arXiv:1805.09313v1 [eess.AS] 23 May 2018

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

Konstantinos Vougioukas[1], Stavros Petridis[1], and Maja Pantic[1]

[1]iBUG Group, Imperial College London

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 doesn't 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 [1]. We achieve this by using a temporal GAN with 2 discriminators, which are capable of capturing different aspects of the video. The effect of each component in our system is quantified through an ablation study. 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.



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.

1 Introduction

Facial animation plays a major role in computer generated imagery because the face is the primary outlet of information. The problem of generating realistic talking heads is multifaceted, requiring high-quality

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

faces, lip movements synchronized with the audio, and plausible facial expressions. This is especially challenging because humans are adept at picking up subtle abnormalities in facial motion and audio-visual synchronization.

Of particular interest is speech-driven facial animation since speech acoustics are highly correlated with facial movements [26]. These systems could simplify the film animation process through automatic generation from the voice acting. They can also be applied in movie dubbing to achieve better lip-syncing results. Moreover, they can be used to generate parts of the face that are occluded or missing in a scene. Finally, 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 [12]. Some methods avoid the use of CG by selecting frames from a person-specific database and combining them to form a video [3, 21]. Regardless of which approach is used these methods are subject dependent and are often associated with a considerable overhead when transferring to new speakers.

Subject independent approaches have been proposed that transform audio features to video frames [5] 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.

Some methods generate frames based solely on present information [5], 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 [24] or in small batches [19]. 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 (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. This is due to the use of an RNN-based generator and sequence discriminator, which also gives us the added advantage of handling variable length sequences.

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.

The third contribution of this paper is a comprehensive assessment of the performance of the proposed method. An ablation study is performed on the model in order to quantify the effect of each component in the system. We measure the image quality using popular reconstruction and sharpness metrics, and compare it to a non-temporal approach. Additionally, we propose 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. Evaluation is performed in a subject independent way on the GRID [6] and TCD TIMIT [11] datasets, where our model achieves truly natural results.

2 Related Work

2.1 Speech-Driven Facial Animation

The problem of speech-driven video synthesis is not new in computer vision and has been the subject of interest for decades. Yehia *et al.* [26] first examined the relationship between acoustics, vocal-tract and facial motion, showing a strong correlation between visual and audio features and a weak coupling between head motion and the fundamental frequency of the speech signal [27].

Some of the earliest methods for facial animation relied on hidden Markov models (HMMs) to capture the dynamics of the video and speech sequences. Simons and Cox [20] used vector quantization to achieve a compact representation of video and audio features, which were used as the states for their HMM. The HMM was used to recover the most likely mouth shapes for a speech signal. A similar approach is used in [25] to estimate the sequence of lip parameters. Finally, the *Video Rewrite* [3] method relies on the same principals to obtain a sequence of triphones that are used to look up mouth images from a database.

Although HMMs were initially preferred to neural networks due to their explicit breakdown of speech into intuitive states, recent advances in deep learning have resulted in neural networks being used in most modern approaches. Like past attempts, most of these methods aim at performing a feature-to-feature translation. A typical example of this is [22], which uses a deep neural network (DNN) to transform a phoneme sequence into a sequence of shapes for the lower half of the face. Using phonemes instead of raw audio ensures that the method is subject independent but fails to capture co-articulation effects. Similar deep architectures based on recurrent neural networks (RNNs) have been proposed in [9, 21], producing realistic results but are subject dependent and require retraining or re-targeting steps to adapt to new faces.

Convolutional neural networks (CNN) are used in [12] to transform audio features to 3D meshes of a specific person. This system is conceptually broken into sub-networks responsible for capturing articulation dynamics and estimating the 3D points of the mesh. Finally, Chung *et al.* [5] proposed a CNN applied on Mel-frequency cepstral coefficients (MFCCs) that generates subject independent videos from an audio clip and a still frame. The method uses an L_1 loss at the pixel level resulting in blurry frames, which is why a deblurring step is also required. Secondly, this loss at the pixel level penalizes any deviation from the target video during training, providing no incentive for the model to produce spontaneous expressions and resulting in faces that are mostly static except for the mouth.

