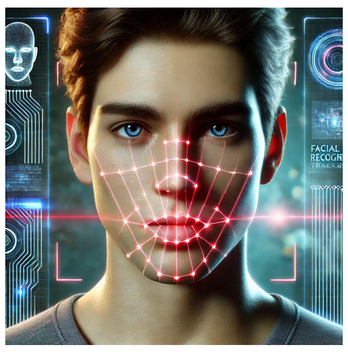
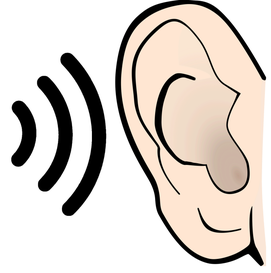
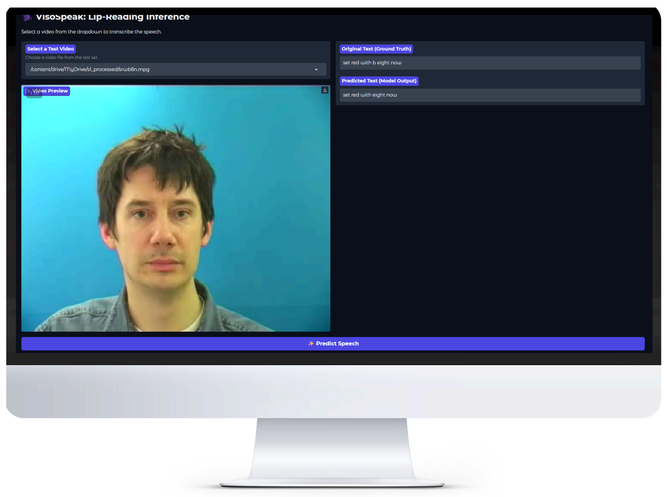


VisoSpeak



Optimizing Lip-Reading Accuracy through Advanced Data Pipelines and Automated Video Processing.



BRAUDE
College of Engineering, Karmiel

AUTHORS

Morad Asakli
Majd Salameh

ADVISOR

Mr. Ilya Zeldner

PROBLEM

Why VisoSpeak really matters?

Speech recognition systems that rely on sound don't help in noisy places, silent environments, or for people who are deaf or hard of hearing. There's a need for a system that can understand speech by reading lips, in any situation.

SOLUTION

Build a complete deep learning system that reconstructs spoken sentences using only visual information from the speaker's lip movements without any audio input. This enables silent speech recognition, accessibility for the hearing impaired and in noisy conditions.

METHOD

We used, modified, and built two models: one based on LipNet, and another based Transformer encoder-decoder, enabling accurate predictions only using visual speech recognition.

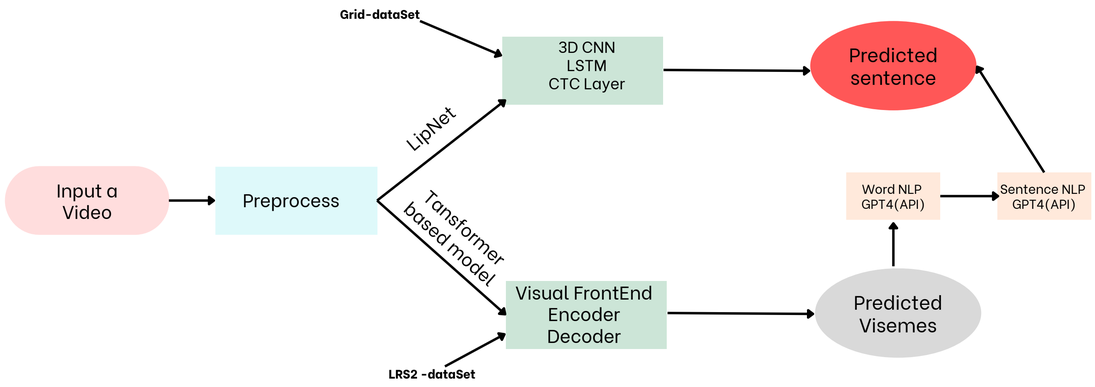
DIFFICULTIES & SOLUTIONS

- LRS2 dataset was private and required special access from Oxford University
  - contacted the relevant team at Oxford and received access; meanwhile, we worked on LipNet.
- Too many technologies and unclear which to use
  - We experimented with various tools and kept the most effective ones
- No access to a powerful GPU for training
  - We used Google Colab Pro with an A100 GPU
- I/O overhead on Colab slowed training and used GPU units
  - We limited training to small sets for testing and optimization
- War during the development phase
  - We stayed strong and focused; the presentation was postponed to support our time

Technolgies Used:

