# Lip Reading

Computer Science, Sapienza Università di Roma
Natural Language Processing Course

Adrian Alexa
alexa.2160731@studenti.uniroma1.it

Adriano Semerano
semerano.2160754@studenti.uniroma1.it

**Abstract**

Lipreading is the task of decoding text from the movement of a speaker's mouth. Traditional approaches separated the problem into two stages: designing or learning visual features, and prediction. More recent deep lipreading approaches are end-to-end trainable (Wand et al., 2016; Chung & Zisserman, 2016a). However, existing work on models trained end-to-end perform only word classification, rather than sentence-level sequence prediction. Studies have shown that human lipreading performance increases for longer words (Easton & Basala, 1982), indicating the importance of features capturing temporal context in an ambiguous communication channel.

## 1 Introduction

This paper addresses the task of lip reading by first delineating the inherent problem, followed by a rigorous presentation of the selected and empirically validated solutions.

Lipreading plays a crucial role in human communication and speech understanding. In fact, it is a notoriously difficult task for humans, especially in the absence of context. Most lipreading cues—besides the lips, and occasionally the tongue and teeth—are subtle and difficult to disambiguate without contextual information (Fisher, 1968; Woodward & Barber, 1960). Consequently, human lipreading performance is poor. Hearing-impaired individuals achieve an accuracy of only $17 \pm 12\%$ even for a limited set of 30 monosyllabic words, and $21 \pm 11\%$ for 30 compound words (Easton & Basala, 1982). An important goal, therefore, is to automate lipreading and aim to achieve good performance where humans often fail.

Machine lip readers have enormous practical potential, with applications in improved hearing aids, silent dictation in public spaces, biometric identification, silent film processing, security, and speech recognition in noisy environments. They can be used in combination with acoustic speech recognizers to compensate for the degraded performance caused by noise.

## 2 Examples of input and expected output

The core input to a lip-reading system is visual data of a person speaking. This usually takes the form of a video sequence which is processed in such a way to detect the face of the subject, pinpoint and extract specific points across the face, crop the area around the mouth, normalize the obtained points and then process the normalized frame sequence based on the underlying structure of the system.

Some systems, as we will discuss later in the paper, are audio-visual speech recognition systems, meaning they combine lip-reading with traditional speech recognition. In such cases, the corresponding audio track is also an input to the systems.

The primary output is the corresponding textual representation, i.e., the predicted text captions or sentences that the person in the video is speaking.

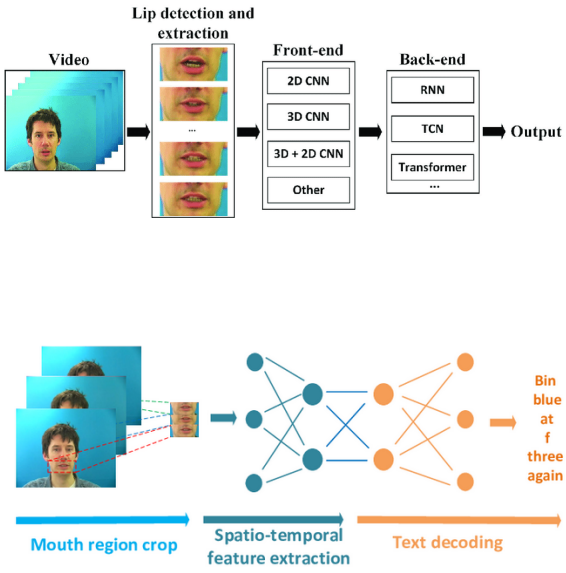


Figure 1: Examples of lip reading tasks.

