Visual Speech Recognition (Lip Reading)

Gitansh Gera[1], Himanshi Jindgar[2], Saanvi Jain[3], Sachin Parashar[4],
Mr. Amit Kumar Pandey[5]
[1,2,3,4] Students, [5]Assistant Professor,
Department of Information Technology
[1, 2,3,4,5] Dr. Akhilesh Das Gupta Institute of Technology & Management

ABSTRACT

Lip Reading is a task of understanding what a person is saying by looking at the movement of his/her lips. There are approximately 10 million people with hearing loss and not everyone can read signs and even sign language as suggested by studies is 10-15 % efficient. Therefore in many places and events accommodation for the deaf and hard hearing person is not provided. In this paper, we are going to discuss the model for lip reading that is based on assistive augmentation. Assistive augmentation helps allow these people to work around their challenges and allow seamless and enhanced interaction. For this purpose, we are using a dataset of diverse videos to have different dialects and ways of speaking covered and then we have extracted the videos and trained them for our model.
Keyword: Visual speech, Lip reading, CNN, Ai, Information Technology

INTRODUCTION

Lip reading is a subcategory of human action recognition and it has become a popular topic recently, being used in various applications such as interaction with deaf people, intelligence services, detection of swearing people in a football stadium.

Lipreading is a very difficult task. Most things that cause lipreading besides the tongue and teeth are existing but are not very well developed and are difficult to remove its ambiguity without context. The human ability to do lip reading is not up to the mark with studies that conclude that people with hearing difficulties can achieve accuracy up to around 20%. This makes it important to realize that we need an automated model of lip-reading trained in AI.

The most important aspect to consider when it comes down to lip reading is the shape or the appearance of the picture made while talking is the key to understanding the letters of the words that the person is going to speak. Lip Segmentation and Detection becomes the prominent step and one full of challenges too. But sometimes in case of poor choice of lights, red blops in the clothing of the speaker, it won't provide us with desired results.

Lip reading has enormous potential in the future and it can help voice recognition in a noisy public or crowded places. It can also help in enhancing security with help of the biometric applications. It can also aid in reducing human effort in getting an education about lip language and different strata of lip movement and also in improvement of the ability of a human to improve the ability to recognize different kinds of speech. With the advancement of technology in all of the different strata of life the automatic lip-reading system can also help us making an

assistive driving system with lip-reading supported GPS and functioning. It also reduces redundancy in information and complements the information passed by speech.

In this project, we have used lipnet which understands the speech by visually interpreting the movements of the lips, tongue, and face when the normal sound is out of bound. It is fairly efficient because it relies on the information gathered by the context of speech and layman knowledge of the diverse language.

Related Works

In the following section, We are going to outline the existing work that has already been done in this field.

The early works in lipreading credit Yuhas et al. (1989) which recognized acoustic spectra and static images for vowel recognition. Pentland And MAse(1989) made use of an estimation technique where optical flow methods were used to estimate the motion of our four lip regions (without using any acoustic system). Bregler et al.(1993)instead used direct pixels instead of the motion of lips. Whereas threshold pixel-based representation of speakers was used by Bischoff and Bodoff (1988). Goldschen et al. (1997) notably were the first to do visual-only sentence-level lipreading using hidden Markov models (HMMs) in a very limited dataset, and with the use of hand-segmented phones.

Most approaches to lipreading have involved machine learning and do not employ deep learning. It has only a recent trend where we have seen deep learning methods have emerged in the field of lip reading.

The usage of deep learning in audio-video speech recognition was prominent in the works of (Srivastava and Salakhutdinov, 2012, Huang and Kingsbury, 2013). In the studies conducted by them different deep architecture lip-reading systems were reviewed and the rate of their speaking dependency was highlighted.

Most of the approaches of deep learning mirror early progress while applying neural networks for acoustic processes in speech recognition(Hinton et al,2012). An audiovisual max-margin matching model was trained by Chung and Zisserman (2016b) which was for learning pre-trained model features, which was used as input for a 10 phase classification on the OuluVS2 dataset. LSTM was introduced by Wand Et al(2016) for recurrent neural networks for lipreading but address neither sentence-level sequence prediction nor speaker independence.

The best viewing angle for automated lip-reading was studied under Lan et al(2012b), it used a purpose based audio-video database which happens to have multi-camera recordings of a speaker who recites 200 sentences with an enormous vocabulary size of 1000 words. A convolutional neural network (CNN) acts as the feature extraction block for a lip-reading system. For training purposes, the speaker's mouth area images along with few phoneme labels are used upon. It is done with 6 speakers each of which is modeled in accordance with an independent CNN in itself. The database comprises 300 Japanese words.

