{\sum_j^{Conv} \log(|det(W_j)|) \times 40}{c}, \end{aligned}$$

where z represents the output of the flow training process. Act, $Couple$, and $Conv$ represent all actnorm, affine coupling and invertible 1x1 conv layers respectively. The first term in $\mathcal{L}_{mle}$ comes from the log-likelihood of a spherical Gaussian. And the remaining terms account for the Jacobian log-determinant of actnorm, affine coupling and invertible 1x1 conv layers.

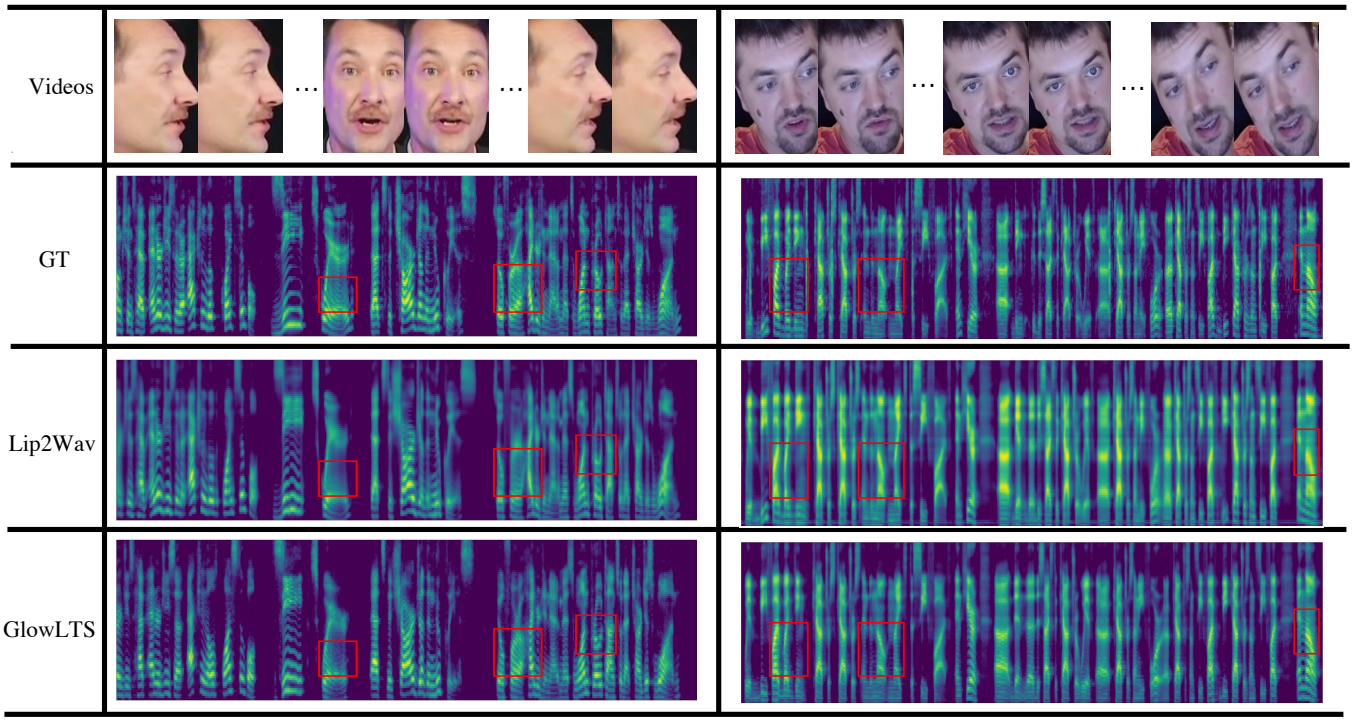


Figure 3: Visualization of the input videos with the corresponding mel-spectrogram. Input videos are visualized in the first column. Ground-truth (GT) mel-spectrogram, mel-spectrogram generated by Lip2Wav and our GlowLTS are visualized in the following columns respectively.

Experiments

Datasets and Preprocessing

In this section, we introduce datasets and preprocessing methods used in our experiments in detail.

