15th European Signal Processing Conference (EUSIPCO 2007), Poznan, Poland, September 3-7, 2007, copyright by EURASIP

USING AUDITORY SALIENCY TO UNDERSTAND COMPLEX AUDITORY SCENES

Varinthira Duangudom and David V. Anderson

School of Electrical and Computer Engineering, Georgia Institute of Technology
Atlanta, GA 30332
phone: + (1) 404-385-6414, email: tira@gatech.edu

该论文提出了一个预测前注意，自下而上的听觉显著性的计算模型。该模型感知上决定了在一个场景中什么对观察者来说是突出的，并且可以确定听觉场景中那一部分是重要的。

ABSTRACT

In this paper, we present a computational model for predicting pre-attentive, bottom-up auditory saliency. The model determines perceptually what in a scene stands out to observers and can be used to determine what part of a complex auditory scene is most important. The vision equivalency of this is visual saliency as defined by Koch and others [1]. The model is based on inhibition of features obtained from auditory spectro-temporal receptive fields (STRFs) and produces results that match well with preliminary psychoacoustic experiments. The model does well in predicting what is salient for some common auditory examples and there is a strong correlation between scenes chosen as salient by the model and scenes that human subjects selected as salient.

1. INTRODUCTION

Sounds can simultaneously change along many dimensions, including, changes in pitch, loudness, timbre, or location. Despite this, humans are able to successfully integrate these cues and use them to identify, categorize, and group sounds [2]. Human performance on these tasks far exceeds that of computational models, especially in noisy conditions. In particular, with speech, most models have problems dealing with speaker variability and also face problems with noise robustness. Humans, on the other hand, are able to account for speaker variability along with changes in the other auditory percepts, such as, pitch, loudness, and timbre, mentioned above.

One reason for studying the salience of different auditory stimuli is to improve the performance of current computational models by being able to determine perceptually what in a particular scene stands out to an observer. Auditory saliency can be defined from either a bottom-up processing perspective or from a top-down processing perspective. From a bottom-up perspective, which is addressed in this paper, salient sounds are defined as those sounds that can be noticed without attention. It is pre-attentive and deals with sounds that grab a listener's attention. From a top-down perspective, humans use previously learned models to understand complex auditory scenes and salient sounds are sounds that violate these models.

Auditory saliency can help observers sort the information present in a complex auditory scene. Since resources are finite, not all information can be processed equally, and we must be able to identify from the scene what is most important. Additionally, a quick response or decision may be required from the observer. In order to respond, the observer must quickly determine what is important in the scene. The issue of saliency is closely related to many areas, including scene analysis.

In this paper, we propose a bottom-up computational model for auditory saliency. The model uses inhibition of features obtained from auditory spectro-temporal receptive fields to compute a saliency map identifying what is most salient in a complex scene. This model is then evaluated by comparing scenes the model chooses as salient to scenes human subjects selected as salient.

2. COMPUTATIONAL AUDITORY SALIENCY MODEL

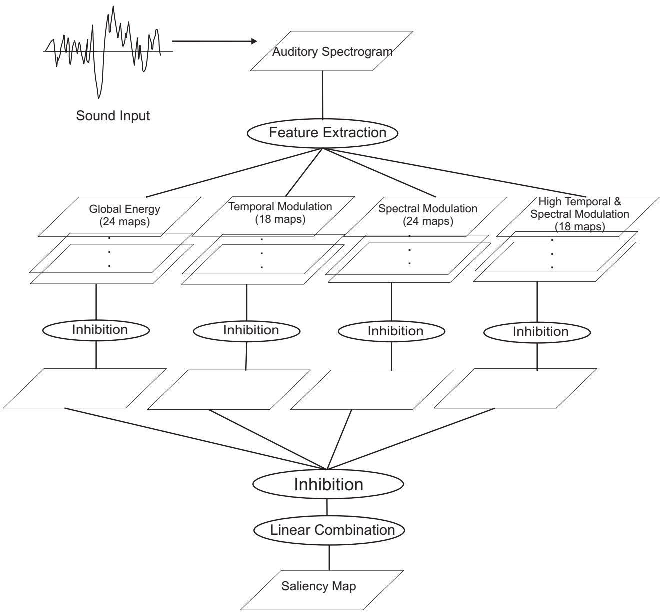


Figure 1: *General architecture for proposed computational auditory saliency model.*