# 3 Challenges in Automatic Lipreading

This task is subject to numerous challenges and limitations which affect the performance of the automatic lipreading systems. Some of these challenges can be addressed by applying preprocessing steps on the input data, with the objective of reducing the impact of noise on the model's performance. The most common limitations are:

- **Visual and Audio Ambiguity:** Many phonemes share indistinguishable visual patterns (visemes), making it hard to differentiate between sounds like /p/, /b/, and /m/ based on lip movements alone.
- **Dependence on Linguistic Context:** Accurate recognition often requires strong language models to resolve ambiguities, as lip movements alone provide limited information about word identity.
- **Speaker and Pose Variability:** Differences in speaking style, facial features, head orientation, facial hair, and expressiveness lead to high inter-speaker variability that degrades model generalization.
- **Video Quality and Recording Conditions:** Low resolution, poor lighting, motion blur, occlusions, and low frame rates reduce the visibility of lip movements and hinder recognition performance.
- **Limited and Biased Datasets:** Most available datasets are small, language-specific (typically English), and collected under controlled conditions, limiting the diversity and robustness of trained models.
- **Lack of Real-World Robustness:** Models that perform well on benchmark datasets often fail in real-world scenarios where speakers are partially visible, lighting is uncontrolled, or videos are compressed.

# 4 Applications of Lip Reading

Lip reading technology, powered by advanced AI and deep learning, has a wide range of practical applications across various sectors, significantly enhancing communication and offering solutions in challenging environments. There are many key areas and so let's give some examples:

- **Accessibility and Assistive Technology for the Hearing Impaired:**
  - *Communication Aids:* Lip-reading systems can convert silent lip movements into text or even synthesized speech, allowing individuals with hearing loss or deafness to understand spoken conversations in real-time.
  - *Support for Speech Impairments:* For individuals who have lost the ability to speak due to medical conditions (e.g., stroke, throat cancer), lip-reading technology can offer a means to communicate silently, providing a dignified and efficient way to express needs and thoughts.
- **Security and Surveillance:**
  - *Forensic Analysis:* In cases where surveillance footage lacks clear audio, or the audio is deliberately obscured, lip-reading technology can be used by law enforcement to decipher conversations, gather evidence, and understand motives in criminal investigations.
- **Media and Entertainment:**
  - *Automated Subtitling and Dubbing:* Improving the accuracy and speed of subtitling for live broadcasts, online videos, and archival silent films. This is especially useful for content where the original audio is poor or missing.
- **Language Learning and Education:**
  - *Speech Therapy:* Assisting individuals in learning proper pronunciation by providing visual feedback on lip and mouth movements.

The applications for this task are numerous and diverse, particularly given the significant impact it has within the sensory domain it operates. However, a critical boundary must be recognized regarding the scope of its applications. While these applications hold considerable promise, it is crucial to acknowledge the inherent ethical and privacy implications, especially in areas such as surveillance. The capacity to 'listen' without auditory input raises serious concerns about consent, data security, and the potential for misuse. Consequently, the responsible development and deployment of this technology are paramount.

# 5 Datasets for Lip Reading

Here we list the datasets taken into consideration for this review, as well as a small introduction for each one. An important note which we want to highlight is that most of these datasets are English-based, directly indicating that the systems are also English-based:

- **LRW (Lip Reading in the Wild):** A large-scale audio-visual dataset primarily used for lip-reading research. It consists of short video clips (29 frames, 1.16 seconds) extracted from BBC programs, each centered around a single spoken word. The dataset features 500 distinct words spoken by over 1,000 different speakers, making it challenging due to variations in head pose and illumination. This dataset was not used for testing purposes given that it was previously used to train both the models we chose to evaluate.

Figure 3: Miracl-VC1 dataset example.

- **GRID Dataset:** A popular audio-visual speech corpus designed for various speech-related tasks, including lip-reading, speech enhancement, and speech separation. It features synchronized audio and video of speakers uttering simple, distinct commands (e.g., "bin red at G 9 now"). Known for its relatively controlled environment, it is suitable for foundational research in audio-visual speech processing. The dataset is available and served as a benchmark for the models.

Figure 2: GRID dataset example.

