Woven Artificial Profile (WARP): Face Video Synthesis from Audio and Profile

By- B21SK05

DEMO

Demo of the fully functioning application.

Web application is created using FastAPI and React JS.

Link to Sample Output 1: https://drive.google.com/file/d/1QUofEJGhsWNXOKQJqPo7k6zstDtvkwJO/view?usp=sharing

1

Introduction

- Motivation
- Problem Statement
- Demographic
- Literature Work

Motivation

Indirect vs. Direct Communication

- Indirect modes of communication lack the essence of Direct means- tone, expressions, gestures and intent.
- Information can be easily misinterpreted via indirect communication.
- Video/Audio calls, though a great alternative, are highly demanding of time and availability of all parties involved.
- WARP aims to personalise indirect communication.

Problem Statement-

Given just a face image and text, WARP is a highly versatile model that can generate realistic talking videos with audio.

Demographic

- Entertainment & Broadcasting Industries
- Educational Institutes
- Marketing Sectors
- Virtual Assistants & Avatars
- Remote Officials
- General Public

Literature Work

Neural Voice Puppetry [5]

Audio driven network that takes audio as input and uses an intermediate 3D face model that reconstructs the target person speaking and checks for synchronisation.

Disadvantage- Audio centric.

Temporal GAN [6]

Generative model capable of learning semantic representations unlabelled videos. Generates videos with number of frames equal to number of latent variables.

Disadvantage- Requires a large latent space. Audio inclusion is out of scope.

DualLip [7]

Leverages task duality by lip reading and generation using unlabelled text and lip videos. Synchronizes the lip generation model to the text.

Disadvantage- Replaces the lip on the guide face by vector concatenation.

2

The Product

Modules, Implementation & Results

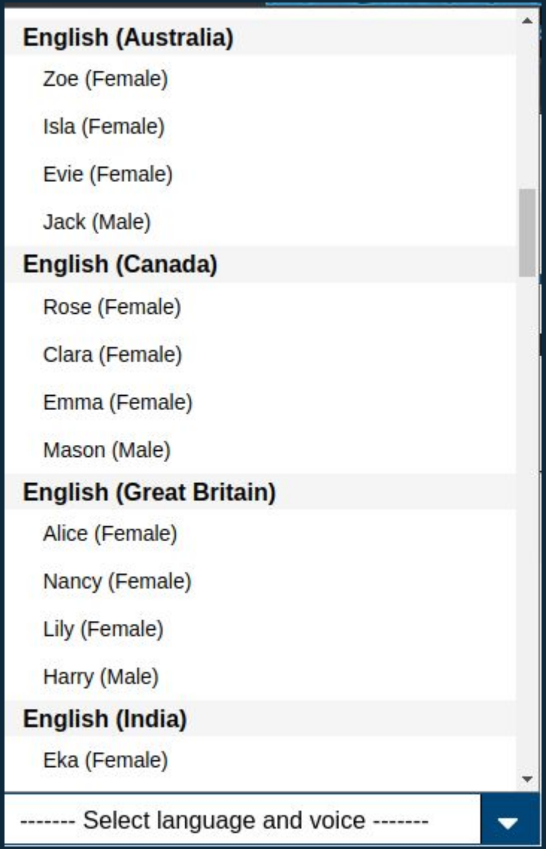
Modules

- **Text-to-Speech**: *audio generation from text*
- **StyleGAN**: *fake face generation*
- **LipGAN**: *lip sync model*

WARP- Ensemble of the above three modules.

Text-to-Speech

- Converts a given text to audio.
- VoiceRSS API service.
- Cloud based API
- Offers several options for language and voice


English (Australia)

Zoe (Female)

Isla (Female)

Evie (Female)

Jack (Male)

English (Canada)

Rose (Female)

Clara (Female)

Emma (Female)

Mason (Male)

English (Great Britain)

Alice (Female)

Nancy (Female)

Lily (Female)

Harry (Male)

English (India)

Eka (Female)

------- Select language and voice -------

http://www.voicerss.org/

StyleGAN [1]

- A style-based generator architecture for Generative Adversarial Networks .
- Generates fake faces.
- Produces high quality Images of size 1024x1024.
- StyleGAN[1] is an improved version of ProGAN [4].
- Generator has 2 networks (mapping and synthesis).
- Adaptive Instance Normalization (AdaIN).
- Discriminator is similar to Synthesis network.
- Celeba HQ dataset.
- FID score of the GAN is 167.26 for a sample of 200 fake images.

