End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs

Konstantinos Vougioukas
Imperial College London
Samsung AI Center
k.vougioukas@imperial.ac.uk

Stavros Petridis
Imperial College London
Samsung AI Center
stavros.petridis04@imperial.ac.uk

Maja Pantic
Imperial College London
Samsung AI Center
m.pantic@imperial.ac.uk

Abstract

Speech-driven facial animation is the process which uses speech signals to automatically synthesize a talking character. The majority of work in this domain creates a mapping from audio features to visual features. This often requires post-processing using computer graphics techniques to produce realistic albeit subject dependent results. We present a system for generating videos of a talking head, using a still image of a person and an audio clip containing speech, that does not rely on any handcrafted intermediate features. To the best of our knowledge, this is the first method capable of generating subject independent realistic videos directly from raw audio. Our method can generate videos which have (a) lip movements that are in sync with the audio and (b) natural facial expressions such as blinks and eyebrow movements. We achieve this by using a temporal GAN with 2 discriminators, which are capable of capturing different aspects of the video. The generated videos are evaluated based on their sharpness, reconstruction quality, and lip-reading accuracy. Finally, a user study is conducted, confirming that temporal GANs lead to more natural sequences than a static GAN-based approach.

1. Introduction

The problem of generating realistic talking heads is multi-faceted, requiring high-quality faces, lip movements synchronized with the audio, and plausible facial expressions. Such systems could simplify the film animation process through automatic generation from the voice acting and generating occluded parts of the face. Additionally, this technology can improve band-limited visual telecommunications by either generating the entire visual content based on the audio or filling in dropped frames.

The majority of research in this domain has focused on mapping audio features (e.g. MFCCs) to visual features (e.g. landmarks, visemes) and using computer graphics (CG) methods to generate realistic faces [7]. Some methods avoid the use of CG by selecting frames from a person-specific database and combining them to form a video [11].

Subject independent approaches have also been proposed that transform audio features to video frames [3] but there is still no method to directly transform raw audio to video. Furthermore, many methods restrict the problem to generating only the mouth. Even techniques that generate the entire face are primarily focused on obtaining realistic lip movements, and typically neglect the importance of generating natural facial expressions.

Figure 1: The proposed end-to-end face synthesis model, capable of producing realistic sequences of faces using one still image and an audio track containing speech. The generated sequences exhibit smoothness and natural expressions such as blinks and frowns.

Some methods generate frames based solely on present information [3], without taking into account the facial dynamics. This makes generating natural sequences, which are characterized by a seamless transition between frames, challenging. Some video generation methods have dealt with this problem by generating the entire sequence at once [13] or in small batches [10]. However, this introduces a lag in the generation process, prohibiting their use in real-time applications and requiring fixed length sequences for training.

We propose a temporal generative adversarial network (GAN), capable of generating a video of a talking head from an audio signal and a single still image [1] (see Fig. 1). First, our model captures the dynamics of the entire face producing not only synchronized mouth movements but also natural facial expressions, such as eyebrow raises, frowns and blinks. Facial gestures are very important since their absence is a telltale sign that can be used to detect synthesized videos [8]. Our model is able to produce such expressions thanks to the use of a sequence discriminator.

Secondly, our method is subject independent, does not rely on handcrafted audio or visual features, and requires no post-processing. To the best of our knowledge, this is the first end-to-end technique that generates talking faces directly from the raw audio waveform.