- **LRS2 (Lip Reading Sentences 2):** An extensive dataset derived from BBC programs (news and talk shows), containing full sentences up to 100 characters in length, spoken by thousands of different speakers. It introduces considerable variation in head pose and provides a realistic benchmark for sentence-level lip-reading and audio-visual speech recognition. Due to data-sharing agreements with the BBC, we were not allowed to include images or references from the dataset.
- **LRS3-TED:** A large-scale multi-modal dataset comprising face tracks from over 400 hours of TED and TEDx videos, along with subtitles and word alignment boundaries. It provides a robust benchmark for research in natural, unconstrained settings. The dataset is not currently available, as it was made private by the BBC.
- **AVICAR (Audio-Visual Speech in a Car):** A large in-car speech corpus containing multi-channel audio and video recordings captured in automotive environments. It includes speech data for digits, letters, phone numbers, and sentences under various noise conditions (e.g., idling, windows open/closed). We attempted to download the dataset through the provided link, but without success.
- **MIRACL-VC1:** A publicly available dataset designed for lip-reading and visual speech recognition, including both depth and color images. It features 15 speakers (5 men, 10 women) uttering 10 words and 10 phrases, each repeated 10 times, for a total of 3000 instances. This dataset was available and served as a benchmark for the models.

# 6 Current state-of-the-art and Evaluation

Deep learning models have dramatically elevated lip-reading accuracy in recent years. State-of-the-art systems achieved up to 94.1% accuracy on large datasets such as Lip Reading in the Wild (LRW) in 2023, marking a substantial improvement from 66.1% in 2016.

The SyncVSR (Word Boundary) model, published in 2024, currently holds the top rank on the LRW dataset leaderboard, achieving a remarkable Top-1 Accuracy of 95.0%. Its counterpart, SyncVSR (without word boundary), also demonstrated strong performance at 93.2%.

Another high-performing model, *3D Conv + ResNet-18 + DC-TCN + KD (Ensemble & Word Boundary)*, achieved 94.1% accuracy on LRW in 2022, showcasing the effectiveness of hybrid architectures and ensemble methods.

The Antoniocolapso/Lip-Reader project reported exceptional accuracy in its final submission, achieving 98.23% Word-level Accuracy and 99.33% Sentence-level Accuracy, with corresponding Mean Word Error Rates (WER) of 1.77% and Sentence Error Rates (SER) of 0.67%. This model's robustness was attributed to training on a diverse dataset encompassing various skin tones, accents, and genders.

A pre-processed lip reading system leveraging Conv3D and LSTM layers demonstrated over 90% accuracy rates on its test dataset, notably outperforming traditional audio-based speech recognition methods in challenging auditory environments.

Finally, a 2025 study achieved a peak validation accuracy of 98.18% using an optimized architecture (combining ResBlock3D, Conv3D, Conv2D, TimeDistributed, attention mechanism, and LSTM) on a newly constructed dataset, *DATAV1*.

Here we show a representation of the different lipreading systems, the benchmarks on which the performance has been achieved:

| **Dataset Name** | **Best Model** | **Top-1 Accuracy / WER** | **Paper / Year** |
|---|---|---|---|
| Lip Reading in the Wild (LRW) | SyncVSR (Word Boundary) | 95.0% Accuracy | SyncVSR: Data-Efficient Visual Speech Recognition with End-to-End Cross-modal Audio Token Synchronization (2024) |
| Lip Reading in the Wild (LRW) | 3D Conv + ResNet-18 + DC-TCN + KD (Ensemble & Word Boundary) | 94.1% Accuracy | Training Strategies for Improved Lip-reading (2022) |
| Lip Reading in the Wild (LRW) | SyncVSR | 93.2% Accuracy | SyncVSR: Data-Efficient Visual Speech Recognition with End-to-End Cross-modal Audio Token Synchronization (2024) |
| LRS3-TED | V-ASR Approach (Data-Efficient) | 18.7% WER | – |
| LRS3 | Streaming AV-ASR (Offline) | 2.0% WER | Streaming Audio-Visual Speech Recognition with Alignment Regularization (Meta AI) |
| LRS3 | Streaming AV-ASR (Online) | 2.6% WER | Streaming Audio-Visual Speech Recognition with Alignment Regularization (Meta AI) |
| Antoniocolapso Project | Bi-GRU LSTM (Final Submission) | 98.23% Word Acc. | Antoniocolapso/Lip-Reader (GitHub) |
| Antoniocolapso Project | Bi-GRU LSTM (Final Submission) | 99.33% Sentence Acc. | Antoniocolapso/Lip-Reader (GitHub) |
| LRW-1000 | SwinLip | SOTA Performance | SwinLip: An Efficient Visual Speech Encoder for Lip Reading Using Swin Transformer (2025) |

Table 1: Summary of datasets, models, and main results.