2.2 GAN-Based Video Synthesis

The recent introduction of GANs in [10] has shifted the focus of the machine learning community to generative modelling. GANs consist of two competing networks: a generative network and a discriminative network. The generator's goal is to produce realistic samples and the discriminator's goal is to distinguish between the real and generated samples. This competition eventually drives the generator to produce highly realistic samples. GANs are typically associated with image generation since the adversarial loss produces sharper, more detailed images compared to L_1 and L_2 losses. However, GANs are not limited to these applications and can be extended to handle videos [15, 14, 24, 23].

Straight-forward adaptations of GANs for videos are proposed in [24, 19], replacing the 2D convolutional layers with 3D convolutional layers. Using 3D convolutions in the generator and discriminator networks is able to capture temporal dependencies but requires fixed length videos. This limitation can be overcome in [19] but constraints need to be imposed in the latent space to generate consistent videos.

The *MoCoGAN* system proposed in [23] uses an RNN-based generator, with separate latent spaces for motion and content. This relies on the empirical evidence shown in [17] that GANs perform better when the latent space is disentangled. *MoCoGAN* uses a 2D and 3D CNN discriminator to judge frames and sequences respectively. A sliding window approach is used so that the 3D CNN discriminator can handle variable length sequences.

GANs have also been used in a variety of cross-modal applications, including text-to-video and audio-to-video. The text-to-video model proposed in [14] uses a combination of variational auto encoders (VAE) and GANs in its generating network and a 3D CNN as a sequence discriminator. Finally, Chen *et al.* [4] propose a GAN-based encoder-decoder architecture that uses CNNs in order to convert audio spectrograms to frames and vice versa.

3 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.

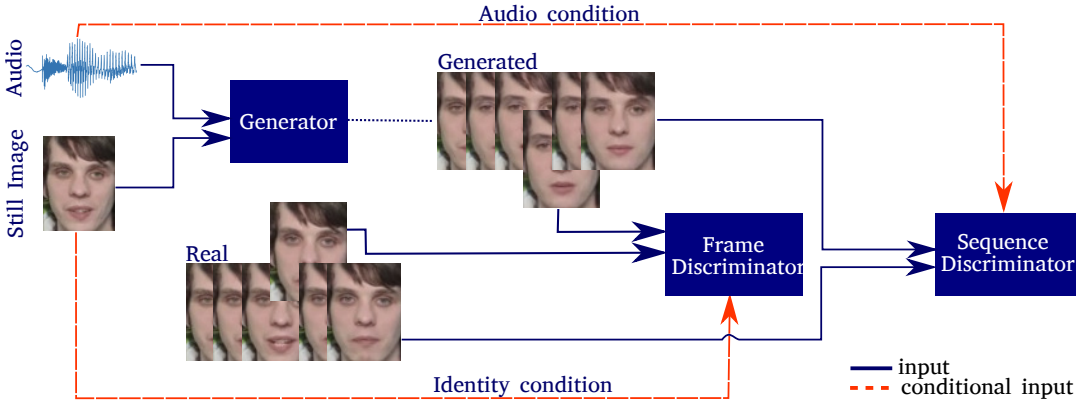


Figure 2: The deep model for speech-driven facial synthesis. This uses 2 discriminators to incorporate the different aspects of a realistic video. Details about the architecture are presented in the supplementary material.

3.1 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 network in this architecture can be conceptually divided into subnetworks as shown in Fig. 3.

3.1.1 Identity Encoder

The speaker's identity is encoded using is 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}.

