



Investigating Contributions of Speech and Facial Landmarks for Talking Head Generation

Ege Kesim, Engin Erzin

KUIS-AI Lab, Koç University, Istanbul, Turkey

ekesim19@ku.edu.tr, eerzin@ku.edu.tr

Abstract

Talking head generation is an active research problem. It has been widely studied as a direct speech-to-video or two stage speech-to-landmarks-to-video mapping problem. In this study, our main motivation is to assess individual and joint contributions of the speech and facial landmarks to the talking head generation quality through a state-of-the-art generative adversarial network (GAN) architecture. Incorporating frame and sequence discriminators and a feature matching loss, we investigate performances of speech only, landmark only and joint speech and landmark driven talking head generation on the CREMA-D dataset. Objective evaluations using the peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) and landmark distance (LMD) indicate that while landmarks bring PSNR and SSIM improvements to the speech driven system, speech brings LMD improvement to the landmark driven system. Furthermore, feature matching is observed to improve the speech driven talking head generation models significantly.

Index Terms: talking head generation, speech driven animation

uated using the peak signal to noise ratio (PSNR), structural similarity index (SSIM) [2] and the landmark distance (LMD) [3]. Networks are able to capture emotion from speech or landmarks and synthesize facial expressions accordingly. In landmark models ground truth landmarks are used. The contribution of this work is two fold: i) We setup a common state-of-the-art GAN based architecture to assess contributions of speech only, landmark only and joint speech and landmark for the performance of talking head generation task. We expect to acquire a deeper insight on the driving factors talking head generation. ii) We experimentally show that even the ground truth landmarks are in use, speech signal keeps improving the LMD performance in the joint system, especially for the synthesis of mouth region.

The organization of the paper is as follows: First, related work on talking head generation is presented in section 2. Then, the proposed architectures are given in section 3. Section 4 introduces the dataset in use and presents experimental evaluations and discussion. Finally, section 5 concludes the paper.

1. Introduction

Generation of talking heads is important for the production of films, computer games and virtual avatars. Despite the recently developed methods there is still lack of fully automatic systems. Availability of this technology will ease animation production effort at lower costs and give an opportunity for amateur filmmakers and game developers to produce higher quality products in fast and cost effective cycles. Apart from maturing the talking head generation technology, it is also needed to understand deeper how speech, facial expressions and facial feature continuity are related to each other. So that various aspects, such as speaker identity, manner of facial expression articulation or affective states, can be modeled and controlled better.

In recent literature, talking head generation research focuses on speech driven facial synthesis. For this problem, two main approaches appear as a single or two stage solution. In the single stage solution, speech is directly mapped to video frames through deep generative models. On the other hand two stage approaches break the problem into two by first synthesizing facial landmarks from speech and then mapping facial landmarks to video frames for the talking head generation.

In this paper, we investigate individual and joint contributions of the speech and facial landmarks to the talking head generation quality through a state-of-the-art generative adversarial network (GAN) architecture. For this purpose, we construct three architectures addressing roles of speech only, landmark only and joint speech and landmark driven talking head generation. Model learning incorporates frame and sequence discriminators and a feature matching loss, runs on the CREMA-D dataset, which includes affective speech and facial expressions [1]. The proposed architecture settings are objectively eval-

2. Related Work

In recent literature, single stage solutions addressing the direct mapping from speech to talking head animations have been studied with increasing interest. One early study on single stage solution investigates a neural network architecture to generate talking faces based on the mel-frequency cepstral coefficient (MFCC) features of speech [4]. They use L_1 loss with the ground-truth that results blurry synthesized images at the output. They solve this problem by training a separate deblurring module and by post-processing the output of the first module to enhance the image quality.

Recent research focusing on GAN [5] architectures handles blurry output generation problems better. A conditional recurrent adversarial network architecture attempts to generate the talking face video directly from speech with accurate lip synchronization and with smooth video frame transitions [6]. They utilize three discriminators enhancing the realness of frames, video and lip movements. An end-to-end system with a similar multi discriminator setup has been studied in [7]. They use a synchronization discriminator, which classifies synchronization of a sequence of frames with the audio. The misaligned pairs are generated by shifting the audio frames. The synchronization discriminator is trained over the lower half of the face images, which do carry audio-visual synchrony.

Another recent end-to-end system for talking face generation from noisy speech has been studied with image quality and mouth-shape synchronization, which is attained by a mouth region mask loss (MRM) [8]. In their following work [9], an end-to-end talking face generation system receives a reference face image, a speech utterance, and a categorical emotion label to generate a talking face video in sync with the speech and

expressing the conditioned emotion. They discard the synchronization discriminator from their previous work and keep only the MRM loss for the mouth movements. They also add an emotion discriminator, which classifies the emotion and the realness of a video. Compared to [7], their system discards emotion expressed in the audio and only the emotion class is used as another input. Note that this may result in expressing non-matching emotions across speech and face.

