Auditory Saliency Using Natural Statistics

Tomoki Tsuchida (ttsuchida@ucsd.edu)

Garrison W. Cottrell (gary@ucsd.edu)
Department of Computer Science and Engineering
9500 Gilman Drive, Mail Code 0404
La Jolla CA 92093-0404 USA

Abstract

In contrast to the wealth of saliency models in the vision literature, there is a relative paucity of models exploring auditory saliency. In this work, we integrate the approaches of (Kayser, Petkov, Lippert, & Logothetis, 2005) and (Zhang, Tong, Marks, Shan, & Cottrell, 2008) and propose a model of auditory saliency. The model combines the statistics of natural soundscapes and the recent past of the input signal to predict the saliency of an auditory stimulus in the frequency domain. To evaluate the model output, a simple behavioral experiment was performed. Results show the auditory saliency maps calculated by the model to be in excellent accord with human judgments of saliency.

Keywords: attention; saliency map; audition; auditory perception; environmental statistics

Introduction

In general, attention plays a very important role in the survival of an organism, by separating behaviorally relevant signals from irrelevant ones. One approach to understanding how attention functions in the brain is to consider the "saliency map" over the sensory input space, which may determine subsequent motor control targets or selectively modulate perceptual contrast thresholds. The brain's putative saliency maps can be thought of as interest operators that organisms use to enhance or filter sensory signals.

Many visual saliency models have been investigated, but relatively little attention has been paid to modeling auditory saliency. However, since the fundamental necessity for perceptual modulation remains the same regardless of modality, the principles of visual saliency models should apply equally well to auditory saliency with appropriate sensory input features. Two representative visual saliency models are the center-surround contrast model (Itti, Koch, & Niebur, 2002) and the SUN (Salience Using Natural Statistics) model (Zhang et al., 2008). Itti et al.'s model is neurally-inspired, with the response of many feature maps (e.g., orientation, motion, color) combined to create a salience map. The SUN model uses a single feature map learned using Independent Components Analysis (ICA) of natural images, and the salience at any point is based on the rarity of the feature responses at that point - novelty attracts attention. Here, rarity is based on statistics taken from natural images, so the model assumes experience is necessary to represent novelty.

Previous works that apply the visual saliency paradigm to the auditory domain include the models of (Kayser et al., 2005) and (Kalinli & Narayanan, 2007). Both adapt the visual saliency model of (Itti et al., 2002) to the auditory domain by using spectrographic images as inputs to the model. Although this is a reasonable approach, these models fail to capture several important aspects of the auditory modality. First, this approach treats time as simply another dimension within the spectrographic representation of the sound. Even though these models utilize asymmetric temporal filters, the resulting saliency map at each time point is contaminated by information from the future. Second, spectrographic features are not the most realistic representations of human auditory sensations, since the cochlea exhibits complex nonlinear responses to sound signals (Lyon, Katsiamis, & Drakakis, 2010). Finally, Itti et al.'s model determines the saliency values from the current input signal, with no contribution from the lifetime experience of the organism. This makes it impossible for the model to account for potential perceptual differences induced by differences in individual experience.

The Auditory Saliency Model

In this work, we propose the Auditory Salience Using Natural statistics model (ASUN) as an extension of the SUN model. The extension involves (1) using realistic auditory features instead of visual ones, and (2) combining long-term statistics (as in SUN) with short-term, temporally local statistics. Although the SUN model has both a top-down, task-based component and a bottom-up, environmentally driven component, here we restrict ourselves to just the bottom-up portion of SUN. SUN defines the bottom-up saliency of point x in the image at time t as:

$$s_x(t) \propto -\log P(F_x = f_x) \tag{1}$$

Here, f is a vector of feature values, whose probability is computed based on prior experience. This is also known as the "self-information" of the features, and conveys that rare feature values will attract attention. In the SUN model, this probability is based on the lifetime experience of the organism, meaning that the organism already knows when feature values are common and when they are rare. Assuming the primary purpose of attention is to separate remarkable events from the humdrum, it is logical to equate the rarity of the event with the saliency of it. For example, a loud bang may be salient not only because of its physical energy content, but also because of its relative rarity in the soundscape. An organism living under constant noise may not find an explosion to be as salient as another organism acclimated to a quieter environment.

