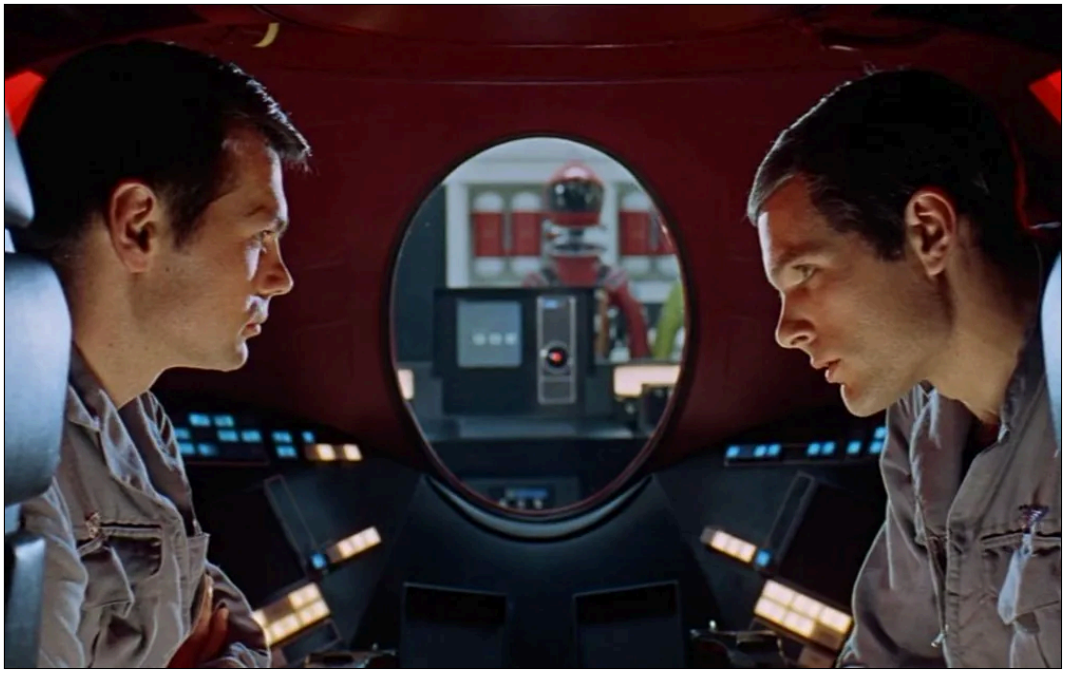
Lip-Reading With Visemes and Machine Learning

Originally published at https://www.unite.ai/lip-reading-with-visemes-and-machine-learning/



New research from the School of Computer Engineering at Tehran offers an improved approach to the challenge of creating machine learning systems capable of reading lips.

The paper, entitled *Lip Reading Using Viseme Decoding*, reports that the new system achieves a 4% improvement in word error rate over the best of similar previous models. The system addresses the general lack of useful training data in this sector by mapping *visemes* to text content derived from the six million samples in the OpenSubtitles dataset of translated movie titles.

A viseme is the visual equivalent of a phoneme, effectively an audio>image mapping that can constitute a 'feature' in a machine learning model.

Visemes in action. Source: https://developer.oculus.com/documentation/unity/audio-ovrlipsync-viseme-reference/

The researchers began by establishing the lowest error rate on available datasets, and developing viseme sequences from established mapping procedures. Gradually, this process develops a visual lexicon of words – though it's necessary to define probabilities of accuracy for different words that share a viseme (such as 'heart' and 'art').

Ground Truth	Predicted
AND SO THIS IS WHAT I DID	AND SO THIS IS WHAT I DID
SO I SHOULD TALK ABOUT ART	SO I SHOULD TALK ABOUT YOU
WELL THE RANGE IS QUITE A BIT	WELL THERE HAD CHASE QUITE A BIT
THIS IS NELL REMMEL	THIS IS THE TROUBLE
NEXT IS SYLVIA SLATER	NEXT IS SILVER IS NOT HER
THAT IS MY MOM AND DAD	THAT IS MY PEOPLE THEN

Visemes extracted from text. Source: https://arxiv.org/pdf/2104.04784.pdf

Where two identical words result in the same viseme, the most frequently-occurring word is selected.

