INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Speech2Video: Cross-Modal Distillation for Speech to Video Generation

Shijing Si, Jianzong Wang, Xiaoyang Qu, Ning Cheng, Wenqi Wei, Xinghua Zhu and Jing Xiao

Ping An Technology (Shenzhen) Co., Ltd., China
jzwang@188.com

Abstract

This paper investigates a novel task of talking face video generation solely from speeches. The speech-to-video generation technique can spark interesting applications in entertainment, customer service, and human-computer-interaction industries. Indeed, the timbre, accent and speed in speeches could contain rich information relevant to speakers' appearance. The challenge mainly lies in disentangling the distinct visual attributes from audio signals. In this article, we propose a light-weight, cross-modal distillation method to extract disentangled emotional and identity information from unlabelled video inputs. The extracted features are then integrated by a generative adversarial network into talking face video clips. With carefully crafted discriminators, the proposed framework achieves realistic generation results. Experiments with observed individuals demonstrated that the proposed framework captures the emotional expressions solely from speeches, and produces spontaneous facial motion in the video output. Compared to the baseline method where speeches are combined with a static image of the speaker, the results of the proposed framework is almost indistinguishable. User studies also show that the proposed method outperforms the existing algorithms in terms of emotion expression in the generated videos.

Index Terms: Video generation, generative adversarial network, distillation, unsupervised learning, representation learning

1. Introduction

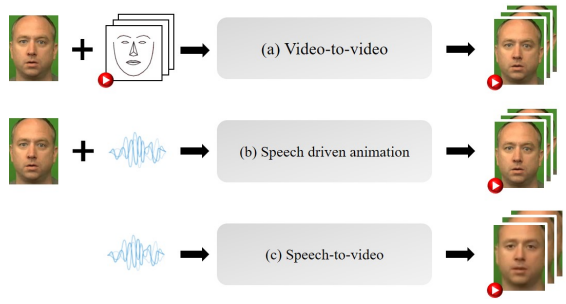

Figure 1: *Comparison of different tasks of talking face video synthesis. a) Video-to-video synthesis with face landmark video and reference image as inputs [1, 2]. b) Speech driven animation leveraging given speeches and reference image [3, 4]. c) The proposed speech-to-video generation solely from speeches.*

Speeches are audio signals carrying affluent information of various domains [5, 6]. From a piece of speech, words and sentences are the linguistic contents it conveys. Emotions can be detected from the tone of a speech [7, 8]. Also, the speakers' identities are engraved in their voices, their accents and ways they talk [9, 10, 11]. Human brain can extract information on different aspects from a single piece of speech and recall or form imaginations of the identity of the potential speaker and also the speaker's facial movements when talking [12, 13, 14, 15, 16, 17]. In this paper, we propose an innovative task of realistic talking face generation from a single audio input.

Prior to this paper, numerous works have been proposed to integrate visual and audio inputs to compose talking face videos of target identity [18, 19, 20, 21]. Different settings of input sources is depicted in Fig. 1. In contrast to prior techniques, the proposed speech-to-video generator takes a segment of speech as the sole input. By "recollecting" a face from learned audio features, the proposed framework produces a video with a face observed in the training process. The results of speech2video is similar to that of speech driven animation.

The proposed task is challenging as well as intriguing. Information about the speaker's identity and his/her spontaneous emotion is entangled in the speech signal, together with the linguistic contents. The system must disentangle the identity and emotional attributes from the audio signal, and transfer these audio features into visual ones. In this paper, we design a two-stage framework, speech2video, to its solution. First, a cross-modal distillation module extracts identity and emotional features from unlabelled talking face videos. Secondly, a generative adversarial network (GAN) [22, 23, 24] with visual identity guidance is trained to compose the talking face video from the audio feature vectors. Experimental results verifies the viability of the proposed task and solution. Our generated videos of observed persons are realistic and emotionally accurate, similar to those of speech driven animation, even without any visual cues as input. Results of unobserved persons do not have the same facial appearance as the ground truth, but is roughly consistent with respect to gender and age.

The contributions of this paper are:

- Propose a multi-modal distillation network to disentangle identity and emotional features from speech signals.
- Devise a GAN structure with carefully crafted discriminators to generate talking face videos from speeches.
- Demonstrate the viability of the proposed task, and evaluate the performance of the proposed video generator framework with extensive experiments.

