A MODEL OF ATTENTION-DRIVEN SCENE ANALYSIS

Malcolm Slaney[⋆], *Trevor Agus*[†], *Shih-Chii Liu*[‡], *Merve Kaya*[¶], *Mounya Elhilali*[¶]

[⋆]Yahoo! Research, malcolm@ieee.org; [†]CNRS & Paris Descartes & ENS, Trevor.Agus@ens.fr;
[‡]Institute of Neuroinformatics, Zurich, shih@ini.phys.ethz.ch;
[¶]Johns Hopkins University, {merve,mounya}@jhu.edu

ABSTRACT

Parsing complex acoustic scenes involves an intricate interplay between bottom-up, stimulus-driven salient elements in the scene with top-down, goal-directed, mechanisms that shift our attention to particular parts of the scene. Here, we present a framework for exploring the interaction between these two processes in a simulated cocktail party setting. The model shows improved digit recognition in a multi-talker environment with a goal of tracking the source uttering the highest value. This work highlights the relevance of both data-driven and goal-driven processes in tackling real multi-talker, multi-source sound analysis.

Index Terms— Attention, Saliency, Auditory Scene Analysis, Cognition, Digit Recognition

1. INTRODUCTION

This paper describes a model of attention-driven auditory scene analysis (ASA). ASA is the process of listening to a complicated auditory environment, a cocktail party is the canonical example, and being able to select and understand a single talker. Due to its significance in both perceptual and engineering sciences, interest in tackling the ASA phenomenon has prompted multidisciplinary efforts spanning the engineering, artificial intelligence and neuroscience communities. In general, most current work on auditory scene analysis takes one of two simplified approaches. The first approaches to computational auditory scene analysis (CASA) use an exclusively bottom-up approach. Low-level perceptual signals are grouped using simple rules such as common onsets or modulations [1]. These systems rely heavily on the conspicuity and salience of stimulus elements; and perform reasonably well in simple and well controlled scene analysis conditions. More recent systems take a more sophisticated approach by including expectations in the analysis. These systems have simple models of what a talker sounds like [2], or what was said before [3]. In this paper we describe a third approach based on a user's goals.

We are interested in saliency and attention because of the interplay between bottom-up and top-down processes. We define saliency to be a bottom-up signal that tells the brain that something novel or interesting has happened. Its role is to alert the subject, perhaps shifting the focus of attention. Top-down attention, on the other hand, is a goal-directed signal based on the desires, needs, and limitations of the animal.

Fig. 1. The goal of the attention-driven scene analysis system is to recognize the highest value sound.

The role of attention in auditory scene analysis is unsettled. It is certainly true that attention can affect stream segregation. For instance, the ability to switch at will between hearing certain tone sequences as one or two streams can be thought of as an effect of attention, but that leaves the question of whether attention is necessary for streaming (e.g., [4]). The bulk of the literature suggests that at least some forms of streaming occur in the absence of top-down attention, in what is termed "primitive" stream segregation [5], and bottom-up saliency may play a role here. Streaming may also be thought of as a process that facilitates attention (rather than only vice versa) in that it only becomes possible to pay exclusive attention to tones of a single frequency if they are successfully segregated from other tones in the sequence. In the case of alternating tone sequences, van Noorden [6] defines two boundaries, the fission boundary and the temporal coherence boundary. The fission boundary defines the frequency difference (or other dimension) below which segregation is not possible; the temporal coherence boundary defines the point above which integration is not possible. The area in between these two boundaries could be the region in which attention plays a role in determining whether we hear one or two streams. These kinds of models have also been studied in vision [7].

We postulate that attention has two purposes: selection and efficiency. We want to attend to a single talker because he is telling us something that we need to know—this is selection. Furthermore, attention is important because our brains have limited computational ability and (ignoring divided at-

tention) we can only process the speech from one talker at a time. We are not arguing, however, that computers should be limited in the same way. It might be desirable to use a computer to simultaneously separate and decode all talkers [8]. But it is important to note the different approaches.

This paper describes in Section 2 our theory for attention-driven auditory scene analysis. Section 3 describes the model we built and we describe its performance in Section 4. As this is an initial attempt to build a model of attention and saliency, we emphasize open questions in Section 5.

2. THEORY AND MODEL

In everyday acoustic scenes, some sounds are important, but most are not. In general the information content of a signal is hard to judge. There are many different factors, most of which we can't hope to understand or model. Here, we postulate a goal of an organism is to select and understand the highest-value audio in a complex auditory environment. Figure 1 describes a simple scenario explored in this study. Here a subject listens to speech signals from two different talkers, a male and a female at different positions. Our goal is not to model the semantic content in natural language, but rather to explore the role of bottom-up saliency and top-down goals in mediating scene analysis. Our talkers utter simple two-digit numeric "sentences" that have a direct semantic meaning (98, the higher value, is more important than 32). Our goal is to separate and understand the highest value sentence over time. We use saliency to tell us when something interesting is happening, perhaps new sound from a different talker. It is up to a "cognitive" layer to decide, based on expectations, whether we should shift our attention to this new signal.