Evaluation is performed in a subject independent way on the GRID [4] and TCD TIMIT [6] datasets, where our model achieves truly natural results. We measure the image quality using popular reconstruction and sharpness metrics, and compare it to a non-temporal approach. Additionally, we propose

[1] Videos are available here: https://sites.google.com/view/facialsynthesis/home

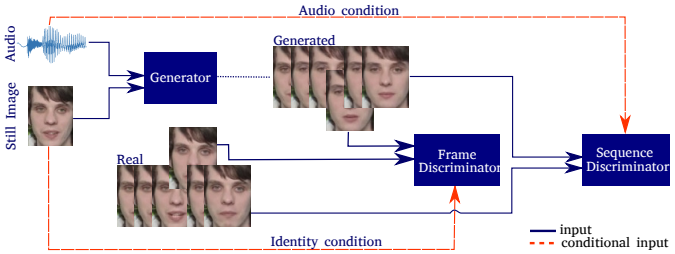


Figure 2: The deep model for speech-driven facial synthesis. This uses 2 discriminators to incorporate the different aspects of a realistic video. Details about the architecture are presented in the supplementary material.

using lip reading techniques to verify the accuracy of the spoken words and face verification to ensure that the identity of the speaker is maintained throughout the sequence.

2. End-to-End Speech-Driven Facial Synthesis

The proposed architecture for speech-driven facial synthesis is shown in Fig. 2. The system is made up of a generator and 2 discriminators, each of which evaluates the generated sequence from a different perspective. The capability of the generator to capture various aspects of natural sequences is directly proportional to the ability of each discriminator to discern videos based on them.

2.1. Generator

The inputs to the generator networks consist of a single image and an audio signal, which is divided into overlapping frames each corresponding to 0.16 seconds. The generator can be conceptually divided into subnetworks as shown in Fig. 3. Using an RNN-based generator allows us to synthesize videos frame-by-frame, which is necessary for real-time applications.

2.1.1 Identity Encoder

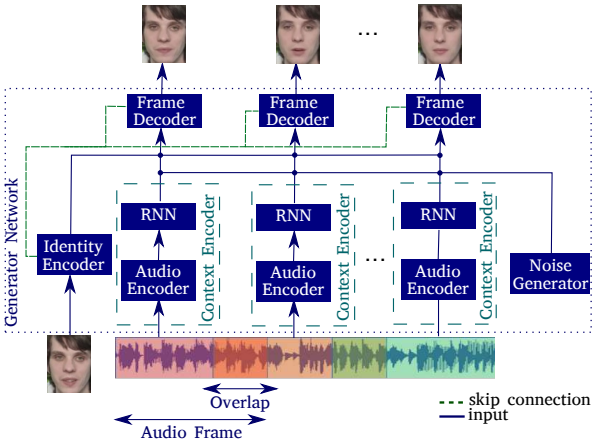
The speaker's identity is encoded using a 6 layer CNN. Each layer uses strided 2D convolutions, followed by batch normalization and ReLU activation functions. The *Identity Encoder* network reduces the input image to a 50 dimensional encoding z_{id}.

2.1.2 Context Encoder

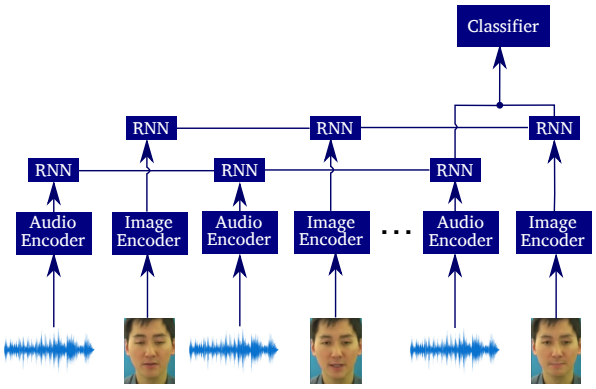
Audio frames are encoded using a network comprising of 1D convolutions followed by batch normalization and ReLU activations. The initial convolutional layer starts with a large kernel which helps limit the depth of the network while ensuring that the low-level features are meaningful. Subsequent layers use smaller kernels until an embedding of the desired size is achieved. The audio frame encodings are input into a 2 layer GRU, which produces a context encoding z_c with 256 elements.

2.1.3 Frame Decoder

The identity encoding z_{id} is concatenated to the context encoding z_c and a noise component z_n to form the latent representation. The 10-dimensional z_n vector is obtained from a *Noise Generator*, which is a 1-layer GRU that takes Gaussian noise as input. The *Frame Decoder* is a CNN that uses strided transposed convolutions to produce the video frames from the latent representation. A U-Net architecture is used with skip connections between the *Identity Encoder* and the *Frame Decoder* to help preserve the identity of the subject.