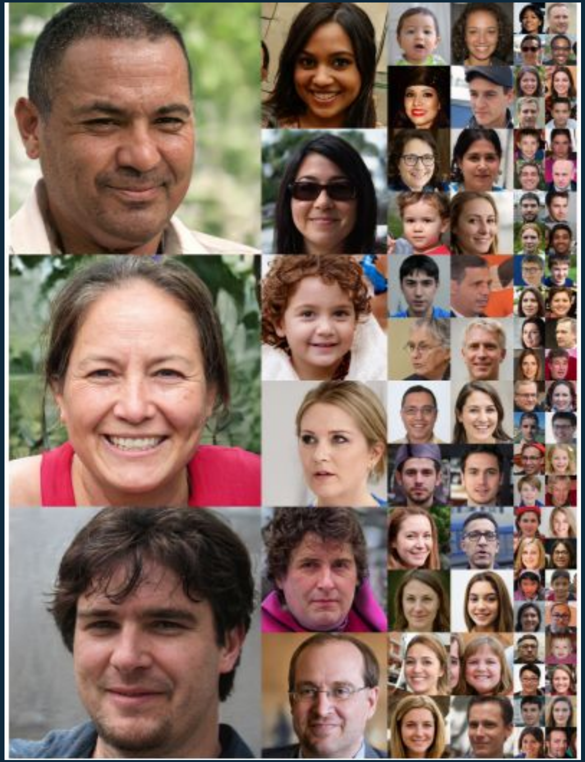


Image credits [1]

LipGAN: Description

- [2]
- Generates lip synced transitional videos given an audio and source-destination images.
- Dataset- LRS2[3] dataset consisting of very short 1-10 second annotated videos.
- Generator: Audio Encoder, Face Encoder. For a given frame, output is a modified frame to sync with the corresponding audio segment.
- Discriminator: Audio and Face Encoder. Binary class output describing the level of synchronization.

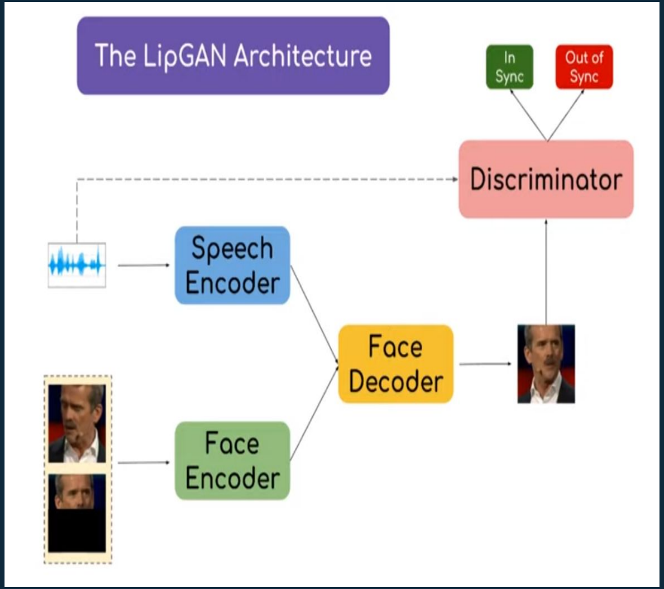
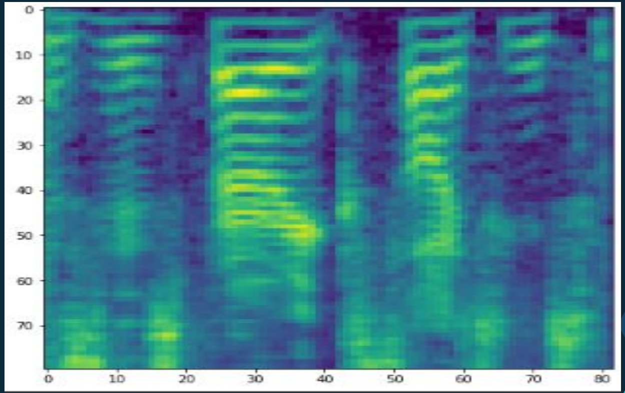


Image Source- [3]

LipGAN: Preprocessing for Training

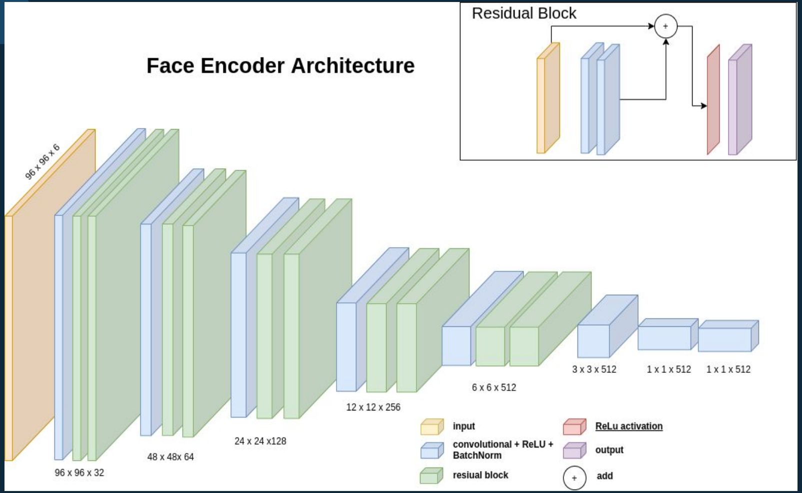
- Original data is just simple videos.
- `Valid` frames: frames containing face views only.
- Use ***dlib***'s Frontal Face Detector (FFD) to extract *valid* frames and keep the ones with the maximum bounding box area of face view.
- Extract audio using ***ffmpeg***.
- Generate the Mel Spectrogram* and save the Mel Frequency Cepstral Coefficients (MFCC) features using ***Librosa***.

