Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

Deciphering Lip Movement A Comparative Study

Pooja Singari, Ch Naganoshith, Samiksha Mehta, K Sujan Rao, Prof. Yashaswini B M
Department of Artificial Intelligence and Machine Learning,
Dayananda Sagar College of Engineering.

Abstract - This paper aims to explore the various methodologies and advances in the field of visual lip reading. With particular emphasis on the influential work of LipNet, which served as a major milestone in the field, the paper explores the various approaches and architectures proposed in the literature. Analyzing the differences and similarities between these architectures will provide deep insight into the current state of lip reading.

The study traces the evolution of lip reading models over time, starting in 2016 with the introduction of "Spoken English Phrase Recognition Using Visual Feature Extraction and Classification" and progressing to the latest developments such as the recently improved Auto AVSR. By examining this timeline, we will trace how researchers have influenced each other, leading to the creation of new methodologies or improvements to existing methodologies.

1. INTRODUCTION

Lip reading has historically presented challenges for humans, with low accuracy in understanding speech based on lip movements alone. However, advances in deep learning have made it possible to recognize common patterns of lip movements, leading to improved speech understanding. While audio speech recognition has already reached near-human accuracy, similar progress is expected to be demonstrated in lip reading.

The applications of lip reading technology are vast and varied. In addition to its tracking potential, it can be of great benefit to the hearing impaired by providing transcriptions based on lip movements, enabling them to understand spoken speech. Notably, the accuracy of reading human lips is approximately 52.3 percent, while the accuracy achieved by most of the discussed works exceeds this limit. For example, LipNet has demonstrated 93.4 percent accuracy in certain tests.

This paper will begin by providing an overview and summary of the relevant papers under consideration. Subsequently, a comparative study will be presented that analyzes the strengths and weaknesses of each approach.

2. VARIOUS APPROACHES

This section will cover some state-of-the-art approaches and architectures in the field of lip reading. Starting from our main focus, LipNet, we will discuss approaches which were there earlier and which came after LipNet.

2.1 LipNet

Lipnet, the first end-to-end sentence-level lipreading model has set the benchmark for all the succeeding Automatic Speech Recognition models with its impressive performance. It acquired a 95.2% sentence-level word accuracy with overlapped speakers and an accuracy of 88.6% on unseen speakers.

2.1.1 Architecture

Lipnet's architecture as shown in Fig 1 below comprises Spatio Temporal convolutional neural networks (STCNN), Gated Recurrent Unit (GRU), and Connectionist Temporal Classification (CTC).

STCNN functions to extract spatial and temporal information from the video, enabling meaningful insights to be derived from lip movements and predicting the corresponding transcript.GRU, known for its ability to retain information over longer sequences, is utilized to predict complete sentences based on individual words. CTC loss is used to eliminate the need for training data that aligns inputs to their corresponding outputs; It computes the probability of the sequences by marginalizing over all sequences that are defined as equivalent to this sequence.

Recent advancements in saliency visualization [1][2] employed provided insights into the regions that the model considers phonologically significant while generating the transcript.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

2.1.2 Results

Due to lack of resources, our reproduction of Lipnet's results was done by training on one speaker's data for analysis. It was observed that the reproduced results had a 0 % word error rate (wer) on the small subset of the original gird corpus dataset.

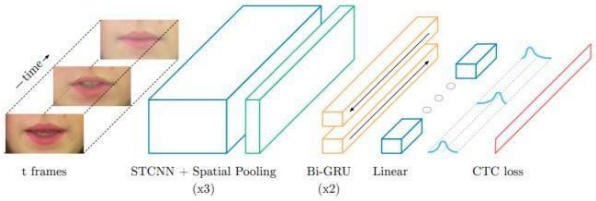

Figure 1: LipNet Architecture.

2.2 Auto AVSR

2.2.1 Architecture

Using the out-of-the-box method suggested in [7], this paper achieves the performance of LRS2 and LRS3 datasets without using external data. The architecture is shown below Figure 1. In particular, the VSR frontend is based on modified ResNet-18 [8, 9]; where the first layer is a spatio-temporal convolutional layer with a core size of 5 × 7 × 7 and pitch 1 × 2. × 2. Spring on back of front end Compatible [6]

Similarly, the ASR encoder consists of 1D ResNet-18 [10] and followed by Conformer. The ASR and VSR encoder outputs are combined via a multilayer perceptron (MLP). The rest of the network consists of a projection system and a Transformer decoder for integrated CTC/monitoring [11].

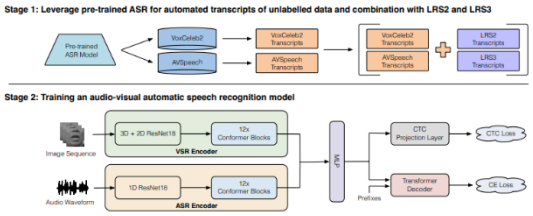

Figure 2: Auto AVSR Architecture.