For features, SUN uses ICA features learned from natural images, following Barlow's efficient coding hypothesis (Barlow, 1961). This provides a normative and principled rationale for the model design. While ICA features are not completely independent, they justify the assumption that the features are independent of one another, making the computation of the joint probability of the features at a point computationally simple. This is the goal of efficient coding: By extracting independent features, the statistics of the visual world can be efficiently represented. Although the saliency filters used in Kayser et al.'s model have biophysical underpinnings, exact shape parameters of the filters cannot be determined in a principled manner. More importantly, their model does not explain *why* the attention filters should be the way they are. In contrast, by using filters based on the efficient coding hypothesis, the SUN and ASUN models make no such assumptions; the basic feature transformation used (Gammatone filters) reasonably approximate the filters learned by the efficient encoding of natural sounds (Lewicki, 2002), and the distributions of filter responses are learned from the environment as well. Assuming that the attentional mechanism is modulated by a lifetime of auditory experience is neurologically plausible, as evidenced by the experience-induced plasticity in the auditory cortex (Jääskeläinen, Ahveninen, Belliveau, Raij, & Sams, 2007).

Here, we extend this model to quickly adapt to recent events by utilizing the statistics of the recent past of the signal (the "local statistics") as well as the lifetime statistics. Denoting the feature responses of the signal at time t as F_t, saliency at t can be defined as the rarity in relation to the recent past (from the input signal) as well as to the long-term past beyond suitably chosen delay k:

$$s(t) \propto -\log P(F_t = f_t | \underbrace{F_{t-1}, ..., F_{t-k}}_{\text{recent past}}, \underbrace{F_{t-k-1}, ...}_{\text{long past}})$$

In this paper, we simply define $t-k$ as the onset of the test stimulus. Under the simplifying assumption of independence between the lifetime and local statistics, this becomes

$$\begin{aligned} s(t) &\propto -\log P(F_t = f_t | F_{t-1}, ..., F_{t-k}) \\ &\quad -\log P(F_t = f_t | F_{t-k-1}, ...) \\ &= s_{local}(t) + s_{lifetime}(t), \end{aligned}$$

where $s_{local}(t)$ and $s_{lifetime}(t)$ are the saliency values calculated from the local and lifetime statistics, respectively. By using the local statistics at different timescales, the model can simulate various adaptation and memory effects as well. In particular, adaptation effects emerge as the direct consequence of dynamic information accrual, which effectively suppresses the saliency of repeated stimuli as time proceeds. With such local adaptation effects, the model behaves similarly to the Bayesian Surprise model (Baldi & Itti, 2006), but with asymptotic prior distributions provided by lifetime experience.

Feature Transformations

A model of auditory attention necessarily relies upon a model of peripheral auditory processing. The simplest approach to modeling the cochlear transduction is to use the spectrogram of the sound, as was done in (Kayser et al., 2005). More physiologically plausible simulations of the cochlear processing require the use of more sophisticated transformations, such as Meddis' inner hair cell model (Meddis, 1986). However, the realism of the model comes at a computational cost, and the complexity of the feature model must be balanced against the benefit. Given these considerations, the following feature transformations were applied to the audio signals in the ASUN model:

1. At the first stage, input audio signals (sampled at 16 kHz) are converted to cochleagrams by applying a 64-channel Gammatone filterbank (from 200 to 8000 Hz.) Response power of the filters are clipped to 50dB, smoothed by convolving with a Hanning window of 1 msec and downsampled to 1 kHz. This yields a 64-dimensional frequency decomposition of the input signal.

2. At the second stage, this representation is further divided into 20 frequency bands comprised of 7 dimensions each (with 4 overlapping dimensions,) and time-frequency patches are produced using a sliding window of 8 samples (effective temporal extent of 8 msec). This yields 20 bands of $7 \times 8 = 56$-dimensional representation of 8 msec patches.

3. Finally, for each of the four sound collections (described below), a separate Principal Components Analysis (PCA) is calculated for each of the 20 bands separately. Retaining 85% of the variance reduces the 56 dimensions to 2 or 3 for each band.

This set of transformations yield a relatively low-dimensional representation without sacrificing biological plausibility. The result of these transformations at each time point, f_t, provides input for subsequent processing. Figure 1 illustrates this feature transformation pipeline.

