END-TO-END GENERATION OF TALKING FACES FROM NOISY SPEECH

Sefik Emre Eskimez[§], *Ross K. Maddox*[†], *Chenliang Xu*[‡], *and Zhiyao Duan*[§*]

[§]Electrical and Computer Engineering, [†]Biomedical Engineering & Neuroscience, [‡]Computer Science
University of Rochester, 500 Wilson Blvd, Rochester, NY 14627, USA

ABSTRACT

Acoustic cues are not the only component in speech communication; if the visual counterpart is present, it is shown to benefit speech comprehension. In this work, we propose an end-to-end (no pre- or post-processing) system that can generate talking faces from arbitrarily long noisy speech. We propose a mouth region mask to encourage the network to focus on mouth movements rather than speech irrelevant movements. In addition, we use generative adversarial network (GAN) training to improve the image quality and mouth-speech synchronization. Furthermore, we employ noise-resilient training to make our network robust to unseen non-stationary noise. We evaluate our system with image quality and mouth shape (landmark) measures on noisy speech utterances with five types of unseen non-stationary noise between -10 dB and 30 dB signal-to-noise ratio (SNR) with increments of 1 dB SNR. Results show that our system outperforms a state-of-the-art baseline system significantly, and our noise-resilient training improves performance for noisy speech in a wide range of SNR.

Index Terms— generative models, talking face from speech, speech animation, generative adversarial networks

1. INTRODUCTION

Visual cues provided in a talking face are important in speech communication. They help improve speech comprehension for the hearing impaired population and when the acoustic signal is corrupted by channel distortion or background noise [1–4]. When not available, automatically generating a talking face from speech will thus have the potential to improve speech comprehension and communication. It will facilitate the hearing impaired population to access abundantly available audio-books, podcasts, and educational talks online beyond reading text or captions. It will also find applications in areas such as staff training, virtual assistants, video games, animated movies, and human-machine interfaces.

These motivations have driven researchers to develop systems that can generate talking faces from speech [5–10]. Among these work, only [10, 11] are end-to-end, while the others include pre- and/or post-processing. Similar to other audio processing tasks, end-to-end approaches often learn better representations of the audio signal and yield improved results compared to hand-crafted features such as Mel Frequency Cepstral Coefficients (MFCCs). In addition, most of these methods were not designed for or evaluated in noisy conditions, while noise resilience is critical in practice.

In this paper, we propose a system that can generate a talking face video with a frame rate of 25 frames-per-second (FPS) from an arbitrarily long noisy speech utterance and a single face image. The speech utterance and the face image do not need to belong to the same person, and neither identity is exposed to the system beforehand. Compared to existing work, our contributions are: 1) We implement this system in a Generative Adversarial Network (GAN) framework and propose a novel *pair discriminator* to ensure the match between the generated mouth shape and the speech utterance for each individual frame, a *frame discriminator* to ensure the image quality of the generated talking face, in addition to an L_1 reconstruction loss during training. 2) We apply a mouth region mask to guide the reconstruction loss and pair discriminator to focus on the mouth region for better image quality and mouth-speech match. 3) Following our prior work [12], we employ noise-resilient training in this end-to-end system to improve its robustness against background noise in the speech input.

Experiments on two different datasets show that our system performs better than a state-of-the-art baseline in terms of image quality and mouth-speech synchronization. The evaluation against five unseen types of non-stationary noise also shows that our noise-resilient training improves face generation performance for a wide range (-10 dB to 30 dB) of Signal-to-Noise Ratio (SNR). Generated samples can be found here[1]. We recognize the societal risks of this technique, and to prevent it being misused, source code and pre-trained models are only available for research purposes upon request.

2. RELATED WORK

Suwajanakorn et al. [5] proposed a system that can generate videos of President Barack Obama from his speech. In the first stage, a long short-term memory (LSTM) network predicts the PCA coefficients of the mouth landmarks from speech features (13 MFCCs plus the energy). In the second stage, the system retrieves the texture according to the predicted PCA coefficients by selecting a few nearest candidate frames from the dataset that contains the images of the target identity. The candidate frames are stitched together by applying the weighted mean. This method works for a single person and requires a substantial amount of data to render realistic visuals.

