LingoLens: Lipreading Algorithm

1st Tamerlan Ormanbayev
Faculty of Science, Department of Computer Science
University of Calgary
Calgary, Canada
tamerlan.ormanbayev@ucalgary.ca

2nd Yurii Bezborodov
Faculty of Science, Department of Computer Science
University of Calgary
Calgary, Canada
yurii.bezborodov@ucalgary.ca

***Abstract*—This paper presents a substantial approach to the exploration of lip-reading machine learning (ML) algorithms. It provides an overview of some existing work in the field, as well as discusses differences and the benefits of each approach. The paper then details the implementation of the presented algorithm, LingoLens, along with suggestions for enhancing its efficiency. Finally, a summary of findings and conclusions is offered.**

I. Introduction

Millions of people globally face communication barriers due to speech or hearing impairments. Approximately 60% of all adults across the planet face certain hearing challenges, out of which 23% have some level of hearing loss [1]. Lipreading in itself is known to be a rare ability to find in a person, because of how difficult the task is. Hearing-impaired people achieve an accuracy of only 12-17% even for a limited subset of 30 monosyllabic words and 11-21% for 30 compound words [2].

For machines, lipreading is a challenging task for multiple reasons. While people mainly struggle with understanding the context, a machine requires accurate data for both lips' motion and position. Traditional speech-to-text systems also fall short in capturing nuanced visual expressions, making it imperative to design an algorithm that accurately interprets and translates lip movements into coherent written text.

In this paper, we present LingoLens, a machine-learning lipreading algorithm that aims to assist individuals with hearing impairments by enabling them to understand spoken language without relying on interpreters or others' ability to understand sign language. This independence targets helping them to navigate various environments more freely. Integrating such a translator into human-machine interfaces can also enhance the interaction experience. Voice commands in noisy environments silent communication with devices, and calls between people could also see great improvements in the clarity of the caller's words.

This paper also has an overview of the work done in the field already, where we discuss several other papers and articles to outline relevant ideas and emphasize the benefits of some over others.

We will describe the methodology for LingoLens by particularly trying to explain how it works and which technologies are behind it. Suggestions on improving the subject will also be provided.

Finally, we will summarize the paper.

II. Overview

1) ”End-to-End Audiovisual Speech Recognition” by Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Feipeng Cai, Georgios Tzimiropoulos, Maja Pantic [3]

This paper affords a give-up-to-quit method to audiovisual speech recognition, which incorporates lip reading as a component. This approach directly methods each modality collectively to improve popularity accuracy. The structure of the proposed system consists of neural community components for audio and visual processing. The audio part usually entails spectrogram-based features extracted from the speech signal, at the same time as the visual part makes a specialty of lip motion evaluation using video frames. These capabilities are then blended and processed with the aid of a joint community that learns to companion audio and visual cues for accurate speech reputation. The paper reviews large enhancements in speech reputation accuracy in comparison to standard systems that depend entirely on audio or visual data. “End-to-End Audiovisual Speech Recognition” gives a complete framework for combining audio and visual cues in a unified version, leading to progressed accuracy and robustness in speech popularity structures.

2) ”Lip Reading Sentences in the Wild” by Joon Son Chung, Andrew Senior, Oriol Vinyals, Andrew Zisserman [4]

The paper addresses the challenge of lip-reading sentences in real-world situations, where conditions such as different lighting, background noise, and loudspeaker changes can affect the performance of convolutional neural networks (CNNs) to extract features from a video image of the types of specimen lips used. These features are then processed by recurrent neural networks (RNNs) or transformer-based architectures to capture time dependence and context in speech The system is trained with large datasets of audio and visual speech. The paper reports promising results from lipreading sentences in scenarios with complicated conditions with different lighting, noise and other obstructions. The model shows its potential applications in speech recognition that are competitive. ”Lip Reading Sentences in the Wild” presents an innovative approach to lip reading in real-world situations, demonstrating the effectiveness of deep learning techniques in extracting meaningful information from visual and verbal cues.

3) “Deep Lip Reading: a comparison of models and an online application” by Triantafyllos Afouras, Joon Son Chung, Andrew Zisserman [5]

The paper presents an approach to lip reading using deep neural network architectures. Three different models are proposed, each consisting of a spatiotemporal visual front-end and a sequence processing module. The visual front end involves a 3D convolution followed by a 2D ResNet to extract features from lip region images. The sequence processing module decodes these features into sentences character by character. The models are trained in stages, with pre-training on word-level datasets and subsequent training of the sequence processing module. The best model achieved a significant improvement in word error rate on the LRS2 benchmark dataset. The study also explores online lip reading, enabling real-time transcription of continuous speech. The findings demonstrate the effectiveness of the proposed architectures in enhancing lip reading accuracy and performance.