# 7 Our Evaluation on the Models

The models taken into consideration for testing purposes are the following:

For the LipNet system, we examined multiple implementations available online. Unfortunately, most of these implementations relied on outdated versions of key libraries, which required us to introduce several modifications in order to address compatibility issues. All of the implementations we considered ultimately derive from the original LipNet paper, which is referenced in the Bibliography of this work. Due to technical challenges and time constraints, we redirected our efforts toward another model cited in the same paper. Consequently, the LipNet model and its various implementations were set aside.

The Temporal CNN system was implemented using an existing codebase from the GitHub repository cited in the bibliography. This implementation was downloaded and used without issue.

As for the second model, the Deep Lip Reading model was selected as an alternative to LipNet. We utilized an implementation of this model, derived from its original paper and found on GitHub, for our testing on the datasets. The specific details and considerations regarding the utilization of these models are discussed in a later section of this report.

## 7.1 Deep Lip Reading

We tested three datasets using the pre-trained Transformer-based model found on the GitHub repository of the Deep Lip Reading project. Here are the details of the implementation and its author:

@InProceedingsAfouras18b,
author = "Afouras, T. and Chung, J. S. and Zisserman, A.",
title = "Deep Lip Reading: a comparison of models and an online application",
booktitle = "INTERSPEECH",
year = "2018",

The model is based on the Transformer architecture and it is one of the three models analyzed and introduced in the Deep Lip Reading paper. The author of the implementation made it available through the GitHub repository, providing:

- the requirements for setting up the environment,
- the architecture of the model,
- the language model used,
- the test labels for evaluation,
- and the results obtained when testing the model on the original dataset it was trained on.

### 7.1.1 Main Features of the Model

The main features of the Deep Lip Reading model are:

- It is based on the Transformer architecture discussed in the Deep Lip Reading paper and is the best at the task of visual speech recognition among the three different models.
- Its goal is to predict sentences from given videos, thus the metrics used for attesting its capabilities are not accuracy but rather WER and CER:
  - **WER** stands for Word Error Rate, measuring the proportion of word-level transcription errors made by the system compared to the ground truth.
  - **CER** stands for Character Error Rate, a finer-grained version of WER that operates at the character level instead of words.
- The datasets used to train it were LRW and LRS, as stated both by the paper and the author of the implementation.
- The dataset used to evaluate the model is LRS2.
- Several modifications to the source code were required in order to adapt the testing procedure to different datasets and to our local machines:
  - Using AMD GPUs caused some issues with CUDA, so CPU support was enabled to run the tests. Logging features were also added to monitor the execution of preprocessing and testing steps.
  - The requirements file that the author instructed downloading, along with the lip and language models, produced errors. We manually modified and installed the required libraries in batches, resulting in a separate environment containing all necessary libraries and the correct Python version.
  - The resolution used by the model to process videos is set as 160x160. The `load_video.py` file was modified to preprocess videos to follow this resolution.

Here are displayed the values obtained from testing the model on the different datasets:

| Dataset Name | Accuracy / WER |
|---|---|
| LRS2 | 58% WER |
| MIRACL-VC1 | 100–150% WER |
| GRID | 86–100% WER |

Table 2: Results of testing the model on different datasets.

### 7.1.2 LRS2

The LRS2 dataset consists of thousands of spoken sentences taken from BBC television. As mentioned in the Dataset section, we are not allowed to share any content from this dataset. However, we can note that the videos are clearly taken from BBC television shows, where the speakers are often moving or performing some activity while speaking. In some cases, the speaker looks directly at the camera, while in others they look elsewhere.

The model implementation relies on a pre-trained lip model and a language model, which were sourced directly from the BBC archives. The documentation within the GitHub repository provided the necessary links to download and utilize these models, allowing for an initial validation on the LRS2 dataset followed by subsequent benchmarking on other datasets.

