INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

UniFLG: Unified Facial Landmark Generator from Text or Speech

Kentaro Mitsui, Yukiya Hono, Kei Sawada

rinna Co., Ltd., Tokyo, Japan
{kemits,yuhono,keisawada}@rinna.co.jp

Abstract

Talking face generation has been extensively investigated owing to its wide applicability. The two primary frameworks used for talking face generation comprise a text-driven framework, which generates synchronized speech and talking faces from text, and a speech-driven framework, which generates talking faces from speech. To integrate these frameworks, this paper proposes a unified facial landmark generator (UniFLG). The proposed system exploits end-to-end text-to-speech not only for synthesizing speech but also for extracting a series of latent representations that are common to text and speech, and feeds it to a landmark decoder to generate facial landmarks. We demonstrate that our system achieves higher naturalness in both speech synthesis and facial landmark generation compared to the state-of-the-art text-driven method. We further demonstrate that our system can generate facial landmarks from speech of speakers without facial video data or even speech data.

Index Terms: audiovisual speech synthesis, facial animation, facial landmark, text-to-speech, multimodal interaction

1. Introduction

In recent years, there has been growing interest in virtual humans and the metaverse, leading to an increased focus on the generation of natural talking faces [1]. The applications of talking face generation can be broadly categorized into two groups, as depicted in Fig. 1. The first group involves generating talking faces based on text inputs, which can be used for video production or multimodal chatbots [2–8]. In most cases, this group also requires simultaneous generation of speech synchronized with talking faces. The second group involves generating talking faces synchronized with speech inputs, which can be used to animate characters' faces or to act like someone else [9–18]. While this group usually requires removing the speaker identity to accommodate arbitrary speakers, it is also important to generate talking faces according to the speech emotions. Although existing research has targeted only one of these groups, integrating them can produce a single versatile model for a variety of applications. A current approach is to train a speech-to-face model for the second group of applications, and combine it with an external text-to-speech (TTS) model for the first group of applications [3, 12]. However, this approach has several drawbacks; linguistic information cannot be utilized in face generation, the quality of generated talking faces is affected by the TTS quality, and additional inference time is required for TTS.

In this paper, we propose a unified facial landmark generator (UniFLG) to integrate text- and speech-driven talking face generation. UniFLG has a TTS module based on the variational autoencoder (VAE) [19], which makes it possible to acquire a time-aligned common representation of text and speech during TTS training. Then, a landmark decoder, which is another module of UniFLG, generates facial landmarks from the intermediate representation. This is beneficial for both text- and speech-driven generation; during text-driven generation, speech and facial landmarks can be generated in parallel, resulting in faster inference and no error propagation from TTS. During speech-driven generation, speaker identity is removed because the speech-based representation is learned to be shared with text-based representation. To preserve speech emotions, we further introduce an utterance-level VAE to extract emotion embeddings and condition the landmark decoder on it. Another important feature of this study is that we regard the facial landmark, a widely used low-dimensional representation of faces [2–7, 9, 11, 12, 18, 20, 21], as common among speakers due to its little speaker dependence. This assumption enables the landmark decoder to be trained with paired text, speech and facial videos of just one speaker, while the TTS module can be trained with existing multi-speaker corpora.