Chung et al. [6] proposed a network to generate talking faces from a single reference face image and MFCC features of an utterance from an unheard speaker. The generated images are blurry, and the authors trained a separate deblurring module (post-processing) to sharpen the images. To eliminate the post-processing requirement, Chen et al. [7] proposed a lip generation system with an adversarial loss function to sharpen the blurry outputs in addition to a reconstruction loss. However, this system can only generate the lip region rather than the entire face.

In a follow up work, Chen et al. [8] proposed a method that can generate an entire face. Similar to [5], a two-stage approach is adopted: speech is first converted to face landmarks using an LSTM

*This work is funded by National Science Foundation grant No. 1741472.

[1] http://www.ece.rochester.edu/projects/air/projects/end2endtface.html

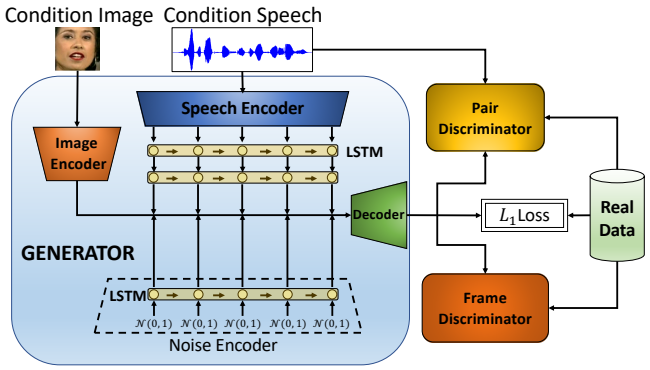

Fig. 1. Overview of the proposed system. The network takes a reference face image (condition image), a speech waveform, and a set of random vector from the standard normal distribution (noise) as input and generates a face video "speaking" the speech in an online fashion. During training, an image reconstruction loss as well as two adversarial losses for improving the mouth-shape-speech synchronization and image quality are adopted.

network, and another neural network then converts these landmarks to face images conditioned on a single reference face image and its landmarks. Experiments show that this method yields better results than [6, 7]. Therefore, we choose it as our baseline.

While not the focus of this paper, generating talking faces with natural movements such as blinking and gestures has drawn attention recently. There are three notable methods [9, 10, 13]. They use recurrent adversarial networks (temporal discriminator) to add natural movements. It is worth to note that this temporal discriminator can be employed in our system to generate natural movements.

3. PROPOSED METHOD

Figure 1 shows the system overview. Specifically, the generator contains an image encoder, a speech encoder, and a noise encoder. It concatenates feature outputs from the encoders and uses a decoder to generate a talking face video. We utilize generative adversarial networks [14] in this system; specifically, we propose two discriminators. The *pair discriminator* evaluates the match between the speech input and the generated face video, while the *frame discriminator* evaluates the validity of every single frame of the video. In the following, we describe each module in detail.

3.1. Network Architecture

Speech Encoder: In this work, we aim to generate 25 FPS videos. Therefore, we designed the speech encoder to take an arbitrary length speech waveform, without any pre-processing, to output 25 feature vectors per second. We realize this by employing five 1-D convolutional layers operating in the time domain. The number of filters, filter sizes and strides for these convolutional layers are (64, 63, 4), (128, 31, 4), (256, 17, 2), (512, 9, 2), (16, 1, 1), respectively. The convolutional layers are followed by a context layer that concatenates the past (136 ms) and future (136 ms) output feature vectors for further processing. This introduces a 136 ms internal delay in a real-time deployment of the system. The context layer then keeps every fifth feature vector and discards the rest. Finally, the resulting features are fed to a fully connected layer. For 1 second of speech in 8 kHz, the input size is 8000, and after these convolutional layers, it is reduced to 125. The context layer further reduces