The general architecture of the proposed computational model for auditory saliency model can be seen in (Fig. 1). The sound input is first converted into an auditory spectrogram. The feature maps generated from the spectrogram are categorized into 4 broad feature classes, where individual feature maps within each class respond to varying rates of change (depending on the filter used). After the feature maps have been generated, each individual map undergoes inhibition, resulting in the demotion of maps with no features that stand out locally. These maps then provide minimal contribution to the overall saliency map. The individual feature maps in each of the 4 categories are then combined into a 'global' feature map for each class. Finally, the 4 global feature maps are again subjected to inhibition and summed to form the final saliency map.

The proposed computational model for auditory saliency is similar in structure to models for visual saliency and auditory saliency [1], [3]. There are two main stages to the model: the feature extraction stage where the features are generated and the inhibition (or suppression) stage where certain features are promoted or suppressed determining their overall contribution to the final saliency map. Each stage will be explained in more detail in the next sections.

The model presented by Kayser et al. [3] mainly differs from visual saliency models in its interpretation, as the features used are generated and processed in very similar ways. The model proposed here relies on inhibition of feature maps generated from auditory spectro-temporal receptive fields (STRFs). This model differs from the model in [3] by the features chosen and the processing of the features to form the saliency map. Additionally, this model is physiologically motivated and may more typically represent human audio analysis. The experiment discussed later, where subjects are asked to choose the most salient scene from a pair of scenes, resulted in slightly higher correlations between subject and model responses than the correlations found in [3] for a similar experiment, but there is no way to make a direct comparison as different sets of stimuli are used.

2.1 Generation of Feature Maps

We want features similar to those the brain potentially uses in processing auditory sounds. Instead of using simple features, we choose to use a cortical model that works in the spectrotemporal domain and has various scales of temporal and spectral resolution, as both spectral and temporal modulations are relevant to auditory perception. The cortical model, proposed by Wang and Shamma in [4], models spectral shape analysis in the primary auditory cortex. It is based on both psychoacoustical and neurophysiological results in the early and central stages of the auditory pathway, and physiological experiments in the primary auditory cortex of ferrets and cats have confirmed that cortical cells have response properties matching those of this model [5, 6]. A schematic of the cortical model is shown in Fig. 2. According to this model, neurons in the primary auditory cortex (A1) are organized along three mutually perpendicular axes. Along the tonotopic axis, the response field of neurons are tuned to different center frequencies. Along the scale axis, the bandwidth of the response field of neurons lined is decreasing. At the center of A1, the response field has an excitatory center, surrounded by inhibitory side bands. As one moves further away from the center of A1, the response field tends to become more asymmetrical. It has been argued that the tonotopic axis is akin to a Fourier transform and the presence of different scales over which this transform is performed leads to a multi-scale Fourier transform. It has been shown that performing such an operation on the auditory spectrum leads to the extraction of spatial and temporal modulation information [7].

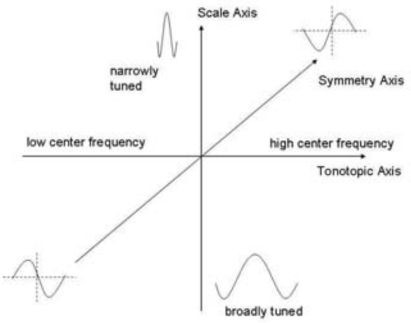


Figure 2: *Schematic of the cortical model. Adapted from [4].*

As seen in (Figure 1), the sound input is first converted into an auditory spectrogram, which is a time frequency representation modelling early auditory processing.The auditory spectrogram is then decomposed into its spectral and temporal components using a bank of spectro-temporally selective filters. From this, we get the STRF (spectro-temporal receptive field) features which estimate spectral and temporal modulations.