Density Estimation Method

In order to calculate the self-information described in equation 1, the probability of feature occurrences $P(F = f_t)$ must be estimated. Depending on the auditory experience of the organism, this probability distribution may vary. To assess the effect of different types of lifetime auditory experiences, 1200 seconds worth of sound samples were randomly drawn from each of the following audio collections to obtain empirical distributions:

1. "Environmental": collection of environmental sounds, such as glass shattering, breaking twigs and rain sounds obtained from a variety of sources. This ensemble is expected to contain many short, impact-related sounds.

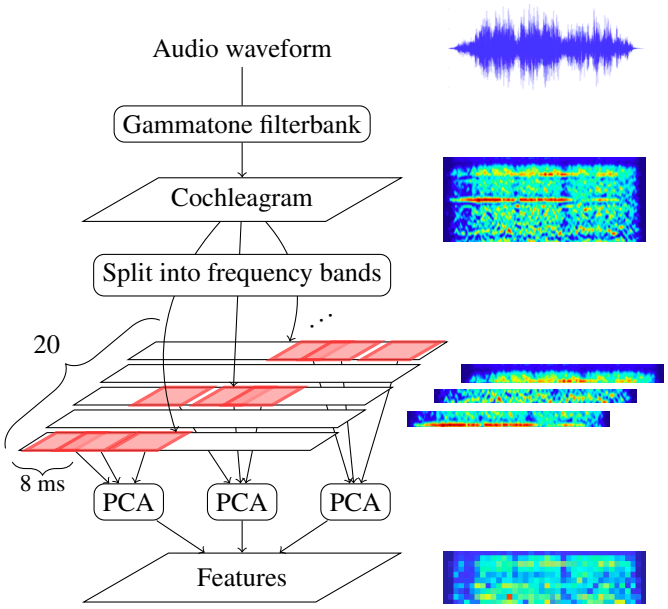


Figure 1: Schematics for the feature transformation pipeline. Input signals are first converted to smoothed cochleagram. This is separated into 20 bands of 8 msec patches. The dimensions of each band are reduced using PCA.

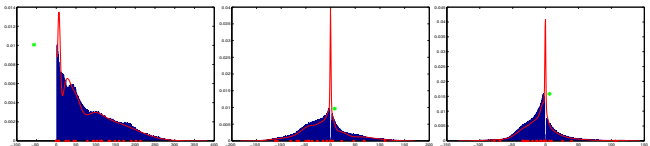
2. "Animal": collection of animal vocalizations in tropical forests from (Emmons, Whitney, & Ross, 1997). Most of the vocalizations are relatively long and repetitious.

3. "Speech": collection of spoken English sentences from the TIMIT corpus (Garofolo et al., 1993). This is similar to the animal vocalizations, but possibly with less tonal variety.

4. "Urban": this is a collection of sounds recorded from a city (van den Berg, 2010), containing long segments of urban noises (such as vehicles and birds), with a limited amount of vocal sounds.

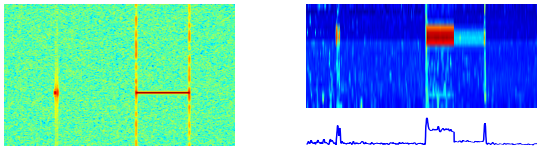
In the case of natural images, ICA filter responses follow the generalized Gaussian distribution (Zhang et al., 2008). However, the auditory feature responses from the sound collections did not resemble any parameterized distributions. Consequently, a Gaussian mixture model with 10 components was used to fit the empirical distributions for each band from each of the collections. Figure 2 shows examples of density model fits against empirical distributions. The distributions from each collection represent the lifetime statistics portion of ASUN model, and each corresponds to a model of saliency for an organism living under the influence of that particular auditory environment.

The local statistics of the input signal were estimated using the same method: at each time step t of the input signal, the probability distribution of the input signal from 0 to $t-1$ was estimated. For computational reasons, the re-estimation of the local statistics were computed every 250 msec. Unfortunately, this leads to a discontinuity in the local probability distribution every 250 msec. This will be improved in future work, where we plan to apply continually varying mixture models to eliminate such transitions.

