Blog Post 1: Blake Buckner

Lip reading is a remarkable human skill that allows for communication in places where sound fails. Some of the primary places lip reading is most useful for are the hearing-impaired, Noisy environments, video captioning, long-distance video, and security footage analysis. Because of this, there are actually professional lip readers who specialize in trying to decipher language for a variety of purposes. Online personalities like Jomboy Media and Legendz make a living as professional lip readers for baseball and basketball, respectively, and their work often goes viral since fans want to know what is being said by athletes. It has gotten to the point where athletes often cover their mouths while talking to each other. As more uses for lip reading become profitable, it becomes a more desirable task. This is a field that has recently grown with the usage of neural networks, but there are still lots of open situations that need to be addressed. It is clearly a coveted skill, and I think it is also a good challenge for applying what I have learned in this class. I definitely have worries that my goals are too ambitious, but I hope to at least have some good results to show, even if my model isn't very accurate.

There are a few levels at which lip reading can be applied, which have varied results. The 3 primary scales are a **sound** level, a **word** level, and a **sentence** level. The sound level involves putting together each sound based on the actual positions of every part of the mouth. Figure 1 shows every possible sound, and Figure 2 shows specific parts of the mouth that may account for the control of different sounds. While this method has some strengths, certain aspects of the mouth just aren't visible, so this makes it more difficult to separate sounds, especially when they move fast. For sentences, this has uses, but it relies on a larger set of data and has a much more complicated structure of ways due to how many possible sentences there are. Because of this, I intend to focus on lip reading on a word-based level, but still include some of the sound-based methods when those are useful.

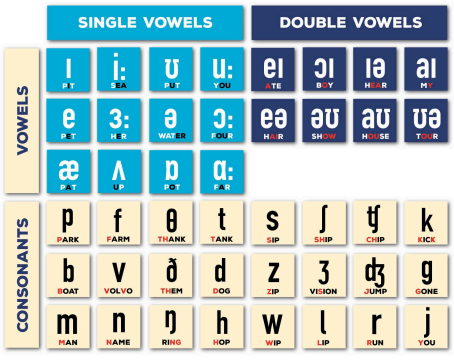

Figure 1: IPA Phonetics Chart (u/Huge_Cut_8327 2022)

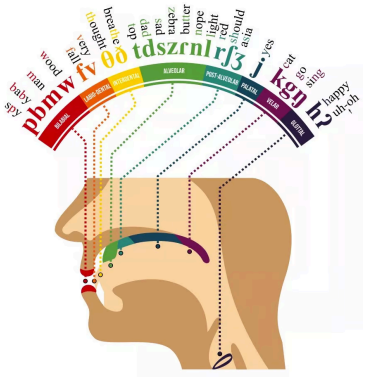

Figure 2: IPA Phonetics Chart (Kotke 2019 via Language Base Camp)

Since this itself remains a difficult task I would like to not use explicit word detection based only on the single most likely, but instead provide a set of confidence levels for different

words, and then from there use and AI model to analyze the possible sentences from words of significant confidence, and choose likely words to fit the context of the rest of the sentence.

Researchers have used Recurrent and Convolutional neural networks, with each having their own uses and benefits. For the purpose of this project, I intend to use CNNs for a few reasons. RNNs are more useful for sentence-based lip reading since they use more extended context-dependent information at the level of the neural network, but CNNs are better at the word level, and as such are the primary method other similar research has used (Mathew et al. 2024). Figure 3 below is my work-in-progress pipeline for this project. It is nowhere near final, but should serve as a basis to start with.

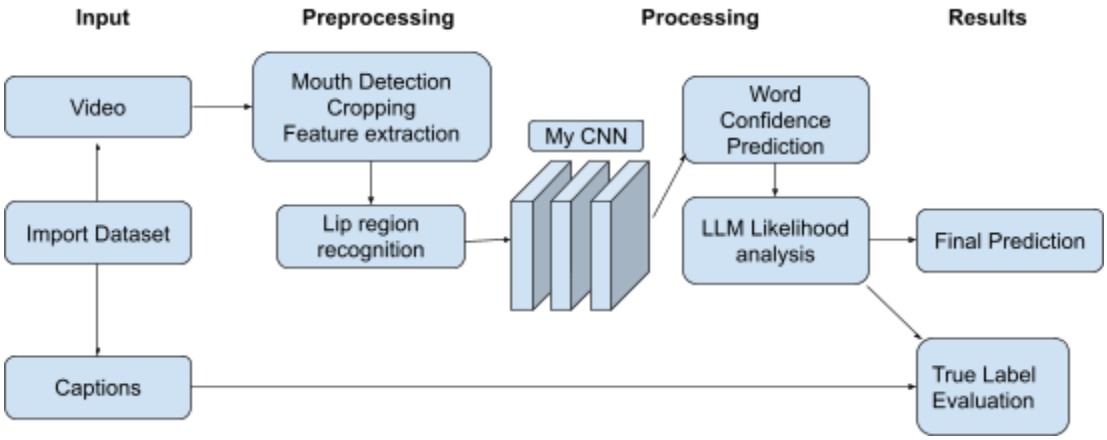

Figure 3: Proposed WIP Pipeline

One of the key research papers in the field of lip reading is “Lip Reading in the Wild” (LRW), a 2016 paper that included the largest available dataset of lip reading videos. This dataset is supposedly public with permission, and I have reached out to get permission, but have not received a response yet. Other datasets available are GRID (Cookie et al. 2006) and Google's AVSpeech (Ephrat et al. 2018). Last year, Ugale et al. released an excellent review of the previous research, so that has been an amazing source for exploring the field and finding open questions.

Lip reading stands at an exciting intersection of human skill and technological advancement. With growing interest from both the hearing-impaired community and professional sectors like sports analysis and security, the demand for accurate lip reading solutions continues to rise. Although this field presents considerable challenges, ranging from the complexity of mouth movements to the ambiguity of visual signals, recent progress in neural networks, especially CNNs, opens up promising avenues for research and application. While I recognize the ambitious nature of this work, I believe the insights I uncover can contribute to accessibility and expand the field. By focusing on word-level recognition and integrating contextual confidence, my project aims to contribute meaningful results to this rapidly developing area. Ultimately, the journey to more precise and accessible lip reading technologies holds great promise, both for technological innovation and for making the world more inclusive.

Works Cited

Alford, C. (2014, November 18). *Speech production*. SlideServe.

https://www.slideserve.com/carol-alford/speech-production

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

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

Huge_Cut_8327. (2022, September 10). The IPA [International Phonetic Alphabet]

Pronunciation Guide. *Reddit*. https://www.reddit.com/r/coolguides/comments/xggpef/the_ipa_international_phonetic_alphabet/

Kottke, J. (2019, March 22). A phonetic map of the human mouth. *Kottke.org*.

https://kottke.org/19/03/a-phonetic-map-of-the-human-mouth

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

Ugale, M., Pole, A., Desai, S., Gupta, H., & Khan, S. (2024). Speech recognition based on lip

movement using deep learning models - A review. *IOSR Journal of Computer Engineering, 26*(4), 10-18.https://www.iosrjournals.org/iosr-jce/papers/Vol26-issue4/Ser-3/B2604031018.pdf