In our simulated cocktail party a male and a female talker speak two-digit sentences. The ten possible words from each talker were drawn from the TIMIT database and do not change during the experiment. Each sentence ends with an even digit so that we can more easily parse the sentence structure. The two streams of digits overlap approximately 59% of the time. Played binaurally, human subjects with native English skills can attend to either talker and perfectly understand either talker, albeit with some difficulty. Played monaurally, the task is very difficult, if not impossible.

In this paper we demonstrate an attention-driven model and measure its performance using several different cognitive strategies. One simple strategy is to always pay attention to a single talker. A second possibility is to always shift attention to the new talker whenever a salient event occurs—we call this the distracted model. Finally, the best approach we describe is a "smart" approach which looks at the speech received so far, and then judges whether the new talker is likely to be more valuable, by virtue of giving us a higher-value sentence.

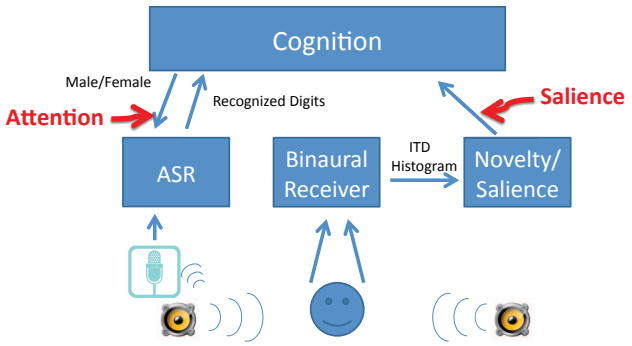

Fig. 2. A block diagram of the attention-driven scene analysis system.

3. IMPLEMENTATION

We built the system shown in Fig. 2. It has five primary components: saliency, cognition, attention, separation and recognition. The system is fully developed, although there are many simplifications, because we want to emphasize the attention/saliency tradeoff in CASA.

Saliency is defined as something that is noticeable or important. There are models of auditory saliency [9, 10] but they are largely straightforward transformations of visual saliency models to audition. The visual saliency models have been successful because they have been validated with eye-tracking data, which is a good indication of attention. Unfortunately, there is no similar indication of auditory attention. While the auditory salience models have been validated with simple sounds (tones or speech) neither was sufficient for the two-talker situation we are testing. Thus we implemented a very simple model of binaural saliency.

Our binaural saliency model is based on a VLSI spike-based implementation of cochlear interaural-time delay (ITD) processing [11]. We play the male and female speech through two different speakers separated by 53 cm. The baseline of the cochlear board's two microphones is 25.5cm. from the speaker baseline, giving a 90 degree difference between the two talkers. We analyze the sound from each analog-digital VLSI cochlea into 64 bandpass channels. Neural spikes for each channel, c, and from each ear are generated by the VLSI system and sent to a computer for analysis. These spikes are cross-correlated and summed across channels, in software, as a function of relative time delay over time, t, to obtain the ITD signal, $R(\tau, t)$. This ITD signal tells us from which directions we are receiving sound energy. We turn this into a binaural saliency signal by taking the temporal derivative of the cross-correlation: $S(\tau, t) = \partial R(\tau, t)/\partial t$. Peaks in this saliency signal represent binaural onsets, which is a simple representation of salience.

Results from our binaural salience model are shown in the three images of Fig. 3. The top image shows the ITD signal for 16 different directions (or time delays) over time. There