4) “LipNet: End-to-End Sentence-level Lipreading” by Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, Nando de Freitas [6]

The paper introduces LipNet, an end-to-end lip reading model at the sentence level that transfers video images to text. The proposed method includes deep neural network architecture combining variable and iterative layers to extract spatial-temporal features and provide efficient temporal aggregation LipNet outperforms traditional methods and achieves a 4.8% word error rate (WER), 2.8 times lower than word-level state-of-the-art in GRID corpus The success of model development proves to be effective in real-world applications requiring strong handwriting capabilities.

5) “Listen, Attend and Spell” by William Chan, Navdeep Jaitly, Quoc V. Le, Oriol Vinyals [7]

This paper introduces a ”Watch, Attend and Spell” (WAS) neural network architecture for large-word conversational speech recognition. Although the focus of the paper is on speech recognition, the methodology and techniques described are also applicable to lipreading, particularly in terms of understanding spoken speech and converting it into text consumption by machine learning approaches.

III. Discussion

The works described above all propose their solutions to a vast range of challenges faced by machine lipreading. While directly comparing the approaches proposed in these works may be inaccurate, as each work brings unique strengths to the table, we will still look at the main benefits and critique each.

In ”End-to-End Audiovisual Speech Recognition”, the authors integrate both audio and visual modalities directly for improved recognition accuracy. They also utilize neural network components for audio and visual processing, then combine them in a joint network. This approach allows for the simultaneous processing of audio (via spectrogram-based features) and visual (lip motion analysis from video frames) cues, leading to improved recognition accuracy by capturing complementary information from both modalities. While this leads to solidified accuracy numbers, this approach may require significantly greater and more powerful computational resources, to process both modalities simultaneously.

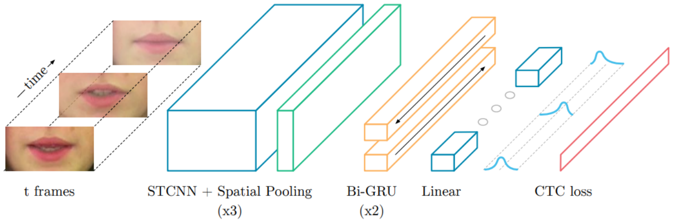


Fig. 1. LipNet’s architecture [6]

An important ability for any lipreading algorithm is to be able to work in unclear conditions. ”Lip Reading Sentences in the Wild” shows substantial results in complicated conditions involving different lighting and other obstructions. The results of the paper show significant performance in scenarios not explicitly seen in training data. While its use of CNNs and RNNs to extract features together is advantageous, the effectiveness of the model relies heavily on the quality and diversity of the training data. If the dataset used for training is biased towards certain conditional characteristics or lacks sufficient variability, the model’s ability to handle real-world scenarios may be insufficient to be completely stable.

The authors of “Deep Lip Reading: a Comparison of Models and an Online Application”, on the other hand, propose multiple deep-learning neural network models with a focus on spatiotemporal features and sequence processing. The inclusion of a spatiotemporal visual front-end with 3D convolutions and a sequence processing module for decoding features into sentences character-by-character leads to significant improvements in accuracy, especially on benchmark datasets. The exploration of real-time transcription further enhances its practical applicability. Yet, such an approach also proposes potential trade-offs between accuracy and computational power efficiency, especially when applied to real-time conditions.

LipNet, a model from ”LipNet: End-to-End Sentence-level Lipreading” at its time was the best and potentially the first end-to-end sentence-level lipreading model that simultaneously learned spatiotemporal visual features and a sequence model. It made significant improvements in recognition accuracy, in comparison to other methodologies at the time. LipNet also required neither hand-engineered spatiotemporal visual features nor a separately trained sequence model (fig. 1). While it is challenging to immediately come up with a critique for such an impressively performing model, its performance may suffer from under-represented accents in the training datasets. The authors of the paper found some confusion in the results presented by LipNet when testing on data that involved different accents, like the British accent. In particular, uncertainties could be found in vowels, bilabial stops and alveolar stops.

Finally, the methodology proposed in “Listen, Attend and Spell” introduces a neural network architecture for large-word conversational speech recognition, which could be adapted for lipreading tasks, especially in understanding spoken speech and converting it into text. However, its adaptation for solving the lipreading task requires additional, if not extensive, modifications and training due to differences in visual and auditory features compared to speech recognition.

