This CVPR 2020 workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Robust One Shot Audio to Video Generation

Neeraj Kumar
neerajku@hike.in

Srishti Goel
srishtig@hike.in

Ankur Narang
ankur@hike.in

Mujtaba Hasan
mujtaba@hike.in

Abstract

Audio to Video generation is an interesting problem that has numerous applications across industry verticals including film making, multi-media, marketing, education and others. High-quality video generation with expressive facial movements is a challenging problem that involves complex learning steps for generative adversarial networks. Further, enabling one-shot learning for an unseen single image increases the complexity of the problem while simultaneously making it more applicable to practical scenarios.

In the paper, we propose a novel approach OneShotA2V to synthesize a talking person video of arbitrary length using as input: an audio signal and a single unseen image of a person. OneShotA2V leverages curriculum learning to learn movements of expressive facial components and hence generates a high-quality talking head video of the given person.

Further, it feeds the features generated from the audio input directly into a generative adversarial network and it adapts to any given unseen selfie by applying few-shot learning with only a few output updation epochs. OneShotA2V leverages spatially adaptive normalization based multi-level generator and multiple multi-level discriminators based architecture. The input audio clip is not restricted to any specific language, which gives the method multilingual applicability. Experimental evaluation demonstrates superior performance of OneShotA2V as compared to Realistic Speech-Driven Facial Animation with GANs(RSDGAN) [43], Speech2Vid [8], and other approaches, on multiple quantitative metrics including: SSIM (structural similarity index), PSNR (peak signal to noise ratio) and CPBD (image sharpness). Further, qualitative evaluation and Online Turing tests demonstrate the efficacy of our approach.

1. Introduction

Audio to Video generation has numerous applications across industry verticals including film making, multi-media, marketing, education and others. In the film industry, it can help through automatic generation from the voice acting and also occluded parts of the face. Additionally, it can help in limited bandwidth visual communication by using audio to auto-generate the entire visual content or by filling in dropped frames. High-quality video generation with expressive facial movements is a challenging problem that involves complex learning steps.

The audio or speech signal contains rich information about the mood (expression) and the intent of the user. On hearing the audio, one can predict the sentiment or emotion depicted by the user. This expressive power of the audio can be used to generate robust and high-quality talking head videos. The talking-head video aims to handle expressive facial movements and head movement based on the audio content and expression.

Most of the work in this field has been centered towards the mapping of audio features (MFCCs, phonemes) to visual features (Facial landmarks, visemes etc.) [1, 40, 7, 20]. Further computer graphics techniques select frames of a specific person from the database to generate expressive faces. Few techniques which attempt to generate the video using raw audio focuses for the reconstruction of the mouth area only [8]. Due to a complete focus on lip-syncing, the aim of capturing human expression is ignored. Further, such methods lack smooth transition between frames which does not make the final video look natural. Regardless, of which approach we use, the methods described above are either subject dependent [37, 18] or generate unnatural videos [47] due to lack of smooth transition and/or require high compute time to generate video for a new unseen speaker image for ensuring high-quality output [48].

We propose a novel approach that is capable of developing a speaker-independent and language-independent high-quality natural-looking talking head video from a single unseen image and an audio clip. Our model captures the word embeddings from the audio clip using a pre-trained deepspeech2 model [11] trained on Librispeech corpus [27]. These embeddings and the image are then fed to the multi-level generator network which is based on the Spatially-Adaptive Normalization architecture [28]. Multiple multi-level discriminators [46] are used to ensure synchronized and realistic video generation. A multi-scale frame discriminator is used to generate high-quality realistic frames.

A multi-level temporal discriminator is modeled which ensures temporal smoothening along with spatial consistency. Finally, to ensure lip synchronization we use SyncNet architecture [2] based discriminator applied to the lower half of the image. To make the generator input-time independent, a sliding window approach is used. Since, the generator needs to finally learn to generate multiple facial component movements along with high video quality, multiple loss functions both adversarial and non-adversarial are used in a curriculum learning fashion. For fast low-cost adaptation to an unseen image, a few output updation epochs suffice to provide one-shot learning capability to our approach.

Specifically, we make the following contributions:
(a) We present a novel approach, **OneShotA2V**, that leverages curriculum learning to simultaneously learn movements of expressive facial components and generate a high-quality talking-head video of the given person.
(b) Our approach feeds the features generated from the audio input directly into a generative adversarial network and it adapts to any given unseen selfie by applying one-shot learning with only a few output updation epochs.
(c) It leverages spatially adaptive normalization based multi-level generator and multiple multi-level discriminators based architecture to generate video which simultaneously considers lip movement synchronization and natural facial expressions incorporating eye blink and eyebrow movements along with head movement.
(d) Experimental evaluation on the GRID [10] datasets, demonstrates superior performance of OneShotA2V as compared to Realistic Speech-Driven Facial Animation with GANs(RSDGAN) [43], Speech2Vid [8], and other approaches, on multiple quantitative metrics including: SSIM (structural similarity index), PSNR (peak signal to noise ratio) and CPBD (image sharpness). Further, qualitative evaluation and Online Turing tests demonstrate the efficacy of our approach.