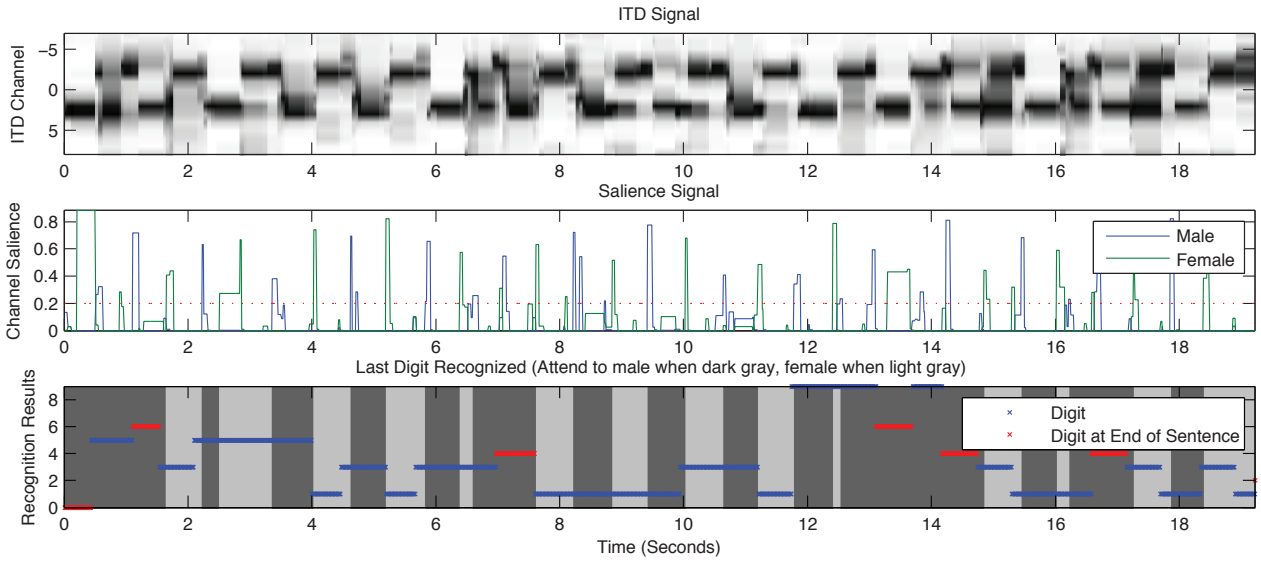

Fig. 3. ITD signals, salience and digit recognition results for one 20-second trial using the "smart" cognition model. Each bar in the bottom graph shows the last recognized digit; the start of the bar represents the end of the spoken digit.

is significant overlap between the talkers. The middle plot shows the output of the binaural saliency model, for just the two most prominent ITD channels. Note there is a salience peak at the start of each digit. Finally the bottom plot summarizes the recognition and attention results from a single trial.

Cognition is a difficult thing to model. As we are only interested in the tradeoff between saliency and attention, we used simple cognitive models to illustrate the concepts. Thus we measured the performance of single-talker, distracted, and smart approaches. Our "smart" approach is still relatively simple. If the sentence received so far is likely to be a low value, we switch as soon as a salient event happens in the opposite channel. This judgement is based on the digits recognized so far. If the first digit of the current sentence is five or higher we make the bet that the sentence from the current talker is likely to be good and we want to keep our attention on the current talker. Conversely, if we have not received anything or we have only recognized a low digit at the start of the sentence we make a bet and switch.

Our top-down attention model is simple. We attend to only one talker at a time, either the male on the left, or the female on the right. Humans have some ability to divide their attention, although dual task experiments often show there are bottlenecks that limit overall performance. In this study we do not attempt to model divided attention. Our subject listens to one talker until told to switch to the other.

Our original goal was to use the ITD signal to select the active binaural channels and then use simple beam forming to select the appropriate audio. Unfortunately this approach is only going to give us at best a 3 dB advantage, when the two signals are constructively added. Thus in this paper we bypass this problem by feeding the recognition stage directly with either the left or the right audio signal and the recognition system gets perfect audio on which to perform its actions.

Our speech recognition system is based on speaker- and digit-dependent template matching. An audio signal from the separation system is analyzed with MFCC and stored. At each frame (100 Hz) we measure the Euclidean distance between the most recently calculated MFCC coefficients and precomputed models of each talker's digits. We recognize a digit when the error (normalized by the length of the target utterance) is lower than a threshold. Recognition of either talker, under these idealized circumstances, is perfect, except when the attention switches in the middle of a digit.

4. TESTING

We tested our system with 200 trials, each trial consisting of 20 two-digit sentences from two overlapping talkers. The primary task was to correctly identify the highest (numerically) valued sentence. We measured both the speech-recognition error rate, and whether the subject got the right answer in our cognitive test.

A sentence is recognized only if the constituent digits are correctly recognized. Given the received sentences it is easy to pick the highest value. Human subjects reported that they could often recognize the highest values but could not recall who said it. We, thus, ignored the gender of the talker in scoring our tests.