Datasets In this paper, we focus on more challenging unconstrained, real-world settings and conduct experiments on Lip2Wav-Chemistry-Lectures and Lip2Wav-Chess-Analysis datasets proposed in Prajwal et al. (2020), which are the currently largest datasets for unconstrained settings. About 20 hours of real lectures videos from Youtube are included in each dataset.

Preprocessing In our experiments, videos and audios are preprocessed before training and inference. For video preprocessing, we extract the facial regions of video frames with a pre-trained face detection model. Facial images are then resized as proposed in Prajwal et al. (2020). Due to the existence of video frames without faces, we filter out these frames and corresponding audios during training and inference. For audio preprocessing, we sample the raw audio at 16kHz and set the window-size, hop-size, and mel-dimension as 800, 200, and 80 respectively for mel-spectrogram extraction.

Model Configurations

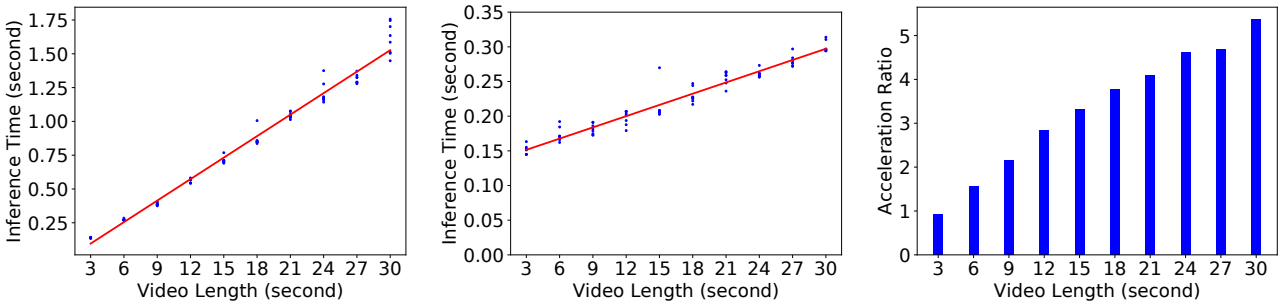
In this section, we introduce the configurations of our proposed model. We use the same network configurations as in Lip2Wav (Prajwal et al. 2020) for the visual encoder. We use 4 feed-forward Transformer blocks with 2 attention heads and a dropout of 0.1 in our condition module. For our flow-based decoder, we use 12 flow blocks in the training and inference process. Each flow block includes 1 actnorm layer, 1 invertible 1x1 conv layer, and 4 affine coupling layers. We optimize our model using Adam (Kingma and Ba 2014) optimizer with an initial learning rate of 2×10^{-4} and weight decay of 1×10^{-6} in both stages. It takes about 200k steps for the first stage of training and about 100k steps for the second stage. Our implementation is based on PyTorch.

Evaluation Methods

We evaluate our method through both objective evaluation and subjective evaluation. During the objective evaluation, we evaluate our lip-to-speech model using STOI and ESTOI, which capture the intelligibility of audios. In addition, we also evaluate the naturalness and intelligibility of generated speeches through direct and subjective human perception.

Objective Evaluation For objective evaluation, we utilize STOI (Taal et al. 2011) and ESTOI (Jensen and Taal 2016) for quantitative evaluation of the speech intelligibility. The higher STOI and ESTOI reflect better speech intelligibility.

Subjective Evaluation Though objective evaluation methods can partially reflect the intelligibility of generated speech, the quality of speech is determined by human perception. None of the existing objective metrics is highly correlated with human perception (Mira et al. 2021). Therefore, subjective human evaluation is the important and decisive criterion. In our experiments, we perform human evaluation



(a) The inference time of Lip2Wav (b) The inference time of GlowLTS (c) The acceleration ratio

Figure 4: The inference time of different models for different input video lengths. (a) represents Lip2Wav, which uses an autoregressive architecture. (b) represents the GlowLTS model we proposed. We adopt a non-autoregressive architecture. (c) shows the acceleration ratio between GlowLTS and Lip2Wav. All measurements are conducted with 1 NVIDIA 2080Ti GPU.

and report the mean opinion scores (MOS) of our GlowLTS and Lip2Wav (current state-of-the-art model), following the similar procedures proposed in Prajwal et al. (2020). 15 participants are asked to rate generated samples from each method. The video content keeps consistent among different systems so that testers only examine the audio naturalness and intelligibility. We also report the mean opinion scores of the corresponding ground-truth for reference.