2. Related Work

A lot of work has been done in synthesizing realistic videos from audio with an image as an input. Speech is a combination of content and expression and there is a perceptual variability of speech that exists in the form of various languages, dialects and accents. The understanding and modeling of the speech are more complicated as compared to the text which is devoid of various aspects of speech.

Various works have been done in this aspect to understand the different aspects of speech to generate realistic speech-driven videos.

The earliest methods for generating videos relied on Hidden Markov Models which captured the dynamics of audio and video sequences. Simons and Cox [33] used the Viterbi algorithm to calculate the most likely sequence of mouth shape given the particular utterances. Such methods are not capable of generating quality videos and lack emotions.

2.1. Phoneme and Visemes based classification of speech

Phoneme and Visemes based approaches have been used to generate the videos. Real-Time Lip Sync for Live 2D Animation [1] has used an LSTM based approach to generate live lip synchronization on 2D character animation.

Some of these methods target rigged 3D characters or meshes with predefined mouth blend shapes that correspond to speech sounds [36, 17, 38, 12, 23, 37], while others generate 2D motion trajectories that can be used to deform facial images to produce continuous mouth motions [4, 5]. These methods are primarily focused on mouth motions only and do not show emotions such as eye blinking, eyebrows movements, etc.

2.2. Video synthesis using deep networks

CNN has been used for generating the videos by giving audio features to the network. Audio2Face [40] model uses the CNN method to generate an image from audio signals. You said That [8](Speech2Vid) has used an encoder-decoder based approach for generating realistic videos. MFCC coefficients of audio signals are being used as an input. L1 loss at the pixel level is used between synthesized image and target image which penalizes any deviation of the generated image from the target one. This disincentivizes the model to generate realistic images without spontaneous expressions except for mouth movement. Our approach uses a spatially adaptive network instead of the encoder as used in [8] to learn the parameters of an image.

Synthesizing Obama: Learning Lip Sync from Audio [37] is able to generate quality videos of Obama speaking with accurate lip-sync. They use RNN based approach to map from raw audio features to mouth shapes. This method is trained on a single target image to generate high-quality videos. Our approach is able to generate videos on a single unseen image using a spatially adaptive network and a one-shot approach. LumireNet: Lecture Video Synthesis from Audio [18] is generating high-quality, full-pose head-shot lecture videos from the instructors new audio narration of any length. They have used dense pose [13], LSTM , variational auto-encoder [29] and GANs based approach to synthesize the videos . The limitation is that they are not able to produce lip-synced video as they are only using dense pose for pose information. Our approach has used a synchronization discriminator for the generation of coherent lip-synced videos. They have used Pix2Pix [15] for the frame synthesis and we are using a spatially adaptive generator for frame generation which is able to generate higher quality videos.

The recent introduction of GANs in [14] has shifted the focus of the machine learning community to generative

modeling. The generators goal is to produce realistic samples and the discriminators goal is to distinguish between the real and generated samples. However, GANs are not limited to these applications and can be extended to handle videos [45].

Temporal Gan [32] and Generating Videos with Scene Dynamics [42] have done the straight forward adaptation of GANs for generating videos by replacing 2D convolution layers with 3D convolution layers. Such methods are able to capture temporal dependencies but require constant length videos. Our approach is able to produce lower word error rate and generate consistent videos of variable length using multi-scale temporal discriminator and synchronization discriminator.

Realistic Speech-Driven Facial Animation with GANs(RSDGAN) [43] used GAN based approach to produce quality videos. They have used identity encoder, context encoder and frame decoder to generate images and used various discriminators to take care of different aspects of video generation. They have used frame discriminator to distinguish real and fake images, sequence discriminator to distinguish real and fake videos and synchronization discriminator for better lip synchronization in videos. We introduce spatially adaptive normalization along with a one-shot approach and implemented curriculum learning to produce better results. This is explained in Sections 3,4,5.

Few-Shot Adversarial Learning of Realistic Neural Talking Head Models [48] have used meta-learning for generating the videos on unseen images taking video as an input. They have used content loss measures the distance between the ground truth image and the reconstruction using the perceptual similarity measure [16] from VGG19 [34] network trained for ILSVRC classification and VGGFace [24] and adversarial loss. For unseen images, the model needs to run for 75 epochs on an unseen image to give better video quality. This is computationally heavy, so we are using a one-shot approach using perceptual loss during inference. Due to the spatially adaptive nature of our generator architecture, we are able to generate good quality video at a low computational cost. Few shot Video to Video Synthesis [44] is able to generate videos on unseen images given a video as an input by using a network weight generation module for extracting the pattern. Such a method is computationally heavy concerning our approach which is one shot approach in video generation.

