Speech to Lip Generation

Mahesh Sawant[1], Keval Senghani[2], Pragati Singh[3], Dr. Shubhangi Vaikole[4]

[1]*Student, Dept. of Computer Engineering, Datta Meghe College of Engineering, Navi Mumbai, India*
[2]*Student, Dept. of Computer Engineering, Datta Meghe College of Engineering, Navi Mumbai, India*
[3]*Student, Dept. of Computer Engineering, Datta Meghe College of Engineering, Navi Mumbai, India*
[4] *Associate Prof., Dept. of Computer Engineering, Datta Meghe College of Engineering, Navi Mumbai, India*

---***---

Abstract -This project involves Lip to Speech Synthesis and Speech to Lip Generation System. In this project the problem of lip-syncing a talking face video of an arbitrary identity to match a target speech segment is invested. Current works excel at producing accurate lip movements on a static image or videos of specific people seen during the training phase. However, they fail to accurately morph the lip movements of arbitrary identities in dynamic, unconstrained talking face videos, resulting in significant parts of the video being out-of-sync with the new audio. Identifying key reasons about this and hence resolving them by learning from a powerful lip-sync discriminator. Also, to explore the task of lip to speech synthesis, i.e., learning to generate natural speech given only the lip movements of a speaker. Focusing on learning accurate lip sequences to speech mappings for individual speakers in unconstrained, large vocabulary settings. To this end, collecting and releasing a large-scale benchmark dataset, the first of its kind, specifically to train and evaluate the single speaker lip to speech task in natural settings. By proposing a novel approach with key design choices to achieve accurate, natural lip to speech synthesis in such unconstrained scenarios for the first time.

Key Words: Speech to Lip, Machine Learning, Audio and Speech Processing, Sound, Computer Vision.

1.INTRODUCTION

With the exponential rise in the consumption of audio-visual content, rapid video content creation has become a quintessential need. At the same time, making these videos accessible in different languages is also a key challenge. For instance, a deep learning lecture series, a famous movie, or a public address to the nation, if translated to desired target languages, can become accessible to millions of new viewers. A crucial aspect of translating such talking face videos or creating new ones is correcting the lip sync to match the desired target speech. Consequently, lip-syncing talking face videos to match a given input audio stream has received considerable attention [1] in the research community.

Machine learning methods have had a great impact on social progress in recent years, which promoted the rapid development of artificial intelligence technology and solved many practical problems. Automatic lipreading technology is one of the important components of human–computer interaction technology and virtual reality (VR) technology. It plays a vital role in human language communication and visual perception. Especially in noisy environments or VR environments, visual signals can remove redundant information, complement speech information, increase the multi-modal input dimension of immersive interaction, reduce the time and workload of humans on learning lip language and lip movement, and improve automatic speech recognition ability. It enhances the real experience of immersive VR. Meanwhile, automatic lip-reading technology can be widely used in the VR system, information security, speech recognition and assisted driving systems. The research of automatic lip-reading involves many fields, such as pattern recognition, computer vision, natural language comprehension, and image processing.

2. BACKGROUND

From the past few surveys about lip to speech generation these are some papers that are referred to as follows.

Pana et al. (2012) discussed the various strategies used for lip segmentation. It remains a difficult problem due to high changeable lip color and low chromatics distinction between the lip and skin. An explicit automatic lip segmentation algorithmic program supports an explicit color transformation in RGB rather than advanced color models. The comparative study with some existing lip segmentation programs has indicated the superior performance of the developed algorithms. [4]

Nasuhal et al. (2013) analyzed that the lip's reading may be widespread application, for example -Audio-Visual Automatic Speech Recognition (AV-ASR), that is speech interface and personal identification. Segmentation of lips is an important part of lip's reading. Lip following may be a way of finding lip to associated lip in successive video outlines. The chan-vese model may be a section-based segmentation rule, which equally is employed as a tracing methodology. This rule can sense the border of an object that is not made public by gradient, where a standard active curve can't be implied [2].