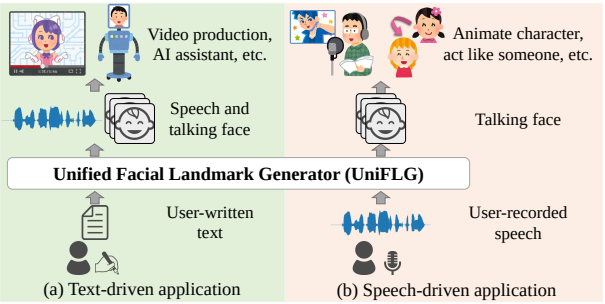


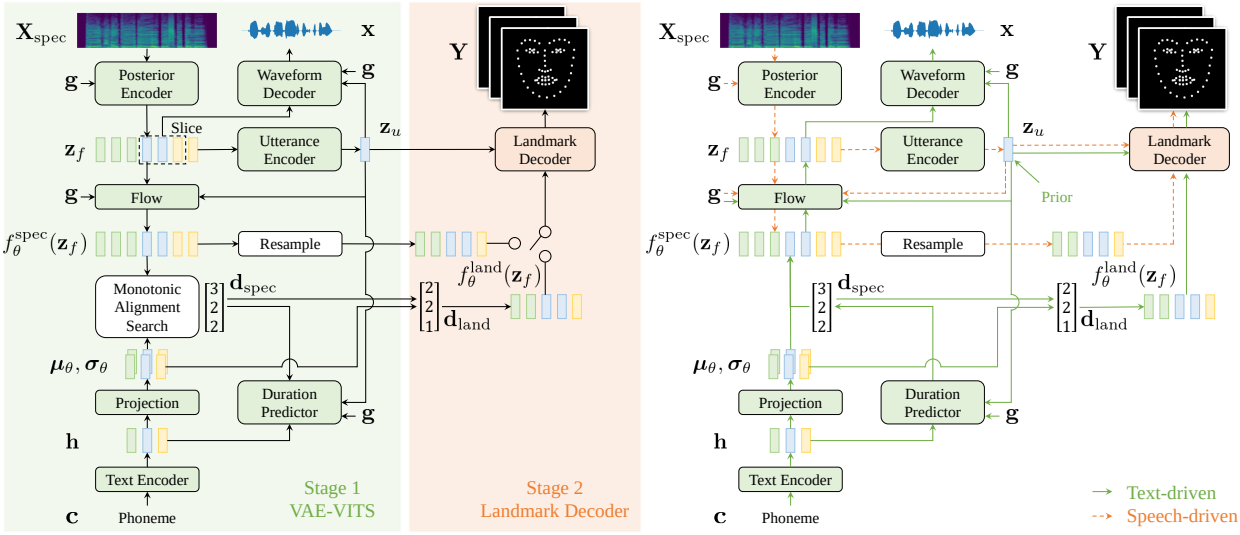
Figure 1: *Major applications of talking face generation.*

2. Related work

Text-driven talking face generation. Most of the text-driven talking face generation methods are accompanied by TTS. Pipelined methods first synthesize speech from text and then generate talking faces in a speech-driven manner [2, 3], and text-based methods utilize temporal alignment between the text and speech obtained via TTS [4, 5]. On the other hand, audiovisual speech synthesis simultaneously generates speech and talking faces [6–8]. AVTacotron2 [8] extends Tacotron2 [22] and demonstrates improved quality than pipelined methods.

Speech-driven facial animation. Early methods targeted a specific speaker or emotion [9, 10]. Recently, several methods that support an arbitrary speaker's voice [11–17] or generate talking faces in accordance with speech emotions [17, 18] have been proposed. Particularly, multimodal methods that consider both text and speech have improved the quality of talking face generation [20, 21, 23]. However, their applicability is limited because they cannot generate talking faces from only text or speech, or they have only been validated on a specific speaker.

10.21437/Interspeech.2023-1067



(a) Training (b) Inference

Figure 2: *Conceptual diagram of the (a) training and (b) inference of UniFLG.*

3. UniFLG

UniFLG consists of two components: (1) VAE-VITS [24], which introduces an utterance-level latent variable into the end-to-end TTS called VITS [25] and (2) a landmark decoder, which generates facial landmarks from the common representation of text and speech extracted by VAE-VITS. Our system is trained in two stages, as depicted in Fig. 2(a). First, VAE-VITS is trained on speech and its transcriptions, and following this, the landmark decoder is trained using paired speech and facial landmarks, as well as its transcriptions with fixed VAE-VITS parameters. UniFLG simultaneously generates speech and facial landmarks during text-driven inference, and it generates facial landmarks without using textual information during speech-driven inference, as illustrated in Fig. 2(b). We provide details regarding UniFLG in the following sections.

