Watch the Whisper: An intersection of speech and computer vision

Shobha Sankar
UCB School of Information
shobha.sankar@
berkeley.edu

Karthik Srinivasan
UCB School of Information
kasri@
berkeley.edu

Jayesh Parikh
UCB School of Information
jparikh@
berkeley.edu

Abstract

Approximately 17.9 million people in the United States have trouble using their voices. In some cases, this affects the vocal folds in the larynx which can result in complete loss of voice. This can affect the basic communication function in their daily life. We propose a novel lip-reading recognition using computer vision and deep curriculum learning. The advantage of the proposed method is that it is a natural extension of their current lifestyle and is very simple to operate. This is a very promising alternative over the currently available external wearable solutions.

1 Introduction

Voice is the sound produced by humans using the lungs and the vocal folds in the larynx, or voice box. Voice is not always produced as speech, however. Infants babble and coo; and adult humans laugh, sing, and cry. When air is pushed past the vocal folds with sufficient pressure, the vocal folds vibrate. If the vocal folds in the larynx did not vibrate normally, speech could only be produced as a whisper. Your voice is as unique as your fingerprint. It helps define your personality, mood, and health. Approximately 17.9 million adults in the United States have trouble using their voices. Disorders of the voice involve problems with pitch, loudness, and quality. Pitch is the highness or lowness of a sound based on the frequency of the sound waves. Loudness is the perceived volume (or amplitude) of the sound, while quality refers to the character or distinctive attributes of a sound. Many people who have normal speaking skills have great difficulty communicating when their vocal apparatus fails. This can occur if the nerves controlling the larynx are impaired because of an accident, a surgical procedure, a viral infection, or cancer. (Clearinghouse, 2017)

Humans express thoughts, feelings, and ideas orally to one another through a series of complex movements that alter and mold the basic tone created by voice into specific sounds. Speech is produced by precisely coordinated muscle actions in the head, neck, chest, and abdomen. The ability to speak and communicate with one's voice is a unique human characteristic and is fundamental to many activities of daily living, such as talking on the phone and speaking to loved ones. When a patients loses their larynx due to illness followed by a total laryngectomy (TL), loss of voice can lead to a devastating decrease in a patient's quality of life, and precipitate significant frustration over their inability to communicate with others effectively. When thinking about the rehab of the patient who has lost their voice, we need to keep a few factors in mind. The speech aid must adjust to the user's lifestyle and must be a natural extension of them. It must be simple to use and easy to operate. (Kaye et al., 2017)

Currently voice restoration is achieved using three speech modalities: electrolaryngeal, esophageal, and tracheoesophageal. The electrolarynx is more straightforward to learn when compared to esophageal speech, but requires patients to purchase and maintain the device. The new silent speech recognition involves face- and neck-worn sensors and signal processing algorithms that are capable of recognizing silently mouthed words and phrases entirely from the surface electromyographic (sEMG) signals recorded from muscles of the face and neck that are involved in the production of speech. Lip reading using image processing and classification of visemes have also been implemented in the past.

In our current solution we have implemented a deep learning algorithm using the curriculum learning and temporal connectivity. This means a sequence of images predict the spoken characters. Using a self-attention model and a Connectionist Temporal Classification (CTC) to train and label the characters, this approach has improved the word error rates compared to the previous implementations.

2 Literature Survey

2.1 Current Wearable Solutions

Over the past 50 years there have been many advances in voice restorations. Of the three main methods of voice restoration: the electrolarynx, esophageal speech, and tracheoesophageal speech through a tracheoesophageal puncture (TEP) with voice prosthesis, the electrolarynx is the most viable and useful option for the patients whose larynx have been removed.

The electrolarynx is an external device that induces vibration of oral or pharyngeal mucosa, generally, at a constant fundamental frequency. The electrolarynx can

function either indirectly by contacting the skin, which induces pharyngeal vibrations or directly through intraoral contact, which induces oral cavity vibrations. Muscles of articulation are, generally, intact after TL and thus amenable to shaping the supplied vibration noise into understandable speech. The drawback of electrolarynx devices includes a mechanical hum radiating from the device that is not filtered by the vocal tract and instead is perceived directly by the listener. Some noise reduction techniques have been developed over the years to reduce this noise and give the human voice richer intonation instead of a robotic monotone.

Figure 1: Silent Speech Recognition using surface Electromyography.

