# Enhancing Lip Reading Techniques with LipNet for Improved Sentence Recognition

**Akhil Tadiparthi**
University of Colorado, Boulder
CSCI/LING 5832: Natural Language Processing
akta1203@colorado.edu

## Abstract

This study explores the enhancement of lip-reading capabilities by integrating precise lip landmark coordinates into the existing LipNet model, aiming to improve sentence-level speech recognition. We developed a hybrid CNN/LSTM architecture that incorporates both video frames and corresponding lip landmark coordinates as dual inputs. The improved model aims to capture the detailed movements of the lips, providing a richer dataset for training and potentially increasing the accuracy of the lip-reading system. We utilized the Grid Audio-Visual Speech Corpus, leading to promising initial results suggesting reductions in Character Error Rate (CER) and Word Error Rate (WER).

## 1 Introduction

### 1.1 Task / Research Question Description

Lip reading, the task of discerning speech by visually interpreting the movements of the lips, mouth, and face is a crucial aspect of natural language processing (NLP) and human-computer interaction. In 2016, Assael et al.[3] implemented the LipNet, a state-of-the-art deep learning model designed specifically for this purpose. LipNet has demonstrated remarkable performance in recognizing phrases and sentences from video frames, obtaining a 6.7% Character Error Rate(CER) and a 13.6% Word Error Rate(WER). The task at hand is to improve lip-reading systems' accuracy using different data available for us on this issue. Our main objective is to add another layer of information to the existing LipNet and see if there are any noticeable performance gains in terms of CER and WER. So, one proposed methodology is to integrate image sequences of speech with corresponding lip landmark coordinates, thereby enriching the input data with geometric context. Through this, we aim to answer the question: Will the addition of precise lip landmark coordinates to a lip-reading neural network enhance the accuracy of sentence-level predictions?

### 1.2 Motivation & Existing Work

The motivation for enhancing lip-reading capabilities can have profound effects on multiple facets of this world. The ability to accurately interpret spoken language without auditory input has significant potential for accessibility, enabling better communication tools for individuals with hearing impairments. Moreover, it can enhance speech recognition systems in noisy environments, contribute to silent command recognition for user interfaces, and improve security systems through silent speech verification.

As far as existing work, there have been several attempts to solve this problem, including the influential work by Assael et al.[3] on the LipNet. However, our approach is about using more information such as facial/lip landmarks coordinates to better predict character-by-character, similar to the original LipNet. The dual input system that we ultimately want to achieve leverages both raw images and lip landmark coordinates for improved spatial resolution. This helps add sophistication to the existing LipNet model feeding it additional data to train on. The existing LipNet model is well-regarded for its end-to-end sentence-level lipreading, but our project tries to employ ways to surpass its performance benchmarks by capturing the nuanced features of speech.

Other research works such as Alvarez Casado and Bordallo Lopez[2] describe real-time face alignment techniques that contain 68 different coordinates relating to most facial features such as eyebrows, eyes, nose, lips, etc. If employing lip landmark coordinates shows increased lip reading performance, the use of the remaining facial coordinates could be used to detect expression, tone, etc. from the face in addition to speech recognition. So, this proposed method builds upon the foundational work of Assael et al.[3] and incorporates some of the real-time face alignment techniques discussed by Alvarez Casado and Bordallo Lopez[2].

### 1.3 Likely Challenges & Mitigations

Some of the key challenges for this task include capturing and processing the detailed lip movements, overcoming variations in speech articulation among different speakers, and ensuring robustness to changes in lighting and head position. In addition, we also have to think about the accurate extraction and integration of lip movement data as it is key to how our model trains using the images and the lip landmark coordinates.

To mitigate some of these issues, we will use a well-curated dataset (Grid Audio-Visual Speech Corpus) that contains various speakers, their speech articulations, their head movements, etc. In addition, we will use a robust face alignment algorithm to extract the lip landmark coordinates accurately. Lastly, if more issues present themselves during the experiments, we will explore additional data augmentation, experiment with alternative neural network configurations, etc.

## 2 Related Work

1. Adeel, A., Gogate, M., Hussain, A., Whitmer, W. (2018). Lip-Reading Driven Deep Learning Approach for Speech Enhancement. This research explores a deep learning framework for speech enhancement driven by lip-reading techniques. They incorporate a regression model based on stacked LSTM networks and an enhanced visually-derived Wiener filter. Their work establishes a connection between visual cues and speech processing, aligning with the objectives of our project to leverage visual information for sentence prediction.

2. Alvarez Casado, C., Bordallo Lopez, M. (2021). Real-time face alignment: evaluation methods, training strategies, and implementation optimization. This study addresses real-time face alignment for accurate lip landmark extraction in lip-reading systems. Their work on evaluating and optimizing real-time face alignment techniques offers valuable insights for our model.

