Speech2Face: Reconstructed Lip Syncing with Generative Adversarial Networks

David Bunker

October 30, 2017

1 Abstract

The purpose of this project is to produce facial reenactment from a target video and provided source audio. Audio frequency information as well as facial identity recovery via non-rigid model-based bundling is derived from video training data. This multimodal training data is processed by a generative adversarial network which produces G, a generator of face shape based on audio information and D, a discriminator of the face shapes produced. G can subsequently be used to produce a reconstructed synchronization.

2 Context

Recent breakthroughs in artificial intelligence and deep learning have revolutionized many areas of machine learning research including text to audio speech [21], audio speech to text [5], video to text lip reading [2], text to machine translated text [23], audio music generation [9] and image to image conversion [10]. Video to video facial matching and reconstruction has been shown to be effective when using model-based bundling and subspace deformations as demonstrated by Face2Face [20]. Combining these techniques with techniques from generative adversarial networks [14] and multi modal deep learning [13] has the potential to allow for convincing resyncing based on audio.

Although it is not possible to get the full range of emotion from audio alone as each phoneme can have many visemes, it is possible to get a good approximation [2, 19]. Audio visual lip syncing is a well explored topic due to its importance in film, game, and virtual reality dubbing. Past techniques

include utilizing database mapping [4, 18], following a probability distribution over possible facial motions [3, 12], using physiological based models focusing on lower mouth movement followed by motion blending [6, 22], as well as trained neural networks [8, 11].

The primary advantage neural networks have over other methods is they do not require specialized knowledge of psycholinguistics or facial structures, however their primary disadvantage is they often require a large amount of labeled training data. Generative adversarial networks help to alleviate this problem by allowing for unsupervised learning [14]. This technique has the potential to greatly simplify audio visual synchronized reconstruction.

3 Methodology

The GAN model consists of a generator G which maps a vector z sampled from a prior distribution P_z, to a set of face shapes and audio frequencies from video training data, and a discriminator D which when provided with face shapes and audio frequencies outputs a probability indicator as to whether the input is as convincing as the training data.

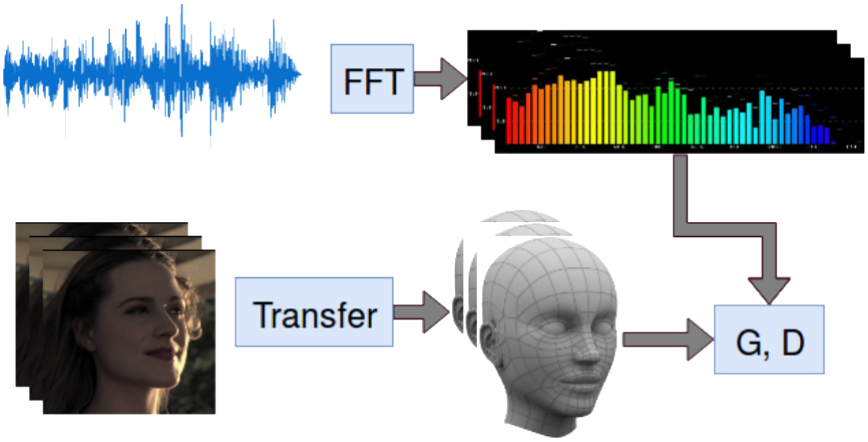

Figure 1: Generator G and discriminator D are trained by audio frequency and face shapes from video data.

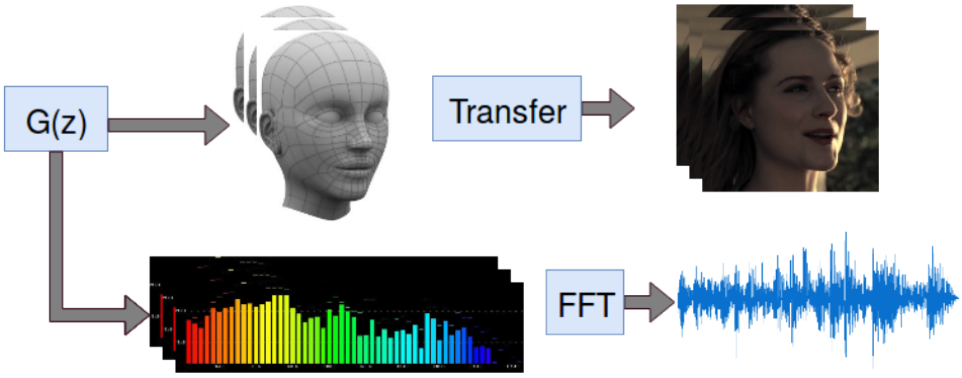

Figure 2: Given a vector z from prior distribution Z, $G(z)$ outputs reasonable face shapes and audio frequencies based on the training data.