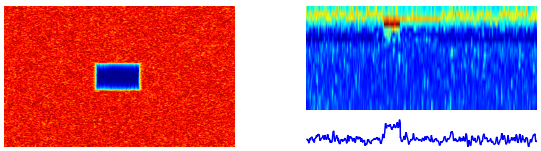


(a) Dimension 1 (b) Dimension 2 (c) Dimension 3

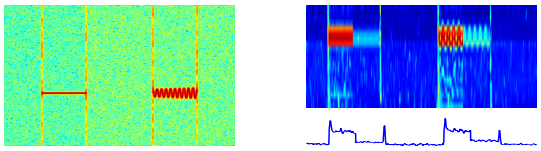
Figure 2: Gaussian mixture model fits (red) against the empirical distribution of feature values (blue). The mixture model is used to estimate $P(F_t = f_t | F_{t-k-1}, \ldots)$.



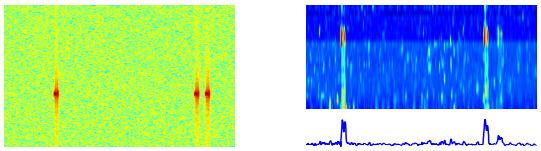
(a) Short and long tones



(b) Gap in broadband noise



(c) Stationary and modulated tones



(d) Single and paired tones

Figure 3: Spectrograms and saliency maps for simple stimuli. Left columns are the spectrograms of the stimuli, and right columns are the saliency maps (top) and saliency values summed over frequency axis (bottom). Due to the nonlinear cochleogram transform, the y-axes of the two plots are not aligned. (a) Between short and long tones, the long tone is more salient. (b) Silence in a broadband noise is salient compared to the surrounding noise. (c) Amplitude-modulated tones are slightly more salient than stationary tones. (d) In a sequence of closely spaced tones, the second tone is less salient.

Qualitative Assessments

In (Kayser et al., 2005), the auditory saliency model reproduces basic properties of auditory scene perception described

in (Cusack & Carlyon, 2003). Figure 3 shows the saliency maps of the ASUN model using the "Environmental" lifetime statistics. These examples demonstrate that the model is capable of reproducing basic auditory salience phenomena.

Human Ratings of Saliency

In order to test the validity of the model in a more quantitative manner, a human rating experiment similar to (Kayser et al., 2005) was performed. In this experiment, seven subjects were asked to pick the more "interesting" of two stimuli. The goal of the experiment was to obtain an empirical rating of the "interestingness" of various audio stimuli, which we conjecture is monotonically related to the saliency. By presenting the same set of stimuli to the saliency models, we can also calculate which of the sounds are predicted to be salient. We assume that the correct model of saliency will have a high degree of correlation with the human ratings of saliency obtained this way.

Materials

Audio snippets were created from a royalty-free sound collection (SoundEffectPack.com, 2011), which contains a variety of audio samples from artificial and natural scenes. In order to normalize the volume across samples, each sample was divided by the square root of the arithmetic mean of the squares of the waveform (RMS). To create snippets used in the experiment, each sample was divided into 1.2-second snippets, and the edges were smoothed by a Tukey window with 500 ms of tapering both sides. Snippets containing less than 10% of the power of a reference sinusoidal signal were removed in order to filter out silent snippets.

From this collection, 50 high-saliency, 50 low-saliency and 50 large-difference snippets were chosen for the experiments. The first two groups contained snippets for which the Kayser and ASUN models agreed on high (or low) saliency. Snippets in the last group were chosen by virtue of producing highest *disagreements* in the predicted saliency values between Kayser and ASUN models.

With these snippets, 75 trial pairs were constructed as follows:

(1) *High saliency difference* trials (50): Each pair consists of one snippet from the high-saliency and another from the low-saliency groups.

(2) *High model discrimination* trials (25): Both snippets were drawn from the large-difference group uniformly.

We expected both models to perform well on *high saliency difference* trials but to produce a performance disparity on the *high model discrimination* trials.

Procedure

In each trial, each subject was presented with one second of white noise (loudness-adjusted using the same method as above) followed immediately by binaural presentation of a pair of target stimuli. The subject would then respond with the left or right key to indicate which stimuli sounded "more interesting" (2AFC.) Each experiment block consisted of 160 such trials: 75 pairings balanced with left-right reversal, plus 10 catch trials in which a single stimulus was presented to one side. Each subject participated in a single block of the experiment within a single experimental session.

Model Predictions

To obtain the model predictions, the same trial stimuli (including the preceding noise mask) were input to the models to produce saliency map outputs. To reduce border effects, 10% buffers were added to the beginning and end of the stimuli and removed after saliency map calculation. The portion of the saliency map that corresponded to the noise mask were also removed from peak calculations.