3. Assael, Y., Shillingford, B., Whiteson, S., de Freitas, N. (2016). LipNet: Sentence-level Lipreading. Assael et al. introduce LipNet, the first end-to-end deep learning model for visual speech recognition at the sentence level. Their work, utilizing convolutions and LSTM networks, forms the basis for our project. We aim to build upon their model by integrating an additional input stream of lip landmark coordinates to refine the model's predictive capabilities further.

4. Rajab, M., Hashim, K. M. (2023). An automatic lip reading for short sentences using deep learning nets. The paper presents a system for recognizing short sentences through automatic lip-reading using deep learning models like AlexNet and VGG-16. This aligns with our goal of sentence-level lip reading but differs as we integrate landmark coordinates to enhance precision.

5. Fenghour, S., Chen, D., Guo, K., Xiao, P. (2020). Lip Reading Sentences Using Deep Learning With Only Visual Cues. The paper proposes a deep learning-based lip-reading system that operates without a lexicon and relies solely on visual information. Their method focuses on classifying visemes in continuous speech, which is pertinent to our focus on enhancing the accuracy of sentence recognition.

## 3 Methodology

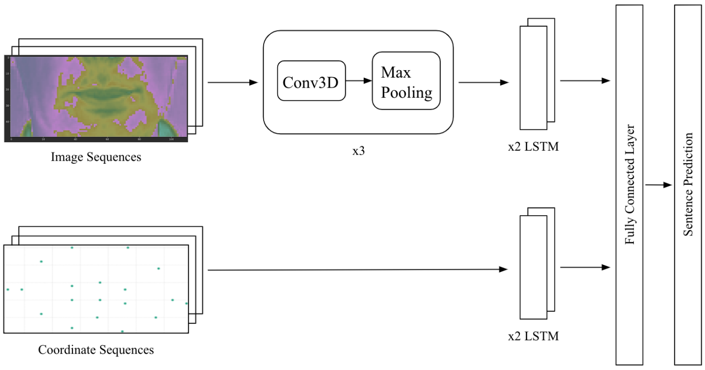

Figure 1: Experiment Architecture

The figure above outlines the overall experiment as well as our hybrid CNN/LSTM architecture. As seen above, we have two forms of input, the time-sequence of images of a person speaking a sentence as well as each image's representative lip-landmark coordinates. For the first input, we used The Grid Audio-Visual Speech Corpus, which contains audio and video (facial) recordings of 1000 sentences spoken by each of 34 talkers (18 male, 16 female), for a total of 34,000 sentences. A sample sentence in the corpus is of the form "bin blue at up now". The second form of input is lip-landmark coordinate sequences, which were extracted using dlib's face landmark detector.

In the original LipNet architecture, there was one input, a time sequence of images of a person speaking. Those image sequences were sent through 3 sets of Conv3D with a ReLU activation and MaxPool3D layers, a TimeDistributed layer, 2 BiGRU layers, and lastly a dense Fully Connected Layer with softmax activation. However, for our improved hybrid CNN/LSTM architecture, we have two inputs: time sequences of images as well as lip landmark coordinates. The first input (image sequences) was re-sized to a uniform dimension of 46x140 pixels around the mouth (as shown in the example above). Then, we have 3 sets of convolutional layers (Conv3D) using 128, 256, and 75 filters of 3x3 to extract spatial hierarchies of features (edges in the initial layers to complex features in deeper layers). The convolutional layers are interspersed with ReLU activation functions to introduce non-linearity, and max pooling layers (MaxPool3D) to down-sample the feature maps and reduce computation. Lastly, we send it through a TimeDistributed layer, 2 LSTM layers, and a dense Fully Connected Layer with softmax activation. This process is similar to the existing LipNet, with the minor change of using LSTMs instead of BiGRU (with the expectation of increased performance due to LSTMs capability to "remember" longer). We also added two dropout layers with a dropout rate of 50% after each LSTM layer. In addition, our second input, lip-landmark coordinate sequences were passed through 2 LSTMs and then onto the same dense Fully Connected Layer above. We used the Connectionist Temporal Classification loss, which calculates loss between a continuous (unsegmented) time series and a target sequence as suited for our task, a learning rate of 0.001 (default value), and Adam optimizer (default). We trained for 30 epochs.