The goal of the generator is to have the discriminator classify its video output as valid while the goal of the discriminator is to discriminate between what is produced by the generator and what is sampled from the training data P_{data}. This can be represented by the following minimax game.

$$\min_G \max_D E_{x\sim p_{\text{data}}} \log D(x) + E_{z\sim p_z}[\log(1 - D(G(z)))] \tag{1}$$

Both G and D are deep neural networks trained with an alternating gradient descent algorithm [7]. Once training is complete, D is able to identify video data that appears fake and D is able to produce convincing video data.

The next step is to create a process such that given a provided audio frequency input and face shape information other than mouth and jaw to convincingly generate the missing mouth and jaw face shape. Here it is possible to apply a similar technique to that used by image inpainting [1,24].

The generator is trained to produce convincing video data from a vector z, however the output would be arbitrary video data of a face shape speaking, not the desired output $x_{reconstructed}$. To produce the requisite missing mouth and jaw face shape information to generate video data specific to the provided data y requires projecting the requisite properties onto the manifold of G, generating $\hat{z}$, $G(\hat{z})$ being the missing mouth and jaw data. Here M is the binary mask for the parts of the image to be retained.

$$x_{\text{reconstructed}} = M \odot y + (1 - M) \odot G(\hat{z}) \tag{2}$$

Creating $x_{reconstructed}$ from the projection onto G requires minimizing the contextual loss $\mathcal{L}_{\text{contextual}}(z)$. This can be represented by the element wise subtraction of the generated values $M \odot G(z)$ from the known values $M \odot y$ and taking the l_1 norm.

$$\mathcal{L}_{\text{contextual}}(z) = ||M \odot G(z) - M \odot y||_1 \tag{3}$$

It is important the discriminator D finds the new video data convincing, this can be represented by $\mathcal{L}_{\text{perceptual}}(z)$.

$$\mathcal{L}_{\text{perceptual}}(z) = \log(1 - D(G(z))) \tag{4}$$

The value $\hat{z}$ that will minimize contextual and perceptual loss is calculated using back propagation. Here λ represents the estimated importance of contextual loss verses perceptual loss. This is expected to be less than 1 as the contextual data represents ground truth.

$$\hat{z} \equiv \arg\min_z(\mathcal{L}_{\text{contextual}}(z) + \lambda \mathcal{L}_{\text{perceptual}}(z)) \tag{5}$$

This value $\hat{z}$ can then be used to construct the final video data $x_{reconstructed}$.

4 Improvements and Challenges

There are many different types of neural networks that could potentially be applied to this problem. Using deep convolutional neural networks [14] could allow for the use of raw image data instead of face shape matching utilizing non-rigid model-based bundling. It is also possible a neural network could be used to better inform the face shape matching. How much preprocessing should be performed can be similarly examined in the use of Fast Fourier Transform versus raw audio files [16].

It may also be effective to apply a recurrent neural network which would allow an LSTM to maintain more past information, as the parameter for how much video time to use and the frame rate would otherwise need to be preset. Unfortunately it is difficult to know what will produce the best

results so some degree of trial and error is required. The combination of specialized knowledge and machine learning techniques is the most likely to result in an effective solution.

Requiring extensive time and CPU and GPU power to train can also reduce the effectiveness of machine learning techniques. It may be possible to improve results and increase speed utilizing techniques drawn from stochastic superoptimization and logic programming [17].

The fact that the information provided is multimodal, involving both face shape and audio frequency data, may also pose a challenge although there are techniques to mitigate this [13]. Deep learning parameters will need to be preset or trained against as well as the use of pre and post processing techniques such as Poisson blending or audio source separation and enhancement. One technique that shows promise is the merging of nonnegative matrix factorization and deep learning to derive meaningful speech separation and enhancement [15].

