NEUR189B PO.01 Final Blog 3 - Blake Buckner

In this project, I set out to use Convolutional Neural Networks to create a system for lip reading using only visual data. Below is my target pipeline, which was slightly altered due to limitations but remains mostly accurate to the actual implementation. I ended up changing a lot through the process and went back and forth with ideas, running into different errors, incompatibilities, and other problems that disrupted my plans.

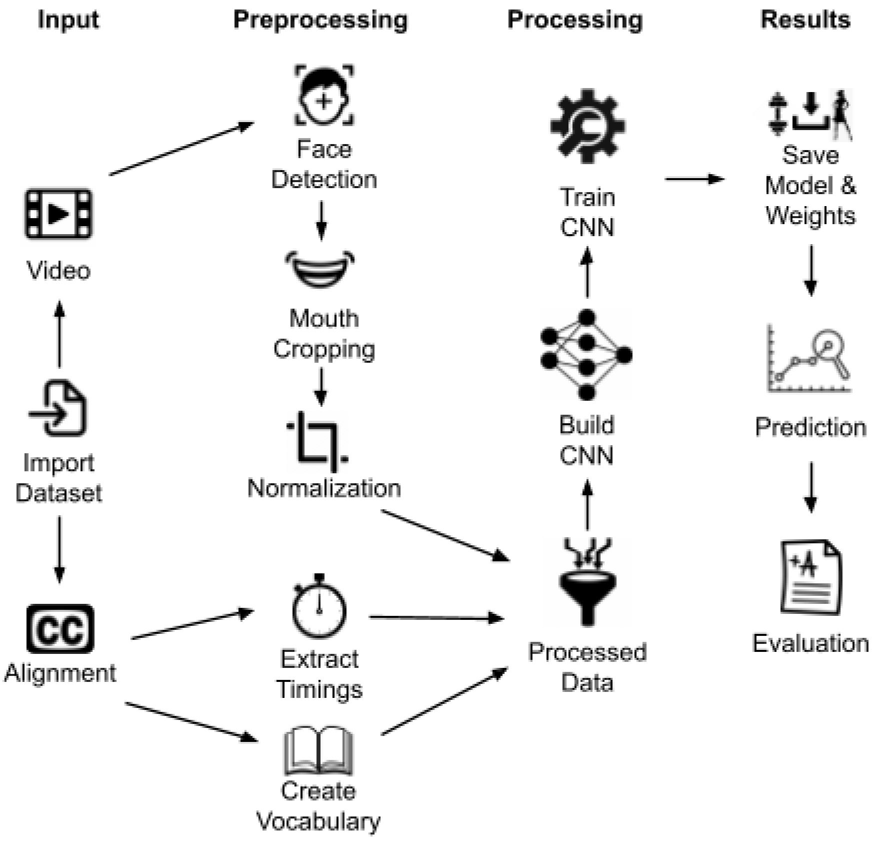

Figure 1: Full Project Pipeline

The concrete results I have ended up with could be considered underwhelming. I could tell with the lack of computing and time, the results wouldn't be the most accurate, but they also just weren't the most substantial either. The loss ended up converging reasonably, but inspecting the actual results shows that not much of value is precisely shown. Despite this I think the methods themselves were actually effective. There were a few steps I was unable to quite integrate properly, but it appears that with them, it could have been very useful.