After obtaining and downloading the LRS2 dataset from the BBC we tested the model following the provided instructions and running the scripts, thus confirming the values for the WER and CER metrics.

The value of WER provided by the author are measured as 58% without the use of beam search and 49% with the use of it while the CER measures 38%.

The results for CER and WER are reported without the application of a beam search due to its substantial computational cost. The author notes that a GPU-based execution of the test would take roughly 10 hours. Consequently, operating in a CPU-only environment due to hardware incompatibility, we deemed the implementation of a beam search to be impractical.

```
=== Final Results ===
global CER: 38.2
global WER: 57.92
========================
```

Figure 4: LRS2 results

### 7.1.3 MIRACL-VC1

The MIRACL-VC1 dataset includes 10 words and 10 sentences spoken by 15 speakers. The recorded videos include the speaker and the external environment, both with color and depth aspects, originally the videos are non-existent since only the frames are present when downloading the dataset. For testing this model only the sentences were chosen to be tested, thus maintaining the intended goal of the model. This was a dataset on which the model was never tested, so as to preprocess it was necessary applying some modifications after analyzing the structure of the dataset and attesting the technical differences between the videos of the LRS2 dataset and this new one:

- The MIRACL-VC1 reconstructed videos have a low frame rate, counting approximately 10-15 frames per video.The sentences spoken inside the videos are thus stretched across the frames.
- The resolution used by the model was 160x160 while the miracl-vc1 was 640x480, indicating that the video frames included external and non useful elements of the environment which could have lowered the performance of the model. In order to bring the videos to the same format of the model we applied cropping techniques on the extracted facial features, centering the newly cropping version around the detected face. We used pre-existing landmarks and applied specific techniques to extract the facial features, all being preprocessed in a python script.
- Lastly we replicated the internal structure of the LRS2 dataset used by the model for the miracl-vc1, this can be summarized with the following image

```
dataset_by_speaker_color/
├── F01/
│   ├── 01_01.mp4
│   ├── 01_01.txt
│   ├── 01_02.mp4
│   ├── 01_02.txt
│   ├── ...
│   ├── 10_10.mp4
│   └── 10_10.txt
├── F02/
│   ├── 01_01.mp4
│   ├── 01_01.txt
│   ├── ...
│   └── 10_10.txt
├── F03/
│   └── ...
└── ...
```

Figure 5: Internal structure of the MIRACL-VC1 dataset adapted to match LRS2 format.

The obtained WER was not what we expected, as displayed in the cell of the results table the WER is set at 150%. This value exceeds the 100% by a half margin, indicating that the number of insertions is high and subsequently the model is generating a significant amount of text that is not present in the ground truth. By modifying slightly the processing technique native that the model used to load and preprocess the videos, the value of WER was lowered to 100%. Still the value for CER didn‘t register any change.The values for WER and CER are shown here below:

```
=== Final Results ===
global CER: 86.6
global WER: 150.0
========================
```

Figure 6: MIRACL-VC1 results

### 7.1.4 GRID

The last dataset on which we tested the model is the Grid Audio-Visual Speech Corpus, a large multitalker audiovisual sentence corpus designed to support joint computational-behavioral studies in speech perception.The dataset contains both audio and video (we used only the video for our tests) and provides word boundaries aligned with the audio track. As we've proceeded with the MIRACL-VC1, we adapted the structure of the GRID dataset to that of the test provided by the author on github:

```
new_GRID_structure/
├── s1_processed/
│   ├── bbab8n.mpg
│   ├── bbab8n.align
│   ├── bbab9s.mpg
│   ├── bbab9s.align
│   └── ...
├── s2_processed/
│   ├── bbaf2n.mpg
│   ├── bbaf2n.align
│   ├── bbaf3s.mpg
│   ├── bbaf3s.align
│   └── ...
├── s3_processed/
│   └── ...
└── ...
```

Figure 7: Internal structure of the GRID dataset adapted to match LRS2 format.

