Building a Listening Eye: Automated Lip Reading

Siddharth Raja
siddharth.raja@gatech.edu

John Turner
jturner65@gatech.edu

BASIC IDEA

Humans integrate visual input with the sounds heard to make speech more intelligible. Researchers in the past have integrated visual cues with audio input to improve speech recognition especially in settings where the audio quality is degraded [1, 2, 3]. On similar lines, our project intends to make a lip-reader that combines these modalities and predicts what a person in a video is saying solely by observing the individual's lip movements.

INTERESTING PREVIOUS WORK

Previous work in this area has largely been focused on using probabilistic graphical models namely HMMs to model the lip movements in the same way as temporal data. Matthws et al. [2] use HMMs to get about 44.6% accuracy on the task. 'Eigenlips' [3] used a hybrid approach of Multi-layered Perceptrons (MLP) and HMMs to get a good accuracy on a small dataset of around 2500 spoken letters. Potamianos et al. [1] mention that although progress on the audio-visual speech recognition has been made, but the systems are nowhere near as good as they should be for practical use.

A very interesting work related to this problem has been done by Ngiam et al. [4]. In their paper 'Multimodal Deep Learning', they employ layer-wise training with Restricted Boltzmann Machines (RBMs) followed by fine-tuning to learn features across an audio-visual training dataset. There is also work done by Hinton and Salakhutdinov to use deep learning to produce useful representations for handwritten digits and text.

OUR INTIAL ANALYSIS

We wish to take a multi-step approach to addressing this problem, using established techniques where appropriate to handle as much of the load as possible, and focusing our research on a novel solution to the difficult unsolved components of this task.

One such difficulty is the inherent under-constrained nature of the underlying task, namely that humans make fewer mouth motions than they make unique sounds – many configurations of the lips and mouth are visually similar while producing different sounds, due to the position of the tongue, for example, and which in turn are used to form entirely different words. While individual mouth "poses" might be difficult to assign to specific component sounds, our hypothesis is that combinations in sequences will have less uncertainty. Our aim is to build on the recent developments in the deep learning especially the deep auto-encoder based architectures [4] and also explore Long Short Term Memory (LSTMs), a kind of Recurrent Neural Networks which are known to have outperformed HMMs on tasks like simple speech recognition [5].

INTIAL STEPS

As a first step along these lines, we plan on accumulating high definition video clips where a speaker is on screen and audio is present. Ideally these clips will have the speaker always visible, oriented directly in front of the camera, and the audio will be clear. We are thinking that sequences from a news channel such as CNN might work for this. We will then isolate images of the mouths of the speakers in every frame of the video to use to train a classifier in conjunction with the audio captured from the video at those frame locations. We hope that this will give us a ground truth from which to explore various techniques for addressing the problem.

DATASETS (OR LACK THEREOF)

In our investigations, we have also noticed that, while there is interest in solving this problem, we could find no large data sets of matched video to audio of multiple different people speaking many words and phrases. The two most popular ones are the CUAVE dataset [6] (roughly 611 MB) and the AVLetters [7] both of which seem extremely small by current standards and also rather old.

One possible alternate project we are considering is the accumulation of a large dataset of videos of students speaking sentences and phrases, to help further the research in this area. We are, naturally, open to guidance and suggestions in this area.

REFERENCES

[1] Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. Proceedings of the IEEE, 91(9), 1306-1326.

[2] Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., & Harvey, R. (2002). Extraction of visual features for lipreading. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24(2), 198-213.

[3] Bregler, C., & Konig, Y. (1994, April). "Eigenlips" for robust speech recognition. In Acoustics, Speech,

and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on (Vol. 2, pp. II-669). IEEE.

[4] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In Proceedings of the 28th international conference on machine learning (ICML-11) (pp. 689-696).

[5] Graves, Alex; Mohamed, Abdel-rahman; Hinton, Geoffrey (2013). "Speech Recognition with Deep Recurrent Neural Networks". Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on: 6645–6649.

[6] Patterson, E. K., Gurbuz, S., Tufekci, Z., & Gowdy, J. N. (2002, May). CUAVE: A new audio-visual database for multimodal human-computer interface research. In Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on (Vol. 2, pp. II-2017). IEEE.

[7] http://www.ee.surrey.ac.uk/Projects/LILiR/datasets/avletters1/index.html