The silent speech recognition can be accomplished using surface electromyography, or sEMG. Electrodes are placed on the skin at key places around the mouth, along the jaw, under the chin or on the neck. Muscle movements in those areas generate neuromuscular signals—essentially an electrical code. Algorithms trained on silent speech then translate the electrical signals, decoding what the user is saying. Delsys' sEMG has eight custom, rigid electrodes that are placed on the face and neck. They are connected to a computer where algorithms perform the decoding. The system can recognize about 2200 mouthed words—about 5 percent of the 42,000 words in the average English speaker's vocabulary. Delsys' device proved to be about 91 percent accurate in translating silent speech. Wearable solutions are complex and the technology that picks up signals using surface EMG can be susceptible to ambient noise. In addition currently it recognize only around 5% of the words in the English dictionary (Kline, 2018).

Figure 2: Bone conduction aural output with bi-directional human machine interface.

MIT lab's Alter-Ego solution is an extension of the sEMG solution mentioned above with a few improvements. In this case Silent speech recognition of the AlterEgo system attempts to open up a unique opportunity to enable personalized bidirectional human-machine interface in a concealed and seamless manner, where the element of interaction is in natural language. This is accomplished through an bone-conduction aural output which helps make the conversation private and personal if need be (Kapur et al., 2018).

2.2 Computer Vision and Lip Reading

Large amounts of work have been carried out in the pre deep learning era for visual speech recognition a.k.a lip reading. A review of these works are summarized in (Zhou et al., 2010). While significant learnings emerged through these studies, researchers faced numerous obstacles in generalizing the methodologies to go beyond lab settings. The advent of deep learning methods have enabled consumption of data from various sources. In particular, researchers in Oxford (Afouras et al., 2018a; Afouras et al., 2018c; Afouras et al., 2018b) have developed and annotated large real-world BBC and TED videos that helped construct complex deep neural networks for improved performance. Interestingly, state of the art deep neural networks outperform professional lip readers by 3-4X.

Inspired by the previous success of using posters and plot summaries, as discussed above, we propose a novel method to combine both posters and plot summaries. The rest of this paper describes the data, the proposed models, their performance and the challenges associated.

3 Datasets

Before the advent of deep learning for visual speech recognition, labeled datasets consisted of small sample sizes. Most of the datasets compiled contained few tens of speakers and were limited to alphabets, numbers or single words, as shown in Table 1.

Name	Output	class	Num. Speakers
AVICAR (**?**)	Digits	10	100
AVLetter	Alphabet	26	10
CUAVE	Digits	10	36
GRID*	Words	8.5	34
OuluVS1	Phrases	10	20
OuluVS2	Phrases	10	52

Table 1: Visual speech recognition datasets in the pre deep learning era.

Researchers from Oxford focused their efforts in developing deep neural networks for visual speech recognition. For this purpose, they compiled large datasets of labeled and segmented videos from real-world video clippings. These datasets also included a pretrain subcollection that contained timestamps of each word uttered in the video. The purpose of this pretrain dataset was to aid in curriculum learning, see Section 5.2. The

datasets compiled by this research group are shown in Table 2. In our work, we use the LRS2 dataset to train our networks.

Name	Output	Words	Num. Videos
LRW	Words	500	400,000
LRS2	Sentences	2,407,827	144,482
LRS3	Sentences	4,268,000	151,819

Table 2: Visual speech recognition datasets in the deep learning era.

The LRS2 dataset was divided into four subgroups; pretrain, train, validation and test. The utterances in the pre-train set correspond to part-sentences as well as multiple sentences, whereas the training set only consists of single full sentences or phrases. There is some overlap between the pre-training and the training sets.

Split	Utterances	Words	Vocab
Pretrain	96,318	2,064,118	41,427
Train	45,839	329,180	17,660
Validation	1,082	7,866	1,984
Test	1,243	6,663	1,698

Table 3: Split of the LRS2 dataset into pretrain, train, validation and test groups.

4 Architecture

4.1 Neural Architecture

In this section, we describe the deep neural network architecture for visual speech recognition. The architecture is similar to the one proposed in (Afouras et al., 2018b), where we restrict the model to video inputs. Since our use case will not include audio inputs, as described in Section 1, we disregard the audio streams within the videos. The model architecture is split into two parts; an encoder network and a decoder network, described below. The overall network architecture is illustrated in Figure 3.

4.1.1 Encoder

The encoder network consists of two parts; a) a visual front end model into which video frames are input and b) a multi-head self attention module that further encodes the dependencies across video frames of the input sequence.