3.1. VAE-VITS

VITS models the distribution of a speech waveform $\mathbf{x}$ conditioned on text $\mathbf{c}$ by introducing a frame-level latent variable $\mathbf{z}_f$. The relationship between $\mathbf{x}$ and $\mathbf{z}_f$ is modeled by a posterior encoder and waveform decoder, and that between $\mathbf{z}_f$ and $\mathbf{c}$ is modeled by a prior encoder. To model the temporal alignment between $\mathbf{z}_f$ and $\mathbf{c}$, they are first converted into latent representations; $\mathbf{z}_f$ is transformed into $f_\theta^{\text{spec}}(\mathbf{z}_f)$ using a normalizing flow f_θ [26], and $\mathbf{c}$ is transformed into $\{\boldsymbol{\mu}_\theta, \boldsymbol{\sigma}_\theta\}$ using text encoder and linear projection. Then, the monotonic alignment search (MAS) [27] algorithm estimates the most probable alignment between $f_\theta^{\text{spec}}(\mathbf{z}_f)$ and $\{\boldsymbol{\mu}_\theta, \boldsymbol{\sigma}_\theta\}$. As the alignment is not available during inference, a duration predictor is simultaneously trained; it uses the text encoder output $\mathbf{h}$ to predict the duration of each phoneme $\mathbf{d}_{\text{spec}}$. During inference, $\{\boldsymbol{\mu}_\theta, \boldsymbol{\sigma}_\theta\}$ is expanded into frame-level using $\mathbf{d}_{\text{spec}}$, and $f_\theta^{\text{spec}}(\mathbf{z}_f)$ is sampled from a Gaussian distibution defined by them. Therefore, $f_\theta^{\text{spec}}(\mathbf{z}_f)$ can be regarded as a speaker-independent representation common to text and speech.

To automatically extract emotion embeddings from speech, UniFLG uses VAE-VITS, which extends VITS with an utterance encoder. By conditioning the entire system with a one-hot speaker embedding $\mathbf{g}$, the utterance encoder extracts utterance-level latent variable $\mathbf{z}_u$ that represents speech emotions [24].

3.2. Landmark Decoder

The landmark decoder generates a series of facial landmarks $\mathbf{Y} \in \mathbb{R}^{T \times N \times 2}$ given $f_\theta^{\text{land}}(\mathbf{z}_f)$, a resampled version of $f_\theta^{\text{spec}}(\mathbf{z}_f)$, where T and N denote the number of frames and 2D keypoints, respectively. A non-causal WaveNet [28] is used for the landmark decoder. Following WaveNet [29], the emotion embedding $\mathbf{z}_u$ is given as the global conditioning.

3.2.1. Mixed-modality training

Although $f_\theta^{\text{spec}}(\mathbf{z}_f)$ is common to text and speech, the ones that come from text and speech do not match exactly. Therefore, the landmark decoder is trained by switching the input between text and speech at each iteration. Given text, the duration $\mathbf{d}_{\text{spec}}$ obtained from the MAS is multiplied by a constant to match the frame rate of $\mathbf{Y}$ to obtain $\mathbf{d}_{\text{land}}$. Thereafter, $\{\boldsymbol{\mu}_\theta, \boldsymbol{\sigma}_\theta\}$ are expanded according to $\mathbf{d}_{\text{land}}$, and $f_\theta^{\text{land}}(\mathbf{z}_f)$ is sampled from a Gaussian distribution defined by them. Given speech, $f_\theta^{\text{spec}}(\mathbf{z}_f)$ extracted from $\mathbf{X}_{\text{spec}}$ is resampled by linear interpolation to obtain $f_\theta^{\text{land}}(\mathbf{z}_f)$. The landmark decoder is trained to minimize the mean squared error between predicted and target facial landmarks.

3.2.2. Inference

The flow of speech-driven inference is exactly the same as that in training. During text-driven inference, $\mathbf{d}_{\text{spec}}$ obtained using the duration predictor is converted into $\mathbf{d}_{\text{land}}$, and the following flow is the same as in training. Additionally, $\mathbf{z}_u$ is extracted from speech during speech-driven inference and sampled from the prior $p(\mathbf{z}_u) = \mathcal{N}(\mathbf{z}_u; \mathbf{0}, \mathbf{I})$ or extracted from reference speech during text-driven inference.

3.3. UniFLG-AS

Although UniFLG eliminates speaker-dependent factors from speech using the posterior encoder and flow, it cannot generate

Table 1: *Objective evaluation of text-driven and speech-driven facial landmark generation in terms of the lip landmark prediction accuracy (D-LL, D-VL, D-A), entire landmark prediction accuracy (D-L, D-V), and real time factor (RTF) of inference.*