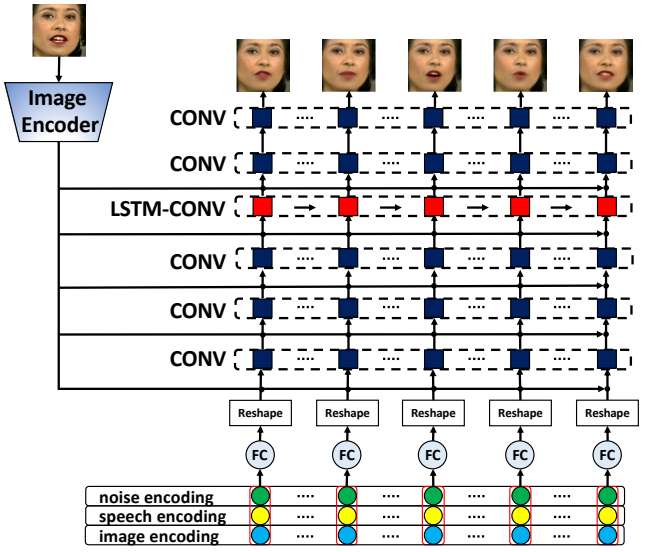

Fig. 2. The decoder architecture. The input is the speech, condition image, and noise features concatenated at each time-step. The features coming from the condition image are concatenated to each layer's output, except for the last layer.

the size to 25 by keeping only every fifth frame. The output of the speech encoder is processed further by two LSTM layers to extract temporal features.

Image Encoder: The image encoder takes a single reference face image as input to encode the target identity. It is a six-layer convolutional network. In each layer, we use padding and down-sampling to reduce the image dimensions by half. However, in the last layer, we do not use padding but a filter size of 4 to reduce the image to 1 in both height and width dimensions, resulting in a flat vector of 512 features along the channel dimension. The number of filters, filter sizes and down-sampling factors for these convolutional layers are (64, 3, 2), (128, 3, 2), (256, 3, 2), (512, 3, 2), (512, 3, 2), (512, 4, 1), respectively. Note that we do not use strides or pooling; instead, we use nearest-neighbor interpolation for down-sampling to avoid image artifacts.

The difference between our image encoder and that of other methods is that we downsample the input image and concatenate it to the next layer's input, meaning all layers have direct access to the input condition image or its down-sized version. This is for improving the gradient flow through the network.

Noise Encoder: In order to make the model robust to movements irrelevant to speech, and promote diverse mouth shapes, we follow [10] to add a noise input to the generator. We generate a 128-d Gaussian noise vector at every time-step with zero mean and unit standard deviation and pass it through an LSTM layer before passing it to the decoder. This noise input also enables the possibility of adding a temporal discriminator, described in Section 2, without changing the architecture.

Decoder: As shown in Figure 2, the decoder concatenates the speech, image, and noise features calculated by their encoders, and feeds them to a fully connected (FC) layer. The resultant vectors are then reshaped to 2D images. They are then passed through 6 convolutional layers that are symmetric to the 6 layers in the image encoder. Before going through each of the first 5 layers, the output of the corresponding layer of the image encoder is concatenated as input to form U-Net [15] style skip connections. To make the gen-

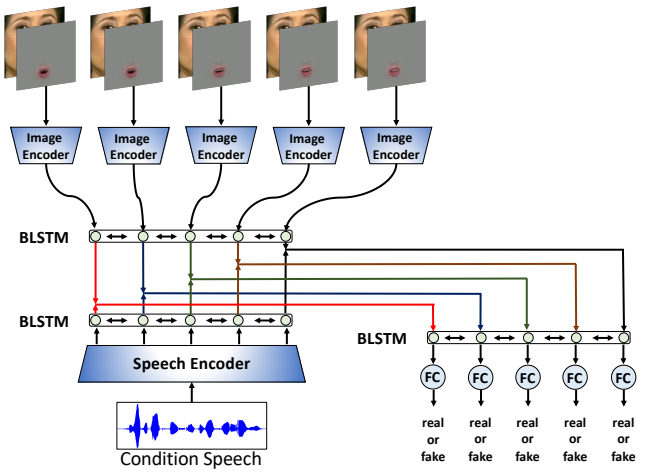

Fig. 3. The pair discriminator architecture. The input is the speech, the condition image, and generated/ground-truth videos.

erated mouth shapes vary smoothly over time, we add LSTM type recurrent connections to the 4th layer, making it an LSTM-CONV layer. The number of filters, filter sizes and up-sampling factors for these convolutional layers are (512, 3, 2), (256, 3, 2), (128, 3, 2), (64, 3, 2), (32, 3, 2), and (3, 3, 1), respectively. Note that we do not use transposed convolutions or strides; instead, we use nearest-neighbor interpolation for up-sampling to avoid image artifacts.