Another potential challenge is that GAN inpainting can have odd results if training data is insufficient to cover all cases [24]. Live action video in particular can result in what is colloquially known as "uncanny valley." Reconstructed lip syncing techniques could be applied to animated media without this challenge, paving the way for future work on live action.

GANs have the ability to infer semantic information from provided environmental data. The generator G is trained to describe a convincing manifold over the probability distribution P_{data} offering the potential of derive meaningful information for reconstructed lip syncing as well as general segmentation without the need for explicitly labeled data.

5 Expected Results

Utilizing methods drawn from deep learning research such as generative adversarial networks and potentially deep recurrent and convolutional networks as well as specialized knowledge in graphics for facial identity recovery such as non-rigid model-based bundling from video footage opens the possibility for a photo-realistic lip-syncing system that could reconstruct lip syncing from a target video given a provided source audio.

References

[1] Brandon Amos. Image Completion with Deep Learning in TensorFlow. Accessed: October, 2017.

[2] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. Lipnet: Sentence-level lipreading. *CoRR*, abs/1611.01599, 2016.

[3] Matthew Brand. Voice puppetry. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, pages 21–28, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[4] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '97, pages 353–360, New York, NY, USA, 1997. ACM Press/Addison-Wesley Publishing Co.

[5] Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. Wav2letter: an end-to-end convnet-based speech recognition system. *CoRR*, abs/1609.03193, 2016.

[6] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: An animator-centric viseme model for expressive lip synchronization. *ACM Trans. Graph.*, 35(4):127:1–127:11, July 2016.

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

[8] Pengyu Hong, Zhen Wen, and T. S. Huang. Real-time speech-driven face animation with expressions using neural networks. *Trans. Neur. Netw.*, 13(4):916–927, July 2002.

[9] Allen Huang and Raymond Wu. Deep learning for music. *CoRR*, abs/1606.04930, 2016.

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.

[11] Samuli Laine, Tero Karras, Timo Aila, Antti Herva, and Jaakko Lehtinen. Facial performance capture with deep neural networks. *CoRR*, abs/1609.06536, 2016.

[12] Gerard Llorach, Alun Evans, Josep Blat, Giso Grimm, and Volker Hohmann. Web-based live speech-driven lip-sync. In *8th International Conference on Games and Virtual Worlds for Serious Applications, VS-GAMES 2016, Barcelona, Spain, September 7-9, 2016*, pages 1–4, 2016.

[13] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 689–696, USA, 2011. Omnipress.

[14] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[15] Jonathan Le Roux, John R. Hershey, and Felix Weninger. Deep NMF for speech separation. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 66–70, 2015.

[16] Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Bo Li, Arun Narayanan, Ehsan Variani, Michiel Bacchiani, Izhak Shafran, Andrew Senior, Kean Chin, Ananya Misra, Chanwoo Kim, Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Bo Li, Arun Narayanan, Ehsan Variani, Michiel Bacchiani, Izhak Shafran, Andrew Senior, Kean Chin, Ananya Misra, and Chanwoo Kim. Multichannel signal processing with deep neural networks for automatic speech recognition. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 25(5):965–979, May 2017.

[17] Eric Schkufza, Rahul Sharma, and Alex Aiken. Stochastic superoptimization. *CoRR*, abs/1211.0557, 2012.

[18] Ingmar Steiner, Sébastien Le Maguer, and Alexander Hewer. Synthesis of tongue motion and acoustics from text using a multimodal articulatory database. *CoRR*, abs/1612.09352, 2016.

[19] Sarah L. Taylor, Barry-John Theobald, and Iain A. Matthews. A mouth full of words: Visually consistent acoustic redubbing. In *ICASSP*, pages 4904–4908. IEEE, 2015.

[20] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nie. Demo of face2face: Real-time face capture and reenactment of rgb videos. In *ACM SIGGRAPH 2016 Emerging Technologies*, SIGGRAPH '16, pages 5:1–5:2, New York, NY, USA, 2016. ACM.

[21] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016.

[22] Li Wei and Zhigang Deng. A practical model for live speech-driven lip-sync. *IEEE Computer Graphics and Applications*, 35(2):70–78, mar 2015.

[23] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

[24] Raymond A. Yeh, Chen Chen, Teck-Yian Lim, Mark Hasegawa-Johnson, and Minh N. Do. Semantic image inpainting with perceptual and contextual losses. *CoRR*, abs/1607.07539, 2016.