Results

We conduct experiments on Lip2Wav-Chemistry-Lectures and Lip2Wav-Chess-Analysis datasets and compare the results of GlowLTS with other lip-to-speech methods. Both objective evaluation results and subjective evaluation results are reported to demonstrate the superiority of generated speeches. We also present the inference speed and show the observable acceleration we achieve compared with the current state-of-the-art method.

Objective Evaluation Results We compute STOI and ESTOI to approximate the intelligibility of generated speeches. We report the scores of Lip2Wav[1] (current state-of-the-art model). Moreover, we also report two constrained lip-to-speech methods, namely, the GAN-based method (Vougioukas et al. 2019) and improved vid2speech (Ephrat, Halperin, and Peleg 2017) as in Lip2Wav. The results can be found in Table 1 and 2. On the whole, our method and Lip2Wav perform much better than the other two methods on both datasets, indicating constrained lip-to-speech methods are hard to achieve comparable results in unconstrained settings. Our method achieve better performance compared with Lip2Wav. This fact shows that our method can better capture the semantic information from facial videos and generate more intelligible speech.

Subjective Evaluation Results For the two datasets used above, we calculate MOS scores to subjectively evaluate the intelligibility and naturalness of generated speeches. According to the results shown in Table 3 and 4, we can find that the speech generated by our proposed model is better than the current state-of-the-art model in terms of intelligibility, which is consistent with the objective evaluation results. More importantly, in terms of naturalness, our model outperforms Lip2Wav by a significant margin, which demonstrates that our model can capture the frequency correlation in the mel-spectrogram more effectively.

[1]The results are obtained using the official pre-trained model.

Method	STOI	ESTOI
GAN-based	0.192	0.132
Improved vid2speech	0.165	0.087
Lip2Wav	0.449	0.321
GlowLTS	**0.470**	**0.328**

Table 1: Objective Evaluation on Chemistry Lectures

Method	STOI	ESTOI
GAN-based	0.195	0.104
Improved vid2speech	0.184	0.098
Lip2Wav	0.377	0.257
GlowLTS	**0.394**	**0.259**

Table 2: Objective Evaluation on Chess Analysis

Method	intelligibility	naturalness
groudtruth+Griffin-Lim	4.16	4.18
Lip2Wav+Griffin-Lim	3.46	3.41
GlowLTS+Griffin-Lim	**3.57**	**3.83**

Table 3: MOS on Test set of Chemistry Lectures

Method	intelligibility	naturalness
groudtruth+Griffin-Lim	4.13	3.93
Lip2Wav+Griffin-Lim	3.45	3.35
GlowLTS+Griffin-Lim	**3.51**	**3.67**

Table 4: MOS on Test set of Chess Analysis

System	intelligibility	naturalness
GlowLTS	0	0
w/o condition module	-0.263	0.045
w/o flow-based decoder	-0.032	-0.292

Table 5: CMOS comparison in the ablation studies

Speedup: To demonstrate that our method can achieve faster speech synthesis, we measure the inference latency required to process different lengths of video on the Lip2Wav-Chemistry-Lectures dataset. Through measurements, we find that our model can achieve up to 5.377x acceleration compared with the autoregressive counterpart. As shown in Figure 4, we respectively measure the inference time of our proposed GlowLTS and Lip2Wav to process 3, 6, 9, 12, 15, 18, 21, 24, 27 and 30 seconds of video.

It can be found that 1) as shown in (a), the inference time required in Lip2Wav increases linearly with the increase of the video length; 2) as shown in (b), GlowLTS is insensitive to video lengths and the inference latency nearly holds a small constant; 3) as shown in (c), the longer video length to process, the greater acceleration our model can achieve.