The rate corresponds to the center frequency of the temporal filters used in the transform and yields temporal modulation information. Scale corresponds to the center frequency of the spatial (frequency) filters used in the transform and yields spatial modulation information.

Receptive fields with varying rates and scales were chosen in order to capture different aspects of the sound, as some filters respond to rapid changes while others respond to slower changes. Also, the filters vary in how wide or narrowly tuned they are. Fourteen temporal filters (rates) from ±0.5 to ± 32 Hz, and 6 spectral filters (scale) from 0.25 cycles/octave to 8 cycles/octave were used. For each scale and rate, we generate 1 feature map, therefore, there are a total of 84, 2-D (time/frequency) feature maps generated. The map size varies depending on the length of the auditory stimulus. Since there are 128 frequency channels, for a 1 second stimulus, maps are of size 125 x 128. Next, the maps were grouped into four different feature classes. The rate and scale determine which grouping each feature map is categorized in. The first set of feature maps (24 feature maps) gives the overall energy distribution. The second set (18 maps) focuses on temporal modulations. Spectral modulations are given by the third set(24 maps). The fourth set (18 maps) looks at areas where there is both high temporal and spectral modulations.

2.2 Formation of Saliency Map

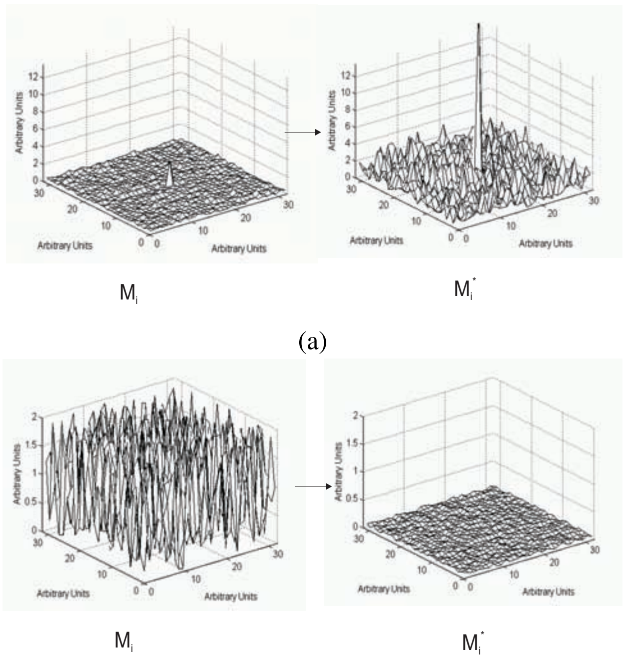
The 2-D (time/frequency) feature maps have now been generated from the different temporal and spectral filters. Inhibition is now performed on each individual feature map. The goal of the inhibition stage is to promote feature maps that contain features which stand out locally on the map and inhibit or suppress feature maps without any prominent peaks. We want to promote maps that have areas of high activity (large global peak) compared to rest of the map. In order to achieve this, each feature map, M_i, was scaled by a factor, D_i. The new scaled feature maps will be referred to as $M_i{}^*$.

$$M_i{}^* = D_i{}^* \, M_i$$

$$D_i{=}(G_i - \overline{L}_i)^2$$

where G_i and $\overline{L}_i$ are defined as follows:

$$G_i = \text{Global peak of map i}$$
$$\overline{L}_i = \text{Average of all other local peaks on map i.}$$



(a)

(b)

Figure 3: Inhibition of feature maps. a) Feature map with single prominent peak is promoted b) Feature map with many peaks and no prominent peaks is suppressed.