The GRID dataset contains for every speaker the videos of speaker's spoken sentence and a .align file with the timestamps for each word of the sentence. For the tests of the model while the microsecond timestamps in the original align files are very useful for general tasks like forced alignment or temporal analysis of speech, they do not directly impact the WER metric We analyzed the common words belonging to both the lrs2_test_sample and GRID word set and found out that the total amounts to 23 unique words

| | | | | |
|---|---|---|---|---|
| A | FIVE | LAY | SET | |
| AGAIN | FOUR | NOW | SEVEN | |
| AT | GREEN | ONE | SIX | TWO |
| BY | I | PLACE | SOON | WHITE |
| EIGHT | IN | PLEASE | THREE | WITH |

Figure 8: words available for testing

After slicing the GRID dataset into a version that had only the videos and the files containing the labels which contained at least one of the words taken from the 23 unique words , we used this new dataset to test the model. The tests revealed the same results as per the original GRID, leading us to take a step further and analyze the situation beforehand, hence the differences we've spotted between the GRID and LRS2 datasets:

- GRID dataset is a controlled, studio-quality dataset with a very limited vocabulary and sentence structure. Sentences follow a strict "command," "color," "preposition," "letter," "digit," "adverb" pattern (e.g., "bin red at G 5 now").
- The LRS2 dataset, on the other hand, is a much larger, "in-the-wild" dataset taken from BBC television. It has a significantly larger and more diverse vocabulary, along with natural, unconstrained sentence structures

We thought of another possible approach, selecting only the labels of GRID that had a minimum percentage(for instance 50%) of unique words in the sentence spoken by the speaker. Of course the percentage was in correlation with the length that each sentence had. This approach would severely limit the number of sentences from the GRID dataset that could be used for testing. This is because GRID has a very limited vocabulary, and it's unlikely that any given sentence would contain a high percentage of words also found in the diverse LRS2 vocabulary. The reported value for WER fluctuates between 86% and 100% as reported in the table, majority of labels tested have an error value of 100% and consequently the average registered value is set to 100%

```
=== FINAL RESULTS ===
global CER : 86
global WER : 100
=======================
```

Figure 9: GRID results.

## 7.2 Temporal Convolutional Network

We tested three datasets using one of the pre-trained models available from the GitHub repository of the Lipreading using Temporal Convolutional Networks project. Among the different models provided, we selected the resnet18_dctcn_video model, which the authors describe as their best-performing model, achieving an accuracy of 89.6% on the LRW dataset. However, among the visual-only models, there is another one, which uses boundaries, that achieves a higher accuracy (92.1%). The project offers several pre-trained models, including both audio-only and visual-only versions, as well as additional visual-only models with lower accuracies and different preprocessing methods. The main features of this model and of the testing procedure that we applied are:

- The selected model was trained on single words, meaning it cannot predict full sentences; it can only output a single word label for each video clip.
- The model was trained on 500 English word labels, since the LRW dataset is a BBC dataset. The limited vocabulary creates compatibility issues with other datasets.

- The project provides its own preprocessing pipeline, which, however, requires significant modifications to adapt new datasets.
- For the LRW dataset, the authors provide precomputed facial landmarks. In our case, since we used different datasets, we had to compute the landmarks ourselves to initiate preprocessing. We used DLib for this purpose, but this implies that our results might slightly differ from those obtained by the authors, depending on how the landmarks were extracted.
- To test the model, it is also necessary to generate an annotation file and a CSV file, used to localize the preprocessing results (NPZ files).
- Several modifications to the source code were required in order to adapt the testing procedure to different datasets and to our local machines. In particular, using AMD GPUs caused some issues with CUDA, so we enabled CPU support to run the tests. We also added logging features to monitor the execution of preprocessing and testing steps.

The following are the accuracy results of the model on the datasets we tested:

| Dataset Name | Accuracy / WER |
|---|---|
| LRS2 | 40.81% Accuracy |
| MIRACL-VC1 | 0.67% Accuracy |
| GRID | 43.61% Accuracy |

Table 3: Results of testing the model on different datasets.