X2face [47] model uses GANs based approach to generate videos given a driving audio or driving video and a source image as an input. The model learns the face embeddings of source frame and driving vectors of driving frames or audio basis which generates the videos. This model is trained on 1fps which can lead to un-natural video synthesis. They have used L1 loss and identity loss for video generation and have not used any loss for temporal coherency. Our approach generates the videos at 25fps which are capable of generating natural realistic videos and have incorporated multi-scale temporal discriminator and lip-sync discriminator for temporal coherency.

The MoCoGAN [41] uses RNN based generators with separate latent spaces for motion and content. A sliding window approach is used so that the discriminator can handle variable-length sequences. This model is trained to generate disentangled content and motion vectors such that they can generate audios with different emotions and contents. Our approach has used deep speech2 features to learn content embeddings such that it is able to produce better videos with low word error rate.

Animating Face using Disentangled Audio Representations [25] has generated the disentangled representation of content and emotion features to generate realistic videos. They have used variational autoencoders [29] to learn representation and feed them into GANs based model to generate videos. Our approach has used deep speech2 features to learn content embeddings instead of variational autoencoders. Instead of their Unet [30] architecture, we are using a spatially adaptive generator to generate high-quality videos.

Audio-driven Facial Reenactment [39] used AudioexpressionNet to generate 3D face model. The estimated 3D face model is rendered using the rigid pose observed from the original target image. Our approach is using a spatially adaptive and one-shot approach to generate 2D videos and are not generating 3D mesh.

Existing methodologies have worked on video generations with lip movement and expressions. Our goal is to create high-quality videos using a one-shot approach, so that we can experience high definition videos along with expressions and multilingual support.

3. Architectural Design

OneShotA2V consists of a single generator and 3 discriminators as shown in Figure 1. Each of the discriminators are used for specific purposes. Different losses are used to make the generator learn better distribution for generating realistic videos.

3.1. Generator

The initial layers of generator, G uses deepspeech2 [11] layers followed by Spatially-Adaptive normalization similar to SPADE architecture [28]. Conditional input of an audio frame and image is fed to the spatially adaptive network. Instead of giving the semantic input to the network as proposed in SPADE architecture, we give aligned images in an upsampling manner. This helps in the prevention of loss of information due to normalization.

An audio input of 200 ms is given along with the image to produce a single frame of the video. The audio input

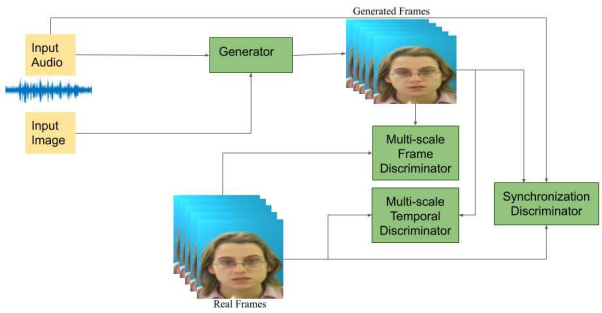

Figure 1. Model for generating robust and high-quality videos. This uses deep speech audio features to be fed into SPADE Generator and 2 discriminators i.e frame discriminator which is a multi-scale discriminator for frame generation and another discriminator for better lip synchronization.

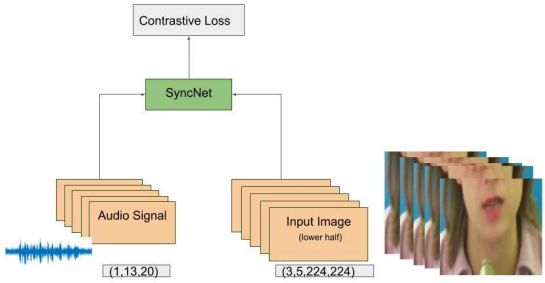

Figure 2. We have used the SyncNet architecture for better lip synchronization which is trained on GRID dataset with contrastive loss and then used its loss in our proposed architecture

is overlapping with the previous audio input with an overlapping interval of 0.16 ms. Every audio frame is centered around a single video frame. To do that, zero padding is done before and after the audio signal and use the following formula for the stride.

$$stride = \frac{\text{audio sampling rate}}{\text{video frames per sec}}$$

3.1.1 Audio features using deepspeech2 model

The MFCC coefficients of audio input is fed into the pre-trained deepspeech2 for extracting the content-related features of audio. We have taken the few layers of deepspeech2 network and fed it as an input to the generator. This helps in improving the lip synchronization aspect for the video generation.

3.2. Discriminator

We have used 3 discriminators namely a multi-scale frame discriminator, a multi-scale temporal discriminator and a synchronization discriminator.