Figure 4 shows our recognition results for several different variations of our experiment. The distracted result is lowest because the attention model switches immediately as

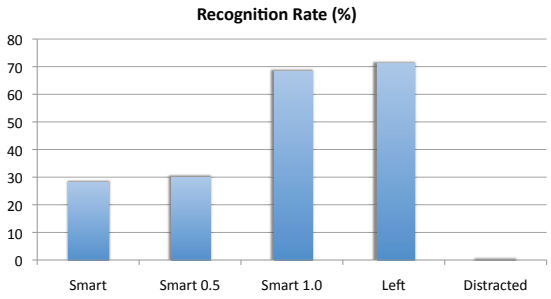

Fig. 4. Recognition results for five different cognitive models. We show the rate at which the highest-value sentenced is identified correctly.

soon as there is a salient event in the unattended direction—the recognition system almost never sees an entire sentence. The smart model works much better because once it has recognized a big digit at the start of the sentence it continues to attend to that direction. The two "smart" alternatives shown in the figure represent variations where we only allow switches to occur after the indicated number of seconds. Thus "Smart 0.5" switches the attentional focus only after at least 0.5 seconds. Finally, the best results are for the single-channel attention model—always pay attention to one speaker. This high result is an artifact of our experimental paradigm. Since there is a very restricted vocabulary, the same high-valued sentence is often spoken by both talkers. With a larger range of possible numbers the best the single-channel model should achieve is 50% accuracy.

5. FUTURE WORK

Our goal is not to show that our brain model is smarter than yours (it might be :-) but instead to describe the features of a model of auditory scene analysis based on saliency and attention. Doing this work we found several scientific areas where more study is needed, and thus our model has shortcomings.

Most importantly, the salience of real-world sounds is difficult to measure and model. Kayser's model [9] picks out single tones well, but our audio environment is much more complicated, including, for example, multiple harmonics in a pitched voice. Kalinli's model includes pitch but is primarily tested by measuring stress in spoken English sentences [10]. The binaural model used here is largely an onset detector, and not nearly as sophisticated as we would like. None of the models we know of allow higher levels of the brain to specify which sounds should be considered most salient.

Separation remains an unsolved problem. We were struck at the difficulty of recognizing the speech when the channels were summed and we tried to listen monaurally to either talker. Furthermore, simple beam forming based on delay and add with the two ear signals was not sufficient. The human auditory system has an amazing ability to understand speech sounds with different directions of arrival.

Perhaps the biggest unknown in our model is how saliency and attention really interact. We chose a simple model to illustrate the concept, but clearly humans use a much more sophisticated mechanism. A list of effects that remain to model includes: divided attention, perhaps based on selective listening; the tradeoff between saliency and attention; expectations that some signals are more likely to be informative; historical information; and the role of visual signals in the saliency and understanding of a cocktail party.

6. ACKNOWLEDGEMENTS

We would like to thank B. Shinn-Cunningham, O. Kalinli, K. Patil, N. Vasconcelos, D. Wang and J. Mitchell for valuable discussions of ideas explored here and T. Delbruck, S. Joshi and B. Verplank for technical support. This collaboration was only possible because of the Telluride Neuromorphic Engineering Workshop. We dedicate this paper to the newly arrived Arianna and Alessia who managed to capture the attention of our friend and colleague J. Martinez-Trujillo.

7. REFERENCES

[1] Martin Cooke, *Modelling auditory processing and organisation*, Cambridge University Press, 1993.

[2] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms and applications*, Wiley Interscience, 2006.

[3] D. Ellis, "The prediction-driven approach to computational auditory scene analysis, and its application to speech/nonspeech mixtures," in *Speech Communication*, M. Cooke and H. Okuno, Eds., April 1999, vol. 27, pp. 281–298.

[4] R. Carlyon, R. Cusack, J. Foxton, and I. Robertson, "Effects of attention and unilateral neglect on auditory stream segregation," *J Exp Psychol Hum Percept Perform*, vol. 27, pp. 115 –127, 2001.

[5] A. Bregman, *Auditory scene analysis: The perceptual organization of sound*, MIT Press, 1990.

[6] L. van Noorden, *Temporal coherence in the perception of tone sequences*, Ph.D. thesis, Eindhoven University, 1975.

[7] V. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision Research*, vol. 45, no. 2, pp. 205 – 231, 2005.

[8] Martin Cooke, John R. Hershey, and Steven J. Rennie, "Monaural speech separation and recognition challenge," *Comput. Speech Lang.*, vol. 24, pp. 1–15, January 2010.

[9] C. Kayser, C. Petkov, Lippert M., and Logothetis N., "Mechanisms for allocating auditory attention: An auditory saliency map.," in *Current Biology*, 2005, vol. 15, pp. 1943–1947.

[10] O. Kalinli and S. Narayanan, "Prominence detection using auditory attention cues and task-dependent high level information," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 5, pp. 1009 –1024, July 2009.

[11] H. Finger and S-C. Liu, "Estimating the location of a sound source with a spike-timing localization algorithm," in *ISCAS*. 2011, pp. 2461–2464, IEEE.