Frame Discriminator: The frame discriminator takes individual video frames (ground-truth or generated) along with the condition image and outputs a probability to determine if the video frame is fake or real. The network contains five layers of convolutional layers followed by 2 FC layers. The number of filters, filter sizes and up-sampling factors for these convolutional layers are (64, 3, 1), (64, 3, 2), (128, 3, 2), (256, 3, 2), and (512, 3, 2), respectively. The fully connected layers outputs 2048 units and 1 unit, respectively. For each batch, we select random indexes from the generated videos instead of applying this discriminator to every frame to improve the training speed. The number of random indexes is equal to the batch size for each video.

Pair Discriminator: The pair discriminator's objective is to improve the synchronization between the mouth shape and the input speech in the generated videos. As shown in Figure 3, its image encoder takes the generated or ground-truth videos that are mouth region masked (Section 3.2) and the condition image as input. It uses the same speech encoder architecture described earlier; however, the weights are not shared. A bidirectional LSTM (BLSTM) processes the output of the speech encoder, where another BLSTM processes the output of the image encoder for each frame. The outputs of the BLSTM's are then concatenated and fed into a third BLSTM followed by an FC layer that outputs the probability score. The hidden neurons for the image, speech and third BLSTMs are 512, 512, and 1024, respectively.

3.2. Loss Functions

We employ both a reconstruction loss and adversarial losses during training. We follow the previous works [7, 8] and choose L_1 loss as our reconstruction loss. For adversarial losses, we use the least squares GAN [16] for both frame and pair discriminators. We find that it yields more stable training compared to the vanilla GAN loss.

However, applying these losses to every region of the face yields blurry results due to speech irrelevant movements of certain parts of the face. To solve this problem, we propose to mask out the regions except for the mouth region before calculating the reconstruction loss and pair-discriminator. To obtain this mask, we run a face landmark estimation method [17] on each video frame, locate the mouth landmarks, and calculate the mean of these points. We put a 2D Gaussian centered at the mean of the mouth landmarks. The mask size is fixed, and we empirically calibrate it on the validation set. Make this mask adaptive to the input frame is our future work. To make the training more stable, we add a constant (0.01) to this mask so that the other regions are not fully ignored by the reconstruction loss. An example of the mouth-region mask is shown in Figure 4.

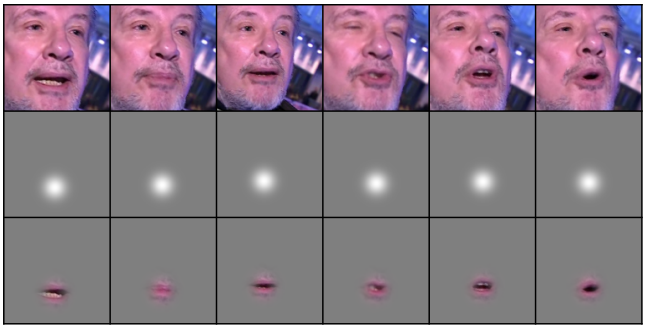

Fig. 4. Examples of mouth region masking during training, showing video frames (first row), 2D Gaussian masks (center row), and masked video frames (bottom row).

Furthermore, to improve the robustness against unseen non-stationary background noises, we include a feature-level loss function described in [12]. First, the clean speech is fed to the network, and the output of the speech encoder is obtained. The same step is applied to the noisy version of the same speech. We calculate the mean-squared error between clean and noisy speech features as the noise-resilience loss.

Therefore, the final loss function for the generator is:

$$J_{GEN} = L_1^{masked} + \tilde{L}_1^{masked} + \alpha J_{FD} + \beta J_{PD} \ , \qquad (1)$$

where L_1^{masked} and $\tilde{L}_1^{masked}$ are the masked reconstruction loss for clean and noisy speech, J_{FD} is the frame discriminator's cost, J_{PD} is the pair discriminator's cost, α and β are their weights.