Two-stage speech-to-landmark-to-video mapping approaches have also been studied in the literature. In an early two-stage approach, first mouth landmarks are generated from the audio, then mouth region texture sequences are generated from the landmarks and embedded on real videos [10]. Since this approach is subject-dependent, it is not easily generalizable to different subjects.

In a recent study, a conditional generative adversarial network is proposed to capture the emotion, lexical content and lip movements relationships [11]. They first use spectral and emotional speech features to condition the generation of realistic movements. Later, emotion-dependent facial synthesis is extracted by adapting the base model to the target emotion.

Another two-stage approach addresses talking head generation from MFCC features of audio to facial landmarks and then to facial animation [12]. The system is able to generate landmarks for the full face. They utilize an attention-based pixel-wise loss for landmark to image mapping. The system can also generate images of unseen identities.

In [13], a multi-stage system is proposed which can preserve the identity of the person in the synthesized videos. Based on DeepSpeech features [14] a network predicts person-independent facial landmarks, which are invariant to voice and accent of the speech. Then, the person-independent landmarks are re-targeted into person-specific landmarks. In an encoder-decoder based generator, person-specific landmarks and a reference image are transformed into the final facial image.

Another recent multi-stage expressive talking head generation system has been investigated for portrait images and 2D cartoon characters [15]. While predicting facial landmarks to capture the speaker-aware dynamics, their main model controls the lower face with the speech and facial expressions are set by the speaker information.

3. Proposed System

In this paper, we investigate and compare the performance of speech-driven and landmark-driven talking head generation systems based on a Generative Adversarial Network (GAN) architecture. The GAN architecture is configured with three different input settings: driven by only speech, only landmarks, and jointly by speech and landmarks. Block diagram of the proposed talking head generator for the speech and landmark-driven model is given in Figure 1. Note that the speech-only and landmark-only architectures are derivatives of the joint model by correspondingly eliminating the landmark or speech input models. The joint speech and landmark-driven architecture receives a single reference image and landmarks of the reference image just to initiate the generation process. The talking head generator is driven by the sequence of mel-spectrogram representation for the speech signal and by the sequence of facial landmarks. In the GAN architecture, frame and sequence discriminators are utilized to attain realistic synthesized videos. The frame discriminator evaluates the realness of the generated frames and the sequence discriminator evaluates the realness of the video. The generator is based on the U-Net architecture and has an encoder-

decoder architecture and skip connections between the image encoder and the decoder [16]. Due to the U-Net type of skip connections, each layer in the image encoder is connected to a layer in the decoder at the same level.

3.1. Image Encoder

The image encoder is built of 5 layers of CNNs, each consisting of convolution, instance normalization, and LeakyRelu layers in order. The number of filters, filter size, and down-sampling size of the layers are: (64,3,2), (128,3,2), (256,3,2), (256,3,2), (256,4,4). At the last layer of the image encoder, the output is an embedding vector with size 256. The image encoder receives the reference image for all models. On the other hand, for the landmark-only and the joint speech and landmark models, the image encoder also receives landmarks of the reference image and landmarks of a given time step that are concatenated with the reference image as heatmaps.

3.2. Speech Encoder

The speech encoder consists of cascaded CNN and LSTM networks. The CNN network receives mel-spectrogram features at each time step and outputs a speech embedding. It has 4 layers, and except for the last layer, each layer consists of convolution, instance normalization, and LeakyRelu layers. The last layer is a single convolution layer. After the CNN layers, the output is fed into a linear layer. Then it is fed into an LSTM network. The LSTM produces the final speech embedding vector. At each time step, the speech encoder receives features from the previous and the next 5 time steps to output the final speech embedding vector. The number of filters, filter size, and down-sampling size of the CNN layers are: (128,3,2), (128,3,2), (128,3,2), (128,3,2), (128,2,1). The output of the linear layer is a 128-dimensional vector and the hidden size of the LSTM is 256.

3.3. Noise Encoder

For every video, a 128-dimensional random noise vector is generated and fed into the LSTM-based noise encoder. The LSTM encoder outputs a 128-dimensional noise embedding at every time step.