(a) Generator



(b) Sequence Discriminator

Figure 3: The architecture of the (a) Generator which consists of a *Context Encoder* (audio encoder and RNN), an *Identity Encoder*, a *Frame Decoder* and *Noise Generator* (b) Sequence Discriminator, consisting of an audio encoder, an image encoder, GRUs and a small classifier.

2.2. Discriminators

Our system has two different types of discriminator. The *Frame Discriminator* helps achieve a high-quality reconstruction of the speakers' face throughout the video. The *Sequence Discriminator* ensures that the frames form a cohesive video which exhibits natural movements and is synchronized with the audio.

2.2.1 Frame Discriminator

The *Frame Discriminator* is a 6-layer CNN that determines whether a frame is real or not. Adversarial training with this discriminator ensures that the generated frames are realistic. The original still frame is used as a condition in this network, concatenated channel-wise to the target frame to form the input as

shown in Fig. 3. This enforces the person's identity on the frame.

2.2.2 Sequence Discriminator

The *Sequence Discriminator* presented in Fig. 3 distinguishes between real and synthetic videos. The discriminator receives a frame at every time step, which is encoded using a CNN and then fed into a 2-layer GRU. A small (2-layer) classifier is used at the end of the sequence to determine if the sequence is real or not. The audio is added as a conditional input to the network, allowing this discriminator to classify speech-video pairs.

2.3. Training

The *Frame discriminator* (D_{img}) is trained on frames that are sampled uniformly from a video x using a sampling function $S(x)$. The *Sequence discriminator* (D_{seq}), classifies based on the entire sequence x and audio a. The loss for each discriminator contributes to the total loss shown in eq. 1.

$$\begin{aligned}\mathcal{L}_{adv}(D_{img}, D_{Seq}, G) = \mathbb{E}_{x\sim P_d}[\log D_{img}(S(x), x_1)] \\ + \mathbb{E}_{z\sim P_z}[\log(1 - D_{img}(S(G(z)), x_1))] \\ + \mathbb{E}_{x\sim P_d}[\log D_{seq}(x, a)] + \mathbb{E}_{z\sim P_z}[\log(1 - D_{seq}(G(z), a))]\end{aligned} \quad (1)$$

An L_1 reconstruction loss is also used to improve the synchronization of the mouth movements. However we only apply the reconstruction loss to the lower half of the image since it discourages the generation of facial expressions. For a ground truth frame F and a generated frame G with dimensions $W \times H$ the reconstruction loss at the pixel level is:

$$\mathcal{L}_{L_1} = \sum_{p\in[0,W]\times[\frac{H}{2},H]} |F_p - G_p| \quad (2)$$

The final objective is to obtain the optimal generator G^*, which satisfies eq. 3. The model is trained until no improvement is observed on the reconstruction metrics on the validation set for 10 epochs. The λ hyperparameter controls the contribution of each loss factor and was set to 400 following a tuning procedure on the validation set.

$$\arg\min_G \max_D \mathcal{L}_{adv} + \lambda\mathcal{L}_{L_1} \quad (3)$$

We used Adam for all the networks with a learning rate of 0.0002 for the *Generator* and 0.001 *Frame Discriminator* which decay after epoch 20 with a rate of 10%. The *Sequence Discriminator* uses a smaller fixed learning rate of $5 \cdot 10^{-5}$.

3. Experiments

3.1. Datasets

The GRID dataset has 33 speakers each uttering 1000 short phrases, containing 6 words taken from a limited dictionary. The TCD TIMIT dataset has 59 speakers uttering approximately 100 phonetically rich sentences each. We use the recommended data split for the TCD TIMIT dataset but exclude some of the test speakers and use them as a validation set. For the GRID dataset speakers are divided into training, validation and test sets with a $50\% - 20\% - 30\%$ split respectively. As part of our preprocessing all faces are aligned to the canonical face and images are normalized. We increase the size of the training set by mirroring the training videos.