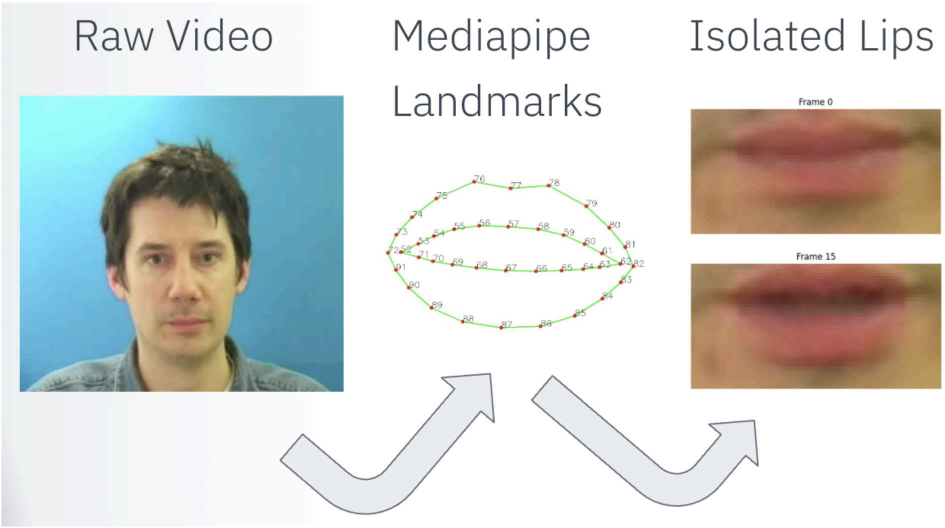

Figure 2: Mediapipe ROI Isolation

My initial plans for isolating the lips used dlib, which, it turns out, was an outdated method that was apparently becoming deprecated. I had to switch to using mediapipe, which worked really well. Mediapipe has lots of points that it detects on the face and lets you select which you want for your region of interest, which you can see in Figure 2 above. This was one of the methods that I would consider especially effective.

Since the last blog post, I fully went in on using visemes. The more I learned about them, the more they seemed effective for various needs I had. I ended up using a chart (chrismbirmingham, 2025) to map visemes to Araphbet phonemes, and those phonemes could be looked up for every word using the CMU Dict (Urban, 2025). Mapping all of the visemes allows me to get all of the specifically visible details without trying to guess at connections that aren't possible to differentiate. They are also very useful because there are only 16 visemes (by this specific mapping) so they are much easier to use for other data. I do not need to actually apply every word possible to whatever is being said, so there aren't problems that would arise from encountering unseen words. Also, having 16 visemes means I could encode each into integers, which are ideal for using in neural networks since they work best with numeric items. This could potentially be even further streamlined by using hexadecimal representations. Using visemes still have other advantages I never got to like in my initial plans to make predictions context dependent.

There are a few limitations and future steps that could be accomplished with more time. One of these is that I really wish I had a more visual representation of my results. Overall, because of the lower processing power I have access to I cannot fully evaluate, but using more manual tests, I found it to be overall not very accurate. This leads on to one of the other limitations, which was the continuity between frames. My plans were to use a system to allow continuity between frames because they are obviously directly influencing each other, but I just ran out of time to effectively implement it. The full results section is definitely lacking, but I unfortunately just couldn't get to it as well as I wanted.

Despite those limitations, I am happy with this project as a whole. I knew it would be difficult, so I am not that surprised I didn't quite reach all my goals.

Works Cited

Assael, Y. M., Shillingford, B., Whiteson, S., & Freitas, N. de. (2016, November 4). *LipNet: End-to-end sentence-level lipreading*. OpenReview. https://openreview.net/forum?id=BkjLkSqxg

Aym98. (2025). *LipNet* [GitHub Repository]. GitHub. https://github.com/Aym98/LipNet

Bear, H. L., & Harvey, R. (2018). Phoneme-to-viseme mappings: the good, the bad, and the ugly. arXiv. https://arxiv.org/pdf/1805.02934

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). The Grid Audio-Visual Speech Corpus (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3625687

chrismbirmingham. (2025). *hci-face* [GitHub Repository]. GitHub. https://github.com/chrismbirmingham/hci-face

Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2017.367

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., &

Rubinstein, M. (2018). Looking to listen at the Cocktail Party. *ACM Transactions on Graphics*, *37*(4), 1–11. https://doi.org/10.1145/3197517.3201357

H, Ramya, Sundararajan G, & Kumaran M. (2023). LipNet: Bridging Communication Gaps

through Real-time Lip Reading and Speech Recognition. *International Journal of Advanced Research in Computer and Communication Engineering, 12*(9), 100-104. https://ijarcce.com/wp-content/uploads/2023/10/IJARCCE.2023.12917.pdf

liukuangxiangzi. (2025). *audio2viseme* [GitHub Repository]. GitHub.

https://github.com/liukuangxiangzi/audio2viseme

Mathew, A., Saldanha, A. & Babu, C.N. Audio–video syncing with lip movements using

generative deep neural networks. *Multimed Tools Appl* 83, 82019–82033 (2024). https://doi.org/10.1007/s11042-024-18695-x

Urban, E. (n.d.). *Get facial position with Viseme - Azure AI services*. Azure AI services | Microsoft

Learn.https://learn.microsoft.com/en-us/azure/ai-services/speech-service/how-to-speech-synthesis-viseme?tabs=visemeid&pivots=programming-language-python

https://github.com/chrismbirmingham/hci-face : Viseme map csv

https://arxiv.org/pdf/1611.01599

https://zenodo.org/records/3625687#:~:text=The%20Grid%20Corpus%20is%20a%20large%20multitalker%20audiovisual,to%20support%20joint%20computational-behavioral%20studies%20in%20speech%20perception.