In (Kayser et al., 2005), saliency maps for each stimuli pair were converted to scores by comparing the peak saliency values. It is unclear what the best procedure is to extract a single salience score from a two-dimensional map of salience scores over time. Following (Kayser et al., 2005), we also chose the peak salience over the snippet. To make predictions, the score for the left stimulus was subtracted from that of the right stimulus in each trial pair. This yielded values between -1 and 1, which were then correlated against the actual choices subjects made (-1 for the left and 1 for the right.)

Seven different candidate models were evaluated in this experiment. (1) The *chance* model outputs -1 or 1 randomly. This model serves as the baseline against which to measure the chance performance of other models. (2) The *intensity* model outputs the Gammatone filter response intensity. This model simply reflects the distribution of intensity within the sound sample. (3) The *Kayser* model uses the saliency map described in (Kayser et al., 2005). Finally, *ASUN* models with different lifetime statistics were evaluated separately: (4) "Environmental" sounds, (5) "Animal" sounds, (6) "Speech" sounds, and (7) "Urban" sounds.

Results

To quantify the correspondence between the model prediction and the human judgments of saliency, Pearson product-moment correlation coefficients (PMCC) were calculated between the model predictions and human rating judgment results (N=7) across all 75 trials. All subjects responded correctly to the catch trials, demonstrating that they were paying attention to the task. Figure 4 shows the correlation coefficient values for the ASUN models for each type of dataset from which lifetime statistics were learned. The correlation between the ASUN model predictions and the human subjects ($M = 0.3262, SD = 0.0635$) was higher than the correlation of the Kayser model predictions ($M = 0.0362, SD = 0.0683$). The result shows that the ASUN model family predicted the human ratings of saliency better than the Kayser model ($t(6) = 7.963, p < 0.01$.)

To evaluate the model performance in context, across-subject correlation was also calculated. Since the models

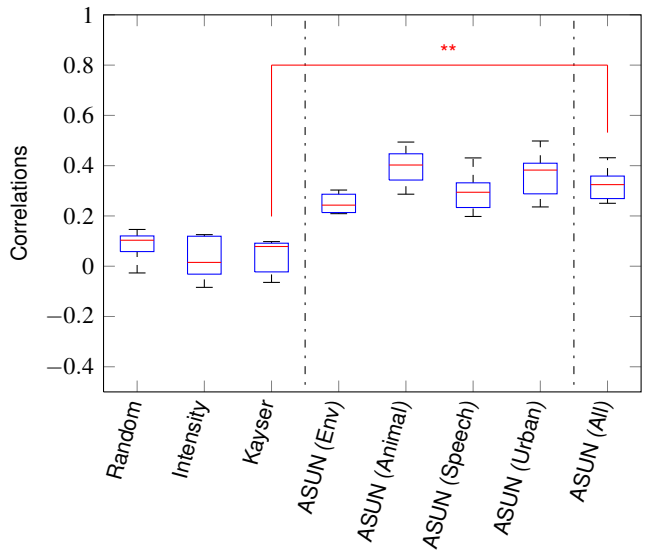


Figure 4: Correlation coefficient between various models and human ratings of saliency (N=7.) ASUN models correlated with the human ratings of saliency significantly better than the Kayser model.

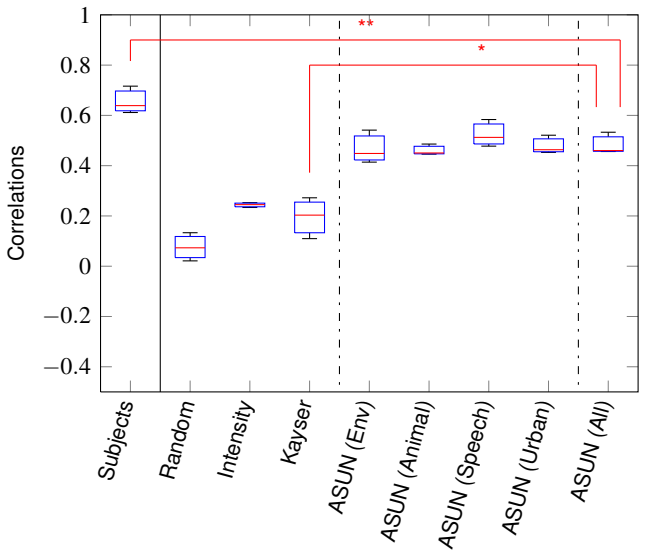


Figure 5: Correlation coefficient between various models and human ratings of saliency. A subset of data for which the same trial pairs were presented was analyzed (N=3). Across-subject performance was estimated using the correlation coefficients for all possible pairs from the three subjects.