We first extract visual features from the video frame input sequence using a 3D ResNet network similar to the one used in (**?**). The input video frames input at 224×224 resolution, 25 frames per sec, are grayscaled and centrally cropped to a 112×112 region surrounding the lip. In this work, we did not use any face detectors to identify the face locations for frame cropping but rather relied on the user to be centrally located within the video. This is a small shortcoming that can

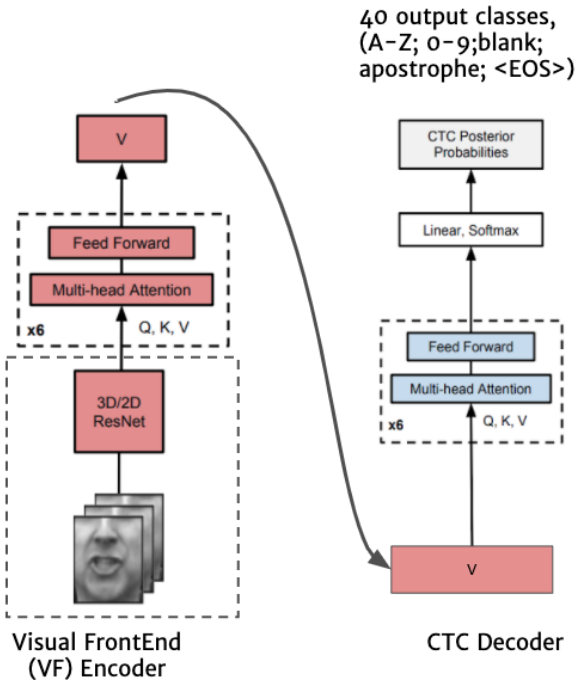

Figure 3: End-to-end deep learning network for visual speech recognition.

be easily addressed in the future. For any video input sequence of $T \times H \times W$ frames, the output is a $T \times H/32 \times W/32 \times 512$ tensor, where the temporal resolution is preserved. Average pooling is performed on this tensor resulting in a compact $T \times 512$ dimensional feature vector for the input sequence.

The second part of the encoder network is the multi-head self attention model, where the input tensor serves as the query, key and value for the attention at the same time, as shown in Figure 4. The information about the sequence order of the inputs is fed to the model via fixed positional embeddings in the form of sinusoid functions (). The multi-head attention block (Vaswani et al., 2017) receives a query (Q), a key (K) and a value (V) tensor as inputs and produces h context vectors, one for every attention head i:

$$Att_i(Q,\ K,\ V) = softmax\left(\frac{(W_i^q Q^T)^T (W_i^k K^T)}{\sqrt{d_k}}\right) \cdot (W_i^v V^T)^T \quad (1)$$

where Q, K, and V have size d_{model} and $d_k = \frac{d_{model}}{h}$ is the size of every attention head. The h context vectors are concatenated and propagated through a feed-forward block that consists of two linear layers with ReLU non-linearities in between. For the self attention layers it is always $Q = K = V$. In this work we use $d_{model} = 512$ and $h = 8$ attention heads everywhere. The sizes of the two linear layers in the feed-forward block are F1 = 2048, F2 = 512.

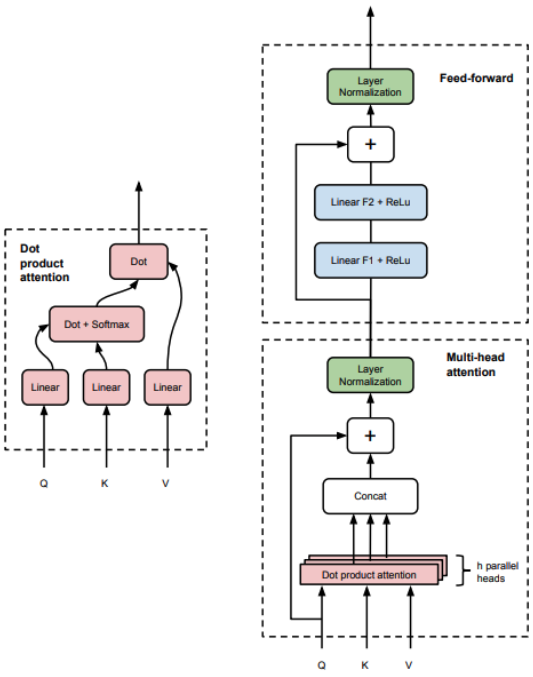

Figure 4: Self attention encoder network (). The illustration on the left is a detailed view of the Dot product attention block within the multi-head attention module.

4.1.2 Decoder

The decoder network takes in the encoded vector and is pushed through a multi-head attention network similar to that described in Section 4.1.1. A softmax layer, as shown in Figure 3, is attached to the multi-head attention layer which computes the character probabilities.