Mardiyantol and Sardjono (2015) discussed the main points of lip shape and offered necessary signals of lip form tracing that were applied for speech detection, lip analysis, and plenty of transmission application. A special acceptable threshold segmentation was given for 6 key points lip's feature abstraction rules. Color transformation in the Red-Green-Blue house and adaptive threshold were applied for the lip segmentation. Curves of the segmental lip were outlined and stuffed along with bounding color. Lastly, the 6 main points that are corners of right or left, minor purpose, 3 points of the cupid on the bow of the lips are observed. Presentation of projected techniques was appropriated and related to offering strategies to achieve a necessary enhancement in the correctness [3].

Rathee (2016) analyzed recognition of the speech supporting the outline of lip actions while talking. Audio speech detection system was the standard for decades and has attained some achievement, however recently the visual's speech detection created curiosity in the minds of researchers for lip's reading. Lip's reading has a spare benefit of elevated correctness and sound freedom. The author gave an associate degree algorithmic program for repeated lip reading. The algorithmic program consisted of 2 major steps: feature extraction and classification of word recognition. Lip's info is derived using lip's geometric and lip's appearance. Correctness obtained from the projected method was ninety-seven percent.

3. PROPOSED SOLUTION

3.1 Constrained Talking Face Generation from Speech.

First reviewed the works on talking face generation that are either constrained by the range of identities they can generate or the range of vocabulary they are limited to. Realistic generation of talking face videos was achieved by a few recent works [5] on videos of Barack Obama. They learn a mapping between the input audio cvit.iiit.ac.in/research/projects/cvit-projects/a-lip-sync-expert-is-all-you-need-forspeech-to-lip-generation-in-the-wild and the corresponding lip landmarks. As they are trained on only a specific speaker, they cannot synthesize for new identities or voices. They also require a large amount of data of a particular speaker, typically a few hours. Recent work along this line [6] proposes to seamlessly edit videos of individual speakers by adding or removing phrases from the speech. They still require an hour of data per speaker to achieve this task. Very recently, another work [7] tries to minimize this data overhead by using a two-stage approach, where they first learn speaker-independent features and then learn a rendering mapping with ≈ 5 minutes of data of the desired speaker. However, they train their speaker-independent network on a significantly smaller corpus and also have an additional overhead of requiring clean training data of each target speaker to generate for that speaker. Another limitation of existing works is in terms of vocabulary. Several works [4] train on datasets with a limited set of words such as GRID (56 words), TIMIT, and LRW (1000 words) which significantly hampers a model from learning the vast diversity of phoneme-viseme mappings in real videos [8]. The proposed work focuses on lip-syncing unconstrained talking face videos to match any target speech, not limited by identities, voices, or vocabulary.

3.2 Unconstrained Talking Face Generation from Speech.

Despite the rise in the number of works on speech-driven face generation, surprisingly, very few works have been designed to lip-sync videos of arbitrary identities, voices, and languages. They are not trained on a small set of identities or a small vocabulary. This allows them to, at test time, lip-sync random identities for any speech. To the best of our knowledge, only two such prominent works [8,9] exist in the current literature. Note that is an extended version of [3]. Both these works [8,9] formulate the task of learning to lip-sync in the wild as follows: Given a short speech segment S and a random reference face image R, the task of the network is to generate a lip-synced version Lд of the input face that matches the audio. Additionally, the LipGAN model also inputs the target face with the bottom-half masked to act as a pose prior. This was crucial as it allowed the generated face crops to be seamlessly pasted back into the original video without further post-processing. It also trains a discriminator in conjunction with the generator to discriminate in-sync or out-of-sync audio-video pairs. Both these works, however, suffer from a significant limitation: they work very well on static images of arbitrary identities but produce inaccurate lip generation when trying to lip-sync unconstrained videos in the wild. In contrast to the GAN setup used in LipGAN [8], by using a pre-trained, accurate lip-sync discriminator that is not trained further with the generator. observing that this is an important design choice to achieve much better lip-sync results.

3.3 Accurate Speech-Driven Lip-Syncing For Videos In The Wild

The core architecture can be summed up as "Generating accurate lip-sync by learning from a well-trained lip-sync expert". To understand this design choice, by first identify two key reasons why existing architectures produce inaccurate lip-sync for videos in the wild. Argued that the loss functions, namely the L1 reconstruction loss used in both the existing works [8,9] and the discriminator loss in LipGAN [8] are inadequate to penalize inaccurate lip-sync generation.