2. Methods

The overall framework of speech2video is illustrated in Fig. 2. The comprehensive procedure composes of two stages. In the first stage, the personal identity and emotional features are disentangled from the speech representation. The representation disentangling module is trained from unlabelled talking face videos, using distillation learning techniques. In the second stage, an adversarial composer gathers the temporal speech representation as well as the identity and emotional features, and

http://dx.doi.org/10.21437/Interspeech.2021-1996

generates the talking face frames. The representation disentangler and adversarial composer are trained separately. Detailed description of their structures is elaborated in the following subsections.

2.1. Speech Representation Disentangling

Speeches carry affluent information about speakers' identity. Setting the linguistic contents aside, human receivers have the natural ability to recognize the speakers' identity and emotion from the speeches alone. However, for a neural network, these attributes are too entangled to understand. In order to delineate and preserve the identity and emotional characteristics in a speech, we must first disentangle these features from the original speech.

The disentangling task is non-trivial. For one thing, the input signal is in the audio domain, while target features are visual, requiring cross-domain information encoding. For another, the disentangling module must extract multiple features, namely the identity feature and emotional feature, at the same time. Furthermore, in order to maximize the applicable source of training samples, the disentangling module is optimized in an unsupervised manner.

In this section, a cross-modal distillation network is proposed as the disentangling module (Fig. 2a). At the training stage, an unlabelled talking face video $V = \{(\alpha_t, \phi_t), t = 1, ..., T\}$ is provided as input. The targeted representation extractor learns only from the audio signals $\alpha = \{\alpha_1, ..., \alpha_T\}$. The visual segments $\phi = \{\phi_1, ..., \phi_T\}$ in the video provide supervisory information to the audio representation through a distillation network. By assigning different teacher networks, the identity and emotional features, ν and μ, are delineated through the respective distillation process simultaneously.

Specifically, a comprehensive representation of audio features is extracted by a pre-trained contrastive prediction coding (CPC) module [25] from the speech. In CPC, the audio sequence α is first recurrently encoded by a nonlinear encoder g_{enc} into latent embedding $z_t, t = 1, ..., T$. Then an autoregressive model g_{ar} summarizes all latent embedding up to time t and produces context latent representation $c_t = g_{\text{ar}}(z_{\leq t})$. CPC has been shown to excel in downstream tasks like speaker classification and phone classification. The temporal vector $C = \{c_1, ..., c_T\}$ is used to generate frame sequences, as elaborated in the next section. Additionally, we take $\omega = c_T$ as a dimensionality-reduced interpretation of the intact information in the audio signal.

Two student networks, S_{emo} and S_{id}, read from ω and distill emotional and identity features, respectively. Let $\nu := S_{\text{id}}(\omega)$ and $\mu := S_{\text{emo}}(\omega)$ denote the distilled feature vectors. In accordance with the student networks, two teacher networks, T_{emo} and T_{id}, are defined to supervise the student networks on knowledge extraction . The T_{id} module adopts a VGGNet structure pre-trained on the VGGFace dataset for face recognition. Through the VGGNet, a 4,096-dimensional identity feature vector is extracted from the first frame of the video, $\nu^* = T_{\text{id}}(\phi_0)$. On the other hand, as emotional motion changes from frame to frame, it is insufficient to observe a single frame. The T_{emo} module takes K randomly sampled frames $\pi_1, ..., \pi_k$ from ϕ, and extracts the corresponding emotion feature vectors through a Squeeze and Excitation Network (SENet). The target feature for μ is the average pooling of these samples, i.e.,

$$\mu^* = \frac{1}{K} \sum_{k=1}^{K} T_{\text{emo}}(\phi_{\pi_k}). \tag{1}$$

The choice of the two teacher networks can be tuned to dataset characteristics and computation power. In this paper, we adopt SENet and VGGNet for simplicity and extensive proofs of their performance in the respective fields. The teacher networks are pre-trained on open-source facial image dataset VGGFace [26, 27] and FERplus [28]. Parameters of S_{id} and S_{emo} are optimized simultaneously, to minimize the joint multi-task distillation loss [29],

$$\begin{aligned} L_1 =& \lambda \|\bar{\mu} - \bar{\mu}^*\|^2 + l_{\text{distill}}(\mu, \mu^*) + \\ & \lambda \|\bar{\nu} - \bar{\nu}^*\|^2 + l_{\text{distill}}(\nu, \nu^*), \end{aligned} \tag{2}$$

where $\lambda = 0.025$ and $\bar{(\cdot)}$ denotes the normalized vectors. l_{distill} stands for the distillation loss, i.e., the softmax cross-entropy

$$l_{\text{distill}}(\mathbf{x}, \mathbf{y}) = -\sum_i \text{softmax}(\mathbf{x})_i \log(\text{softmax}(\mathbf{y})_i). \tag{3}$$

2.2. Adversarial Video Composition

To alleviate the complexity of speech-to-video generation, the proposed framework is divided into two separate stages. The cross-modal encoding of speech signal has been covered in the last section. In this section, an adversarial decoder is proposed to transform the encoded vectors into a talking face video.