3.2.1 Multi-scale Frame Discriminator

Multi-scale discriminator [46], D is used in the proposed model to distinguish the coarser and finer details between real and fake images. Adversarial training with the discriminator helps in generating realistic frames. To have high resolution generated frames, we need to have an architecture with better receptive field. A deeper network can cause overfitting, to avoid that, multi-scale discriminators are used. Multi-scale frame discriminator consists of 3 discriminators that have an identical network structure but operate at different image scales. These discriminators are referred to as D1, D2 and D3. Specifically, we downsample the real and synthesized high-resolution images by a factor of 2 and 4 to create an image pyramid of 3 scales. The discriminators D1, D2 and D3 are then trained to differentiate real and synthesized images at the 3 different scales, respectively. The discriminators operate from coarse to fine level and help the generator to produce high-quality images.

3.2.2 Multi-scale Temporal Discriminator

Every frame in a video is dependent on its previous frames. To capture the temporal property along with a spatial one, we have used a multi-scale temporal discriminator [18]. This discriminator is modeled to ensure a smooth transition between consecutive frames and achieve a natural-looking video sequence. The multi-scale temporal discriminator is described as

$$L(T, G, D) = \sum_{i=t-L}^{t} [\log(D(x_{\mathrm{i}}))] + [\log(D(1 - G(x_{\mathrm{i}})))]$$

where t is the time instance of an audio and L is the length of the time interval for which the adversarial loss is computed.

3.2.3 Synchronization Discriminator

To have coherent lip synchronization, the proposed model uses SyncNet architecture proposed in Lip Sync in the wild [9]. As shown in Figure 2 the input to the discriminator is an audio signal of 200ms time interval(5 audio signals of 40ms each) and 5 frames of the video. The lower half of the frame of resized to (224,224,3) is fed as an input.

4. Curriculum Learning

We have trained OneShotA2V in multiple phases so that it can produce better results. In the first phase we have used a multi-scale frame discriminator and applied the adversarial loss, feature matching loss and perceptual loss to learn the higher-level features of the image. When these losses stabilize, we move to the second phase in which we have

added a multi-scale temporal discriminator and synchronization discriminator and used reconstruction loss, Contrastive loss and temporal adversarial loss to get a better quality image near mouth region and coherent lip synchronized high-quality videos. After the stabilization of the above losses, we have added blink loss in the third phase to generate a more realistic image capturing emotions such as eye movement and eye blinks.

4.1. Losses

OneShotA2V is trained with different losses to generate realistic videos as explained below.

4.1.1 Adversarial Loss

Adversarial Loss is used to train the model to handle adversarial attacks and ensure generation of high-quality images for the video. The loss is defined as:

$$L_{\text{GAN}}(G, D) = E_{\text{x}\sim\text{P}_\text{d}}[\log(D(x))] + E_{\text{z}\sim\text{P}_\text{z}}[\log(D(1 - G(x)))]$$

where G tries to minimize this objective against an adversarial D that tries to maximize.

4.1.2 Reconstruction loss

Reconstruction loss [21] is used on the lower half of the image to improve the reconstruction in mouth area. L1 loss is used for this purpose as described below:

$$L_{\text{RL}} = \sum_{n\epsilon[0,W]*[H/2,H]} (R_\text{n} - G_\text{n})$$

where, R_n and G_n are the real and generated frames respectively.

4.1.3 Feature Loss

Feature-matching Loss [46] ensures generation of natural-looking high-quality frames. We take the L1 loss of between generated images and real images for different scale discriminators and then sum it all. We extract features from multiple layers of the discriminator and learn to match these intermediate representations from the real and the synthesized image. This helps in stabilizing the training of the generator. The feature matching loss, $L_{FM}(G,D_k)$ is given by:

$$L_{\text{FM}}(G, D_\text{k}) = E_{(\text{s,x})} \sum_{n=1}^{T} [\frac{1}{N_\text{i}} ||{D_\text{k}}^{(i)}(x) - {D_\text{k}}^{(i)}(G(x))||_1]$$

where, T is the total number of layers and N_i denotes the number of elements in each layer.

4.1.4 Perceptual Loss

The perceptual similarity metric is calculated between the generated frame and the real frame. This is done by using features of a VGG19 [34] model trained for ILSVRC classification and VGGFace [24] dataset.The perceptual loss [16],(L_{PL}) is defined as:

$$L_{\text{PL}} = \lambda \sum_{n=1}^{N} [\frac{1}{M_\text{i}} ||F^{(i)}(x) - F^{(i)}(G(x))||_1]$$

where, λ is the weight for perceptual loss and $F^{(i)}$ is the ith layer of VGG19 network with M_i elements of VGG layer.

4.1.5 Contrastive Loss

For coherent lip synchronization, we use the Synchronization Discriminator with Contrastive loss. The training objective is that the output of the audio and the video networks are similar for genuine pairs, and different for false pairs.