4. EXPERIMENTS

4.1. Dataset

In our experiments, we used the Lip Reading in the Wild (LRW) dataset [18]. The dataset contains short duration videos (1.16 s) of a single speaker uttering a single word. There are a total of 500 words in the dataset, and for each word, there are 1000 videos spoken by different people. The videos are provided in 25 FPS. We detected the face landmarks for each frame and registered each image to a template image by using a similarity transform using three points: the mean of the left eye, right eye, and nose landmarks. This way, in all images, the eye and nose locations are aligned. The final size of the images is 128 by 128 pixels. We followed the same train (90%), validation (5%), and test (5%) splits as [18]. For evaluation, we randomly selected 1000 samples from the test set. To measure the generalization capability, we evaluated our method against 1000 randomly selected samples from the Grid dataset [19].

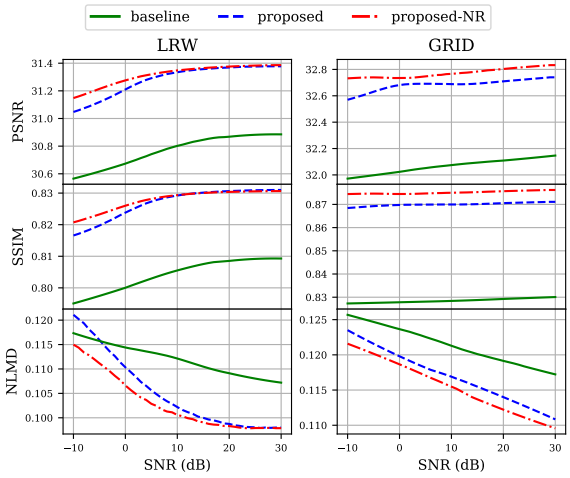

Fig. 5. Average results of the *baseline* [8], our *proposed*, and our proposed noise-resilient (*proposed-NR*) on speech utterances from LRW and Grid datasets mixed with five types of unseen noises with 1 dB SNR increments between -10 dB and 30 dB. For PSNR and SSIM, higher is better; for NLMD, lower is better.

For noise-resilient training, we used 138 types of non-stationary noise obtained from the Sound Ideas [20] corpus. We mixed the speech utterances with these noises randomly during training between -6 dB to 30 dB SNR with 3 dB increments. For evaluation, we selected five types of noise that were not included in the training set, namely babble, cafeteria, motorcycle, speech-shaped, and factory noises. We mixed speech utterances with these noises between -10 dB to 30 dB SNR with 1 dB increments.

4.2. Training Details

We implemented our network with PyTorch. We used Adam optimizer with β_1 parameter of 0.5. For the generator parameters, we set learning rate to 1e-04, and for the discriminators, we set learning rate to 1e-05 for stable training. The α and β parameters described in Section 3.2 were set to 1e-03. We first trained the network with only the reconstruction losses and added the discriminator losses after it converged. The total training time was approximately one week on a single NVIDIA GTX 1080 TI GPU.

4.3. Results

We employ Peak SNR (PSNR) and Structural Similarity (SSIM) [21] to evaluate the image quality of the generated videos. In order to evaluate the mouth synchronization, similar to [7], we extract the face landmarks of the generated and ground-truth videos and calculate their L_2 distance. Differently, however, we align the predicted and ground-truth landmarks to a template face landmarks for each frame using Procrustes analysis [22] before calculating the distance between them. In [7], this alignment is performed by matching the mean points of the mouth landmarks only. Simply matching the means of the mouth landmarks, however, is prone to errors caused by facial movements such as rotation and scaling. Procrustes analysis removes the translation, rotation, and scaling factors without changing the relative positions of the landmarks, resulting in a more meaningful evaluation metric. We call this metric normalized landmarks distance (NLMD).

The results are shown in Figure 5. We can see that our method significantly outperforms the baseline method for all metrics, both on LRW and Grid datasets. This shows the advances of our method