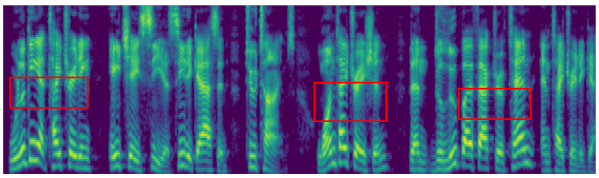
Ablation Study

In our experiments, we perform the comparison mean opinion score (CMOS) to quantify the magnitude of the difference in preference between our GlowLTS and ablation settings. The same 15 participants are asked to compare the performance of different ablation settings. To demonstrate the effectiveness of the condition module, we remove the condition module and feed the aligned visual features directly to the flow-based decoder. Similarly, we evaluate the influence of the flow-based decoder by removing the flow-based decoder and utilizing the output of the condition module as the generated speech. We conduct CMOS evaluation for these ablation studies. To be mentioned, all ablation results are based on the Lip2Wav-Chemistry-Lectures dataset.

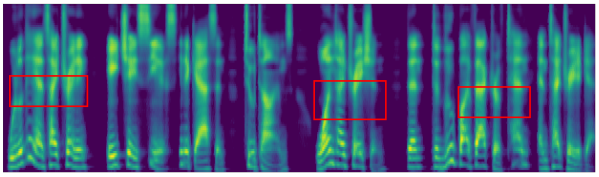
According to the results of ablation experiments, without the condition module, the naturalness is not much affected, but the intelligibility is decreased, which indicates that the condition module proposed by us contributes to effectively improving the intelligibility of the generated speech. At the same time, we find that without the flow-based decoder, the intelligibility is similar to that of GlowLTS, but there is a decrease in naturalness, indicating that the flow-based decoder can convert the coarse speech output of the condition module into more natural speech.

Qualitative Visualization

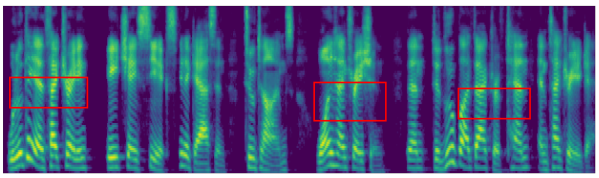
In order to show the superiority of our model more intuitively, we perform a visualization analysis. As shown in Figure 3, we visualize in turn the input face video, the ground-truth mel-spectrogram, the mel-spectrogram generated by Lip2Wav and the mel-spectrogram generated by GlowLTS. It can be found that the result of Lip2Wav suffers from mel-spectrogram over-smoothness due to the ignorance of correlation among frequency bins, especially in regions framed by red boxes. However, the result of our model is more detailed and sharp, indicating our method can model the correlation more effectively, and our model can generate more natural speech.



(a) GT



(b) w/o flow-based decoder



(c) GlowLTS

Figure 5: Visualization of different mel-spectrogram. (a), (b) and (c) represent the ground-truth mel spectrogram, the mel-spectrogram generated without the flow-based decoder and the mel-spectrogram generated by GlowLTS respectively.

In addition, we also perform the visualization analysis and demonstrate the efficiency of the flow-based decoder more intuitively. As depicted in Figure 5, without the flow-based decoder, the generated speech suffers from over-smoothness in mel-spectrogram, especially in red box regions. This is because the condition module also uses the MSE loss as the constraint and ignores the frequency bins dependency. The flow-based contributes a lot in modeling the dependency and eliminating mel-spectrogram over-smoothness.

Conclusion

In this paper, we present GlowLTS, the first flow-based non-autoregressive lip-to-speech model in unconstrained settings, which can generate more intelligible and natural speech with a faster speed. Instead of directly utilizing the flow-based decoder, we propose a condition module, which markedly improves the intelligibility of the generated speech. We demonstrate through qualitative and quantitative evaluation that our GlowLTS can achieve faster inference and generate high-quality speech.