Input	Method	D-LL ↓ [%]	D-VL ↓ [%]	D-A ↓ [%]	D-L ↓ [%]	D-V ↓ [%]	RTF ↓
Text	AVTacotron2 [8]	10.2	**1.65**	18.1	4.50	**0.572**	0.088
	TTL	9.16	1.80	16.7	3.91	0.633	0.027
	UniFLG-P	8.69	1.83	16.5	3.70	0.653	0.038
	UniFLG-T	**8.41**	1.86	**16.2**	**3.55**	0.660	0.027
Speech	STL	9.00	1.86	11.5	3.83	0.721	0.014
	STL-D	9.59	2.08	12.1	4.11	0.803	0.004
	UniFLG-S	**8.82**	**1.66**	**11.1**	**3.76**	**0.629**	0.014

facial landmarks from the speech of unseen speakers because these modules require speaker embeddings. To overcome this limitation, we propose a variant of our system, UniFLG for an arbitrary speaker (UniFLG-AS), which uses a one-hot emotion embedding as $\mathbf{g}$ instead of the speaker embedding, and $\mathbf{z}_u$ to represent speakers instead of emotions. The landmark decoder uses not $\mathbf{z}_u$ but $\mathbf{g}$ as a global condition.

4. Experiments

4.1. Experimental conditions

4.1.1. Datasets

Although the datasets such as VoxCeleb2 [30] and MEAD [31] have been widely used for talking face generation, they have relatively small amounts of data per speaker and no transcriptions exist. Thus, we recorded 3,359 (1,499 normal, 860 happy, and 1,000 sad) utterances of paired speech and facial videos from a female Japanese speaker according to predefined transcripts for training the landmark decoder. These will be hereinafter referred to as the *Paired* dataset. The development and evaluation sets comprised 45 emotion-balanced utterances, respectively, and the remaining utterances were used as the training set. For training the VAE-VITS module, we further recorded 30,542 (14,508 talk, 7,771 happy, and 8,263 sad) utterances of speech uttered by 26 speakers (eighteen females and eight males) according to predefined transcripts. These will be hereinafter referred to as the *Unpaired* dataset because the facial videos are not included. The development and evaluation sets comprised 225 speaker- and emotion-balanced utterances, respectively, and the remaining utterances were used as the training set. All the experiments were conducted using $24\,\mathrm{kHz}/16\,\mathrm{bit}$ speech signals and 30 frames per second 1280×720 videos. We used 50-dimensional linguistic features for $\mathbf{c}$[1], which contains phonemes, accents, and whether the current accent phrase is interrogative extracted using Open JTalk[2]. We extracted 70 points of facial landmarks from each frame of the facial videos using OpenPose [33].

4.1.2. Model architecture and training

The model architecture and training scheme of VAE-VITS were the same as those in a previous study [24]. For the landmark decoder, we used 16-layer non-causal WaveNet, where each layer had 192 filters with a kernel size of five and dilation factor of one, and 1×1 convolution layers placed before and after the non-causal WaveNet. A 16-dimensional $\mathbf{z}_u$ was fed to each layer of the non-causal WaveNet as a global condition. The landmark decoder was trained using an AdamW optimizer [34] with $\beta_1 = 0.8, \beta_2 = 0.99$, and a weight decay of 0.01. The initial learning rate was set to 2×10^{-4} and was multiplied by 0.999875 every epoch. The training was conducted over 10,000 iterations (approximately 150 epochs) with a batch size of 48, which took $2.5\,\mathrm{h}$ on a single NVIDIA Tesla P40 GPU.

[1]This is because prior research has demonstrated that incorporating accent information can enhance speech synthesis quality, particularly in pitch-accent languages such as Japanese [32]. We believe that it is sufficient to use phonemes as $\mathbf{c}$ for many other languages.

[2]`https://open-jtalk.sourceforge.net/`

4.2. Results

4.2.1. Facial landmark prediction accuracy

To evaluate the prediction accuracy of text- and speech-driven inference using UniFLG (hereinafter referred to as UniFLG-T and UniFLG-S, respectively), facial landmarks over the evaluation set of the *Paired* dataset were predicted. For comparison, three text-driven and two speech-driven methods (presented sequentially) were used: (1) AVTacotron2 [8], the state-of-the-art method that generates speech and facial landmarks jointly from text, similar to the proposed method, (2) TTL (text-to-landmark) that uses the same architecture as UniFLG but trains the landmark decoder only in a text-driven manner, (3) UniFLG-P (P stands for pipelined) that uses the proposed system, but it first synthesizes speech and then generates facial landmarks in a speech-driven manner, (4) STL (speech-to-landmark) that uses the same architecture as UniFLG but trains the landmark decoder only in a speech-driven manner, and (5) STL-D (D stands for direct) that does not use the posterior encoder and flow of VAE-VITS but directly feeds $\mathbf{X}_{\mathrm{spec}}$ to the landmark decoder. Following previous studies [3, 12], we evaluated the accuracy of lip movements using the landmark distance for lips (D-LL), landmark velocity difference for lips (D-VL), and difference in the open mouth area (D-A) and the accuracy of entire face movements using the landmark distance (D-L) and landmark velocity difference (D-V). Because the sequence length of facial landmarks predicted using text-driven methods is not always the same as the target, we aligned them using dynamic time warping [35].