From the encoder, we have obtained 3 feature vectors, namely the frame feature sequence C, the emotion feature μ and the identity feature ν. For every frame $c_t \in C$, the emotion feature is directly concatenated for transposed convolution. On the other hand, the identity feature ν is separately handled for enhanced feature supervision.

Specifically, ν goes through 3 transposed convolution layers, each doubling its width and height. For every intermediate feature map, a 1×1 convolution aligns its shape to the corresponding layer in a VGGNet. The pretrained VGGNet features of a single facial image is adopted to supervise the generation of high-resolution identity feature maps by an adversarial loss, to be detailed in the remainder of this section. The inflated identity features are concatenated to the corresponding layers of the frame generator, and input to the following transposed convolutions. We have found that processing the identity feature separately greatly enhanced the individual distinction in the final video frames. The frame decoder F is trained to optimize the combination of a number of target functions, as elaborated below.

Adversarial losses. The frame decoder F is a generator that maps audio features to video frames. Generative Adversarial Networks (GANs) supplement the generator with a competing discriminator, to train the generator models unsupervisedly. In the proposed framework, 3 discriminators are devised to audit the decoder F from different perspectives, namely, identity preservation, frame authenticity, and synchronization. These discriminators are denoted as D_{ID}, D_{fr}, D_{sync}, respectively. The adversarial losses used in our experiments are Least Squares GAN (LSGAN) losses.

D_{ID} discriminates the inflated identity feature maps $\tilde{\nu} = \{\tilde{\nu}^{(i)}, i = 1, 2, 3\}$ against the VGG features $\mathbf{V} = \{\text{VGG}^{(i)}(I), i = 1, 2, 3\}$ of a static facial image I of the speaker, i.e.,

$$l_{\text{adv}}^{\text{ID}} = \frac{1}{2}\mathbb{E}\left[(D_{\text{ID}}(\tilde{\nu}) - 1\right]^2 + \frac{1}{2}\mathbb{E}\left[D_{\text{ID}}(\mathbf{V})\right]^2. \tag{4}$$

D_{fr} discriminates between the generated and real video

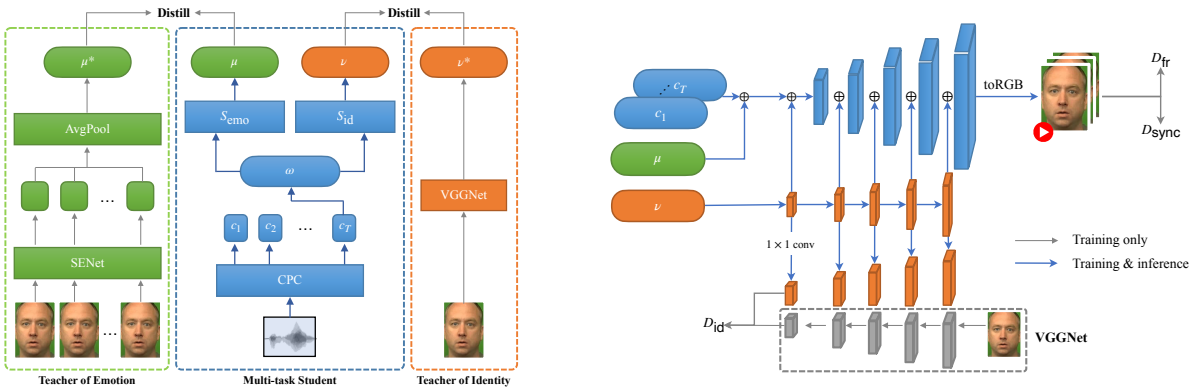

Figure 2: *The speech2video framework. Left: Proposed feature disentangling model. Right: Adversarial frame generator with disentangled intermediate features. The* $\oplus$ *symbol represents feature concatenation and transposed-convolution, followed by two* 1×1 *convolution layers.*

frames. The corresponding adversarial loss is

$$l_{\text{adv}}^{\text{fr}} = \frac{1}{2}\sum_{t=1}^{T}\left(\mathbb{E}\left[(D_{\text{fr}}(F(c_t,\nu,\mu)-1\right]^2 + \mathbb{E}\left[D_{\text{fr}}(\phi_t)\right]^2\right). \quad (5)$$