2.2.2 Pre-processing

For visual flow, we first adjusted the dataset after previous work in [7]. We cut the oral region of interest (ROI) using a 96 × 96 bounding box and normalized each square by subtracting the mean and dividing by the standard deviation of the training.

For audio streams, we only z-normalize each expression.

2.2.3 Results

When trying out the model we saw an wer of 0.9% on the available video which is not a part of the dataset. The model is able to recognise and predict the sentence even when there is noise in the video.

2.3 Audio-Visual Efficient Conformer for Robust Speech Recognition

The Classification (CTC)-based architecture by processing both audio and visual modalities was the central notion for choosing this model. It improves upon the previous lip-reading methods by the inclusion of a Conformer back-end on top of a ResNet-18 visual front-end and by adding intermediate CTC losses between blocks.

2.3.1 Architecture

The architecture comprises of 4 main components: An audio encoder, a visual encoder, an audio-visual fusion module and an audio-visual encoder. The audio front end is responsible for transforming the raw audio signals into mel-spectrograms which are further processed by a 2D convolution stem to extract local temporal-frequency features. The visual front-end transforms the frames from the input video into temporal sequences which are then fed to the back-end network. The back-end network comprises of the Efficient Conformer encoder [3] wherein the temporal sequence is repeatedly down-sampled to lower the number of computations. The audio-visual fusion module functions to concatenate the acoustic and visual features from back-end networks into a joint feed-forward network. Audio Visual encoder is a single-stage back-end network composed of 5 Conformer blocks without downsampling. The architecture is shown in the figure below.

The innovative approach presented in this model is the introduction of Patch Multi-Head Self-Attention.In this method, the input sequence is downsampled using an average pooling before applying multi-head self-attention. It reduces the complexity to O((n/k) 2 · d) where k denotes the pooling/upsampling kernel size.

2.3.2 Preprocessing

Bounding boxes of 96 x 96 pixels are used to crop the lip regions and eliminate the differences in rotation and scale. The RetinaFace [5] face detector and Face Alignment Network (FAN) [4] are used to detect 68 facial landmarks. These cropped images are then converted to a gray scale and normalized between −1 and 1.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

2.3.3 Results

Three different models were evaluated to check the wer against the ground truth. The models performed respectively:

Visual-only: wer=30.77%
Audio-only: wer=3.85%
Audio-visual: wer=0.00%

These performances indicate that a fusion of audio and visual outperforms individual models

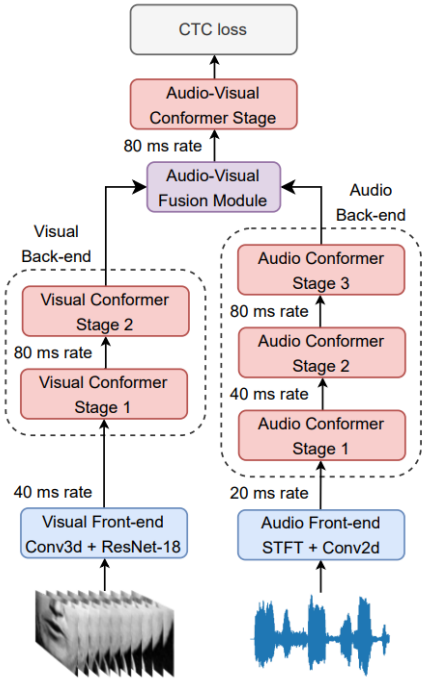

Figure 3: Audio-Visual Efficient Conformer Architecture

3. Comparison

Table 1: A comparison of all the discussed papers.

Name	Architecture used	Dataset used	Accuracy
Lip Net	Spatio Temporal convolutional neural networks (STCNN), Gated Recurrent Unit (GRU), and Connectionist Temporal Classification (CTC).	Grid dataset	Training on one speaker's data and 0 % word error rate (wer) on the small subset of the original gird corpus dataset.
Auto AVSR	VSR frontend is based on modified ResNet-18 ASR encoder consists of 1D ResNet-18 and followed by Conformer.	LRS3	wer of 0.9% on the available video which is not a part of the dataset.
Audio-Visual Efficient Conformer	An audio encoder, a visual encoder, an audio-visual fusion module and an audio-visual encoder.	Lip Reading in the Wild (LRW) data set	Visual-only: wer=30.77% Audio-only: wer=3.85% Audio-visual: wer=0.00%

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

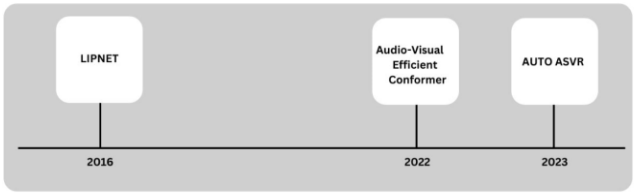

Figure 4: Lip reading journey

4. DATASETS

This section introduces the various datasets used while developing the benchmarked models mentioned above.