3.2. Metrics

We use common reconstruction metrics such as the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) index to evaluate the generated videos. However, it is important to note that reconstruction metrics penalize videos for any spontaneous expression. Frame sharpness is evaluated using the cumulative probability blur detection (CPBD) measure [9], which determines blur based on the presence of edges in the image and the frequency domain blurriness measure (FDBM) proposed in [5], which is based on the spectrum of the image. For these metrics larger values imply better quality.

The content of the videos is evaluated based on how well the video captures identity of the target and on the accuracy of the spoken words. We verify the identity of the speaker using the average content distance (ACD) [12], which measures the average Euclidean distance of the still image representation, obtained using OpenFace [1], from the representation of the generated frames. The accuracy of the spoken message is measured using the word error rate (WER) achieved by a pre-trained lip-reading model (LipNet [2]). For both content metrics lower values indicate better accuracy.

3.3. Qualitative Results

Our method is capable of producing realistic videos of previously unseen faces and audio clips taken from the test set. The same audio used on different identities is shown in Fig. 4. From visual inspection it is evident that the lips are consistently moving similarly to the ground truth video. Our method not only produces accurate lip movements but also natural videos that display characteristic human expressions such as frowns and blinks, examples of which are shown in Fig. 5.

We compare our model to a static method that produces video frames using a sliding window of audio samples like that used in [3]. This is a GAN-based method that uses a combination of an L_1 loss and an adversarial loss on individual frames. We use this method as the baseline for our quantitative assessment in the following section. This baseline produces sequences characterized by jitter, which becomes worse in cases where the audio is silent. This is likely due to the fact that there are multiple mouth shapes that correspond to silence and since the model has no knowledge of its past state generates them at random.

3.4. Quantitative Results

We measure the performance of our model on the GRID and TCD TIMIT datasets using the metrics proposed in section 3.2 and compare it to the static baseline. Additionally, we present the results of a 30-person survey, where users were shown 30 videos from each method and were asked to pick the more natural ones. The results in Table 1 show that our method outperforms the static baseline in both frame quality and content accuracy. Although the difference in performance is slight for frame-based measures (e.g. PSNR, ACD) it is substantial in terms of user preference and lipreading WER, where temporal smoothness of the video and natural expressions play a significant role.

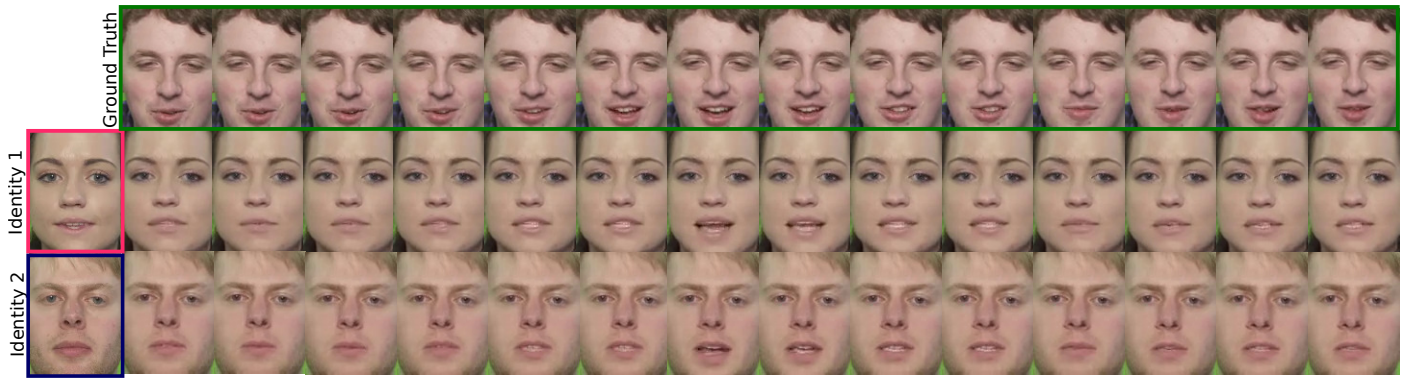
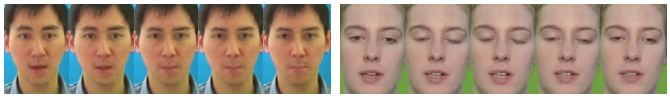