将每个特征图经过抑制后，每个类别合成一张特征图，
四个类别再抑制以后线性组合

Scaling each feature map by this factor, D_i, promotes maps with a large global peak compared to the rest of the activity on that map. This is demonstrated in Figure 3 (a). Conversely, feature maps with high activity everywhere on the map are suppressed (Figure 3 (b)). Using this method, we preserve the general shape of each feature map, while ensuring that feature maps with prominent peaks make a larger contribution to the final saliency map. After this scaling, the feature maps in each respective category are combined to form 1 global feature map for each category. Each of the 4 global feature maps are again scaled by D_i before being summed to form the final saliency map.

One variation of the model is to perform the inhibition locally. This was done, since auditory percepts are often more influenced or affected by other auditory events or cues closer in time or frequency. Here, the feature extraction stage is the same as discussed above. Once the feature maps have been obtained, the feature map is divided into non-overlapping 2-dimensional areas, covering about 200 ms in time and 1/3 octave in frequency. In order to retain the peaks, for each local area, the mean of the signal is determined and subtracted from the signal. This is then followed by the previously described method of scaling used to promote the maps with prominent peaks. The differences between the two versions of the model presented are summarized below.

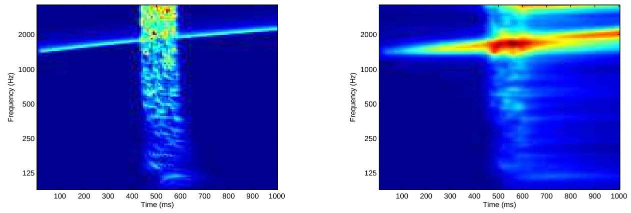
3. SALIENCY MAP EXAMPLES

In this section, we show the saliency map for several different examples of common auditory stimuli. The saliency maps for these examples match what is expected from known psychoacoustic experimental results.

The auditory system is well-versed in change detection. From an auditory scene analysis perspective, older sounds that are relatively constant or unchanging tend to become background, while changes in a sound or new sounds will stand out from the background and are more salient.

Model	Description
1	Promotes or inhibits entire feature maps using scaling by D_i
2	Uses local inhibition and then scaling by D_i

Table 1: Summary of model 1 and model 2 differences.



(a)

(b)

Figure 5: Auditory saliency map for the continuity illusion. a) Auditory spectrogram of gliding tone interrupted by a noise burst b) Auditory saliency map of gliding tone interrupted by a noise burst.

Two change detection examples are presented where new sounds stand out perceptually from the other unchanging "background" components. In the first example, we have a tone complex where part of one of the four tones in the complex is amplitude modulated. We would expect the modulation to be salient, since the modulated tone should stand out perceptually from the other unchanging tones. In Fig. 4, (a) and (b), the auditory spectrogram and saliency map for this example are shown. From the saliency map, the modulated part is, as expected, what is most salient and the rest of the tone complex except for the onsets are suppressed.

The second change detection example shows the well-known experimental result of hearing out a mistuned harmonic from the rest of the complex. In this example, the 4th harmonic of 200 Hz tone is mistuned by 48 Hz for 50 ms causing the mistuned harmonic to pop-out. This mistuned harmonic is heard separate from the rest of the complex, since it stands out perceptually. The saliency map confirms this experimental result and Fig. 4, (c) and (d), shows that the mistuned harmonic does pop-out from the rest of the complex.

A third example shows two 250 ms, 2 kHz tones in white noise. In Fig. 4, (e) and (f), the auditory spectrogram and saliency map for this example show that the noise is suppressed and the two tones are emphasized. From this, there may be several applications of auditory saliency in the area of noise suppression.

In one final example shown in Fig. 5 (a) and (b), a gliding tone is interrupted by a noise burst. From the continuity illusion, it is expected the tone will be perceived as continuing through the noise. The saliency map for this example (Fig. 5 (b)), reflects the perceptual continuity of the tone through the noise, and this continuity is indicated as being salient.