4.1 Grid Corpus

The Grid Corpus which is a large multi-speaker audiovisual sentence corpus designed to support joint computational-behavioral studies in speech perception. In brief, the corpus consists of high-quality audio and video (facial) recordings of 1000 sentences spoken by each of 34 talkers (18 male, 16 female), for a total of 34000 sentences.

The Grid Corpus specifically focuses on the visual aspect of speech by capturing the movements of the speaker's lips and face during the utterance. It is designed to support research on automatic lip-reading, audio-visual speech recognition, and related applications.

Combining the visual cues with the audio signal, aims to enhance speech recognition performance, especially in challenging conditions such as noisy environments or when the audio signal is degraded.

4.2 Lip Reading in the Wild(LRW)

The Lip Reading in the Wild (LRW) dataset is a widely accepted benchmark for lip reading and visual speech recognition. Developed by researchers at Oxford College, it provides a comprehensive collection of video sequences for training and evaluating lip reading algorithms.

Focusing on the challenging task of reading in real-world scenarios, the LRW dataset covers a wide range of backgrounds, lighting conditions and speaker characteristics. It contains over 100,000 video clips from BBC TV programs in which different people pronounce individual words from a vast vocabulary of over 1,000 classes.

4.3 LRS-3

LRS2 [12] and LRS3 [13], are the two largest publicly available datasets for audio-visual speech recognition in English. LRS2, collected from BBC programs, contains 144 482 video clips with a total of 225 hours. Specifically, the pre-training, training, validation and test set contains 96 318 (195 hours), 45 839 (28 hours), 1 082 (0.6 hours) and 1 243 (0.5 hours) video clips, respectively. LRS3 consists of 151 819 video clips from TED talks with a total of 439 hours. It contains 118 516 (408 hours), 31 982 (30 hours) and 1 321 clips (0.9 hours) in the pretraining, training-validation, and test set, respectively. For training, we also use the English-speaking videos from AVSpeech (1 323 hours) and VoxCeleb2 (1 307 hours) as the additional training data together with automatically-generated transcriptions.

5. CONCLUSION

Lip Reading, the process of decoding spoken words through mere lip movements, has traditionally been challenging, particularly in the absence of audio. However, recent advancements in deep learning models have led to significant progress in this field. By conducting a comparative analysis of various models developed over the years, insights into the latest technologies and advancements have been obtained.

6. FUTURE WORK

While reproducing the results of these benchmarked models, it has been observed that the previous environments are no longer compatible with today's updated libraries and technologies. It has become essential to adapt to the evolving times and build models by leveraging the functionalities offered by the updated algorithms and architectures to enhance the performance and accuracy of future work.

REFERENCES

[1] Z. Zhou, G. Zhao, X. Hong, and M. Pietikainen. A review of recent advances in visual speech decoding. ¨ Image and Vision Computing, 32(9):590–605, 2014.

[2] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualizing image classification models and saliency maps. arXiv preprint arXiv:1312.6034, 2013.

[3] Maxime Burchi and Valentin Vielzeuf. Efficient conformer: Progressive downsampling and grouped attention for auto[1]matic speech recognition. In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 8– 15. IEEE, 2021

[4] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In Proceedings of the IEEE International Conference on Computer Vision, pages 1021–1030, 2017.

[5] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multilevel face localisation in the wild. In Proceedings of the IEEE/CVF conference on computer

[6] A. Gulati et al., “Conformer: Convolution-augmented transformer for speech recognition,” in Interspeech, 2020, pp. 5036– 5040.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 06, June-2023

[7] P. Ma et al., "End-to-end audio-visual speech recognition with conformers," in ICASSP, 2021, pp. 7613–7617
[8] K. He et al., "Deep residual learning for image recognition," in CVPR, 2016, pp. 770–778.
[9] T. Stafylakis et al., "Combining residual networks with LSTMs for lipreading," in Interspeech, vol. 9, 2017, pp. 3652– 3656.
[10] S. Petridis et al., "End-to-end audiovisual speech recognition," in ICASSP, 2018, pp. 6548–6552.
[11] S. Watanabe et al., "Hybrid CTC/attention architecture for end-to-end speech recognition," IEEE J. Sel. Top. Signal Process., vol. 11, no. 8, pp. 1240–1253, 2017.
[12] J. S. Chung et al., "Lip reading sentences in the wild," in CVPR, 2017, pp. 3444–3453.
[13] T. Afouras et al., "LRS3-TED: A large-scale dataset for An audio encoder, a visual encoder, an audio-visual fusion module and an audio-visual encoder.visual speech recognition," arXiv preprint arXiv:1809.00496, 2018.