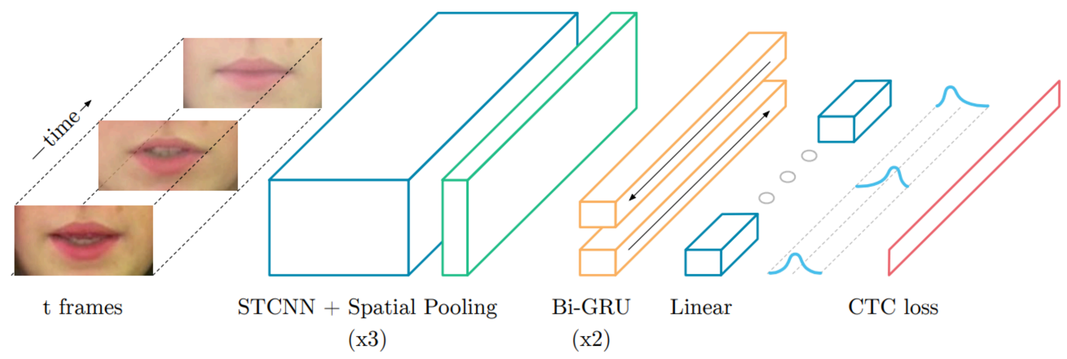
Category	Transformer-Based Model	LipNet Model
Deep Learning	torch, torch.nn, torch.nn.functional, torch.optim, torch.utils.data, torch.cuda.amp	tensorflow, tensorflow.keras, tensorflow.keras.models, tensorflow.keras.layers, tensorflow.keras.optimizers, tensorflow.keras.callbacks
Computer Vision	cv2, mediapipe	cv2
Data Handling	numpy, pandas, json, os, sys, pathlib, glob, re, math, random, datetime, gc	numpy, os, csv, glob
Visualization	matplotlib.pyplot, tqdm, tqdm.auto	matplotlib
Concurrency	threading, queue, concurrent.futures, multiprocessing.pool	—
Notebook Tools	google.colab, IPython.display, ipywidgets	google.colab
Custom Modules	src.dataset	—
Language Tools	inflect, g2p_en	—
External APIs	Google Drive API, ChatGPT-4 API	—
Evaluation	Levenshtein, Cosine similarity,	jiwer (for WER)

END TO END WORKFLOW:



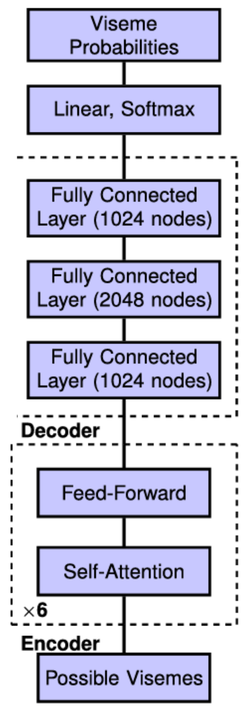
Model Architecture :

LipNet



Transformer based Model

Layer Type	Filter	Output Dimensions
3D Convolution	$[5 \times 7 \times 7, 64]/(1,2,2)$	$180 \times 56 \times 56 \times 64$
3D Max Pooling	(1,2,2)	$180 \times 28 \times 28 \times 64$
Residual 2D Convolution	$[3 \times 3, 64] \times 2/(1,1)$	$180 \times 28 \times 28 \times 64$
Residual 2D Convolution	$[3 \times 3, 64] \times 2/(1,1)$	$180 \times 28 \times 28 \times 64$
Residual 2D Convolution	$[3 \times 3, 128] \times 2/(2,2)$	$180 \times 14 \times 14 \times 128$
Residual 2D Convolution	$[3 \times 3, 128] \times 2/(1,1)$	$180 \times 14 \times 14 \times 128$
Residual 2D Convolution	$[3 \times 3, 256] \times 2/(2,2)$	$180 \times 7 \times 7 \times 256$
Residual 2D Convolution	$[3 \times 3, 256] \times 2/(1,1)$	$180 \times 7 \times 7 \times 256$
Residual 2D Convolution	$[3 \times 3, 512] \times 2/(2,2)$	$180 \times 4 \times 4 \times 512$
Residual 2D Convolution	$[3 \times 3, 512] \times 2/(1,1)$	$180 \times 4 \times 4 \times 512$



The architecture of transformer for the Viseme Classifier.

TESTS

Functional Tests:

The system successfully recognized words and full sentences from silent videos using only lip movements, without relying on audio.

Non-Functional Tests:

The system processed videos efficiently without crashes, even under high load or with long input sequences.

Accessibility Tests:

Deaf and hard-of-hearing users confirmed that the system provided an accurate and helpful way to understand spoken content visually.

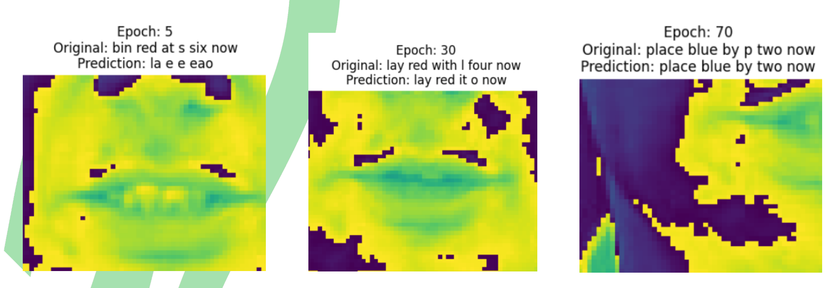
CONCLUSION

VisoSpeak is a deep learning-based system that understands speech using only lip movements, without audio.

By modifying LipNet and building a Transformer-based model, we achieved accurate visual speech recognition across various conditions.

The system was trained on large datasets and tested thoroughly, proving its reliability and adaptability.

RESULTS




```
Predictions for Epoch 78

 Predicted    : <sow> W IH N T <eow> <space> <sow> W OW <eow> <space> <sow> W AH <eow> <space> <sow> W UW <eow>
<space> <sow> K AH N <eow> <space> <sow> AH T <eow> <space> <sow> DH ER <eow> <space> <sow> F AH N T ER S EY SH AH
N <eow> <eow> <space>
 Ground truth : <sow> D OW N T <eow> <space> <sow> N OW <eow> <space> <sow> W AY <eow> <space> <sow> Y UW <eow>
<space> <sow> JH OY N <eow> <space> <sow> IH N <eow> <space> <sow> AW ER <eow> <space> <sow> K AA N V ER S EY SH
AH N Z <eow> <eos>
```