Although the flow-based method achieves promising performance, the size of the flow-based model is relatively larger. In the future, we will try to compress the footprint of the flow-based decoder. We will also make an attempt on utilizing a more advanced vocoder to generate more natural speech.

Ethics Statement

Our method has the potential to be misused for fooling people with synthesized speech. Therefore, we plan to mitigate against such use cases (e.g., generating fake speech) by adding some constraints: 1) people who apply our code and pre-trained model are strongly suggested to present generated results as synthetic. 2) people who want to apply for the pre-trained model are required to guarantee that they would not use the model in illegal cases.

Acknowledgments

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

References

Akbari, H.; Arora, H.; Cao, L.; and Mesgarani, N. 2018. Lip2audspec: Speech reconstruction from silent lip movements video. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2516–2520. IEEE.

Chen, N.; Watanabe, S.; Villalba, J.; and Dehak, N. 2019. Listen and fill in the missing letters: Non-autoregressive transformer for speech recognition. *arXiv preprint arXiv:1911.04908*.

Cooke, M.; Barker, J.; Cunningham, S.; and Shao, X. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5): 2421–2424.

Dinh, L.; Krueger, D.; and Bengio, Y. 2014. Nice: Nonlinear independent components estimation. *arXiv preprint arXiv:1410.8516*.

Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2016. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

Ephrat, A.; Halperin, T.; and Peleg, S. 2017. Improved speech reconstruction from silent video. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 455–462.

Ephrat, A.; and Peleg, S. 2017. Vid2speech: speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5095–5099. IEEE.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*.

Griffin, D.; and Lim, J. 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2): 236–243.

Gu, J.; Bradbury, J.; Xiong, C.; Li, V. O.; and Socher, R. 2017. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Jensen, J.; and Taal, C. H. 2016. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11): 2009–2022.

Ji, S.; Xu, W.; Yang, M.; and Yu, K. 2012. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1): 221–231.

Kim, J.; Kim, S.; Kong, J.; and Yoon, S. 2020. Glow-TTS: A generative flow for text-to-speech via monotonic alignment search. *arXiv preprint arXiv:2005.11129*.

Kim, S.; Lee, S.-G.; Song, J.; Kim, J.; and Yoon, S. 2018. FloWaveNet: A generative flow for raw audio. *arXiv preprint arXiv:1811.02155*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; and Dhariwal, P. 2018. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*.

Kirichenko, P.; Izmailov, P.; and Wilson, A. G. 2020. Why normalizing flows fail to detect out-of-distribution data. *arXiv preprint arXiv:2006.08545*.

Kumar, M.; Babaeizadeh, M.; Erhan, D.; Finn, C.; Levine, S.; Dinh, L.; and Kingma, D. 2019a. Videoflow: A flow-based generative model for video. *arXiv preprint arXiv:1903.01434*, 2(5).

Kumar, Y.; Jain, R.; Salik, K. M.; Shah, R. R.; Yin, Y.; and Zimmermann, R. 2019b. Lipper: Synthesizing thy speech using multi-view lipreading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2588–2595.

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.

Mira, R.; Vougioukas, K.; Ma, P.; Petridis, S.; Schuller, B. W.; and Pantic, M. 2021. End-to-End Video-To-Speech Synthesis using Generative Adversarial Networks. arXiv:2104.13332.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13796–13805.

Prenger, R.; Valle, R.; and Catanzaro, B. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3617–3621. IEEE.

Pumarola, A.; Popov, S.; Moreno-Noguer, F.; and Ferrari, V. 2020. C-flow: Conditional generative flow models for images and 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7949–7958.

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. *arXiv preprint arXiv:1905.09263*.

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.

Sheng, L.; and Pavlovskiy, E. N. 2019. Reducing oversmoothness in speech synthesis using Generative Adversarial Networks. In *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, 0972–0974. IEEE.

Taal, C. H.; Hendriks, R. C.; Heusdens, R.; and Jensen, J. 2011. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7): 2125–2136.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Vougioukas, K.; Ma, P.; Petridis, S.; and Pantic, M. 2019. Video-driven speech reconstruction using generative adversarial networks. *arXiv preprint arXiv:1906.06301*.