Methodology

LIP NET

Lipnet is known to be the first end-to-end sentence level lip reading model. It is a neural network for lip reading that is trained end-to-end by mapping variable-length sequences of video frames to text sequences. It operates at the character level with providing 88.6% accuracy to unseen speakers and uses the concept of convolutional neural network(CNN), Gated Recurrent Unit (GRU) And Connectionist Temporal Classification (CTC). The model attends only the phonologically important areas in the videos only due to the use of saliency visualization techniques.

CNN

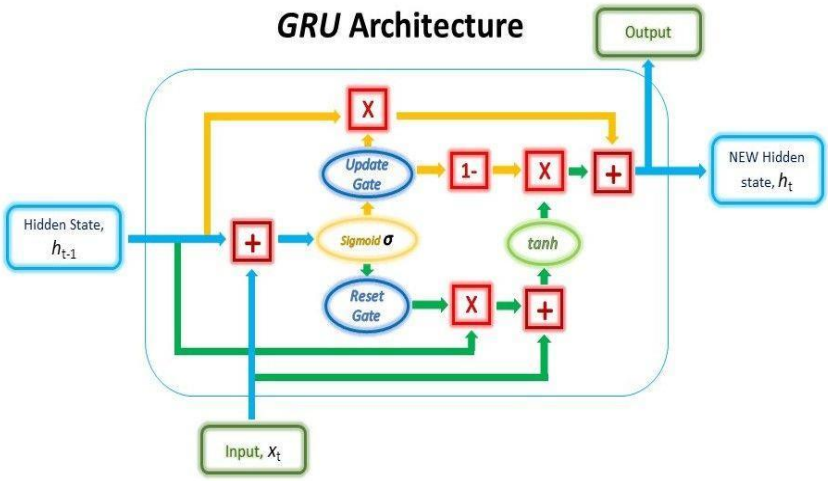
CNN is related to much known STCNN. CNN can process the given video by convolving at the spatial dimensions as well as time.

ARCHITECTURE OF CNN

CNN Contains stacked convulsions operating spatially over an image, it has been prominent in advancing performances in computer vision tasks such as object recognition.

GRU

The most effective variation that is used to solve vanishing exploding gradients problems that is often encountered during the operation of basic Recurrent Neural Network is known as Gated Recurrent Network(GRU). It improves previous performances of Earlier RNNs by adding more cells and gates so that information.
Can propagate at more time steps and flow could be controlled.

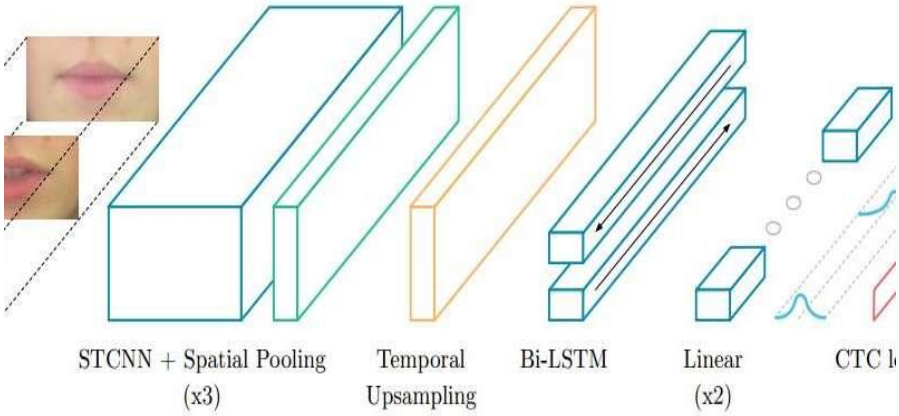


CTC

A CTC's main application is to eliminate the need of training data that aligns inputs to target outputs(Amodei et al., 2015; Graves & Jaitly, 2014; Maas et al., 2015). CTC, by computing the probability of the sequences, eliminates the need to address variable length sequences and the need of alignments.

Lip Net Architecture

The lipnet architecture consists of 3 CNNs, 2 GRUs, linear transformation which is applied at each time-step, the importance of GRU is such that it makes output of CNN more effective and efficient and the system is followed by SoftMax over vocabulary augmented with CTC. Then CTC loss. For the activation function layers use rectified linear unit (ReLu).



Architecture of Lip Net

Assistive Augmentation

Assistive augmentation is perception of Artificial intelligence which focuses on AI's supportive role i.e AI's role is not to replace human intelligence rather enhance it.