Last but not least, D_{sync} ensures the synchronization between audio and video frames. For this discriminator, time steps τ and τ' are sampled from $\{1,...,T\}$, $\tau \neq \tau'$. D_{sync} takes a pair of audio feature and video frame as input. The synchronous adversarial loss encourages synchronized audio-video pairs while punishing the asynchronous ones, i.e.,

$$\begin{aligned} l_{\text{adv}}^{\text{sync}} =& \mathbb{E}\left[(D_{\text{sync}}(c_\tau,\phi_\tau)\right]^2 + \\ & \frac{1}{2}\mathbb{E}\left[(D_{\text{sync}}(c_{\tau'},\phi_\tau)-1\right]^2 + \\ & \frac{1}{2}\mathbb{E}\left[D_{\text{sync}}(c_\tau, F(c_\tau,\nu,\mu))-1\right]^2. \end{aligned} \quad (6)$$

The overall adversarial loss is given by

$$L_2^{\text{adv}} = l_{\text{adv}}^{\text{ID}} + l_{\text{adv}}^{\text{fr}} + l_{\text{adv}}^{\text{sync}}. \quad (7)$$

Frame similarity loss. In addition to the adversarial losses, a pixel-wise similarity loss is imposed on the generated frames. Particularly, the facial appearance of a talking person remains largely identical from the nose up, that occupies roughly the upper half of the frame image. Therefore, the frame similarity loss is defined as

$$L_2^{\text{sim}} = \sum_{p\in[0,W]\times[H/2,H]} |F^p(c_t,\nu,\mu) - \phi_t^p|, \quad (8)$$

where W and H are the width and height of the video frames, respectively. F^p and ϕ_t^p stands for the pixel value of the corresponding frame image at position p.

Gradient loss. As suggested by [30], a gradient loss is also applied to alleviate the blurriness caused by L1 frame similarity function.

$$L_2^{\text{grad}} = \left|\nabla\psi(F(c_t,\nu,\mu) - \nabla\psi(\phi_t))\right|, \quad (9)$$

where $\psi(\cdot)$ is a smoothing filter, ∇ is the gradient filter.

In summary, the total loss for the frame decoder F is

$$L_2 = L_2^{\text{adv}} + L_2^{\text{sim}} + L_2^{\text{grad}}. \quad (10)$$

3. Experiments

3.1. Implementation Details

The proposed speech2video framework is implemented on PyTorch with a single NVIDIA Tesla V100 GPU. The learning rate of the generator, identity discriminator, frame discriminator and naturalness discriminator is set to 3×10^{-4}, 3×10^{-4}, 1×10^{-4} and 1×10^{-5}, respectively. RMSProp optimizer is adopted for all the training. The disentanglement and frame decoder modules are trained for 100 epochs respectively.

The proposed methods are experimented with two open datasets of talking face videos, Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) [31] and VoxCeleb2 [32]. *CREMA-D* contains 7,442 clips uttered by 91 ethnically-diverse actors(48 male, 43 female). Each speaker utters 12 sentences in 6 different emotions (Anger, Disgust, Fear, Happy, Neutral, Sad). This dataset is challenging because facial movements of the speaker under certain emotions are expressive or even exaggerated. The audio sample rate and video sample rate of this dataset is 16kHz and 30fps respectively. The dataset is divided by the proportion 70%,15%, and 15% for training, validation and testing, respectively. *VoxCeleb2* contains more than 1 million utterances of 6,112 celebrities extracted from YouTube videos. The training and testing sets are given by the authors of the dataset.

In the following experiments, performance of Speech Driven Facial Animation (SDFA) [3, 4] is compared with the proposed framework as a baseline. SDFA utilizes a static facial image and a piece of raw waveform speech to synthesize a talking face video and has achieved realistic reconstruction of facial features in the generated results. Furthermore, for a baseline of facial video generation with only a voice input, we combine the voice encoder of speech2face [33] with the video generator of SDFA, to form the S2F+SDFA model.

