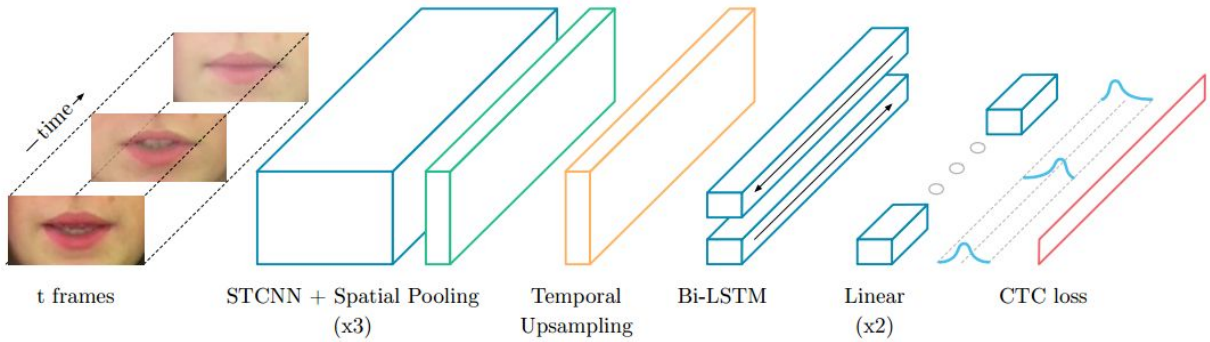
Lipreading system is focus of research team at University of Oxford

November 10 2016, by Nancy Owano



LipNet architecture. A sequence of T frames is used as input, and is processed by 3 layers of STCNN, each followed by a spatial max-pooling layer. The features extracted are temporally up-sampled and are processed by a Bi-LSTM; each timestep of the LSTM output is processed by a 2-layer feed-forward network and a softmax. This end-to-end model is trained with CTC. Credit: arXiv:1611.01599 [cs.LG]

(Tech Xplore)—Another take-a-bow for research at University of Oxford: A Department of Computer Science research team has developed a system for automatic lipreading. Using machine learning, their goal has been to help those who are hard of hearing.

The good news is that it can surpass the performance of human lip readers and even previous automatic lip reading systems, according to a university news release.

How so? The researchers define lipreading as the task of decoding text from the movement of a speaker's mouth.

The software reads lips faster and more accurately than was previously possible. The team used deep learning AI to create LipNet. The BBC described their materials and methods: "They said that the AI system was provided with whole sentences so that it could teach itself which letter corresponded to which lip movement."

The team fed it nearly 29,000 videos, labeled with the correct text, to train. The BBC said, "Each video was three seconds long and followed a similar grammatical pattern."

A video was posted earlier this month on their work, LipNet: Sentence level lipreading, by Yannis Assael, Brendan Shillingford, Shimon Whiteson and Nando de Freitas. Their paper is on arXiv.

The video asked, How easy do you think lipreading is? A female with sound off says something then slows it down but hard to guess what she said. Then the video shows LipNet's prediction, turns out, is right, something like place blue in m1 soon. Two male voices also say things and LipNet gets their sentences right.

The video said the average experienced lipreader performance is 52% while that of LipNet goes up to 93%.

The video notes included text from the abstract. "To the best of our knowledge, LipNet is the first lipreading model to operate at sentence-level, using a single end-to-end speaker-independent deep model to simultaneously learn spatiotemporal visual features and a sequence model. On the GRID corpus, LipNet achieves 93.4% accuracy, outperforming experienced human lipreaders and the previous 79.6% state-of-the-art accuracy."

(The authors wrote in their paper that the end-to-end model "eliminates the need to segment videos into words before predicting a sentence. LipNet requires neither hand-engineered spatiotemporal visual features nor a separately-trained sequence model.")

The authors in the video post thanked CIFAR, Google DeepMind and NVIDIA for financial support.

What would be its potential use in the real world? Would surveillance personnel use it to spy on others? In its current form, said the school department news release, it is unsuitable to be used for lip reading as a surveillance tool. "But the team is keen to develop it further, especially as an aid for people with hearing disabilities."

What's next? The question remains how can LipNet's AI actually take lip reading into the future. The university department news release said "it is still at a relatively early stage of development. It has been trained and tested on a research dataset of short, formulaic videos that show a well-lit person face-on."

Indeed. The BBC reported that "experts said the system needed to be tested in real-life situations. Lip-reading is a notoriously tricky business with professionals only able to decipher what someone is saying up to 60% of the time."

More information: LipNet: Sentence-level Lipreading, arXiv:1611.01599 [cs.LG] arxiv.org/abs/1611.01599

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). All existing works, however, perform only word classification, not 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. Motivated by this observation, we present LipNet, a model that maps a variable-length sequence of video frames to text, making use of spatiotemporal convolutions, an LSTM recurrent network, and the connectionist temporal classification loss, trained entirely end-to-end. To the best of our knowledge, LipNet is the first lipreading model to operate at sentence-level, using a single end-to-end speaker-independent deep model to simultaneously learn spatiotemporal visual features and a sequence model. On the GRID corpus, LipNet achieves 93.4% accuracy, outperforming experienced human lipreaders and the previous 79.6% state-of-the-art accuracy.