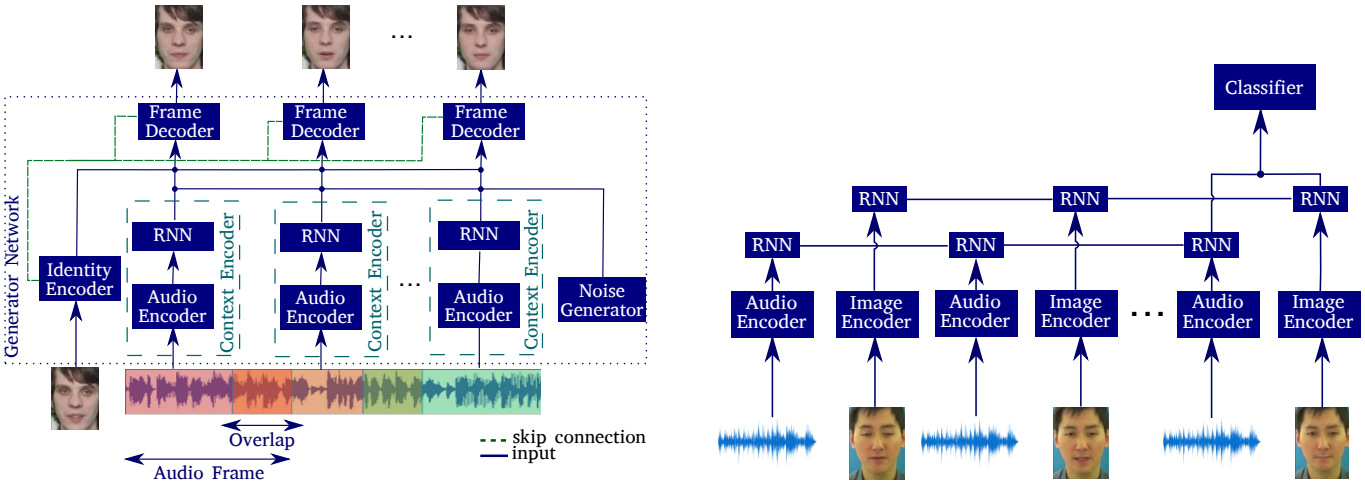
3.1.2 Context Encoder

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, as recommended in [7], 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.

3.1.3 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 [18] 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.

3.2 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.

3.2.1 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.

3.2.2 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.

3.3 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 of our GAN is an aggregate of the losses associated with each discriminator as shown

in Equation 1.

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

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:

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

The final objective is to obtain the optimal generator G^*, which satisfies Equation 3. The model is trained until no improvement is observed (PSNR, SSIM) on the validation set for 10 epochs. The λ hyperparameter controls the contribution of each loss factor and was set to 400 after tuning on validation sets.

$$\arg\min_G \max_D \mathcal{L}_{adv} + \lambda\mathcal{L}_{L_1} \tag{3}$$

We used Adam [13] 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 $5 \cdot 10^{-5}$.

4 Experiments

4.1 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. The amount of data used for training and testing is presented in Table 1.

Dataset	Samples (Tr)	Hours (Tr)	Samples (V)	Hours (V)	Samples (T)	Hours (T)
GRID	31639	26.36	6999	5.83	9976	8.31
TCD	8218	9.13	686	0.76	977	1.09

Table 1: The samples and hours of video in the training (Tr), validation (V) and test (T) sets.

4.2 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. During the evaluation it is important to take into account the fact that reconstruction metrics penalize videos for any spontaneous expression. The frame sharpness is evaluated using the cumulative probability blur detection (CPBD) measure [16], which

determines blur based on the presence of edges in the image and the frequency domain blurriness measure (FDBM) proposed in [8], 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) [23], 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. We use the LipNet model [2], which surpasses the performance of human lipreaders on the GRID dataset. For both content metrics lower values indicate better accuracy.

4.3 Ablation Study

In order to quantify the effect of each component of our system we perform an ablation study on the GRID dataset (see Table 2). We use the metrics from section 4.2 and a pre-trained LipNet model which achieves a WER of 21.4% on the ground truth videos. The average value of the ACD for ground truth videos of the same person is $0.74 \cdot 10^{-4}$ whereas for different speakers it is $1.4 \cdot 10^{-3}$. The L_1 loss achieves slightly better PSNR and SSIM results, which is expected as it does not generate spontaneous expressions, which are penalized by these metrics unless they happen to coincide with those in ground truth videos. This variation introduced when generating expressions is likely the reason for the small increase in ACD. The blurriness is minimized when using the adversarial loss as indicated by the higher FDBM and CPBD scores and Fig. 4. Finally, the effect of the sequence discriminator is shown in the lip-reading result achieving a low WER.