Contrastive loss,(L_{CL}) is given by following equation

$$L_{\text{CL}} = \frac{1}{2N} \sum_{n=1}^{N} (y_\text{n}){d^2}_\text{n} + (1 - y_\text{n}) max(margin - d_\text{n}, 0)^2$$

$$d_\text{n} = ||v_\text{n} - a_\text{n}||_2$$

where, v_n and a_n are fc_7 vectors for video and audio inputs respectively. y ϵ [0,1] is the binary similarity metric for video and audio input.

4.1.6 Blink loss

We have used the eye aspect ratio (EAR) taken from Real-Time Eye Blink Detection using Facial Landmarks [35] to calculate the blink loss. A blink is detected at the location where a sharp drop occurs in the EAR signal. Loss is defined as:

$$m(x) = \frac{||p2 - p6|| + ||p3 - p5||}{||p1 - p4||}$$

$$L_{\text{BL}} = ||m_\text{r} - m_\text{g}||$$

where, p_i is described in Figure 3. We have taken the L1 loss of eye aspect ratio(EAR) between real image m_r and synthesized frame m_g.

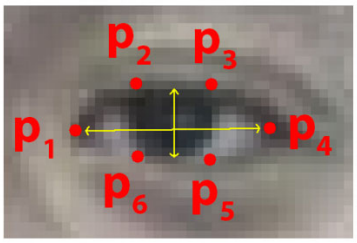

Figure 3. Description of the 6 eye points

5. Few shot learning

To achieve a more sharp and a better image quality for an unseen subject, we have used one shot approach using perceptual loss during inference time. Our approach is computationally less expensive as compared to [48, 44] which we have described in Section 2 and because of the spatially adaptive nature of generator architecture, we are able to achieve high-quality video. We run the model for 5 epochs during inference time to get high-quality video frames.

6. Experiments and Results

6.1. Datasets and Training

We have used the GRID dataset [10] and LOMBARD GRID [26] for the experiment and evaluation of different metrics. GRID dataset is a large multi-talker audiovisual sentence corpus. This corpus consists of high-quality audio and video (facial) recordings of 1000 sentences spoken by each of 34 talkers (18 male, 16 female). LOMBARD GRID dataset is a bi-view audiovisual Lombard speech corpus that can be used to support joint computational-behavioral studies in speech perception. The corpus includes 54 talkers, with 100 utterances per talker (50 Lombard and 50 plain utterances). It consists of 5400 videos generated on 54 talkers comprising 30 female talkers and 24 male talkers.

Our model is implemented in Pytorch and takes approximately 4 days to run on 4 Nvidia V100 GPUs for training. Around 5000 and 1200 videos of the GRID dataset are used for training and testing purposes respectively. We have taken 3000 and 600 videos of the LOMBARD GRID dataset for training and testing purposes. The frames are extracted at 25fps. We have taken 16khz as sampling frequency for audio signals and used 13MFCC coefficients for 0.2 sec of overlapping audio for experimentation.

The aligned face is generated for every speaker using facial landmark detector [3] and HopeNet [31] for calculating the yaw, pitch and roll angles to get the most aligned faces for every speaker as an input.

We take the Adam optimizer [19] with learning rate = 0.002 and β_1= 0.0 and β_2 = 0.90 for the generator and discriminators. The learning rate of the generator and discriminator is constant for 50 epochs and after that it decays to zeros in the next 100 epochs.

6.2. Metrics

1. PSNR- Peak Signal to Noise Ratio: It computes the peak signal to noise ratio between two images. The higher the PSNR the better the quality of the reconstructed image.

2. SSIM- Structural Similarity Index: It is a perceptual metric that quantifies image quality degradation. The larger the value the better the quality of the reconstructed image.

3. CPBD- Cumulative Probability Blur Detection: It is a perceptual based no-reference objective image sharpness metric based on the cumulative probability of blur detection developed at the Image.

4. WER- Word error rate: It is a metric to evaluate the performance of speech recognition in a given video. We have used LipNet architecture [2] which is pre-trained on the GRID dataset for evaluating the WER. On the GRID dataset, Lipnet achieves 95.2 percent accuracy which surpasses the experienced human lipreaders.

5. ACD- Average Content Distance([41]): It is used for the identification of speaker from the generated frames using OpenPose [6]. We have calculated the Cosine distance and Euclidean distance of representation of the generated image and the actual image from Openpose. The distance threshold for the OpenPose model should be 0.02 for Cosine distance and 0.20 for Euclidean distance [22]. The lesser the distances the more similar the generated and actual images.

6.3. Qualitative Results

OneShotA2V is able to produce natural-looking high-quality videos of previously unseen input image and audio signals. The videos are able to do lip synchronization on the sentences provided to them. Videos were generated targeting different languages ensuring the proposed method is language independent and can generate videos for any linguistic community.