The obtained results are presented in Table 1. Among the four text-driven methods, UniFLG-T achieved the lowest D-LL, D-A, and D-L scores. In particular, these values were lower than those of UniFLG-P, which implies that UniFLG-T could reduce the prediction errors caused by TTS. AVTacotron2, on the other hand, achieved the lowest D-VL and D-V scores. This is possibly because these metrics focus on the difference between consecutive frames and AVTacotron2 was the only method that generated facial landmarks autoregressively. Among the three speech-driven methods, UniFLG-S achieved the lowest values for all evaluation metrics. In addition, UniFLG-{T, S} achieved better performance than TTL and STL, respectively. This result demonstrates the effectiveness of the proposed method trained using both text and speech.

Table 2: *Subjective evaluation of text-driven and speech-driven systems in terms of the speech, facial landmark, and lip-sync quality.*

Data	Method	Speech ↑	Landmark ↑	Lip-Sync ↑
Paired	AVTacotron2	1.85±0.20	3.71±0.22	3.79±0.24
	UniFLG-P	4.30±0.16	4.26±0.15	4.26±0.16
	UniFLG-T	4.24±0.15	**4.28±0.15**	**4.31±0.14**
	UniFLG-S	**4.57±0.11**	4.20±0.15	4.20±0.15
Unpaired	AVTacotron2	1.99±0.16	3.52±0.17	3.66±0.19
	UniFLG-P	3.85±0.20	**4.12±0.17**	4.12±0.19
	UniFLG-T	3.88±0.16	**4.12±0.15**	**4.31±0.15**
	UniFLG-S	**4.43±0.16**	4.08±0.17	4.18±0.18

Table 3: *Subjective evaluation of speech-driven facial landmark generation for the Paired, Unpaired, and Unseen datasets.*

Data	Method	Landmark ↑	Lip-Sync ↑
Paired	STL-D	3.95±0.16	4.01±0.16
	UniFLG-AS-S	**4.45±0.12**	**4.44±0.12**
Unpaired	UniFLG-AS-S	4.23±0.14	4.22±0.14
Unseen	STL-D	3.96±0.17	3.68±0.18
	UniFLG-AS-S	**4.26±0.13**	**4.04±0.17**

4.2.2. Inference speed

Fast inference is crucial, especially for real-time applications such as human conversation and live streaming. We measured the real time factor (RTF), the time required to generate 1 s speech and facial landmarks, on one NVIDIA Tesla P40 GPU. For AVTacotron2, we included the time required for waveform generation, for which we used HiFi-GAN [36] trained on all speech data in the *Paired* and *Unpaired* datasets. The results are listed in the right-most column of Table 1. As UniFLG is non-autoregressive, its RTF was generally smaller than that of the autoregressive AVTacotron2. We also confirmed that UniFLG-T improved the RTF by 41% over UniFLG-P because it eliminates the need for generating speech before facial landmark generation. Among the speech-driven methods, UniFLG-S was not as fast as STL-D; however, it was still faster than any text-driven method.

4.2.3. Generated speech and facial landmark quality

Subjective evaluation was conducted based on the following three criteria: speech quality, facial landmark quality, and lip-sync quality[3]. For the second and third criteria, the facial landmarks were plotted frame by frame to construct a facial video and presented to the raters. Four methods (AVTacotron2 and UniFLG-{P, T, S}) were compared over the two datasets (*Paired*, *Unpaired*). Note that the speech quality of UniFLG-S indicates the quality of the recorded speech. Thirty-one raters participated in the evaluation, and each rater evaluated 30 samples on a five-point scale from one (bad) to five (excellent).