The four examples presented above are used to demon-

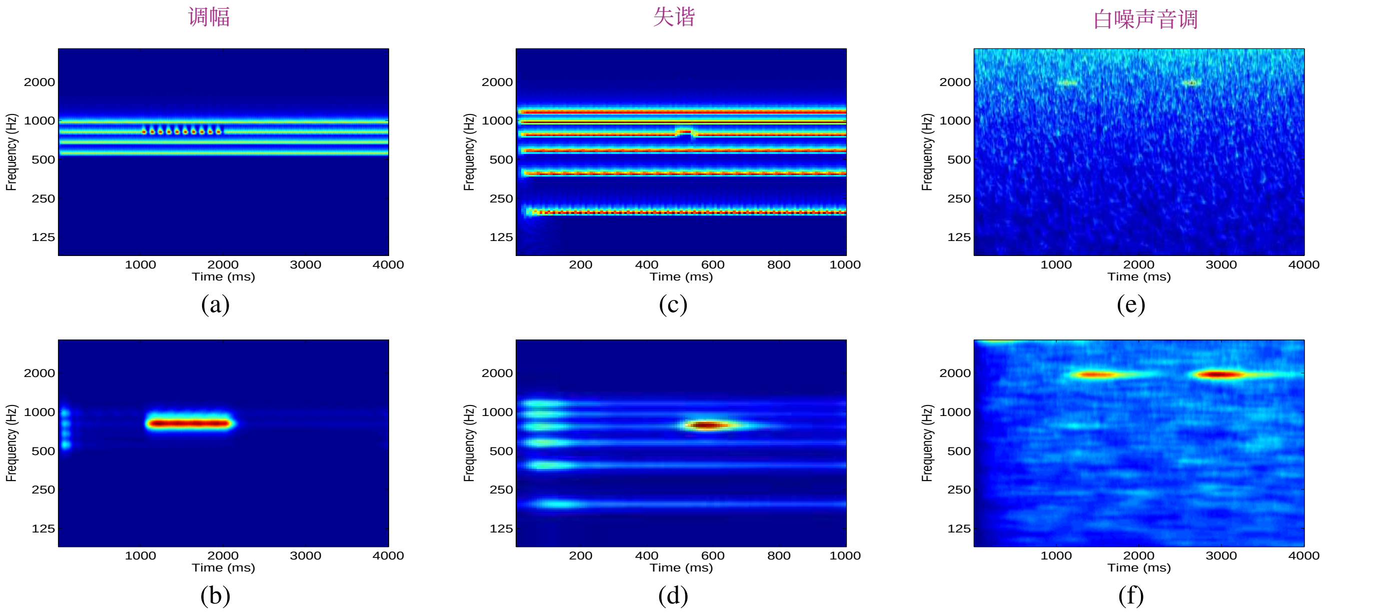


Figure 4: Auditory saliency map for some simple auditory stimuli. The top row shows the auditory spectrograms for each of the three examples. In the second row, below each auditory spectrogram, the saliency map for that example can be found. a) Auditory spectrogram of amplitude modulated tone b) Auditory saliency map of amplitude modulated tone shows the modulated part is salient c) Auditory spectrogram of 200 Hz tone, with 6 harmonics, where the 4th harmonic is mistuned by 48 Hz d) Auditory saliency map of 200 Hz tone, with 6 harmonics, where the 4th harmonic is mistuned by 48 Hz e) Auditory spectrogram of two 2 kHz tones in white noise tones f) Auditory saliency map of two 2 kHz tones in white noise.

strate how the saliency map is able to predict several known experimental results from auditory scene analysis.

4. SALIENCY SCENE COMPARISON

In order to compare the model's performance to human subjects, we did an experiment where subjects were asked to give subjective ratings of the saliency of different scene pairs. The same pairs were then presented to the computational auditory saliency model for comparison.

4.1 Subjects

Results were obtained from 14 university students. Normal hearing was determined by self-report. Subjects were informed about the general aim of the experiment, but were naive to the exact purpose of the study.

Subjects listened through headphones (Sennheiser HD 280 pro) to auditory stimuli presented using MATLAB and then entered their responses into a MATLAB GUI. The sound card on the PC used is AD1981A AC'97 SoundMAX Codec (Full-duplex with variable sampling rates from 7040 Hz to 48 kHz with 1 Hz resolution). The experiment took each subject approximately 50 minutes to complete.