#### 7.2.1 LRS2

The first dataset we tested on the Temporal Convolutional Network model was the LRS2 dataset, another dataset provided by the BBC. It consists of thousands of spoken sentences taken from BBC television. As mentioned in the Dataset section, we are not allowed to share any content from this dataset. However, we can note that the videos are clearly taken from BBC television shows, where the speakers are often moving or performing some activity while speaking. In some cases, the speaker looks directly at the camera, while in others they look elsewhere. After signing the Data Sharing Agreement, we downloaded the dataset, which contains both videos and text files providing word-level boundaries within each sentence. Using these text files, we extracted 127,092 clips of speakers pronouncing a single word. Only the clips corresponding to words included in the 500-word vocabulary used to train the model were kept, since all other words would be unknown to the model and therefore impossible to predict. The results obtained on this dataset differ significantly from those achieved on LRW. Specifically, testing on the 127k clips yielded an accuracy of 40.81% with an average loss of 2.92.

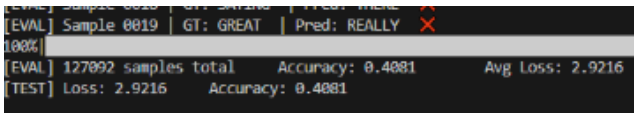


Figure 10: LRS2 Results.

#### 7.2.2 MIRACL-VC1

The Miracl-VC1 dataset contains both single words and full sentences. Unfortunately, we encountered two major compatibility issues that significantly affected the outcome of our tests:

- The dataset includes 10 words and 10 sentences spoken by 15 speakers. However, the sentences are not annotated with word-level boundaries, which prevented us from splitting them into single-word clips.
- Even more limiting, among the 10 words included in the dataset, only one ("START") appears in the 500-word vocabulary used to train the TCN model. As a result, we were only able to test the model on 150 clips, all corresponding to the same word label.

In addition, the dataset has relatively low quality: the frame rate is low, and the speaker is positioned far from the camera. This creates a large discrepancy compared to the high-quality BBC videos used for training the model. As expected, the results of this experiment were very poor. The model achieved an accuracy of just 0.67% on the 150 clips, with an average loss of 6.3. The model produced a large number of incorrect predictions, often confusing the ground-truth word with labels that share similar lip movements, such as "heart", "started", or most frequently "other". In some cases, it even predicted completely unrelated words, such as "government".

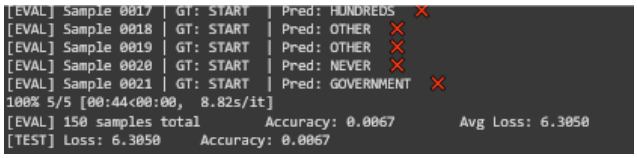


Figure 11: Miracl-VC1 Results.

#### 7.2.3 GRID

Finally, we tested the model on the GRID dataset, which consists of high-quality videos of speakers facing the camera and reading sentences. The dataset contains both audio and video (we used only the video for our tests) and provides word boundaries aligned with the audio track. These boundaries are specified in audio samples rather than milliseconds. By processing the audio and converting these sample-based boundaries into corresponding video segments, we were able to extract the single-word clips required for testing. However, we again faced a compatibility issue due to

the limited overlap between the dataset's vocabulary and the model's 500-word training set. Only 4 words from the GRID dataset were present in the model's vocabulary. Despite this, the number of usable clips was still considerable: a total of 23,038 clips, distributed as follows:

- again: 8,000+ samples
- place: 8,000+ samples
- seven: 3,000+ samples
- three: 3,000+ samples

The results obtained are comparable to those on the LRS2 dataset, though slightly better—likely due to the higher quality of the videos and the fact that speakers consistently face the camera while speaking. The model achieved an accuracy of 43.61% with an average loss of 2.48. Moreover, upon reviewing the predictions, we observed that even when the model produced an incorrect label, the predicted word often corresponded to lip movements visually similar to those of the ground-truth label.

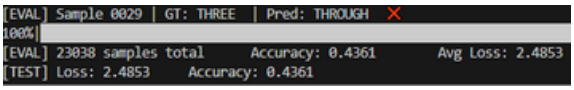


Figure 12: Grid Results.