As some of the mentioned above algorithms perform well under diverse conditions, most of them evidently rely to a great extent on training data quality and diversity. Models that show impressive accuracy with spatiotemporal features may face challenges in real-time efficiency. While LipNet achieves high accuracy without hand-engineered features, its performance might suffer with under-represented accents. The paper ”Listen, Attend and Spell” introduces a useful architecture but requires extensive modifications for lipreading tasks due to differences in visual and auditory features compared to speech recognition. These critiques highlight the need for robustness and adaptation in lipreading algorithms, especially in challenging and diverse real-world scenarios.

IV. Our Implementation

1. Dataset

We will be using “Grid Corpus” [8] dataset by University of Sheffield researchers, it is a comprehensive collection of audiovisual recordings designed to aid research in speech perception through computational and behavioral studies. It includes recordings of 1000 sentences spoken by each of 34 talkers (18 male, 16 female), featuring high-quality audio and video (facial) recordings. The sentences follow a specific format, such as ”put red at G9 now”. The corpus, along with transcriptions, is freely accessible for research purposes. alignments.zip provides word-level time alignments separated by the talker. s1.zip, s2.zip etc contain .mpg videos for each talker. For the purposes of this project we only used speaker #1 from the dataset.

2. Data Loading (Fig. 2)

OpenCV [9] was used for our data loading and processing needs. First, we load a video from the specified path using OpenCV (cv2). It reads each frame, converts it to grayscale, and extracts a specific region of interest. It then calculates the mean and standard deviation of the frames and returns normalized frames as a TensorFlow float32 tensor. Using imageio [10] GIFs made of processed frames can also be generated to display what the models ”sees”.

Furthermore, we load text alignments from the specified path. It reads each line of the file, extracts relevant tokens, and converts them into numerical indices using a StringLookup layer. It returns a list of numerical indices representing the text.

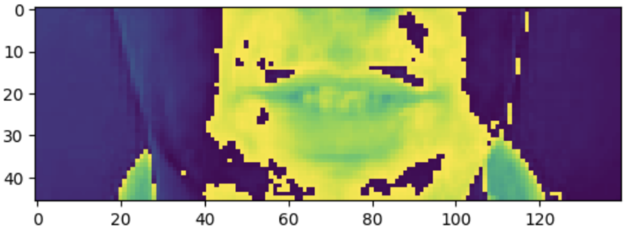


Fig. 2. Processed frame.

3. Model Architecture (Fig. 3)

The proposed model is a Sequential model built using TensorFlow’s [11] Keras API [12]. It consists of the following layers:

1) 3D Convolutional Layers [13]
The model starts with three 3D convolutional layers with ReLU activation and max pooling. These layers are used to extract features from the 3D volumetric data.
2) TimeDistributed Layer
A TimeDistributed layer is added after the convolutional layers to flatten the output of the convolutional layers while preserving the temporal information.
3) Bidirectional LSTM Layers [14]
Two Bidirectional LSTM layers are added with 128 units each and orthogonal kernel initializer. These layers are used to capture the temporal dependencies in the data.
4) Dropout Layers
Dropout layers with a dropout rate of 0.5 are added after each LSTM layer to reduce overfitting.
5) Dense Layer
Finally, a Dense layer with units equal to the vocabulary size + 1 is added with ’softmax’ activation. This layer is used for sequence prediction

4. Model Training

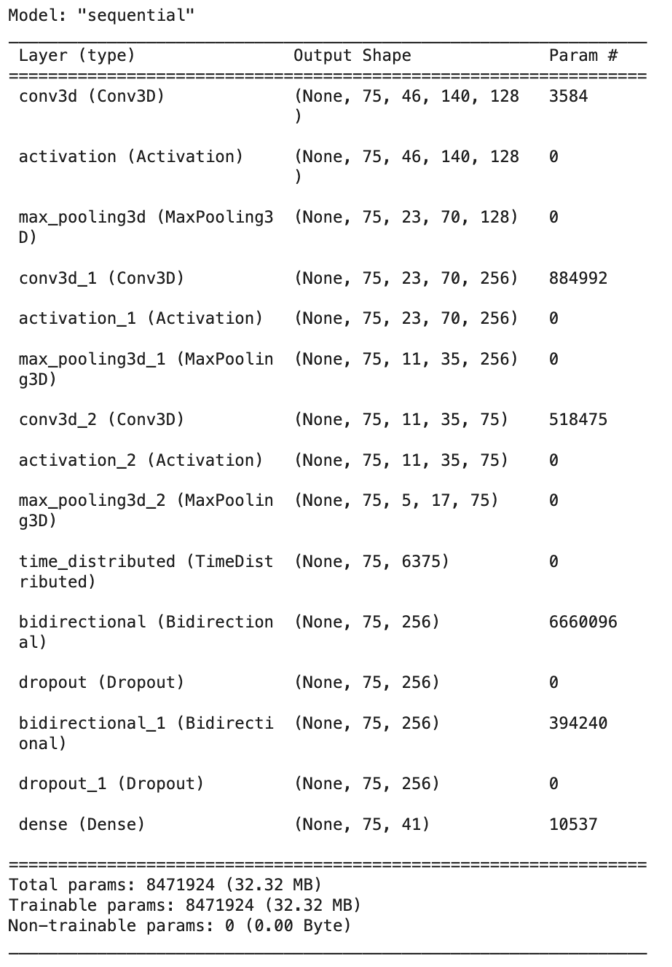
The model is compiled using the Adam optimizer with a learning rate of 0.0001 and the custom CTC loss [16] function. Three callbacks are defined for the model: ModelCheckpoint to save the model weights, LearningRateScheduler to adjust the learning rate, and ProduceExample to generate and print examples.