In lip reading, the concept of assistive augmentation can be used to create a model which can easily be able to understand the context of a given argument or discussion. The model will utilize human intelligence and its previous experiences of speaking to derive context out of

the sentences when two or more similar sounding words appear which can confuse the model such that a machine would identify the words Ex- 'meet her' and 'meteor' in similar fashion but the context the style in which these words are used always differ so in this situation augmented intelligence is used in order to derive relevance out of the sentence

Future Scope

The future efforts on our project would be refining our speech recognition model by working on the larger data sets in various languages and different speaking styles. Another feature that we would like to work upon in future is to recognize the silent dictation and to analyze it for our model.We want to clup our model with car gps system and other upcoming technologies based on speech.The difficulties faced by the people will drastically come down with the model especially in noisy environments.

Conclusions

In this paper we discussed about a model which helps us to detect what a person is trying to say by visually interpreting the movement of the person's lips and in order to understand the context it uses assistive AI (i.e emphasizing the role of AI as a supporter) by using diverse languages and context.The model used is end-to-end and is trained in deep learning the end-to-end feature removes the need of words that is essential to predict a sentence.Our results have indicated that this technology can easily enhance the ability of speech recognition.

REFERENCES

[1] https://www.lipreading.org/
[2] https://en.wikipedia.org/wiki/Lip_reading#Phonemes_and_visemes
[3] E. D. Petajan, B. Bischoff & D. Bodoff. (1988) An improved automatic lipreading system to enhance speech recognition. ACM SIGCHI-88, 19-25.
[4] B. P. Yuhas, M. H. Goldstein, Jr., T. J. Sejnowski & R. E. Jenkins. (1988) Neural network models of sensory integration for improved vowel recognition. Proc. IEEE 78(10), 1658- 1668.
[5] A. Pentland & K. Mase (1989) Lip reading: Automatic visual recognition of spoken words. Proc. Image Understanding and Machine Vision, Optical Society of America, June 12-14.
[6] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.
[7] C. Bregler, S. Manke, H. Hild & A. Waibel. (1993) Bimodal Sensor Integration on the example of "Speech-Reading". Proc. ICNN-93, Vol. II 667-677.
[8] J. Goldschen, O. N. Garcia, and E. D. Petajan. Continuous automatic speech recognition by lipreading. In Motion-Based recognition, pp. 321–343. Springer, 1997.
[9] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N.
[10] Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6):82–97, 2012.
[11] M. Wand, J. Koutnik, and J. Schmidhuber. Lipreading with long short-term memory. In IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6115–6119, 2016.
[12] Srivastava N, Salakhutdinov RR, (2014). Multimodal learning
with deep boltzmann machines. Advances in Neural
Information Processing Systems, 2222–30.
[13] Huang J, Kingsbury B (2013). Audio-visual deep learning for noise robust speech recognition. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7596–9.
[14] Lan Y, Harvey R, Theobald BJ (2012). Insights into machine lip reading. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4825–8

[15] Noda K, Yamaguchi Y, Nakadai K, Okuno HG, Ogata T (2014). Lipreading using convolutional neural networks. 15th Annual Conference of the International Speech Communication Association.
[16] Yannis M. Assael , Brendan Shillingford , Shimon Whiteson1 & Nando de Freitas (2016) LIPNET: END-TO-END SENTENCE-LEVEL LIPREADING
[17] Abiel Gutierrez Zoe-Alanah Robert :Lip Reading Word Classification
[18] Wolff, Prasad, Stork, and Hennecke:Lipreading by neural networks: Visual preprocessing, learning and sensory integration
[19] Hamza mirza ,Saiqa Khan (2017): implication and Utilization of various Lip Reading Techniques
[20] Fatemah Vakhshiteh , Farshad Almasganj,, Ahmad Nickabadi(2018)LIP-Reading Via Deep Neural Networks Using Hybrid Visual Features
[21] https://www.analyticsvidhya.com/blog/2020/02/mathematics-behind-convolutional-neural-network/
[22] https://blog.floydhub.com/gru-with-pytorch/
[23] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, et al. Deep Speech 2: End-to-end speech recognition in English and Mandarin. arXiv preprint arXiv:1512.02595, 2015.
[24] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In International Conference on Machine Learning, pp. 1764–1772, 2014.
[25] A. L. Maas, Z. Xie, D. Jurafsky, and A. Y. Ng. Lexicon-free conversational speech recognition with neural networks. In NAACL, 2015.
[26] Vyom Jain, Srishti Lamba, Shweta Airan : LipNet: A comparative study