4.2 Network Architecture

This section describes the end to end network architecture as shown below in Figure 5. Video was captured using USB webcam attached to Nvidia Jetson TX2. Video was captured using OpenCV and forwarded to the instance on Cloud (P100 GPU) using MQTT broker. On the Cloud, video was received by inference engine. We chose to perform inference on the cloud instead of on the edge due to model latency on Jetson TX2. On the Cloud, inference engine loads the model at startup and subscribe to an incoming messages (videos) from MQTT broker. Lip Reading is performed in real time and output of the inference is displayed on the screen. Video could also be saved on the Object Storage for future reference.

5 Model Training

5.1 Curriculum Learning: Pretraining

As discussed in Section 3, a lot of data is gathered for pretraining the model. We adopt curriculum-based learning to help with the the overall training of the deep neural network. The pretrain dataset contains word endings that help correlate the sequences of visemes

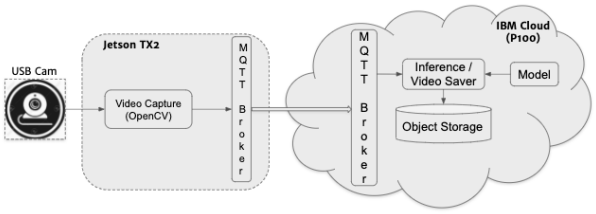

Figure 5: End To End Network Architecture

to the individual words within the sentences. For example, an annotated video pretrain file is shown below,

```
Text:  OH YEAH WASN'T A BIG DEAL
WORD START END ASDSCORE
OH 0.08 0.25 4.9
YEAH 0.25 0.60 7.6
WASN'T 1.03 1.46 4.5
A 1.46 1.54 6.1
BIG 1.54 1.98 6.5
DEAL 2.13 2.28 9.3
```

In this example, we can construct video frame sequences for each word as well as pairs or triplets of words such as "YEAH WASN'T A" or "A BIG DEAL" and so on. In curriculum learning, the deep neural network is trained sequentially on increasing word count. We first pretrain the network one word-at-a-time and use WER scores of the validation sets to conduct early-stopping of the pretrain. The 1-word pretrained model is then used as an initial checkpoint for pretraining two words-at-a-time. This is carried on for 11 separate pretrains that include $(1, 2, 3, 5, 7, 9, 13, 17, 21, 29, 37)$ words-at-a-time. The final pretrained model is then passed on to the training phase.

5.2 Training

In the training phase, the network is fed full sentences of varying lengths. The network is initialized using the pretrained weights from the aforementioned curriculum learning exercise. Separate validation and test sets are used to evaluate the generalized nature of the deep neural network model.

In all, the entire pretraining/training exercise took 11 calendar days to complete. Most of the time was spent in the pretrain phase due to the sequential nature of curriculum training. The training was carried out on P100 IBM cloud GPU virtual machines.

6 Results

In this section, we discuss the accuracy metrics, loss functions and the inferences we gathered from the end-to-end lip read pipeline.

6.1 Accuracy Metrics and Loss Functions

The accuracy of the model is evaluated using the Word Error Rate (WER). We define WER as,

$$WER = \frac{S + D + I}{N} \tag{2}$$

where S, D and I are the number of substitutions, deletions, and insertions needed to match the prediction to the ground truth labels.

We use the Connectionist Temporal Classification (CTC) loss gradient to compute the output character probabilities that are obtained from the final softmax layer. The maximum likelihood principle is used by carrying out gradient descents on the log likelihoods of the target character labels. More details of the CTC loss function can be found in (Graves et al., 2006).

In this study, the trained network resulted in a WER of 57%. Other implementations of the same methodology along with beam search on output characters have yielded WER's of 49%.

6.2 Inference

We observed that model performed better on professional lip reader's video than the authors visual speech samples.

Following table displays inference results

Speaker	Ground Truth	Prediction
Professional	When I make my pastry	When I make my pantrick
Author 2	Morning	More ing
Author 1	It's that simple	It's that simple
Author 3	How are you	On and are you
Author 1	Morning	It's boring
Author 2	It's a wonderful day	It's work of art the

Table 4: Inference Results depicting speaker, what was said and how it was predicted by the mode

We observed challenges with inference. As seen in Figure **??** below, Phonemes and visemes do not share a one-to-one correspondence; often, several phonemes share the same viseme. In other words, several phonemes look the same on the face when produced. This makes it harder to interpret lip movement and make correct prediction.

For example, as seen in the figure 6 below, 'm','b', 'p' and 'x' has the same viseme. Similarly 'TH' and 'DH' can also get confusing which can be observed in the table below where sentence ending with 'day' was predicted to end with 'the'

There are the basic sounds that need visemes to do lip-sync.