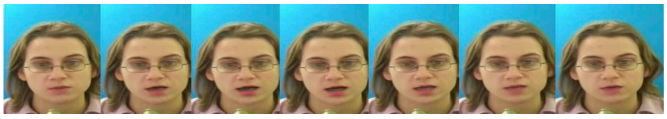

Figure 4. Female uttering the word "now"

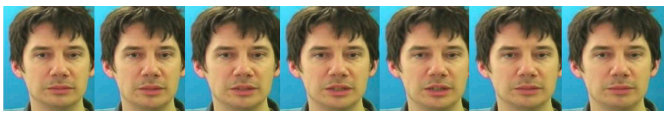

Figure 5. Male uttering the word "bin"

Figure 4 and Figure 5 show different examples of the generated and lip synchronized videos for male and female test cases for the same audio clip and their ground truth frames. As observed the opening and closing of the mouth is in sync with the audio signals. Our method is able to produce synchronized lip movements displaying facial expressions such as forehead lines and eye blinks ensuring a natural-looking aesthetic output.

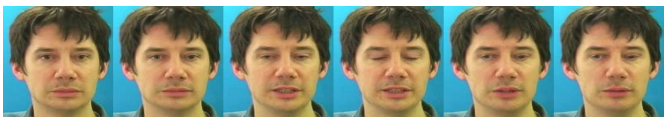

Figure 6. Movement of eyes while speaking

Figure 6 show the the movement of eyes of speaker while speaking . Such frames are able to generate natural videos capturing the eye's movement while speaking. s

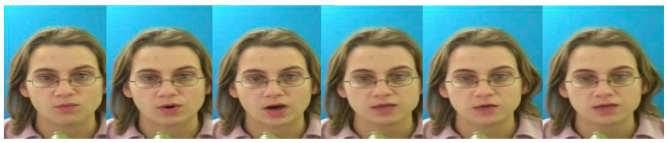

Figure 7. Speaker uttering a hindi male name "Modi"

Figure 7 show the generated output for a hindi audio clip ("Modi"). As observed, the generated frames are able to produce the expected lip movements and provide multi-lingual support.

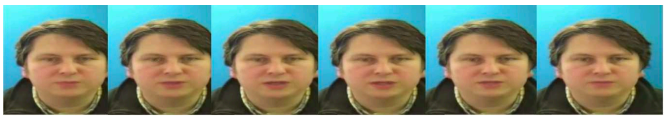

Figure 8. Speaker uttering the word "Please" on the GRID dataset

Figure 9. Speaker uttering the word "Please" on the LOMBARD GRID dataset

Figure 8 show the generated output with the model trained on the GRID dataset and Figure 9 show the generated output with the model trained on the LOMBARD GRID dataset.

For video clips, see the supplementary data.

6.4. Quantitative Results

The Proposed Model has performed better on image reconstruction metrics such as peak signal to noise ratio(PSNR) and Structural Similarity Index(SSIM) as compared to Realistic Speech-Driven Facial Animation with GANs(RSDGAN) [43] and Speech2Vid [8] Model as shown in Table 1. We also display the comparison with OneShotA2V trained on the LOMBARD GRID dataset [26]. This is achieved with the use of spatially adaptive normalization in the generator architecture along with training of the proposed model in curriculum learning fashion.

Method	SSIM	PSNR	CPBD
OneShotA2V	**0.881**	**28.571**	0.262
OneShotA2V(lombard)	0.922	28.978	0.453
RSDGAN	0.818	27.100	**0.268**
Speech2Vid	0.720	22.662	0.255

Table 1. Comparision of OneShotA2V with RSDGAN and Speech2Vid for SSIM, PSNR and CPBD

Method	WER	ACD-C	ACD-E
OneShotA2V	27.5	0.005	0.09
OneShotA2V(lombard)	26.1	0.002	0.064
RSDGAN	23.1	-	1.47×10^{-4}
Speech2Vid	58.2	-	1.48×10^{-4}

Table 2. The above comparison is on lip synchronizing metric i.e word error rate(WER) and average content distance(ACD) by calculating cosine distance(ACD-C) and euclidean distance(ACD-E) between the actual image and the generated image.

Table 2 shows the comparison of OneShotA2V with the RSDGAN and Speech2Vid [8] models against the metrics such as word error rate (WER) to see the lip synchronizing performance of the generated videos. For this, we have used the pre-trained LipNet model whose accuracy is 95.2% on GRID datasets. We find that OneShotA2V performed better than Speech2Vid but lagged behind RSDGAN. We have used the pre-trained OpenFace model to calculate the Cosine distance and Euclidean distance for average content distance. Experiments on OpenFace show that the distance threshold for the model should be 0.02 for cosine distance and 0.20 for euclidean distance [22].

6.5. Psychophysical assessment

Results are visually rated (on a scale of 5) individually by 25 persons, on three aspects, lip synchronization, eye blinks and eyebrow raises and quality of video. The subjects were shown anonymous videos at the same time for the different audio clips for side-by-side comparison. Table 3 clearly shows that OneShotA2V performs significantly better in quality and lip synchronization which is of prime importance in videos.