Method	PSNR	SSIM	FDBM	CPBD	ACD	WER
Ground Truth Videos	N/A	N/A	0.121	0.281	$0.74 \cdot 10^{-4}$	21.4%
L_1 loss	28.47	0.859	0.101	0.210	$0.90 \cdot 10^{-4}$	27.90%
$L_1 + Adv_{img}$	27.71	0.840	0.114	0.274	$1.04 \cdot 10^{-4}$	27.94%
$L_1 + Adv_{img} + Adv_{seq}$	27.98	0.844	0.114	0.277	$1.02 \cdot 10^{-4}$	**25.45**%

Table 2: Assessment of each model in the ablation study performed on the GRID dataset



(a) L_1 loss on entire frame

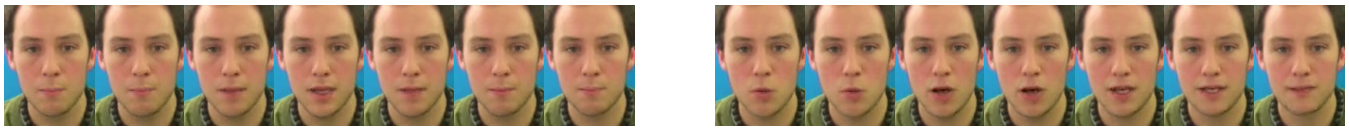
(b) Proposed loss on frames

Figure 4: Frames using (a) only an L_1 loss on the entire face compared to (b) frames produced using the proposed method. Frames generated using an L_1 loss on the entire face (a) are blurrier than those produced from the proposed method (b).

4.4 Qualitative Results

Our method is capable of producing realistic videos of previously unseen faces and audio clips taken from the test set. The examples in Fig. 5 show the same face animated using sentences from different subjects (male and female). The same audio used on different identities is shown in Fig. 6. 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. 7.



(a) Female voice uttering the word "bin" (b) Male voice uttering the word "white"

Figure 5: Generated sequences for (a) the word "bin" (b) the word "white" from the GRID test set. Coarticulation is evident in (a) where "bin" is followed by the word "blue".

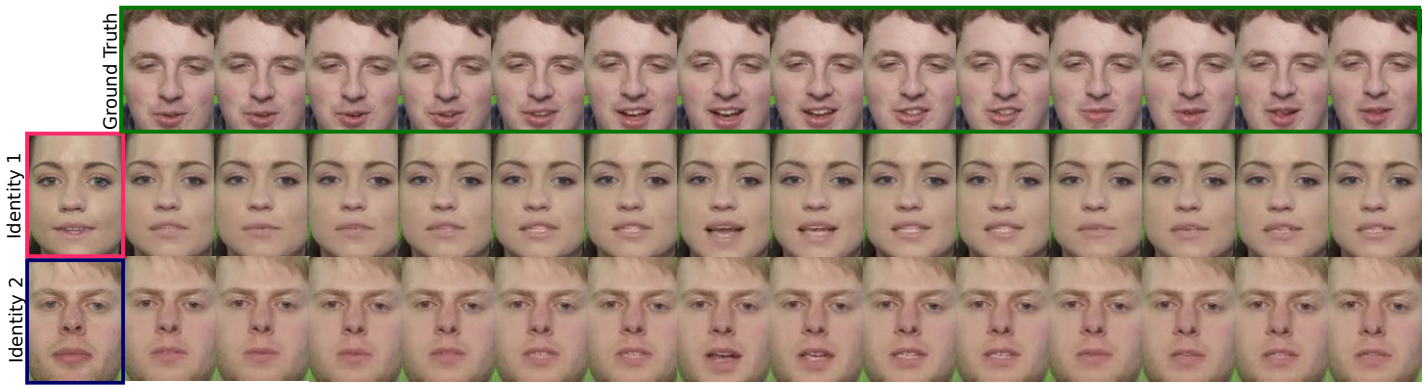
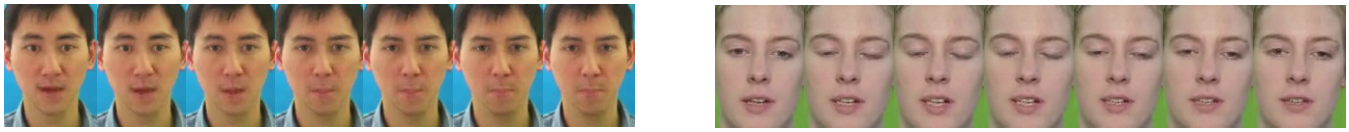


Figure 6: Animation of different faces using the same audio. The movement of the mouth is similar for both faces as well as for the ground truth sequence. Both audio and still image are unseen during training.