Figure 4: Animation of different faces using the same audio. The movement of the mouth is similar for both faces as well as for the ground truth sequence. Both audio and still image are unseen during training.

	Method	PSNR	SSIM	FDBM	CPBD	ACD	User	WER
GRID	Proposed Model	27.98	0.844	0.114	0.277	$1.02 \cdot 10^{-4}$	79.77%	25.4%
	Baseline	27.39	0.831	0.113	0.280	$1.07 \cdot 10^{-4}$	20.22%	37.2%
TCD	Proposed Model	23.54	0.697	0.102	0.253	$2.06 \cdot 10^{-4}$	77.03%	N/A
	Baseline	23.01	0.654	0.097	0.252	$2.29 \cdot 10^{-4}$	22.97%	N/A

Table 1: Performance comparison of the proposed method against the baseline. The pretrained LipNet model is not available for the TCD TIMIT so the WER metric is omitted.



(a) Example of generated frown (b) Example of generated blink

Figure 5: Facial expressions generated using our framework include (a) frowns and (b) blinks.

We further evaluate the realism of the generated videos through an online Turing test. In this test users are shown 10 videos randomly chosen from GRID and TCD-TIMIT databases and are asked to label them as real or fake. Responses from 316 users were collected with the average user labeling correctly 63% of the videos.

4. Conclusion and Future Work

In this work we have presented an end-to-end model using temporal GANs for speech-driven facial animation. Our method produces more coherent sequences and more accurate mouth movements compared to the static approach and also produces facial expressions like blinks and frowns. We believe that these improvements are not only a result of using a temporal generator but also due to the use of the conditional *Sequence Discriminator* which encourages spontaneous facial gestures. Moving forward, we would like to capture and reflect the mood of the speaker in the facial expressions.

References

[1] B. Amos, B. Ludwiczuk, and M. Satyanarayanan. OpenFace: A general-purpose face recognition library with mobile applications. Technical Report 118, 2016. 3

[2] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas. LipNet: End-to-End Sentence-level Lipreading. *arXiv preprint arXiv:1611.01599*, 2016. 3

[3] J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? In *BMVC*, pages 1–12, 2017. 1, 3

[4] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006. 1

[5] K. De and V. Masilamani. Image Sharpness Measure for Blurred Images in Frequency Domain. *Procedia Engineering*, 64:149–158, 1 2013. 3

[6] N. Harte and E. Gillen. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615, 2015. 1

[7] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM TOG*, 36(94), 2017. 1

[8] Y. Li, M. Chang, and S. Lyu. In Ictu Oculi : Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. *arXiv preprint arXiv:1806.02877*, 2018. 1

[9] N. D. Narvekar and L. J. Karam. A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection. *International Workshop on Quality of Multimedia Experience (QoMEx)*, 20(9):87–91, 2009. 3

[10] M. Saito, E. Matsumoto, and S. Saito. Temporal Generative Adversarial Nets with Singular Value Clipping. In *ICCV*, pages 2830–2839, 2017. 1

[11] S. Suwajanakorn, S. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing Obama: Learning Lip Sync from Audio Output Obama Video. *ACM TOG*, 36(95), 2017. 1

[12] S. Tulyakov, M. Liu, X. Yang, and J. Kautz. MoCoGAN: Decomposing Motion and Content for Video Generation. *arXiv preprint arXiv:1707.04993*, 2017. 3

[13] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating Videos with Scene Dynamics. In *NIPS*, pages 613–621, 2016. 1