3.4. Image Decoder

The first layer of the image decoder network is a linear layer and receives the concatenated audio, noise, and image embedding vectors. Then the output of the linear layer and the image encoder is concatenated as a skip connection. The concatenated vector is fed into 5 CNN layers that consist of up-sampling, convolution, instance normalization, and LeakyRelu layers as in the encoder. Rather than the transposed convolution, we use nearest neighbour interpolation for the up-sampling to avoid artifacts. Except for the first layer of the network, at each layer prior to the up-sampling layers, we use U-NET type skip connections. First, the output at the same level in the encoder and the output from the previous layer are concatenated. Then it is fed into the skip-connection layer. After the skip-connection layer, the output is fed into the up-sampling layer. Skip-connection layers consist of convolution, instance normalization, and LeakyRelu layers. At the last layer of the network, a Tanh layer is used as the activation function. The number of filters, filter size, and up-sampling size of the CNN layers are: (256,3,2), (256,3,2), (128,3,2), (64,3,2), (3,2,2). The number of filters and filter size of the skip-connection layers are: (256,3), (256,3), (128,3), (64,3,2).

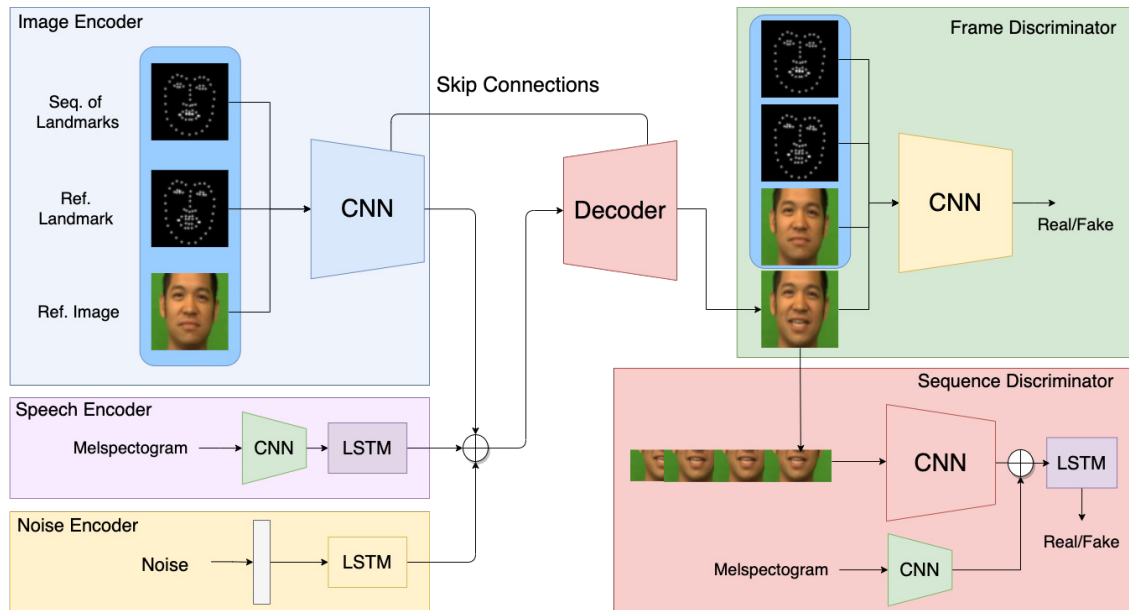


Figure 1: Block diagram of the generator for the speech and landmark driven model. The model takes a single reference image, landmarks of the reference image, speech and the sequence of landmarks.

3.5. Frame Discriminator

The frame discriminator takes a batch of real/fake images concatenated with the reference image. For landmark models target and reference landmark heatmaps are also concatenated with reference and real/fake image. All the concatenations are channel-wise while the landmarks are single channel in grayscale. The reference image conditions the discriminator helping to preserve the identity of the person in the generated frames. Architecture of the discriminator is similar to the image encoder. It has 5 layers of CNNs with a sigmoid activation function at the last layer. Frame discriminator takes randomly sampled image frames from a set of sampled videos and classifies them independently. Half of the frames of a sampled video are selected randomly for this purpose. The number of filters, filter size and down-sampling size of the layers are: (64,3,2), (128,3,2), (256,3,2), (512,3,2), (512,4,4).

3.6. Sequence Discriminator

Sequence discriminator takes a sequence of real/fake images and additionally speech mel-spectrogram features for the models with the speech encoder. Sequence discriminator ensures realistic movement of the lips and synchronization of the audio and the mouth. A four-layer CNN network takes all frames in the video, processes them separately and outputs an embedding for each image. Another CNN network takes the speech mel-spectrogram and outputs an audio embedding. At each time step, the image and audio embedding vectors are concatenated and then the resulting vector is fed into an LSTM followed by a linear layer. At every 5 frames, the sequence is classified as real or fake. To limit the data and ensure discriminator pay attention to the mouth region, we only use the lower half of the image as in [7]. The parameters of the CNN layers and linear layer for the mel-spectrogram is same with the speech encoder. The hidden size of the LSTM is 512. The number of filters, filter size and down-sampling size of the CNN layers are: (64,3,2), (128,3,2), (128,3,2), (128,3,2).

