Pay Attention, Please: Attention at the Telluride Neuromorphic Cognition Workshop

Malcolm Slaney

SLTC Newsletter, November 2012

The role of attention is growing in importance as speech recognition moves into more challenging environments. Attention is not a factor when we speak into a close-talking microphone with a push-to-talk button. But the acoustic world in which we live is not so simple; multiple sources add together in a confusing mixture of sound, making it difficult to analyze any one source. In such settings, attention serves two important purposes: 1) It helps us to focus our computational resources on the problem at hand, and 2) it helps us to piece together portions of our environment to solve a single task. Understanding how attention works may be the critical advance that allows us to build speech-recognition machines that cope with complex, everyday settings robustly and flexibly, just as a human listener does.

My interest in attention stems from a realization that we cannot solve the speech recognition problem using a purely bottom-up engineering approach. As an engineer I am trained to build a system starting with fundamental principles, a solid foundation. This tendency by the entire field leads to models where all the information flow is directed in one direction, from the ears or microphone to the final linguistic output. This approach has served us well, so far....

Yet, our acoustic world is too complicated for such a simple approach. Many years ago I attended a conference on binaural hearing. Most of the world's experts on binaural perception were in attendance and there was much serious discussion about how we localize sounds and the cues necessary to make this work [1]. But whenever the discussion got difficult, somebody would stand up and say "We can't do that, it's the c-word!!!" The c-word was either cognition or cortex, which at the time seemed like an impossibly hard problem to tackle. We no longer have this luxury. We must understand the role cognition plays in speech understanding, if for no other reason than our machines must "play" in this world. Attention is one important part of this puzzle.

I've had the privilege to help lead three recent projects on attention at the Neuromorphic Cognition Workshop [2]. These projects have studied different parts of attention in a short, focused, working workshop. During the summer of 2011 we studied and built a complete cocktail-party system, with both top-down and bottom-up signals. In the summer of 2012 we used EEG signals to "listen" to a subject's brain and decode which of two speech signals he was attending. I will describe the Neuromorphic Engineering Workshop, and then the two latest attention projects.

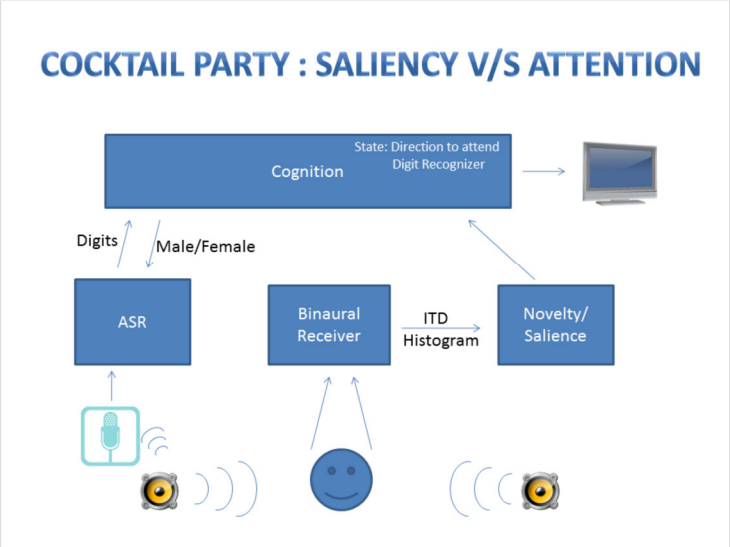
WHAT IS NEUROMORPHIC COGNITION?

The Telluride Neuromorphic Cognition Workshop brings together engineers, biologists, and neurophysiologists to study problems related to human cognition. It is held for three weeks every summer in the western Colorado mountain town of Telluride. It is a real working workshop, in the sense that people bring their VLSI chips, oscilloscopes, prototype robots, and measurement equipment and do real work. In past years we have had scientists measuring the neural spikes from

the visual cortex of a (locally caught) dragonfly; sitting next to the technologist that designed the Robosapien [3] line of walking robotic toys; sitting next to modelers that are building models of learning in cortical systems, using both computational and VLSI models.

The biggest strength of the workshop is bringing a range of faculty and young researchers together, from many different institutions around the world, to actually work on problems of joint interest. There are usually about 30 faculty members presenting their research and leading projects. About 40 young researchers, Ph.D. students and post-docs, are in Telluride to learn and work. Some of the projects are related to their area of expertise, while other projects are not. A typical day starts with lectures and discussions in the morning, and project meetings and work from noon till late at night. The small town and surrounding mountains give us enough room to explore, but keep us close enough so we are always working.

A COCKTAIL PARTY



Our project during the summer of 2011 aimed to model the cocktail party effect by building a real-time system for understanding speech with two speakers [4]. A key aspect of this problem is having a way to describe and build both bottom-up and top-down signal-processing pathways.