3.2. Qualitative Results

Figure 3 demonstrates the performance of designed model comparing to the baseline methods, SDFA and S2F+SDFA. It can be seen that the S2F+SDFA network yields the worst performance with large identity error and blurry frames. This is because unlike the designed model with adversarial training in the identity reconstruction process, modified speech2face model only leverages a normalization loss function on high dimensional

embedding to supervise the identity which is insufficient. Furthermore, the SDFA model is designed to have a reference image as input. However, in S2F+SDFA network, the reference image is substituted by high dimensional facial features. This lacks of detailed information for the decoding process which results in bad quality frames. On the other hand, the proposed model obtains competitive results comparing to the state-of-art SDFA model, though leveraging no reference images. Identity and facial details of the speaker are accurately reconstructed. Though, the SDFA model contains more texture details, such as light and shade, which are carried in the reference image. In summary, the experiments well demonstrated the competitive performance and the feasibility of the proposed methodology.

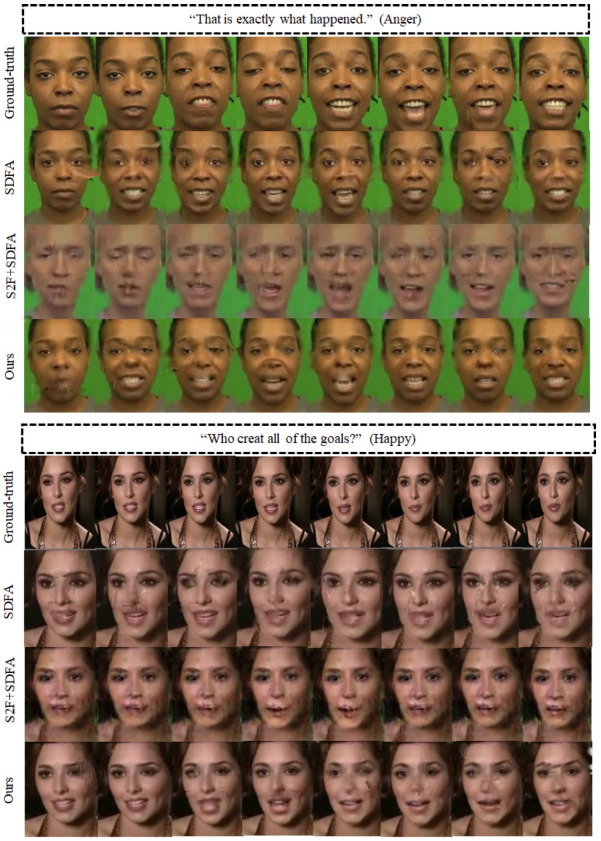

Figure 3: *Qualitative result comparison of our system with two baseline models and the ground-truth on two input speeches. For each input utterance, Ground-truth is in the first row, followed by SDFA and S2F+SDFA, with our method in the last row. Our result not only possesses accurate general features but also abundant facial details.*

In the above experiments, the proposed framework is tasked to "recollect" faces from observed voices. It would be also interesting to see what we can achieve if an unobserved voice is given to the speech2video framework. Fig. 4 illustrates two cases of un-observed voice inputs. As it shows, the appearance of the generated faces are different from the ground truth. But the speakers' gender, complexion and age are roughly preserved in the results.

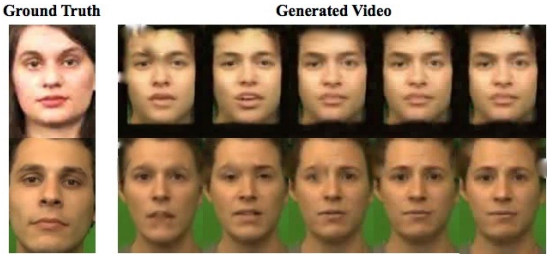

Figure 4: *Imaginary facial video generation by our method from unobserved voice inputs.*

Table 1: *Quantitative comparison of the proposed method with baseline methods.*

Dataset	Method	SSIM	PSNR	Confidence
CREMA-D	SDFA	0.705	23.568	5.4
	S2F+SDFA	0.519	20.104	4.9
	SDFA(mismatch)	0.521	20.024	5.0
	speech2video	0.541	20.540	5.1
VOXCELEB2	SDFA	0.7342	25.121	5.8
	S2F+SDFA	0.5643	20.413	4.9
	SDFA(mismatch)	0.5612	20.126	5.0
	speech2video	0.5707	20.891	5.1

3.3. Quantitative Results

Quantitatively, we evaluate proposed model based on the following aspects: correctness of identity reconstruction, the quality of generated videos and the audio-visual synchronization. The corresponding metrics are structural similarity (SSIM), peak signal-to-noise ratio (PSNR) and AV confidence [34] indices, respectively.