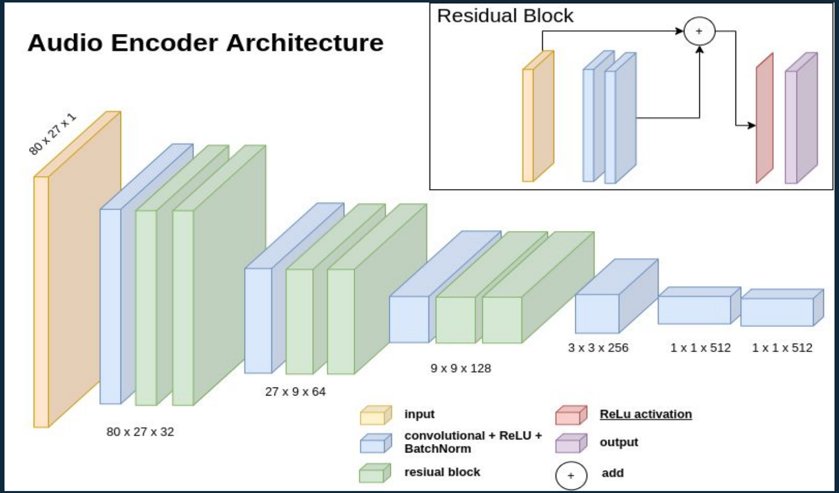


* Spectrogram** frequencies transformed to Mel Scale***.
** FFTs of audio signals stacked on top of each other.
*** scale of human audio perception rather than audio itself.

LipGAN: Training

- The model takes a random masked frame (other than the ground truth), target audio segment and ground truth frame as input.
- The training objective is to recreate a frame matching the ground truth frame for a given audio segment. Similar to Conditional GAN.
- A joint embedding is formulated by concatenating the face and audio encoder outputs.
- The decoder has a Sigmoid head for classifying the recreated frame as real/fake wrt. the ground truth.





Architecture Diagrams for Face and Audio Encoders generated using https://app.diagrams.net/

WARP: Video Generation

- The model gets the text as input and generates the audio using TTS technique.
- If no user image is given, it generates a fake face for the video.
- Extracts the Mel Spectrogram from the audio and initializes the frames to match the audio length.
- For each frame, the model predicts the face view with accurate lip pose corresponding to the audio segment.
- Stitches and writes the video together using OpenCV.
- Uses ffmpeg to combine the audio and video for the final output.

3

Extras

- Timeline
- Contributions
- References

Contributions

Nethra Gunti

S20180010061

- LipGAN Preprocessing
- LipGAN Training
- Video Generation

Tarun Teja Obinna

S20180010120

- StyleGAN
- LipGAN Training
- Web Application

Nithin Ramancha

S20170010120

- Text-to-Speech
- Web App UI
- Web Application

References

1. Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4401-4410).
2. Prajwal K R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C V Jawahar. 2019. Towards Automatic Face-to-Face Translation. In proceedings of the 27th ACM International Conference on Multimedia (MM '19).
3. T. Afouras, J. S. Chung, A. Senior, O. Vinyals, A. Zisserman, Deep Audio-Visual Speech Recognition, arXiv:1809.02108.
4. Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196.
5. Thies, J., Elgharib, M., Tewari, A., Theobalt, C., & Nießner, M. (2020, August). Neural voice puppetry: Audio-driven facial reenactment. In European Conference on Computer Vision (pp. 716-731). Springer, Cham.
6. Saito, M., Matsumoto, E., & Saito, S. (2017). Temporal generative adversarial nets with singular value clipping. In Proceedings of the IEEE international conference on computer vision (pp. 2830-2839).
7. Chen, W., Tan, X., Xia, Y., Qin, T., Wang, Y., & Liu, T. Y. (2020, October). DualLip: A System for Joint Lip Reading and Generation. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 1985-1993).

Thanks!

Nethra Gunti (S20180010061)
Tarun Teja Obbina (S20180010120)
Nithin Ramancha (S20170010120)

Credits

Special thanks to all the people who made and released these awesome resources for free:

- ◇ Presentation template by SlidesCarnival
- ◇ Photographs by Unsplash