The bottom-up pathway is easy. Our current recognizers look at an acoustic signal, and under the constraint of a language model, recognize the speech that is heard. Creating a useful top-down pathway is harder. What does our cortex tell the speech recognizer to make it easier to understand speech and reach our goals?

We built a real-time cocktail party analyzer by combining bottom-up processing for speech recognition and binaural localization, with a "cognitive" model that directs attention to the most important signal. In a real cocktail party we shift our attention, sampling different audio streams, looking for the signal with the highest "information" content. With a tip of the hat to Claude Shannon, information content is difficult to define in this scenario. You might be nodding pleasantly as the person in front of you talks about their vacation to Kansas, while listening to the woman behind you talk about where to go for dinner after the party. Both are forms of information.

Both exogenous and endogenous processes are important to a model of attention. Exogeneous (or external) signals redirect our attention when we hear new sounds. We are especially sensitive to such signals, perhaps because sudden auditory onsets warned us of approaching predators. Now exogenous signals (bottom-up) are used to alert us to new speakers and to higher-level information such as changes in topics.

Modeling endogenous (or internal) processes is harder. Endogenous signals come from within, and are goal directed. I really need to listen to Bob, but if Susan starts talking about me I'll listen to that instead. This is where the informational goals of the listener meet the road. Do I keep listening to Bob drone on, or do I switch to Susan? Our "cognitive" model embodied this decision process.

In the real world, we use lots of information to judge the potential value of a signal. We know the past history of a speaker and his or her value to us. We know that some subjects are more interesting than others. And our priorities change with time. We operate in an information-foraging mode [5], sampling different parts of the cocktail party, as we learn and change from our environment.

In our Telluride model, we replaced the dirty real world with two-digit numerical strings from two different speakers. This made it trivial to assign information content to the signal---the value of the signal is the numerical value of the two-digit string. Our cocktail party analyzer's problem was then to choose the speaker that is most likely to give the highest value digit strings. We had a binaural

cochlear model, implemented in VLSI, and we used the output from this chip, along with a simple energy-change detector to prime exogeneous attention. When new energy is detected from some direction in the sound field, we sent a signal to the cognition model to get its attention. The cognition model also had access to the output from a speech recognizer that was listening to just one direction at a time. The cognition model's goal was to direct the speech-recognition system to listen to the one stream, over time, that would maximize thereceived information. In our case, that was the sum of the values of the correctly received two-digit sentences.

To demonstrate the value of a cognitive model, even a primitive one, we implemented several different information-gathering strategies. In the simplest model that we labeled as "distracted," the cognition module switched the speech recognizer to a new direction every time a new speaker started speaking. The "smart" cognitive model received the same inputs, but only switched to the new speaker when the information content of the current speaker was judged to be low. This was based on the value of the first digit received. If the first digit of the sentence is less than 5, then the entire signal is likely to be low value, and we might more profitably use our resources to listen to the new direction.

This model is valuable because it demonstrates all the pieces of a real cocktail party analyzer in one place. While we "cheated" in many ways---for example using digit strings instead of semantically rich signals, or using a DTW-based recognizer instead of a full HMM---the system has all the right pieces, and demonstrates both exogenous and endogenous, both top-down and bottom-up processing [6]. We hope future cocktail-party analyzers will build upon this model of attention to do even better.

DECODING ATTENTION IN REAL TIME

Our project during the summer of 2012 aimed to decode human attention [7]. Attention is a latent signal. While your eyes might be looking at me, you are really thinking about lunch---we can't always tell from external signals. Thus, we want to measure a signal from a human brain, decode it, and then decide to which signal the subject is listening. From a scientific point of view, this will tell us more about how the auditory system decodes speech in a noisy environment. From a practical point of view, having a machine that knows what you are attending to will make it easier for the machine to complement human behavior.

An auditory attention-decoding experiment has been done using electrodes implanted over auditory cortex [8]. These patients are being monitored for seizures, but in the mean time some are available for experiments. Mesgarani and Chang use system-identification techniques to correlate the speech information with a cortical neuron's output. They invert this model to estimate the speech that generated a given spike train. At the locations from which they are recording spikes, they discovered that the neural responses best correspond to the speech signal to which the subject was told to attend.

We can do a lot of research in Telluride, but surgically implanting electrodes in human subjects is difficult. Instead, we used EEG (electroencephalography) to measure brain responses, and thus infer the direction of attention. We believe this is the first time this has been done in real time, on any kind of subject.

EEG is a very rough measurement technique compared to direct neural recordings. The received signal represents the combined response to hundreds of thousands of neurons over centimeters of space. But yet over time, and with sufficient averaging, EEG signals do contain information about the internal state of a subject.