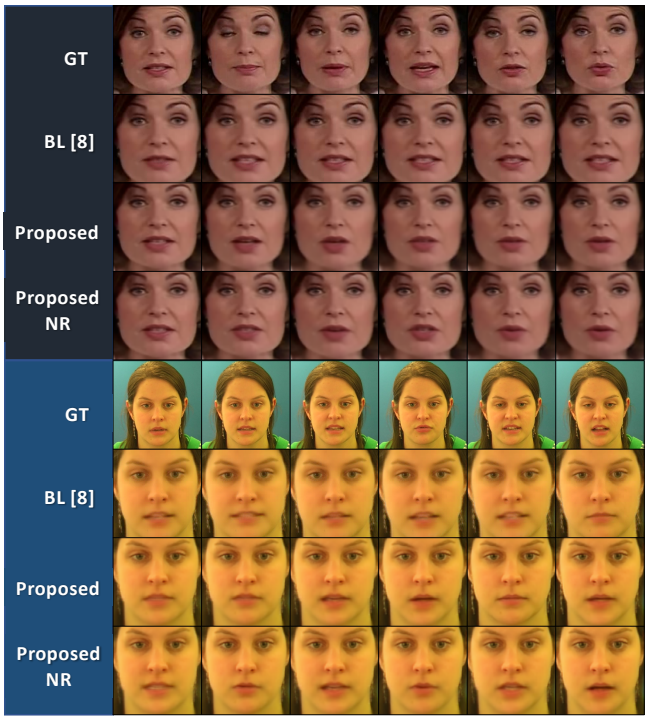

Fig. 6. Example ground-truth and generated talking faces from noisy speech utterances. The upper example is from LRW with speech-shaped noise in 0 dB SNR, and the lower example is from Grid with babble noise in 5 dB SNR. For each example, the four rows are corresponding video frames from the ground truth (GT), baseline model (BL) [8], the proposed base model with NR training (Proposed), and proposed model with NR training (Proposed-NR).

on both image quality and mouth shape validity. On the LRW dataset, the noise-resilient (NR) version of our model outperforms our base model without NR training when the SNR is lower than 10 dB. On the Grid dataset, this outperformance is across all SNRs. This demonstrates the noise resilience brought by the NR training strategy.

Figure 6 shows two examples obtained in challenging noise conditions. The first part shows an example from the LRW dataset with 0 dB SNR speech-shaped noise, where the second part shows an example from the Grid dataset with 5 dB SNR babble noise. Compared to the baseline and our *proposed* methods, our *proposed-NR* method is more robust to noise and yields the best mouth shape match with the ground truth (GT).

5. CONCLUSION

In this work, we proposed an end-to-end system for talking face generation from a noisy speech utterance of an unheard talker and a single image of an unseen face. We proposed a mouth region mask that can improve the mouth movements. The network utilizes two discriminators to improve image quality and mouth-speech synchronization. Experiments on noisy speech inputs showed that the proposed system outperforms a state-of-the-art baseline significantly on all evaluation metrics across a wide range of SNR. Experiments also showed that our noise-resilient training improves the noise resilience of the model to unseen non-stationary noise types. Future work includes a subjective study in the presence of background noise and an ablation study of the proposed system.

6. REFERENCES

[1] Carl A Binnie, "Bi-sensory articulation functions for normal hearing and sensorineural hearing loss patients," *Journal of the Academy of Rehabilitative Audiology*, vol. 6, no. 2, pp. 43–53, 1973.

[2] Karen S Helfer and Richard L Freyman, "The role of visual speech cues in reducing energetic and informational masking," *The Journal of the Acoustical Society of America*, vol. 117, no. 2, pp. 842–849, 2005.

[3] Joshua GW Bernstein and Ken W Grant, "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3358–3372, 2009.

[4] Ross K Maddox, Huriye Atilgan, Jennifer K Bizley, and Adrian KC Lee, "Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners," *eLife*, vol. 4, 2015.

[5] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 95, 2017.

[6] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman, "You said that?," in *British Machine Vision Conference*, 2017.

[7] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu, "Lip movements generation at a glance," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 520–535.

[8] Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] Yang Song, Jingwen Zhu, Dawei Li, Andy Wang, and Hairong Qi, "Talking face generation by conditional recurrent adversarial network," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 7 2019, pp. 919–925.

[10] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic, "End-to-end speech-driven facial animation with temporal gans," in *BMVC*, 2018.

[11] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic, "Realistic speech-driven facial animation with gans," *International Journal of Computer Vision*, Oct 2019.

[12] Sefik Emre Eskimez, Ross Maddox, and Zhiyao Duan, "Noise-resilient training method for face landmark generation from speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. PP, pp. 1–1, 10 2019.

[13] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 9299–9306.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[16] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2794–2802.

[17] Davis E King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[18] Joon Son Chung and Andrew Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 87–103.

[19] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[20] Sound-ideas.com, *https://www.sound-ideas.com/*, 2019.

[21] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al., "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[22] John C Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975.