3.3.1 Pixel-level Reconstruction loss is a Weak Judge of Lip-sync

The face reconstruction loss is computed for the whole image, to ensure correct pose generation, preservation of identity, and even background around the face. The lip region corresponds to less than 4% of the total reconstruction loss (based on the spatial extent), so a lot of surrounding image reconstruction is first optimized before the network starts to perform fine-grained lip shape correction. This is further supported by the fact that the network begins morphing lips only at around half-way (≈ 11th epoch) through its training process (≈ 20 epochs [8]). Thus, it is crucial to have an additional discriminator to judge lip-sync, has also done in LipGAN [8]. But how powerful is the discriminator employed in LipGAN?

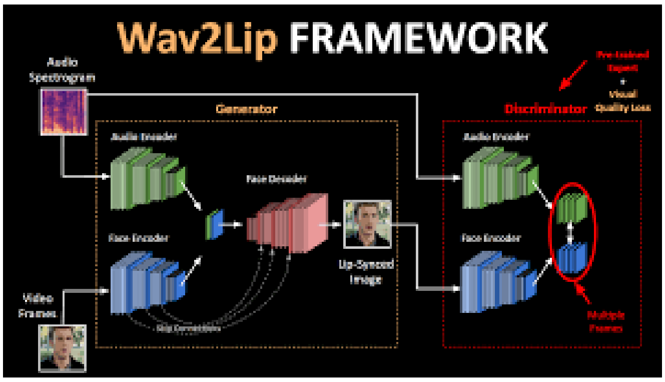

Figure - 1: Wav2Lip Framework

3.3.2 A Weak Lip-sync Discriminator

Found that the LipGAN's lip-sync discriminator is only about 56% accurate while detecting off-sync audio-lip pairs on the LRS2 test set. For comparison, the expert discriminator that is used in this work is 91% accurate on the same test set. As hypothesize two major reasons for this difference. Firstly, LipGAN's discriminator uses a single frame to check for lip-sync. In figure 1, it shows that a small temporal context is very helpful while detecting lip-sync. Secondly, the generated images during training contain a lot of artifacts due to the large scale and pose variations. Arguing that training the discriminator in a GAN setup on these noisy generated images, as done in LipGAN, results in the discriminator focusing on the visual artifacts instead of the audio-lip correspondence. This leads to a large drop in off-sync detection accuracy. By arguing and show that the "real", accurate concept of lip-sync captured from the actual video frames can be used to accurately discriminate and enforce lip-sync in the generated images.

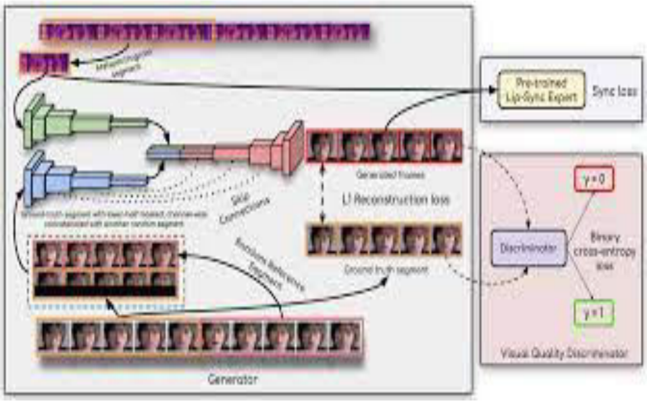

Figure – 2: Accurate lip-sync by learning from an "already well-trained lip-sync expert"

3.3.3 A Lip-sync Expert Is All You Need

Based on the above two findings, it was proposed to use a pre-trained expert lip-sync discriminator that is accurate in detecting sync in real videos. Also, it should not be fine-tuned further on the generated frames like it is done in LipGAN. One such network that has been used to correct lip-sync errors for creating large lip-sync datasets [1] is the SyncNet model. proposed to adapt and train a modified version of SyncNet for the task.