We used a system-identification scheme to model the auditory to EEG pathway. Our subject first listened to audio, while we recorded 32 channels of EEG signals. We could then build models that correlate the attended audio and EEG signals, to create forward and the inverse models. In the online experiment we moved the subject to a different room and presented two speech signals to the subject, both diotically and dichotically. The subject was told to attend to one or the other signal, and then we decoded the resulting EEG signals. We could use the inverse system model to predict the incoming

audio, and then compare the prediction to the two known input speech signals. In the best case, we could determine the attended signal with 95% accuracy using a 60-second long decision window.

CONCLUSIONS

Telluride and the Neuromorphic Cognition Workshop are a wonderful venue to foster new collaborations on cutting edge topics related to human perception and cognition. Project proposals are submitted towards the end of the year, and a small number of them are chosen as focus areas for the coming summer. The project leaders choose their faculty. Student applications are accepted in the early part of the year, and about 40 students are selected and given housing assistance for the three-week workshop. I enjoy the collaboration and chance to meet and work with new people. But the experience is intense---I have pulled more all-nighters, preparing for the final Telluride demos, than I have for any other reason since graduate school. I hope to meet you in Telluride.

ACKNOWLEDGEMENTS

The attention work described in this report is due to the contributions of many, many participants, but I do want to thank my co-leaders: Shihab Shamma, Barbara Shinn-Cunningham, Ed Lalor, Adrian KC Lee, Mounya Elhilali and Julio Martinez-Trujillo. The project participants are listed in an appendix to this column. We appreciate the financial support of the NSF, ONR and EU ERC. We are grateful for the equipment support we received from BrainVision and software support from Mathworks.

APPENDIX

A large number of people contributed to these two projects. I would like to acknowledge their contributions, both big and small.

Adam McLeod, Adam O'Donovan, Adrian KC Lee, Andreas Andreou, Asha Gopinathan, Barbara Shinn-Cunningham, Bruce Bobier, Ching Teo, Claire Chamers, Connie Cheung, Daniel B. Fasnacht, Diana Sidtis, Dimitris Pinotsis, Edmund Lalor, Fabio Stefanini, Francisco Barranco, Fred Hamker, Inyong Choi, James Wright, Janelle Szary, Jeffrey Pompe, Jennifer Hasler, Jonathan Brumberg, Jonathan Tapson, Jongkil Park, Julio Martinez-Trujillo, Kailash Patil, Lakshmi Krishnan, Magdalena Kogutowska, Malcolm Slaney , Mathis Richter, Matthew Runchey, Mehdi Khamassi, Merve Kaya, Michael Pfeiffer, Mounya Elhilali, Nai Ding, Nils Peters, Nima Mesgarani, Nuno Vasconcelos, Ozlem Kalinli, Roi Kliper, Ryad Benjamin Benosman, Sahar Akram, Samuel Shapero, Shihab Shamma, Shih-Chii Liu, Siddharth Joshi, Siddharth Rajaram, Sudarshan Ramenahalli, Theodore Yu, Thomas Murray, Timmer Horiuchi, Tobi Delbruck, Tomas Figliolia, Trevor Agus, Troy Lau, Yan Wu, Yezhou Yang, Ying -Yee Kong, Yulia Sandamirskaya

REFERENCES

[1] Robert H. Gilkey and Timothy R. Anderson, eds. Binaural and Spatial Hearing in Real and Virtual Environments. Lawrence Erlbaum Associates, Mahwah, NJ, 1997.

[2] http://ine-web.org/workshops/workshops-overview/index.html

[3] http://en.wikipedia.org/wiki/RoboSapien

[4] http://neuromorphs.net/nm/wiki/2011/att11/AutoAttention

[5] P. Pirolli and S. Card. Information foraging. *Psychological Review,*106, 643-675, 1999.

[6] Malcolm Slaney, Trevor Agus, Shih-Chii Liu, Merve Kaya, Mounya Elhilali. "A Model of Attention-Driven Scene Analysis." *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* Kyoto, Japan, March, 2012.

[7] http://neuromorphs.net/nm/wiki/2012/att12/FinalReport

[8] Nima Mesgarani and Edward F. Chang. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature,* 485, pp. 233-236, 2012.

Malcolm Slaney is a Principal Researcher in the Conversational Systems Research Center at Microsoft Research, and a (Consulting) Professor at Stanford CCRMA. He is interested in all aspects of auditory perception. Email: malcolm@ieee.org

SPS Senior Members
My IEEE
Volunteer Opportunities
Membership Development
SPS Travel Grants

SP Cup

Resources
Volunteer Opportunities
Frequently Asked Questions
Contact