The obtained results are summarized in Table 2. Overall, similar trends were observed for the *Paired* and *Unpaired* datasets. The speech quality of AVTacotron2 was quite low; this is because the usage of multiple datasets, speakers, and emotions made the training difficult, resulting in the failure of alignment and stop token prediction. UniFLG-{P, T} demonstrated significantly improved speech quality and those scores were close to that of UniFLG-S for the *Paired* dataset. The scores slightly decreased for the *Unpaired* dataset, which is possibly because the amount of training data per speaker was approximately one-third of the data in the *Paired* dataset. UniFLG-{P, T, S} achieved similar facial landmark quality and lip-sync quality scores. These scores exceeded 4.0 and were significantly better than those of AVTacotron2. Based on these results, we concluded that the proposed UniFLG can generate high-quality facial landmarks from either text or speech. Furthermore, based on the fact that the scores of UniFLG-T were comparable to those of UniFLG-P, we can conclude that it can be used as a faster alternative to the pipelined counterpart.

4.2.4. Facial landmark generation for unseen speakers

To evaluate the facial landmark generation quality for unseen speakers during training, we considered 240 (80 talk, happy, and sad) utterances of speech uttered by 10 unseen speakers (five males and five females) as the *Unseen* dataset. For each of the *Paired*, *Unpaired*, and *Unseen* datasets, we generated facial landmarks from speech using the UniFLG-AS, described in section 3.3 (hereinafter referred to as UniFLG-AS-S) and conducted subjective evaluation as outlined in the previous section. For the baseline, we used STL-D described in section 4.2.1. As STL-D does not use VAE-VITS, both the *Unpaired* and *Unseen* datasets were unseen during training, and hence, we omitted evaluation for the *Unpaired* dataset.

The obtained results are summarized in Table 3. Both facial landmark and lip-sync quality of UniFLG-AS-S for the *Unseen* dataset outperformed those of STL-D for the *Paired* dataset. This indicates that the proposed system can generate high-quality facial landmarks from the speech of unseen speakers. Although the lip-sync quality score of UniFLG-AS-S for the *Unseen* dataset exceeded 4.0, it was slightly worse than that for the *Unpaired* dataset. The number of speakers seen while training VAE-VITS was 27 in total, which we assume was insufficient to represent all unseen speakers with the learned latent space. Narrowing this performance gap by using more speakers for training VAE-VITS will form a part of future research.

5. Conclusions

This paper proposed UniFLG, which integrates audiovisual speech synthesis and speech-driven facial animation frameworks. The involved experimental evaluation demonstrated that the proposed system could generate higher quality facial landmarks than conventional methods from either text or speech. Moreover, UniFLG-AS, a variant of UniFLG, could generate natural facial landmarks even from the speech of unseen speakers. Future research would involve attempts to extend the proposed system to an end-to-end framework that includes the training of a video generation network. It would also be advantageous to integrate UniFLG and UniFLG-AS to generate facial landmarks from the speech of arbitrary speakers and emotions.

[3] https://rinnakk.github.io/research/publications/UniFLG

6. References

[1] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, "Deep audio-visual learning: A survey," *International Journal of Automation and Computing*, vol. 18, no. 3, pp. 351–376, Apr. 2021.

[2] R. Kumar, J. Sotelo, K. Kumar, A. de Brébisson, and Y. Bengio, "ObamaNet: Photo-realistic lip-sync from text," in *Proc. NIPS 2017 Machine Learning for Creativity and Design Workshop*, California, U.S.A., Dec. 2017.

[3] X. Wang, Q. Xie, J. Zhu, L. Xie, and O. Scharenborg, "AnyoneNet: Synchronized speech and talking head generation for arbitrary persons," *IEEE Transactions on Multimedia*, pp. 1–12, Oct. 2022.

[4] S. Zhang, J. Yuan, M. Liao, and L. Zhang, "Text2video: Text-driven talking-head video synthesis with personalized phoneme - pose dictionary," in *Proc. ICASSP*, Singapore, May 2022, pp. 2659–2663.

[5] L. Li, S. Wang, Z. Zhang, Y. Ding, Y. Zheng, X. Yu, and C. Fan, "Write-a-speaker: Text-based emotional and rhythmic talking-head generation," in *Proc. AAAI*, online, Feb. 2021, vol. 35, pp. 1911–1920.

[6] S. Dahmani, V. Colotte, V. Girard, and S. Ouni, "Conditional variational auto-encoder for text-driven expressive audiovisual speech synthesis," in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 2598–2602.