The quantitative results of the proposed model compared with baseline methods on the aforementioned metrics are shown in table 1. The proposed model scores below the SDFA model but higher than the S2F+SDFA network which follows the expectation. SSIM and PSNR metrics represent the distance between the generated and ground-truth frame in various aspects. The proposed system does not use ground-truth image as reference. Because of the resulting diversity of GAN based approach, it is impossible to reconstruct facial features as perfectly without any additional image as references.

4. Conclusion and Future Work

This paper explores a novel model to synthesize talking face video solely from a single audio. With the help of cross-modal distillation, the model extracts embedding vectors representing emotions and speaker identities, and generates face videos through an adversarial framework. The design of multiple discriminators ensures the naturalness and fluency of the generated video. Through a series of experiments, our designed model shows persuasive results.

5. Acknowledgements

This work is supported by National Key Research and Development Program of China under grant No.2018YFB0204403, No.2017YFB1401202 and No.2018YFB1003500. Corresponding author is Jianzong Wang from Ping An Technology (Shenzhen) Co., Ltd.

6. References

[1] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9299–9306.

[2] G. Mittal and B. Wang, "Animating face using disentangled audio representations," *arXiv preprint arXiv:1910.00726*, 2019.

[3] K. Vougioukas, S. Petridis, and M. Pantic, "End-to-end speech-driven realistic facial animation with temporal gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 37–40.

[4] ——, "Realistic speech-driven facial animation with gans," *International Journal of Computer Vision*, pp. 1–16, 2019.

[5] A. Richard, C. Lea, S. Ma, J. Gall, F. de la Torre, and Y. Sheikh, "Audio-and gaze-driven facial animation of codec avatars," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 41–50.

[6] M. Masood, M. Nawaz, K. M. Malik, A. Javed, and A. Irtaza, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *arXiv preprint arXiv:2103.00484*, 2021.

[7] A. Nagrani, J. S. Chung, S. Albanie, and A. Zisserman, "Disentangled speech embeddings using cross-modal self-supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6829–6833.

[8] X. Wang, T. Qiao, J. Zhu, A. Hanjalic, and O. Scharenborg, "Generating images from spoken descriptions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 850–865, 2021.

[9] Z. Cai, C. Zhang, and M. Li, "From speaker verification to multi-speaker speech synthesis, deep transfer with feedback constraint," *arXiv preprint arXiv:2005.04587*, 2020.

[10] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, and G. Lee, "Speech gesture generation from the trimodal context of text, audio, and speaker identity," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–16, 2020.

[11] P. Lopez-Otero and L. Docio-Fernandez, "Analysis of gender and identity issues in depression detection on de-identified speech," *Computer Speech & Language*, vol. 65, p. 101118, 2021.

[12] S. Emre Eskimez, Y. Zhang, and Z. Duan, "Speech driven talking face generation from a single image and an emotion condition," *arXiv e-prints*, pp. arXiv–2008, 2020.

[13] X. Huang, M. Wang, and M. Gong, "Fine-grained talking face generation with video reinterpretation," *The Visual Computer*, pp. 1–11, 2020.

[14] A. Koumparoulis, G. Potamianos, S. Thomas, and E. da Silva Morais, "Audio-assisted image inpainting for talking faces," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7664–7668.

[15] 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 *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 484–492.

[16] S. Sinha, S. Biswas, and B. Bhowmick, "Identity-preserving realistic talking face generation," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–10.

[17] D. Zeng, H. Liu, H. Lin, and S. Ge, "Talking face generation with expression-tailored generative adversarial network," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1716–1724.

[18] 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 *European Conference on Computer Vision*. Springer, 2020, pp. 700–717.

[19] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, "End-to-end generation of talking faces from noisy speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1948–1952.

[20] C. Yang and S.-N. Lim, "One-shot domain adaptation for face generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5921–5930.

[21] W. Wang, Y. Wang, J. Sun, Q. Liu, J. Liang, and T. Li, "Speech driven talking head generation via attentional landmarks based representation," *Proc. Interspeech 2020*, pp. 1326–1330, 2020.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[23] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[24] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[25] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[26] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 41.1–41.12, 2015.

[27] N. Haque and S. S. Tokey, "Grayscale portrait colorization using cnns and pretrained vgg-face descriptor," in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2019, pp. 1–5.

[28] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 279–283.

[29] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[30] F. Kou, W. Chen, C. Wen, and Z. Li, "Gradient domain guided image filtering," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4528–4539, 2015.

[31] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[32] J. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Interspeech*, 2018.

[33] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2face: Learning the face behind a voice," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7539–7548.

[34] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 251–263.