The model is trained on the training dataset for 100 epochs using the defined callbacks. The ModelCheckpoint callback saves the model weights whenever the training loss improves, the LearningRateScheduler callback adjusts the learning rate as per the defined scheduler function, and the ProduceExample callback generates and prints examples to monitor the model’s performance during training.

5. Results

In the end, the model was highly effective. When subjected to testing on 100 videos from the dataset, the model demonstrated a sentence-level word accuracy rate of 99.96%.

The training process spanned 96 epochs and required approximately 12 hours to complete.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv3d (Conv3D)             (None, 75, 46, 140, 128   3584
                             )

 activation (Activation)     (None, 75, 46, 140, 128   0
                             )

 max_pooling3d (MaxPooling3  (None, 75, 23, 70, 128)   0
 D)

 conv3d_1 (Conv3D)           (None, 75, 23, 70, 256)   884992

 activation_1 (Activation)   (None, 75, 23, 70, 256)   0

 max_pooling3d_1 (MaxPoolin  (None, 75, 11, 35, 256)   0
 g3D)

 conv3d_2 (Conv3D)           (None, 75, 11, 35, 75)    518475

 activation_2 (Activation)   (None, 75, 11, 35, 75)    0

 max_pooling3d_2 (MaxPoolin  (None, 75, 5, 17, 75)     0
 g3D)

 time_distributed (TimeDist  (None, 75, 6375)          0
 ributed)

 bidirectional (Bidirection  (None, 75, 256)           6660096
 al)

 dropout (Dropout)           (None, 75, 256)           0

 bidirectional_1 (Bidirecti  (None, 75, 256)           394240
 onal)

 dropout_1 (Dropout)         (None, 75, 256)           0

 dense (Dense)               (None, 75, 41)            10537

=================================================================
Total params: 8471924 (32.32 MB)
Trainable params: 8471924 (32.32 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Fig. 3. Model Structure

The prediction time ranges from 2 to 4 seconds for a 3-second video. This suggests that, when combined with a facial detection model, our system has the potential to be utilized in real-time applications with minimal delay.

6. Possible Improvements

There are many possible ways to improve our model.

It can be paired with a facial-detection algorithm [15], which will allow a much larger and more varied training dataset to be used, potentially improving the model's accuracy and generalization. This integration could enable the model to focus on recognizing facial expressions within the detected faces, enhancing its ability to interpret emotions in real-world scenarios.

Incorporating data augmentation techniques such as rotation, scaling, and translation can help in artificially increasing the diversity of the training dataset. This can make the model more robust to variations in facial expressions, lighting conditions, and facial orientations, improving its performance on unseen data.

V. Conclusions

We have looked at the problem of hearing loss and how providing a solution to that problem could help a significant amount of people around the Earth. After a deep dive into machine lipreading and establishing reasons why it is a challenging task, we have looked at some of the most popular works in the field and tried to discuss the advantages of each approach, highlighting the strong parts and features, while simultaneously criticizing them and looking for improvements in the field.

We proposed LingoLens, a sentence-level lipreading machine-learning algorithm that converts a video into a sequence of image frames to analyze them and produce text based on the speaker's speech. Despite the minimal testing phase, the methodology that was proposed has managed to show great performance on the GRID dataset. With accuracy numbers being expected to drop with larger testing samples, there is a lot of room for the model's improvement with even larger datasets, and potential real-world application when properly adapted with respective technologies.