3.7. Loss Functions

Generator maximizes the realness score determined by the discriminators. Other than discriminator losses, we also use pair wise feature matching loss, which improves quality of the generated images, preservation of the identity and the mouth synchronization [17]. Since the generated and ground-truth frames are not pixel-wise identical and faces can appear in different locations, a simple L2 norm loss can not help the training. However, extracting features using the frame discriminator and calculating the L2 norm on these features solve the face localization problem. Feature matching loss is calculated using features coming from all layers of the frame discriminator except the last layer, which outputs the binary class prediction. The feature matching loss can be written as

$$L_{FM} = \sum_{i=1}^4 \frac{1}{N} \sum_{j=1}^N \|\phi_i(Q_{gt}) - \phi_i(Q_{syn})\|^2, \quad (1)$$

where Q_{gt} denotes ground-truth frames, Q_{syn} denotes synthesized frames ϕ_i is the extracted feature vector at the i^{th} layers and N denotes the number of sampled frames. Then, the full objective loss function for the generator is given as

$$L_{GEN} = L_{FD} + L_{SD} + \lambda L_{FM}, \quad (2)$$

where L_{FD} is the frame discriminator loss, L_{SD} is the sequence discriminator loss and L_{FM} is the feature matching loss. We set the feature matching weight as $\lambda = 100$ to attain blur-free generated images. Furthermore, both discriminators use binary cross entropy loss.

4. Experimental Evaluations

4.1. Dataset and Preprocessing

Experimental evaluations are executed on the CREMA-D dataset [1]. It consists of over 7000 videos of 91 (48 male/43 female) subjects. In each video, an actor/actress repeat one of the

12 different sentences in an acted emotion. In the CREMA-D, there are six emotion categories, which are angry, happy, sad, disgust, fear, and neutral. Each video clip consists of a single actor representing an emotion while saying a single sentence. Clip lengths vary from one to four seconds. The dataset is divided into three subject independent subsets as train, test and validation. The test set is selected same with the one in [7]. Similarly, we also use mirroring as augmentation by randomly selecting 50% of the videos to mirror.

Faces and landmarks in the video frames are detected using the Dlib [18]. A reference frame is set and faces are aligned with the reference frame. For the face alignment, landmark points corresponding to the eyes, nose and the point between the cheek and temple are used. Since the landmarks are estimated in each frame independently, landmark points oscillate across consecutive frames. This jitter is a problem both for the landmarks and the videos, since the transformation matrix for the face alignment is calculated from the landmarks. The ground-truth landmarks are extracted after a 5-point moving average filter. Furthermore, the face alignment transformations are computed after a second layer of 5-point Gaussian filtering on the smoothed ground-truth landmarks. Aligning the faces and decreasing the jitter on face orientation improve training of the generative models. Lastly, heatmaps of the ground-truth landmarks are extracted as grayscale images where each landmark is marked with a 2D Gaussian centered at the landmark point.

Audio is resampled at 22kHz and mel-spectrum features for 60 mel-frequency bands are extracted over 33 millisecond windows, which are sliding with the half window size.

4.2. Implementation

Batch size in the training is set as 1 to include a single video clip. Adam optimizer is used in training of all the networks. Learning rate is set to 1e-4 for the discriminators and 2e-4 for the generator. All the images are normalized into -1 to 1 value range and resized to 64x64 scale. Networks are trained for 100 epochs.

4.3. Comparison of models

We evaluate three models using PSNR, SSIM and LMD [3]. The landmark distance is computed across the landmarks of the generated and the ground-truth frames. First, the generated and the ground-truth frame landmarks are aligned to have the same mean. Then, LMD is computed as the average L_2 distance, normalized by the number of landmarks, over the mouth region landmarks.

Table 1 presents the objective metric results. The first two rows compare speech only driven talking head generation with and without the feature matching loss in the model training. Results indicate that in the absence of visual cues to drive the generation, the use of feature matching loss improves the generated image quality in terms of PSNR, SSIM and LMD measures. Improvement in the LMD indicates role of the FM loss in better maintaining the synchronization for the mouth region.