are not fit to individual subjects, this value provides the ceiling for any model predictions. Because three of the seven subjects went through the same trial pairs in the same order, these trials were used to calculate the across-subject correlation value, and the model responses. Figure 5 shows the correlation values including the across-subject correlation. The result shows that the difference between the across-subject correlations ($M = 0.6556, SD = 0.0544$) and the ASUN model predictions ($M = 0.4831, SD = 0.0432$) was significant ($t(2) = 16.9242, p = 0.0035$), indicating that the models do not yet predict saliency at the subject-consensus level. Nevertheless, the ASUN model correlations were still significantly higher than the Kayser model ($M = 0.1951, SD = 0.0815$) at ($t(2) = -9.855, p = 0.0101$).

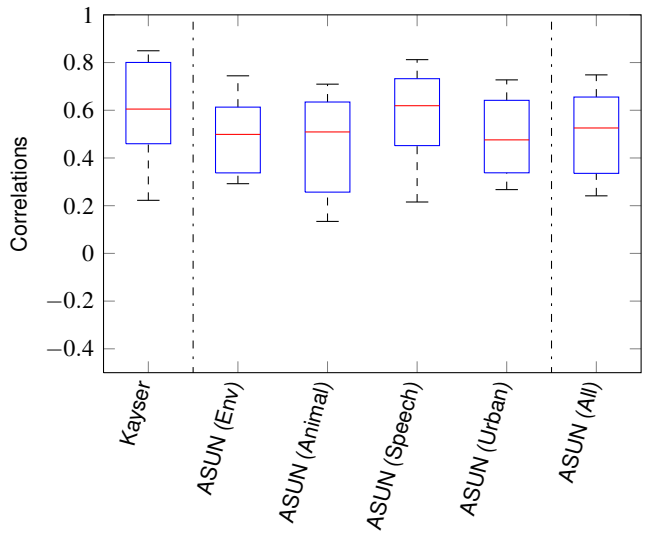
The performance for the Kayser model in this experiment was notably worse than what was reported in (Kayser et al., 2005). There are several possible explanations for this. First, the audio samples presented in this experiment were roughly normalized for the perceived loudness. This implies that a saliency model that derives saliency values from the loudness measure in large part may not perform well in this experiment. Indeed, the intensity model does not predict the result above chance ($t(6) = 0.66, p = 0.528$). Although the Kayser model does combine information other than the intensity image alone, it is possible that the predictive power of the model is produced largely by loudness information.

Second, as described previously, some of the trial pairs were chosen intentionally to produce maximal difference between the Kayser and ASUN models, and this produced the large performance disparity. Figure 6 support this hypothesis: in the *high saliency difference* trials, both models performed equally well ($t(6) = 0.3763, p = 0.7091$.) In contrast, in *high model discrimination* trials, ASUN models performed significantly better than the Kayser model ($t(6) = 17.31, p < 0.01$.) Note that the *high model discrimination* group was not picked based on the absolute value (or "confidence") of the model predictions, but rather solely on the large difference between the two model predictions. This implies the procedure itself does not favor one model or the other, nor does it guarantee performance disparity on average. Nevertheless, the result shows that the ASUN models perform better than the Kayser model in those trials, suggesting the performance disparity may be explained in large part from those trials.