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

Figure 7: Facial expressions generated using our framework include (a) frowns and (b) blinks.

Finally, we compare our model to a static model that produces video frames using a sliding window of audio samples like that used in [5]. This is a GAN-based method that uses a combination of an L_1 loss and an adversarial loss on individual frames. We will also use this method as the baseline for our quantitative assessment in the following section. This baseline produces less coherent sequences, characterized by jitter, which becomes worse in cases where the audio is silent (e.g. pauses between words). 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. Fig. 8 shows a comparison between our approach and the baseline in such cases.

4.5 Quantitative Results

We measure the performance of our model on the GRID and TCD TIMIT datasets using the metrics proposed in section 4.2 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



(a) Audio-visual inconsistency during silence (b) Extraneous frames that break continuity

Figure 8: Examples of consecutive frames showcasing the failures of the static baseline including (a) opening the mouth when words are not spoken (b) producing irelevant frames that do not take into account the previous face state, thus breaking the sequence continuity.

natural ones. The results in Table 3 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.

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

5 Conclusion and Future Work

In this work we have presented an end-to-end model using temporal GANs for speech-driven facial animation. Our model is capable of producing highly detailed frames scoring high in terms of PSNR, SSIM and in terms of the sharpness measures on both datasets. According to our ablation study this can be mainly attributed to the use of a *Frame discriminator*.

Furthermore, our method produces more coherent sequences and more accurate mouth movements compared to the static approach, as demonstrated by a resounding user preference and the difference in the WER. 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*. Unlike previous approaches [5] that prohibit the generation of facial expressions, the adversarial loss on the entire sequence encourages spontaneous facial gestures. This has been demonstrated with examples of blinks and frowns.

This model has shown promising results in generating lifelike videos. Moving forward, we believe that different architectures for the sequence discriminator could help produce more natural sequences. Finally, at the moment expressions are generated randomly by the model so a natural extension of this method would attempt to also capture the mood of the speaker from his voice and reflect it 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.

[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] C. Bregler, M. Covell, and M. Slaney. Video Rewrite. In *Conference on Computer Graphics and Interactive Techniques*, pages 353–360, 1997.

[4] T. Chen and R. R. Rao. Audio-Visual Integration in Multimodal Communication. *Proceedings of the IEEE*, 86(5):837–852, 1998.

[5] J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? In *British Machine Vision Conference (BMVC)*, pages 1–12, 2017.

[6] 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.

[7] W. Dai, C. Dai, S. Qu, J. Li, and S. Das. Very Deep Convolutional Neural Networks for Raw Waveforms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 421–425, 2017.

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

[9] B. Fan, L. Wang, F. Soong, and L. Xie. Photo-real talking head with deep bidirectional lstm. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4884–4888, 2015.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. In *Advances in neural information processing systems (NIPS)*, pages 2672–2680, 2014.

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

[12] 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 Transactions on Graphics (TOG)*, 36(94), 2017.

[13] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Y. Li, M. R. Min, D. Shen, D. Carlson, and L. Carin. Video Generation From Text. *arXiv preprint arXiv:1710.00421*, 2017.

[15] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[16] 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.

[17] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv preprint arXiv:1511.06434*, 2015.

[18] 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*, pages 234–241, 2015.

[19] M. Saito, E. Matsumoto, and S. Saito. Temporal Generative Adversarial Nets with Singular Value Clipping. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2830–2839, 2017.

[20] A. D. Simons and S. J. Cox. Generation of mouthshapes for a synthetic talking head. *Proceedings of the Institute of Acoustics, Autumn Meeting*, 12(January):475482, 1990.

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

[22] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(93), 2017.

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

[24] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating Videos with Scene Dynamics. In *Advances In Neural Information Processing Systems (NIPS)*, pages 613–621, 2016.

[25] E. Yamamoto, S. Nakamura, and K. Shikano. Lip movement synthesis from speech based on hidden Markov Models. *Speech Communication*, 26(1-2):105–115, 1998.

[26] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1-2):23–43, 1998.

[27] H. C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson. Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, 30(3):555–568, 2002.