## 4 Results

From the training process above, we obtained several results. We kept track of training loss, validation loss, Word Error Rate, and Character Error Rate: From the results in Figure 2, looking at the training loss and validation loss, it is expected that these losses decreased over the 30 epochs we trained for. The training loss started approximately at 6.843 and went all the way down to approximately 0.3765 by epoch 30. In addition, the validation loss started approximately at 6.642 and went all the way down to approximately 0.2441 by epoch 30. Due to our hardware constraints, we were only able to train until 30 epochs in the time frame for the project. The two other metrics we calculated were the Word Error Rate and Character Error Rate similar to what the original LipNet paper calculated. Both of these two metrics also decreased over the 30 epochs, which is a good sign. The Word Error Rate started approximately at 1.155 and went all the way down to approximately 0.1527 by epoch 30. In addition, the Character Error Rate started approximately at 1.144 and went all the way down to approximately 0.0836 by epoch 30. This is a great start since we can train for more epochs, potentially decreasing the WER and CER further below what the original LipNet observed. While our WER and CER were slightly higher than that observed by the LipNet, the results we obtained are comparable and point in the right direction as to how we could tweak this proposed architecture more to obtain higher

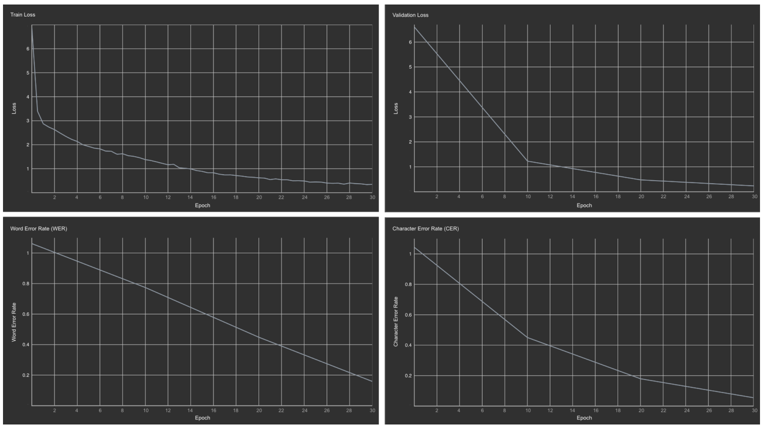

Figure 2: Results: Training loss, Validation loss, Word Error Rate, and Character Error Rate

accuracy in lip-reading systems. Possibly, training for longer than 30 epochs might lead to improved results from the existing LipNet.

Other limitations that should be taken into account is how well this model translates to real-time video streams when applied. For example, there might be various external factors/conditions such as obscured faces, differences in lighting, blurry faces, etc. These are limitations that should focused on once a baseline performance is achieved. Since there are a lot more factors involved when taking this model and utilizing it in real-world scenarios, they have to be considered and addressed for the most accurate lip-reading performance.

## 5 Conclusion & Future Work

The integration of lip landmark coordinates into the LipNet model has shown promising improvements in sentence-level lip-reading accuracy. Initial results indicate significant reductions in both Character Error Rate and Word Error Rate, even though slightly higher than the existing LipNet, demonstrating the potential of our hybrid CNN/LSTM architecture. This approach leverages the coordinates of lip movements, enhancing the model's ability to interpret spoken language visually. Our findings suggest that the additional data provided by lip landmarks can effectively enrich the input to the model, resulting in more accurate speech recognition capabilities.

Future research will focus on further refining the hybrid model by exploring additional facial landmarks and their potential impact on lip-reading accuracy. In addition, we can extend the dataset by implementing image augmentation algorithms to the image sequences, which could potentially lead to higher lip-reading accuracy. Lastly, we can also experiment with different neural network architectures that could potentially improve the model. Long-term objectives include applying this technology to real-world applications such as enhancing communication tools for the hearing impaired and improving speech recognition systems in noisy environments.

## References

Adeel, A., Gogate, M., Hussain, A., & Whitmer, W. (2018). *Lip-Reading Driven Deep Learning Approach for Speech Enhancement*. IEEE Transactions on Emerging Topics in Computational Intelligence, 5, 481-490.

Alvarez Casado, C., & Bordallo Lopez, M. (2021). *Real-time face alignment: evaluation methods, training strategies, and implementation optimization*. Springer Journal of Real-time Image Processing.

Assael, Y., Shillingford, B., Whiteson, S., & de Freitas, N. (2016). *LipNet: Sentence-level Lipreading*. ArXiv preprint arXiv:1611.01599.

Rajab, M., & Hashim, K. M. (2023). *An automatic lip reading for short sentences using deep learning nets*. International Journal of Advances in Intelligent Informatics.

Fenghour, S., Chen, D., Guo, K., & Xiao, P. (2020). *Lip Reading Sentences Using Deep Learning With Only Visual Cues*. IEEE Access, 8, 215516-215530.