- M, B, P
- EEE (long "E" sound as in "Sheep")
- Err (As in "Earth", "Fur", Long-er" - also covers phonemes like "H" in "Hello")
- Eye, Ay (As in "Fly" and "Stay")
- i (Short "I" as in "it", "Fit")
- Oh (As in "Slow")

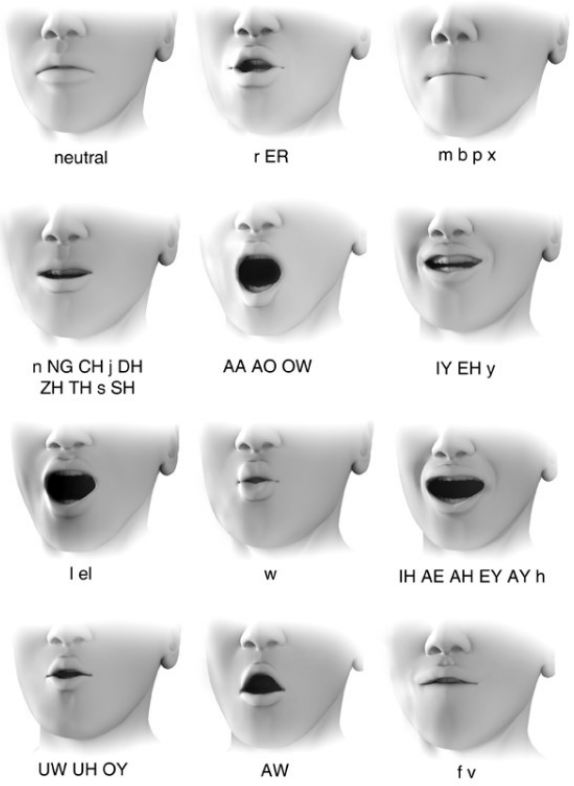

Figure 6: Phonemes and corresponding Visemes (O'Neill, 2010)

- OOO, W (As in "Moo" and "Went")
- Y, Ch, J ("You, Chew, Jalopy")
- F, V

As explained above, 'm' and 'b' phoneme can get confusing which can be seen in the results table 4 above.'Morning' got predicted as 'its boring'

As the training was performed on BBC video dataset (LRS2), we believe this may have introduced bias towards lip movement associated with British accent. Predictions might be different if the dataset included videos of people from different background/accent. Also, the speed of the speech can impact how model would predict the outcome. Some cultures tend to speak faster than others. Getting model to perfect the predictions would require personalized training on users. some user training may also be required so as to interpret the lip movement correctly.

We also observed challenges with predicting long sentences and is something that can be explored further.

7 Conclusion

In this study, we demonstrated the viability of using a lip-reading recognition using computer vision on the edge device. The trained network resulted in a WER of 57% which is comparable with state-of-the-art visual

speech recognition models. The model performed better on a professional lip reader when compared against authors' visual speech samples. However the model performed poorly on long sentences. Adding multimodality, such as audio, will help improve model performance. Another possible improvement to model accuracy can be achieved through personalization, where the models can be fine tuned with user's visual speech samples. Finally, addition of a language model that corrects sentences and implements spell checks will also improve model performance.

Acknowledgements

We would like to thank our instructors Dima Rekesh and Esteban Arias for providing valuable guidance.

References

T. Afouras, J. S. Chung, and A. Zisserman. 2018a. Deep lip reading: a comparison of models and an online application. In *INTERSPEECH*.

Triantafyllos Afouras, Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. 2018b. Deep audio-visual speech recognition. *CoRR*, abs/1809.02108.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018c. LRS3-TED: a large-scale dataset for visual speech recognition. *CoRR*, abs/1809.00496.

NIDCD Information Clearinghouse. 2017. Voice, speech, language. *Digital Art*. https://www.nidcd.nih.gov/health/what-is-voice-speech-language.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. Alterego: A personalized wearable silent speech interface. *Digital Art*. https://dl.acm.org/doi/pdf/10.1145/3172944.3172977.

Rachel Kaye, Christopher G Tang, and Catherine F Sinclair. 2017. The electrolarynx: voice restoration after total laryngectomy. *Digital Art*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5484568/pdf/mder-10-133.pdf.

Joshua C Kline. 2018. Development of semg sensors and algorithms for silent speech recognition. *Digital Art*. https://iopscience.iop.org/article/10.1088/1741-2552/aac965/ampdf.

Rob O'Neill. 2010. 3d character animation. *Digital Art*. http://www.morphometric.com/class/charanim/coursenotes/week12.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Howard Zhou, Tucker Hermans, Asmita V. Karandikar, and James M. Rehg. 2010. Movie genre classification via scene categorization. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 747–750, New York, NY, USA. Association for Computing Machinery.