[7] S. Dahmani, V. Colotte, V. Girard, and S. Ouni, "Learning emotions latent representation with CVAE for text-driven expressive audiovisual speech synthesis," *Neural Networks*, vol. 141, pp. 315–329, Sep. 2021.

[8] A. Hussen Abdelaziz, A. P. Kumar, C. Seivwright, G. Fanelli, J. Binder, Y. Stylianou, and S. Kajareker, "Audiovisual speech synthesis using Tacotron2," in *Proc. ICMI*, Montreal, Canada, Oct. 2021, pp. 503–511.

[9] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: Learning lip sync from audio," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–13, Aug. 2017.

[10] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews, "A deep learning approach for generalized speech animation," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–11, Jul. 2017.

[11] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, "Generating talking face landmarks from speech," in *Proc. LVA/ICA*, Guildford, U.K., Jul. 2018, pp. 372–381.

[12] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "MakeItTalk: Speaker-aware talking-head animation," *ACM Transactions on Graphics*, vol. 39, no. 6, pp. 1–15, Dec. 2020.

[13] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proc. MM*, Seattle, U.S.A., Oct. 2020, pp. 484–492.

[14] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *Proc. CVPR*, online, Jun. 2021, pp. 4176–4186.

[15] B. Liang, Y. Pan, Z. Guo, H. Zhou, Z. Hong, X. Han, J. Han, J. Liu, E. Ding, and J. Wang, "Expressive talking head generation with granular audio-visual control," in *Proc. CVPR*, New Orleans, U.S.A., Jun. 2022, pp. 3387–3396.

[16] Y. Zhang, W. He, M. Li, K. Tian, Z. Zhang, J. Cheng, Y. Wang, and J. Liao, "Meta Talk: Learning to data-efficiently generate audio-driven lip-synchronized talking face with high definition," in *Proc. ICASSP*, Singapore, May 2022, pp. 4848–4852.

[17] K. Deng, A. Bansal, and D. Ramanan, "Unsupervised audiovisual synthesis via exemplar autoencoders," in *Proc. ICLR*, online, May 2021.

[18] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, and F. Xu, "Audio-driven emotional video portraits," in *Proc. CVPR*, online, Jun. 2021, pp. 14080–14089.

[19] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, Banff, Canada, Apr. 2014.

[20] L. Yu, J. Yu, M. Li, and Q. Ling, "Multimodal inputs driven talking face generation with spatial–temporal dependency," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 203–216, Feb. 2020.

[21] L. Yu, H. Xie, and Y. Zhang, "Multimodal learning for temporally coherent talking face generation with articulator synergy," *IEEE Transactions on Multimedia*, vol. 24, pp. 2950–2962, Jun. 2021.

[22] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, Calgary, Canada, May 2018, pp. 4779–4783.

[23] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "Joint audio-text model for expressive speech-driven 3d facial animation," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 5, no. 1, pp. 1–15, May 2022.

[24] K. Mitsui, T. Zhao, K. Sawada, Y. Hono, Y. Nankaku, and K. Tokuda, "End-to-end text-to-speech based on latent representation of speaking styles using spontaneous dialogue," in *Proc. INTERSPEECH*, Incheon, Korea, Sep. 2022, pp. 2328–2332.

[25] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. ICML*, online, Jul. 2021, pp. 5530–5540.

[26] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. ICML*, Lille, France, Jul. 2015, pp. 1530–1538.

[27] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," in *Proc. NeurIPS*, online, Dec. 2020, vol. 33, pp. 8067–8077.

[28] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. ICASSP*, Brighton, U.K., May 2019, pp. 3617–3621.

[29] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *Proc. SSW*, Sunnyvale, U.S.A., Sep. 2016, p. 125.

[30] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 1086–1090.

[31] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "MEAD: A large-scale audio-visual dataset for emotional talking-face generation," in *Proc. ECCV*, online, Aug. 2020, pp. 700–717.

[32] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, "Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language," in *Proc. ICASSP*, Brighton, U.K., May 2019, pp. 6905–6909.

[33] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[34] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, New Orleans, U.S.A., May 2019.

[35] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proc. KDD Workshop*, Seattle, U.S.A., 1994, vol. 10, pp. 359–370.

[36] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, online, Dec. 2020, vol. 33, pp. 17022–17033.