Landmark only driven model attains better performance than the speech only model. This is expected as we are using the ground truth landmarks to drive the talking head generation that sets an upper bound performance for the landmark driven systems. Furthermore, addition of the speech to the landmark based generation improves the PSNR and LMD measures further. This indicates an important observation on the role of speech signal which helps to improve landmark generation on the mouth region as well as improving the overall image quality

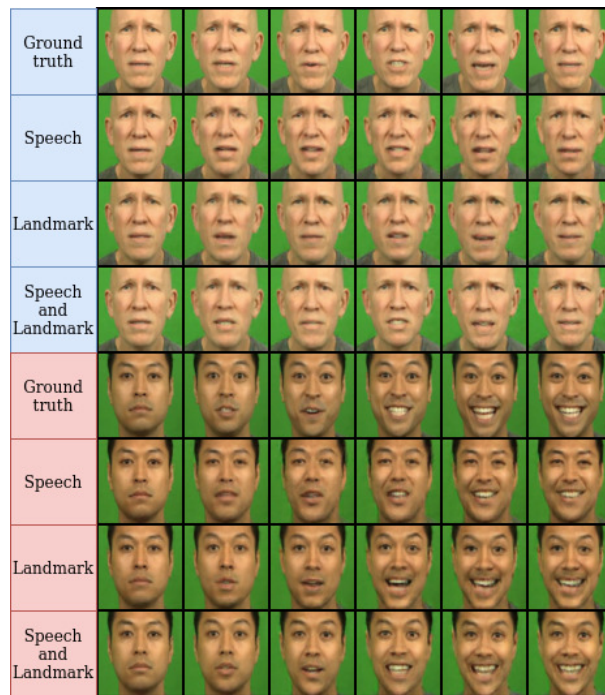


Figure 2: Sample video frames for two subjects. For each subject, the first row is the ground truth video. The following three rows are speech only, landmark only, and joint speech and landmark driven facial outputs. Each video starts with the same reference image, which is the first frame in each row. Every tenth frame is shown in each video.

in terms of PSNR.

Table 1: Objective evaluations of the talking head generation models

Model	PSNR	SSIM	LMD
Speech w/o FM loss	23.23	0.71	1.82
Speech w FM loss	24.96	0.77	1.36
Landmark	27.22	0.87	0.66
Speech and Landmark	27.36	0.87	0.59

5. Conclusions

In this study, we setup a talking head generation framework to assess contributions of the speech only, landmark only and joint speech and landmark inputs. Experimental studies show that even the ground truth landmarks are in use, speech signal keeps improving the PSNR and LMD metrics in the joint system, especially for the synthesis of mouth region.

6. References

- [1] 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.
- [2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity,"

- IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [3] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu, “Lip movements generation at a glance,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 520–535.
- [4] J. S. Chung, A. Jamaludin, and A. Zisserman, “You said that?” *arXiv preprint arXiv:1705.02966*, 2017.
- [5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *arXiv preprint arXiv:1406.2661*, 2014.
- [6] Y. Song, J. Zhu, D. Li, X. Wang, and H. Qi, “Talking face generation by conditional recurrent adversarial network,” *arXiv preprint arXiv:1804.04786*, 2018.
- [7] K. Vougioukas, S. Petridis, and M. Pantic, “Realistic speech-driven facial animation with GANs,” *International Journal of Computer Vision*, vol. 128, pp. 1398–1413, 2020.
- [8] 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.
- [9] S. E. Eskimez, Y. Zhang, and Z. Duan, “Speech driven talking face generation from a single image and an emotion condition,” *arXiv preprint arXiv:2008.03592*, 2020.
- [10] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, “Synthesizing obama: learning lip sync from audio,” *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–13, 2017.
- [11] N. Sadoughi and C. Busso, “Speech-driven expressive talking lips with conditional sequential generative adversarial networks,” *IEEE Transactions on Affective Computing*, 2019.
- [12] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, “Hierarchical cross-modal talking face generation with dynamic pixel-wise loss,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7832–7841.
- [13] S. Sinha, S. Biswas, and B. Bhowmick, “Identity-preserving realistic talking face generation,” *arXiv preprint arXiv:2005.12318*, 2020.
- [14] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, “Deep speech: Scaling up end-to-end speech recognition,” *arXiv preprint arXiv:1412.5567*, 2014.
- [15] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, “MakeItTalk: speaker-aware talking-head animation,” *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–15, 2020.
- [16] O. Ronneberger, P. Fischer, and T. 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.
- [17] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, “CVAE-GAN: fine-grained image generation through asymmetric training,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2745–2754.
- [18] D. E. King, “Dlib-ml: A machine learning toolkit,” *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.