Overview of SyncNet. SyncNet inputs a window V of Tv consecutive face frames (lower half only) and a speech segment S of size Ta × D, where Tv and Ta are the video and audio timesteps, respectively. It is trained to discriminate sync between audio and video by randomly sampling an audio window Ta × D that is either aligned with the video (in-sync) or from a different time-step (out-of-sync). It contains a face encoder and an audio encoder, both comprising a stack of 2D-convolutions. L2 distance is computed between the embeddings generated from these encoders, and the model is trained with a max-margin loss to minimize (or maximize) the distance between synced (or unsynced) pairs. Figure 2: this approach generates accurate lip-sync by learning from an "already well-trained lip-sync expert". Unlike previous works that employ only a reconstruction loss [9] or train a discriminator in a GAN setup [8], using a pre-trained discriminator that is already quite accurate at detecting lip-sync errors. Showing that fine-tuning it further on the noisy generated faces hampers the discriminator's ability to measure lip-sync, thus also affecting the generated lip shapes. Additionally, also employ a visual quality discriminator to improve the visual quality along with the sync accuracy.

The expert lip-sync discriminator. By making the following changes to SyncNet to train an expert lip-sync discriminator that suits the lip generation task. Firstly, instead of feeding grayscale images concatenated channel-wise as in the original model, feeding color images. Secondly, proposed model is significantly deeper, with residual skip connections. Thirdly, inspired by this public implementation2, used a different loss function: cosine-similarity with binary cross-entropy loss. That is, computing a dot product between the ReLU-activated video and speech embeddings v,s to yield a single value between [0, 1] for each sample that indicates the probability that the input audio-video pair is in sync: Psync = v · s max(‖v‖2 · ‖s ‖2, ϵ) (1) by training expert lip-sync discriminator on the LRS2 train split (≈ 29 hours) with a batch size of 64, with Tv = 5 frames using the Adam optimizer [7] with an initial learning rate of 1e −3 .The expert lip-sync discriminator is about 91% accurate on the LRS2 test set, while the discriminator used in LipGAN is only 56% accurate on the same test set.

4. EXPERIMENTAL RESULTS

	LRW [11]			LRS2[10]			LRS3 [1]		
Method	**LSE-D ↓**	**LSE-C ↑**	**FID ↓**	**LSE-D ↓**	**LSE-C ↑**	**FID ↓**	**LSE-D ↓**	**LSE-C ↑**	**FID ↓**
Speech2Vid [9]	13.14	1.762	11.15	14.23	1.587	12.32	13.97	1.681	11.91
LipGAN [8]	10.05	3.350	2.833	10.33	3.199	4.861	10.65	3.193	4.732

Wav2Lip (proposed)	**6.512**	**7.490**	3.189	**6.386**	**7.789**	4.887	**6.652**	**7.887**	4.844
Wav2Lip + GAN (proposed)	6.774	7.263	**2.475**	6.469	7.781	**4.446**	6.986	7.574	**4.350**
Real Videos	7.012	6.931	-	6.736	7.838	-	6.956	7.592	-

Table 1: We propose two new metrics "Lip-Sync Error-Distance" (lower is better) and "Lip-Sync Error-Confidence" (higher is better), that can reliably measure the lip-sync accuracy in unconstrained videos. We see that the lip-sync accuracy of the videos generated using Wav2Lip is almost as good as real synced videos. Note that we only train on the train set on LRS2 [1], but we comfortably generalize across all datasets without any further fine-tuning. We also report the FID score (lower is better), which clearly shows that using a visual quality discriminator improves the quality by a significant margin.

Method	**Video Type**	**LSE-D ↓**	**LSE-C ↑**	**FID ↓**	**Sync Acc.**	**Visual Qual.**	**Overall Exp.**	**Preference**
Unsynced Orig. Videos	Dubbed	12.63	0.896	—	0.21	4.81	3.07	3.15%
Speech2Vid [9]		14.76	1.121	19.31	1.14	0.93	0.84	0.00%
LipGAN [8]		10.61	2.857	12.87	2.98	3.91	3.45	2.35%
Wav2Lip (proposed)		**6.843**	**7.265**	15.65	**4.13**	3.87	4.04	34.3%
Wav2Lip + GAN (proposed)		7.318	6.851	**11.84**	4.08	**4.12**	**4.13**	**60.2%**
Without Lip-syncing	Random	17.12	2.014	—	0.15	4.56	2.98	3.24%
Speech2Vid [9]		15.22	1.086	19.98	0.87	0.79	0.73	0.00%
LipGAN [8]		11.01	3.341	14.60	3.42	3.77	3.57	3.16%
Wav2Lip (proposed)		**6.691**	**8.220**	14.47	**4.24**	3.68	4.01	29.1%
Wav2Lip + GAN (proposed)		7.066	8.011	**13.12**	4.18	**4.05**	**4.15**	**64.5%**
Without Lip-syncing		16.89	2.557	—	0.11	4.67	3.32	8.32%
Speech2Vid [9]		14.39	1.471	17.96	0.76	0.71	0.69	0.00%
LipGAN [8]		10.90	3.279	11.91	2.87	3.69	3.14	1.64%

Wav2Lip (proposed)	TTS	**6.659**	**8.126**	12.77	**3.98**	3.87	3.92	41.2%
Wav2Lip + GAN (proposed)		7.225	7.651	**11.15**	3.85	**4.13**	**4.05**	**51.2%**
Untranslated Videos		7.767	7.047	—	4.83	4.91	—	—

Table 2: Real world evaluation using the newly collected ReSyncED benchmark. Evaluating using both quantitative metrics and human evaluation scores across three classes of real videos. Observed that in all cases, the Wav2Lip model produces high-quality, accurate lip-syncing videos. Specifically, the metrics indicate that the lip-synced videos are as good as the real synced videos. Also noted that human evaluations indicate that there is a scope for improvement when trying to lip-sync TTS generated speech. Finally, it is worth nothing that the proposed lip-synced videos are preferred over existing methods or the actual unsynced videos over 90% of the time.

5. CONCLUSION:

In this paper, a novel approach to generate accurate lip-synced videos in the wild. After discussing that a pretrained, accurate lip-sync expert" can enforce accurate, natural lip motion generation. Before evaluating the model, by re-examining the current quantitative evaluation framework and highlighting several major issues. To resolve them, several new evaluation benchmarks, and metrics, and a real-world evaluation set. Believing that future works can be reliably judged in this new framework. By outperforming the current approaches by a large margin in both quantitative metrics and human evaluations and believe that all efforts and ideas in this problem can lead to new directions such as synthesizing expressions and head-poses along with the accurate lip movements.

Also, investigating the problem of synthesizing speech based on lip movements, will formulate the task at hand as a sequence-to-sequence problem, and show that by doing so, can achieve significantly more accurate and natural speech than previous methods and will evaluate the model with extensive quantitative metrics and human studies.

REFERENCES

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. LRS3-TED: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496 (2018).

[2] Lele Chen, Haitian Zheng, Ross K Maddox, ZhiyaoDuan, and Chenliang Xu. 2019. Sound to Visual: Hierarchical Cross-Modal Talking Face Video Generation. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops.

[3] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. 2017. You said that? arXiv preprint arXiv:1705.02966 (2017).

[4] Lele Chen, Ross K Maddox, ZhiyaoDuan, and Chenliang Xu. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7832–7841.

[5] SupasornSuwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (TOG) 36, 4 (2017), 95.

[6] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, ZeyuJin, Christian Theobalt, and Maneesh Agrawala. 2019. Text-based editing of talking-head video. ACM Transactions on Graphics (TOG) 38, 4 (2019), 1–14.

[7] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2019. Neural Voice Puppetry: Audio-driven Facial Reenactment. arXiv preprint arXiv:1912.05566 (2019).

[8] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. 2019. Towards Automatic Face-to-Face Translation. In Proceedings of the 27th ACM International Conference on Multimedia. ACM, 1428–1436.

[9] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. 2019. You said that?:Synthesising talking faces from audio. International Journal of Computer Vision 127, 11-12 (2019), 1767–1779.

[10] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. 2018. Deep Audio-Visual Speech Recognition. In arXiv:1809.02108.

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