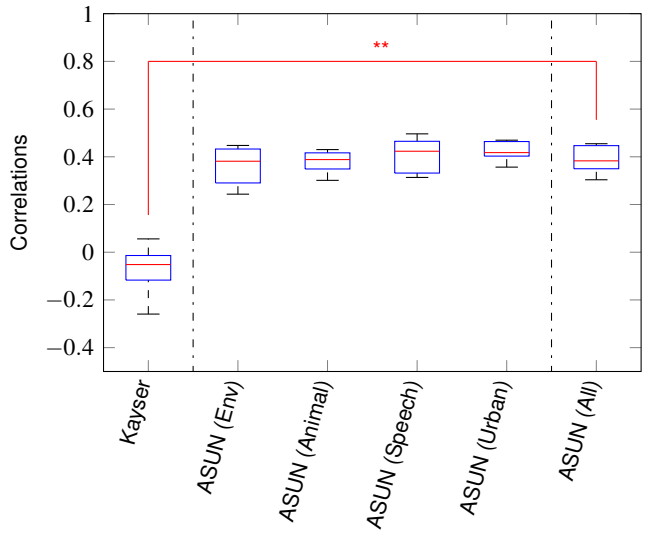
Discussion

In this work, we demonstrated that a model of auditory saliency based on the lifetime statistics of natural sounds is feasible. For simple tone signals, auditory saliency maps calculated by the ASUN model qualitatively reproduce phenomena reported in the psychophysical literature. For more complicated audio signals, assessing the validity of the saliency map is difficult. However, we have shown that the relative magnitudes of the saliency map peaks correlate with human ratings of saliency. The result was robust across different training sound collections, which suggest a certain commonality in the statistical structure of naturally produced sounds.

There are aspects of the saliency model that may be improved to better model human physiology. For example, there is ample evidence of temporal integration at multiple timescales in human auditory processing (Poeppel, 2003). This indicates that the feature responses of the input signal



(a) *High saliency difference* trials



(b) *High model discrimination* trials

Figure 6: Correlation coefficients for the subsets of trials. (a) For *High saliency difference* trials, both Kayser and ASUN models show high correlation to human rating of saliency, and there are no significant differences between them. (b) For *High model discrimination* trials, ASUN models show significantly higher correlation with human ratings of saliency compared to the Kayser model.

may be better modeled by multiple parallel streams of inputs, each convolved with exponentially decaying kernels of varying timescales. This may be especially important for calculating saliency of longer signals, such as music and spoken phrases. In order to accommodate higher-level statistical structure, the model can be stacked in a hierarchal manner as well, with appropriate feature functions at each level. These expansions will provide insights into the nature of attentional modulations in human auditory processing.

Acknowledgments

We thank Dr. Christoph Kayser for kindly providing us with the MATLAB implementation of his model. We also thank Cottrell lab members, especially Christopher Kanan, for insightful feedback. This work was supported in part by NSF grant #SBE-0542013 to the Temporal Dynamics of Learning Center.

References

Baldi, P., & Itti, L. (2006). Bayesian Surprise Attracts Human Attention. In *Nips 2005* (pp. 547–554).

Barlow, H. B. (1961). Possible Principles Underlying the Transformations of Sensory Messages. *Sensory Communication*, 217–234.

Cusack, R., & Carlyon, R. (2003). Perceptual asymmetries in audition. *J Exp Psychol Human Percept Perf*, *29*(3), 713–725.

Emmons, L. H., Whitney, B. M., & Ross, D. L. (1997). *Sounds of the neotropical rainforest mammals.* Audio CD.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., et al. (1993). *Timit acoustic-phonetic continuous speech corpus*. Linguistic Data Consortium, Philadelphia.

Itti, L., Koch, C., & Niebur, E. (2002). A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, *20*(11), 1254–1259.

Jääskeläinen, I. P., Ahveninen, J., Belliveau, J. W., Raij, T., & Sams, M. (2007). Short-term plasticity in auditory cognition. *Trends Neurosci.*, *30*(12), 653–661.

Kalinli, O., & Narayanan, S. (2007). A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. In *Interspeech 2007* (pp. 1941–1944). Antwerp, Belgium.

Kayser, C., Petkov, C., Lippert, M., & Logothetis, N. (2005). Current biology; mechanisms for allocating auditory attention: An auditory saliency map. , *15*(21), 1943–1947.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *nature neurosci*, *5*(4), 356–363.

Lyon, R. F., Katsiamis, A. G., & Drakakis, E. M. (2010). History and future of auditory filter models. In *Iscas* (pp. 3809–3812). IEEE.

Meddis, R. (1986). Simulation of mechanical to neural transduction in the auditory receptor. *JASA*, *79*(3), 702–711.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication*, *41*(1), 245–255.

SoundEffectPack.com. (2011). *3000 sound effect pack.* Retrieved 2011-03-31, from `tinyurl.com/7f4z2wo`

van den Berg, H. (2010). *Urban and nature sounds.* Retrieved 2011-02-27, from `http://tinyurl.com/89mr6dh`

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A bayesian framework for saliency using natural statistics. *Journal of vision*, *8*(7), 1-20.