To test the naturalism of the generated videos we conduct an online Turing test [1]. Each test consists of 25 questions

[1] https://forms.gle/JEk1u5ahc9gny7528

Method	Lip-Sync	Eye-blink	Quality
OneShotA2V	90.8	88.5	**76.2**
RSDGAN	**92.8**	**90.2**	74.3
Speech2Vid	90.7	87.7	72.2

Table 3. Psychophysical Evaluation (in percentages) based on users rating

with 13 fake and 12 real videos. The user is asked to label a video real or fake based on the aesthetics and naturalism of the video. Approximately 300 user data is collected and their score of the ability to spot fake video is displayed in Figure 10.

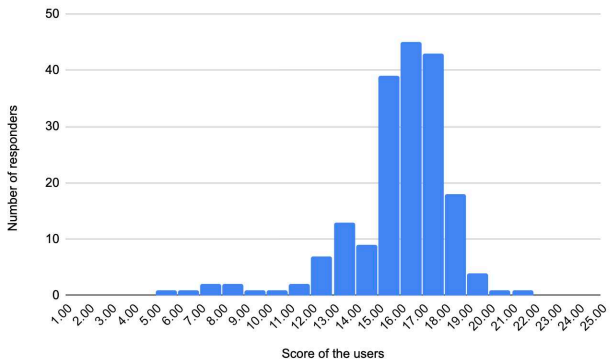

Figure 10. Distribution of user scores for the online Turing test

6.6. Ablation Study

We studied the incremental impact of various loss functions on the LOMBARD GRID dataset and the GRID dataset. We have provided corresponding videos in supplementary data for better visual understanding. As mentioned in section 4 (Curriculum learning) each loss has a different impact on the final output video. Table 4 and Table 5 depicts the impact of different losses on both datasets. The base model mentioned is the includes the adversarial gan loss, feature loss and perceptual loss. The addition of contrastive loss and multi-scale temporal adversarial loss in sequence discriminator helps in achieving coherent lip synchronized videos and improves the SSIM, PSNR and CPBD values. Further addition of Blink Loss, ensures improved quality of the final video.

Method	SSIM	PSNR	CPBD
Base Model(BM)	0.869	27.996	0.213
BM + CL +TAL	0.873	28.327	0.258
BM + CL + TAL+ BL	0.881	28.571	0.262

Table 4. Ablation Study on the GRID dataset where, CL is the contrastive loss ,TAL is the multi-scale temporal adversarial loss and BL is the Blink loss

Method	SSIM	PSNR	CPBD
Base Model(BM)	0.909	28.656	0.386
BM + CL + TAL	0.913	28.712	0.390
BM + CL + TAL+ BL	0.922	28.978	0.453

Table 5. Ablation Study on the LOMBARD GRID dataset where, CL is the contrastive loss, TAL is the multi-scale temporal adversarial loss and BL is the Blink loss

The use of deeepspeech2 to generate audio to content embeddings helped in the improvement of WER and help us reach almost similar performance as RSDGAN.

6.7. Conclusions and Future Work

In this paper, we have considered robust one-shot video generation from audio input. Our approach, OneShotA2V, uses multi-level generator and multiple multi-level discriminators along with curriculum learning and few-shot learning to generate high-quality videos. Spatially adaptive normalization helps to ensure light generator architecture without encoders and also efficient few-shot learning with few updation epochs on the generator. The coherent lip movement and lower word error rate(WER) is attributed to the use of multi-scale temporal discriminator and synchronization discriminator. The use of deep speech features helps the model to learn the content vectors of audio in a better manner which led to lower word error rate.

Experimental evaluation on GRID dataset demonstrates superior performance of OneShotA2V as compared to Realistic Speech-Driven Facial Animation with GANs(RSDGAN) [43], Speech2Vid [8], and other approaches, on multiple quantitative metrics including: SSIM (structural similarity index), PSNR (peak signal to noise ratio) and CPBD (image sharpness). Further, qualitative evaluation and Online Turing tests demonstrate the efficacy of our approach. Moreover, OneShotA2V is able to perform generation of robust high-quality natural-looking videos across multiple languages without any additional requirement of multilingual datasets (audio signals).

In the future, we plan to add emotions so that the generated videos can capture varying degrees of emotional expressions of the speaker. This will make the video output from our approach more realistic and robust. Further, we plan to consider sophisticated curriculum learning techniques to enable the generation of more dynamic talking videos.

References

[1] Deepali Aneja and Wilmot Li. Real-time lip sync for live 2d animation. 2019.

[2] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. Lipnet: End-to-end sentence-level lipreading. *GPU Technology Conference*, 2017.

[3] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Incremental face alignment in the wild. pages 1859–1866, 06 2014.