References

[1] Deafness and hearing loss (no date) World Health Organization. Available at: https://www.who.int/health-topics/hearing-loss (Accessed: 26 March 2024).

[2] R. D. Easton and M. Basala. Perceptual dominance during lipreading. Perception & Psychophysics, 32(6): 562–570, 1982.

[3] Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., & Pantic, M. (2018). End-to-end Audiovisual Speech Recognition. ArXiv. /abs/1802.06424

[4] Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2016). Lip Reading Sentences in the Wild. ArXiv. https://doi.org/10.1109/CVPR.2017.367

[5] Afouras, T., Chung, J. S., & Zisserman, A. (2018). Deep Lip Reading: A comparison of models and an online application. ArXiv. /abs/1806.06053

[6] Assael, Y. M., Shillingford, B., Whiteson, S., & De Freitas, N. (2016). LipNet: End-to-End Sentence-level Lipreading. ArXiv. /abs/1611.01599

[7] Chan, W., Jaitly, N., Le, Q. V., & Vinyals, O. (2015). Listen, Attend and Spell. ArXiv. /abs/1508.01211

[8] M. Cooke, J. Barker, S. Cunningham, and S. Xu, "The Grid Audio-Visual Speech Corpus," Zenodo (CERN European Organization for Nuclear Research), Jan. 2006, doi: 10.5281/zenodo.3625687.

[9] G. Bradski, "The OpenCV library," Journal of Software Tools, vol. 25, pp. 120–125, Jan. 2000, [Online]. Available: https://ci.nii.ac.jp/naid/10028167478/en/

[10] S. Silvester et al., "imageio/imageio v0.9.0," Zenodo, Jul. 2020, doi: 10.5281/zenodo.3931847.

[11] Martín Abadi et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org. 2015. doi: 10. 5281 /zenodo. 4724125. url: https://www.tensorflow.org/.

[12] F. Chollet, "Keras: The Python Deep Learning library," Astrophysics Source Code Library, Jun. 2018, [Online]. Available: http://ui.adsabs.harvard.edu/abs/2018ascl.soft06022C/abstract

[13] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1):221–231, 2013.

[14] A. Garg, J. Noyola, and S. Bagadia. Lip reading using CNN and LSTM. Technical report, Stanford University, CS231n project report, 2016

[15] Sun, Y., Ren, Z., & Zheng, W. (2022). Research on Face Recognition Algorithm Based on Image Processing. Computational Intelligence and Neuroscience, 2022. https://doi.org/10.1155/2022/9224203

[16] "Papers with Code - CTC Loss Explained," paperswithcode.com. https://paperswithcode.com/method/ctc-loss (accessed Apr. 15, 2024).

[17] Rekik, Ahmed & Ben-Hamadou, Achraf & Mahdi, Walid. (2015). An adaptive approach for lip-reading using image and depth data. Multimedia Tools and Applications. 75. 10.1007/s11042-015-2774-3.

[18] “Deep Learning Model that can LIP READ using Python and Tensorflow”, www.youtube.com. https://www.youtube.com/watch?v=uKyojQjbx4c&ab_channel=NicholasRenotte (accessed Apr. 15, 2024).

[19] M. R. Kummitha, “Building a lip reading system with deep learning and tensorflow,” Medium, https://medium.com/@maanideeprkummiitha/building-a-lip-reading-system-with-d eep-learning-and-tensorflow-b6eec196eefe (accessed Mar. 26, 2024).

[20] “Deafness And Hearing Loss Statistics,” Forbes Health, Mar. 14, 2024. https://www.forbes.com/health/hearing-aids/deafness-statistics/#:~:text=Nearly% (accessed Apr. 15, 2024).

[21] “Resources for Patients & Families,” www.hopkinsmedicine.org. https://www.hopkinsmedicine.org/all-childrens-hospital/patient-families#:~:text=Tu (accessed Apr. 15, 2024).

[22] B.-S. Lin, Y.-H. Yao, C.-F. Liu, C.-F. Lien, and B.-S. Lin, “Development of novel lip-reading recognition algorithm,” IEEE Access, vol. 5, pp. 794–801, 2017. doi:10.1109/access.2017.2649838

[23] D. Kalbande, A. A. Mishra, S. Patil, S. Nirgudkar, and P. Patel, “Lip reading using Neural Networks,” SPIE Proceedings, Oct. 2011. doi:10.1117/12.913406

[24] Sooraj, V., M. Hardhik, Nishanth S. Murthy, C. Sandesh, and R. Shashidhar. ”Lip-reading techniques: a review.” Int. J. Sci. Technol. Res 9, no. 02 (2020): 1-6.

[25] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5):602–610, 2005.