## 8 Conclusions

Lip reading remains a highly challenging task, particularly when models are required to generalize beyond the datasets on which they were trained. While state-of-the-art approaches often achieve strong performance on benchmark datasets, such results can be misleading, as the same models may exhibit a severe drop in accuracy when applied to previously unseen data. This degradation is largely driven by the limited vocabulary size of training corpora, differences in language and phonetic structures, as well as variations in illumination, image quality, and speaker pose. In some cases, these factors can cause performance to deteriorate dramatically, even to the point of system failure.

Future work should therefore prioritize strategies aimed at improving robustness and generalization. Promising directions include domain adaptation techniques, the use of larger and more diverse datasets covering multiple languages and conditions, and the integration of multimodal signals such as audio-visual fusion. Furthermore, methods based on self-supervised learning and transfer learning may help mitigate overfitting while enhancing cross-domain adaptability. These approaches will be essential in moving lip reading systems from controlled research environments toward reliable deployment in real-world applications.

***For non-commercial individual research and private study use only. BBC content included courtesy of the BBC.***

**Contribution Statement**

- **Adrian Alexa:** Testing Deep Lip Reading models, state-of-the-art analysis, dataset research and analysis.
- **Adriano Semerano:** Testing Lipreading using Temporal Convolutional Networks models, state-of-the-art analysis, dataset research and analysis.

## References


[1] A. Author, "Survey on automatic lip-reading in the era of deep learning," *ScienceDirect*, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0262885618301276

[2] B. Author, "Advances and Challenges in Deep Lip Reading," 2021. [Online]. Available: https://ar5iv.labs.arxiv.org/html/2110.07879

[3] C. Author, "Visual Speech Recognition for Multiple Languages in the Wild," 2022. [Online]. Available: https://ar5iv.labs.arxiv.org/html/2202.13084

[4] R. Assael et al., "LipNet: End-to-End Sentence-Level Lipreading," 2016. [Online]. Available: https://arxiv.org/pdf/1611.01599

[5] R. Assael et al., "LipNet implementation," GitHub. [Online]. Available: https://github.com/rizkiarm/LipNet

[6] D. Author, "Lipreading using Temporal Convolutional Networks," IEEE, 2020. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9053841

[7] E. Author, "Lipreading using TCN," 2022. [Online]. Available: https://arxiv.org/abs/2209.01383

[8] F. Author, "Lipreading using TCN," 2020. [Online]. Available: https://arxiv.org/abs/2009.14233

[9] G. Author, "Lipreading using TCN," 2020. [Online]. Available: https://arxiv.org/abs/2007.06504

[10] H. Author, "Lipreading using Temporal Convolutional Networks implementation," GitHub. [Online]. Available: https://github.com/mpc001/Lipreading_using_Temporal_Convolutional_Networks

[11] T. Afouras, J.S. Chung, A. Zisserman, "Deep Lip Reading: a comparison of models and an online application," 2018. [Online]. Available: https://www.robots.ox.ac.uk/~vgg/publications/2018/Afouras18b/afouras18b.pdf

[12] T. Afouras et al., "Deep Lip Reading implementation," GitHub. [Online]. Available: https://github.com/afourast/deep_lip_reading

[13] I. Author, "The proposed end-to-end lipreading system," 2018. [Online]. Available: https://www.researchgate.net/figure/The-proposed-end-to-end-lipreading-system-includes-three-major-steps-1-The-mouth_fig1_325638464

[14] J. Author, "Lipreading using deep learning," 2018. [Online]. Available: https://arxiv.org/pdf/1804.03619

[15] VGG, "LRS2 Dataset," [Online]. Available: https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html

[16] "Lip Reading in the Wild Benchmark," [Online]. Available: https://paperswithcode.com/sota/lipreading-on-lip-reading-in-the-wild

[17] K. Author, "VALLR: Visual ASR Language Model for Lip Reading," 2025. [Online]. Available: https://arxiv.org/html/2503.21408v1

[18] Meta AI, "Streaming Audio-Visual Speech Recognition with Alignment Regularization," [Online]. Available: https://ai.meta.com/research/publications/streaming-audio-visual-speech-recognition-with-alignment-regularization/