[4] Matthew Brand. Voice puppetry. *In Proceedings of the 26th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co*, page 2128, 1999.

[5] Yong Cao, Wen Tien, Petros Faloutsos, and Frederic Pighin. Expressive speech-driven facial animation. *ACM Trans. Graph.*, 24:1283–1302, 10 2005.

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018.

[7] Luca Cappelletta and Naomi Harte. Phoneme-to-viseme mapping for visual speech recognition. *Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRAM 2012)*, 2, 05 2012.

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

[9] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[10] Barker J. Cunningham S. Shao X Cooke, M. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):24212424, 2006.

[11] Eric Battenberg Carl Case Jared Casper Bryan Catanzaro-Jingdong Chen Mike Chrzanowski Adam Coates Greg Diamos Erich Elsen Jesse Engel Linxi Fan Christopher Fougner Tony Han Awni Hannun Billy Jun Patrick LeGresley Libby Lin Sharan Narang Andrew Ng Sherjil Ozair Ryan Prenger Jonathan Raiman Sanjeev Satheesh David Seetapun Shubho Sengupta Yi Wang Zhiqian Wang Chong Wang Bo Xiao Dani Yogatama Jun Zhan Zhenyao Zhu Dario Amodei, Rishita Anubhai. Deep speech 2: End-to-end speech recognition in english and mandarin. 2005.

[12] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics*, 35:1–11, 07 2016.

[13] Riza Guler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. pages 7297–7306, 06 2018.

[14] Mehdi Mirza Bing Xu David Warde-Farley Sherjil Ozair Aaron Courville Yoshua Bengio Ian J. Goodfellow, Jean Pouget-Abadie. Generative adversarial nets.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei Efros. Image-to-image translation with conditional adversarial networks. pages 5967–5976, 07 2017.

[16] Alexandre Alahi Justin Johnson and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. 2016.

[17] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics*, 36:1–12, 07 2017.

[18] Byung-Hak Kim and Varun Ganapath. Lumirenet: Lecture video synthesis from audio. 2016.

[19] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[20] Soonkyu Lee and Dongsuk Yook. Audio-to-visual conversion using hidden markov models. pages 563–570, 08 2002.

[21] Yanchun Li, Nanfeng Xiao, and Wanli Ouyang. Improved generative adversarial networks with reconstruction loss. *Neurocomputing*, 323, 10 2018.

[22] Yu Tian Mubbasir Kapadia Long Zhao, Xi Peng and Dimitris Metaxas1. Learning to forecast and refine residual motion for image-to-video generation, 2018.

[23] Wesley Mattheyses and Werner Verhelst. Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, 66, 11 2014.

[24] Wang Mei and Weihong Deng. Deep face recognition: A survey. 04 2018.

[25] Gaurav Mittal and Baoyuan Wang. Animating face using disentangled audio representations, 2019.

[26] Ricard Marxer Jon Barker Najwa Alghamdi, Steve Maddock and Guy J. Brown. A corpus of audio-visual lombard speech with frontal and profile view, the journal of the acoustical society of america 143, el523 (2018); https://doi.org/10.1121/1.5042758, 2018.

[27] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. pages 5206–5210, 04 2015.

[28] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[29] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. 09 2016.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. volume 9351, pages 234–241, 10 2015.

[31] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[32] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. 10 2017.

[33] A. Simons and Stephen Cox. Generation of mouthshapes for a synthetic talking head. *Proceedings of the Institute of Acoustics, Autumn Meeting*, 01 1990.

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.

[35] Tereza Soukupova and Jan Cech. Real-time eye blink detection using facial landmarks, 2016.

[36] Andreea Stef, Kaveen Perera, Hubert Shum, and Edmond Ho. Synthesizing expressive facial and speech animation by text-to-ipa translation with emotion control. pages 1–8, 12 2018.

[37] Supasorn Suwajanakorn, Steven Seitz, and Ira Kemelmacher. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics*, 36:1–13, 07 2017.

[38] Sarah Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. Dynamic units of visual speech. pages 275–284, 07 2012.

[39] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Niener. Neural voice puppetry: Audio-driven facial reenactment, 12 2019.

[40] Guanzhong Tian, Yi Yuan, and Yong Liu. Audio2face: Generating speech/face animation from single audio with attention-based bidirectional lstm networks. 05 2019.

[41] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1526–1535, 2018.

[42] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. 09 2016.

[43] Konstantinos Vougioukas, Stavros Petridi, and Maja Pantic. End-to-end speech-driven facial animation with temporal gans. *Journal of Foo*, 14(1):234–778, 2004.

[44] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[45] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[46] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[47] O. Wiles, A.S. Koepke, and A. Zisserman. X2face: A network for controlling face generation by using images, audio, and pose codes. In *European Conference on Computer Vision*, 2018.

[48] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models, 05 2019.