The model builds on traditional sequence-to-sequence learning by adding a sub-processing stage wherein visemes are predicted from text and modeled in a dedicated pipeline:

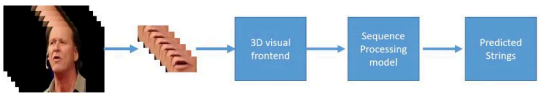
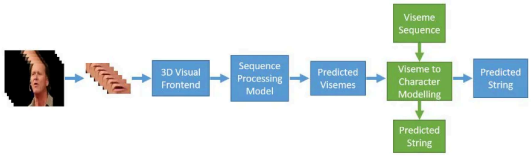


Figure 1: Traditional sequence to sequence methods.



Above, traditional sequence-to-sequence methods in a character model; below, the addition of viseme character modeling in the Tehran research model. Source: https://arxiv.org/pdf/2104.04784.pdf

The model was applied without visual context against the LRS3-TED dataset, released from Oxford University in 2018, with the worst word error rate (WER) obtained a respectable 24.29%.

The Tehran research also incorporates the use of a grapheme-to-phoneme converter.

In a test against the 2017 Oxford research *Lip Reading Sentences In The Wild* (see below), the Video-To-Viseme method achieved a word error rate of 62.3%, compared to 69.5% for the Oxford method.

The researchers conclude that the use of a higher volume of text information, combined with grapheme-to-phoneme and viseme mapping, promises improvements over the state of the art in automated lip-reading machine systems, while acknowledging that the methods used may produce even better results when incorporated into more sophisticated current frameworks.

Machine-driven lip-reading has been an active and ongoing area of computer vision and NLP research over the last two decades. Among many other examples and projects, In 2006 the use of automated lip-reading software captured headlines when used to interpret what Adolf Hitler was saying in some of the famous silent films taken at his Bavarian retreat, though the application seems to have vanished into obscurity since (twelve years later, Sir Peter Jackson resorted to human lip-readers to restore the conversations of WW1 footage in the restoration project *They Shall Not Grow Old).*

In 2017, *Lip Reading Sentences in The Wild*, a collaboration between Oxford University and Google's AI research division produced a lip-reading AI capable of correctly inferring 48% of speech in video without sound, where a human lip-reader could only reach a 12.4% accuracy from the same material. The model was trained on thousands of hours of BBC TV footage.

This work followed on from a separate Oxford/Google initiative from the previous year, entitled LipNet, a neural network architecture that mapped video sequences of variable length to text sequences using a Gated Recurrent Network (GRN), which adds functionality to the base architecture of a Recurrent Neural Network (RNN). The model achieved a 4.1x improved performance over human lip-readers.

Besides the problem of eliciting an accurate transcript in real time, the challenge of interpreting speech from video deepens as you remove helpful context, such as audio, 'face-on' footage that's well-lit, and a language/culture where the phonemes/visemes are relatively distinct.

Though there's currently no empirical understanding of which languages are the most difficult to lip-read in the complete absence of audio, Japanese is a prime contender. The different ways that Japanese natives (as well as certain other West and East Asian natives) leverage facial expressions against the content of their speech already make them a greater challenge for sentiment analysis systems.

However, it's worth noting that much of the scientific literature on the topic is generally circumspect, not least because even well-intentioned objective research in this sphere risks to cross over into racial profiling and the promulgation of existing stereotypes.

Languages with a high proportion of guttural components, such as Chechen and Dutch, are particularly problematic for automated speech extraction techniques, while cultures where the speaker may express emotion or deference by looking away (again, generally in Asian cultures) add another dimension where AI lip-reading researchers will need to develop additional methods of 'in-filling' from other contextual clues.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai