The Interplay of Speech and Lip Movements

Prajwal K R*, Rudrabha Mukhopadhyay*, Sindhu Hegde*, Vinay Namboodiri† and C.V. Jawahar*
*Centre of Visual Information Technology
IIIT Hyderabad, Hyderabad, India - 500032
†University of Bath
Claverton Down
Bath, United Kingdom
Contact: {prajwal.k, radrabha.m, sindhu.hegde}@research.iiit.ac.in

I. Introduction

The interplay of speech and lip movements has received considerable attention in recent years. The strong connection between these two modalities allows us to understand and predict one from the other. In this demo video, we extensively explore this close relationship through several of our state-of-the-art efforts. We will also look at the multitude of applications that can be spawned from each of these works.

II. Lip-syncing any identity to a given speech segment

We start with the practically important task of lip-syncing arbitrary talking face videos to the desired target speech. We demonstrate two models for this task, each addressing a significant limitation of the previous works. The first model LipGAN [1], is the first speaker-independent model to generate lip-synced face images that can be seamlessly pasted back into the video frame. We show that this is a critical feature if we wish to apply a lip-syncing model on real-world videos. Further, it is the first model to (i) employ a lip-sync discriminator in a GAN setup, and (ii) be able to generate lip-synced faces in any pose. Our model produced state-of-the-art results on popular datasets and was widely appreciated in the community. Now that we have a significant improvement in the generation quality, we follow it up with a second model [2], that addresses the next major issue: poor lip-sync accuracy for videos in the wild. By using an even more powerful, accurate lip-sync discriminator and a carefully-designed training setup, we show that we can lip-sync arbitrary talking face videos in the real-world with remarkable accuracy and quality. We will show the following applications in the demo video using both [1] and [2].

1) lip-syncing CGI characters in real animated movie clips,
2) correcting the lip movements in (automatically) dubbed lecture videos and movies,
3) compressing a video call or an online lecture video,
4) lip-syncing translated press conferences, interviews, and speeches, and
5) content creation for social media and gaming.

III. Generating Speech solely from the lip movements

Next, we explore the opposite [3], a more challenging task of generating accurate speech solely from the lip movements. We take inspiration from the fact that deaf individuals or professional lip readers find it easier to lip read people with whom they frequently interact. Consequently, we explore this question from a data-driven learning perspective, "How accurately can we infer an individual's speech style and content from his/her lip movements?" We collect and release a large-scale speaker-specific lip to speech benchmark, and propose a novel sequence-to-sequence architecture for the problem. We demonstrate that we can indeed generate accurate speech from silent lip videos in large vocabulary, unconstrained settings for the first time. In addition to single-speaker lip to speech, we are the first to demonstrate results on the highly challenging task of "word-level multi-speaker lip to speech". We showcase multiple such results in our demo video. We finally show an application of our model where we generate seamlessly generate and fill in missing speech segments during a lecture thus maintaining user experience.

IV. Audio-visual speech enhancement without a real visual stream

While the goal of the previous problem is to generate speech in its entirety, our next effort in this space is to enhance a noisy speech segment. The current state-of-the-art works [4], [5] use the lip movements to perform this task of audio-visual speech enhancement. These works take the lip movements of a speaker along with the corrupted speech to generate impressive clean speech segments. However, an inherent weakness of these works is the limited field of application. Both of these works are applicable only for enhancing corrupted speech in near-frontal talking face videos with clearly visible lips. On the other hand, audio-only works, albeit less accurate are applicable in a wide range of situations without any such special needs. Our goal in this work, is to design a new paradigm termed as "pseudo-visual speech enhancement", which combines the best of both worlds. Using the synthesized lip movements from Wav2Lip [2], we show that we can exploit the benefits of audio-visual works, while still being applicable to all situations. The main contributions of our work lie in the formulation of the pseudo-visual stream and then using it further for enhancing the noisy speech. This work also opens up a myriad of applications such as:

1) Helping reporters report in windy/noisy conditions

2) Improving speech recognition systems by removing background music or noise for automatic subtitling in movies (also applicable for animated movies)
3) Helping vloggers record their videos in highly noisy conditions
4) Removing background noise from public speeches
5) Enhancing historically important recordings and transmissions

Currently, this work is under review in a popular computer vision conference. We will be describing the work in more detail in our final submission as well as make changes in the final demo video showcasing its results.

Our four models together, comprehensively demonstrate the recent breakthroughs in the space of lip movements and speech along with several impactful real-world applications. We have attached publicly available demos, code, and model links below for the already published works. We will be doing the same for the paper under review, once it is accepted.

V. Conclusion

Through this demo, we aim to showcase the wide range of applications and endless possibilities made possible by models that jointly understand the fine-grained relationship between speech and the lip movements. We believe this consolidated demo video will encourage future research efforts in this space.

References

[1] P. K R, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, and C. V. Jawahar, "Towards automatic face-to-face translation," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: ACM, 2019. [Online]. Available: http://doi.acm.org/10.1145/3343031.3351066

[2] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 484–492. [Online]. Available: https://doi.org/10.1145/3394171.3413532

[3] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "Learning individual speaking styles for accurate lip to speech synthesis," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 793–13 802.

[4] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *CoRR*, vol. abs/1804.04121, 2018. [Online]. Available: http://arxiv.org/abs/1804.04121

[5] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *CoRR*, vol. abs/1804.03619, 2018. [Online]. Available: http://arxiv.org/abs/1804.03619