4.2 Experimental Procedures

Each subject was presented with a total of 162 scene pairs made up from 50 unique target sounds. The scenes consisted of a target (1 sec) sound and a background (4 sec). The target sounds consisted of sounds, such as, animal sounds, sirens, noise, music instruments etc. The background consisted of white noise and a random sample of some target sounds.

Subjects were presented with different scene pairs. Each pair consisted of two 4 second scenes with a pause of 1 second between each scene. Subjects then had to indicate whether the first scene they heard or the second scene they heard had the most salient element or if the saliency of the two scenes was equal. After choosing which scene was more salient, subjects were asked to rate from 1 to 5 how much more salient the scene they chose was compared to the other scene.

Catch trials, where the scene pair was made up of the same scene, were presented to the subjects. In addition, there were also catch trials where scene pairs presented earlier in the test were presented again later to get an indication of how consistent a subject was.

4.3 Results and Discussion

Scenes that subjects selected as equally salient were excluded from the analysis. Subject accuracy on the catch trials is presented in Table 2. All subjects had performance on the cache trials greater than 50% and only three subjects (6,7,and 12) had performance lower than 70% correct on these trials. Poor performance on catch trials could indicate that subjects were not focused on the task, were randomly selecting answers, or did not understand the experiment. All subjects correctly identified the catch trials where the same scene was presented twice as being equally salient.

In this experiment, we found a significant correlation between scenes that subjects selected as salient and scenes that the model chose as salient. For model 1, the mean of the correlation coefficients from all subjects was 0.4776 ± 0.2155 and for model 2, where the inhibition was performed locally, it was 0.5302 ± 0.2379. These results for can be found in Table 3.

Saliency can also vary greatly depending on top-down input. Therefore, what is salient to one observer can be very different from what is considered salient to another observer.

Subject	% Correct
1	88.9
2	72.2
3	94.4
4	88.9
5	77.8
6	64.7
7	66.7
8	83.3
9	94.4
10	82.3
11	94.4
12	64.7
13	83.3
14	83.3

Table 2: Subject performance on the catch trials.

	Correlation to	
Subject	Model 1	Model 2
1	0.7324	0.8738
2	0.4536	0.5329
3	0.4138	0.5247
4	0.774	0.7872
5	0.0449	0.0182
6	0.0632	0.1136
7	0.397	0.4725
8	0.4073	0.4178
9	0.5977	0.7033
10	0.5995	0.692
11	0.4234	0.4234
12	0.622	0.678
13	0.6131	0.63
14	0.5447	0.5555
Average	0.4776	0.5302
Std Dev	0.2155	0.2379

Table 3: Correlation between subject and model responses.

The model we propose is a bottom-up processing model, and does not provide any top-down input. This may provide some explanation for why 2 of the subjects (5, 6) had almost no correlation between the model's responses and subject's responses. Additionally, subject 6 also had poor performance on the catch trials, indicating that they may not have been properly attending to the task. It is also interesting to note that subjects 5 and 6 along with subject 1 were the only 3 subjects that performed the experiment in the evening. Since the task does require a subject's attention, performing the experiment later in the day when they are likely tired or lack concentration may affect the results.

Since the model is a bottom-up processing model, we are particularly interested in comparing the model's performance for scenes where there is some agreed salience among the observers. Therefore, we next looked at scene pairs where the majority of the subjects were in agreement that one of the two scenes was more salient than the other. This removes some of the individual variation that may cause certain types of stimuli to be salient to particular observers, since we are looking at scenes where at least half of the observers agree that it is salient. The results for this are shown Table 4. For these pairs, the results show a strong correlation between the subject and model responses. The correlation was 0.7224 (95% CI = 0.6239-0.7983) for model 1 and 0.8043 for model 2 (95% CI = 0.7303-0.8596).

Model	Correlation
1 Scaling by D_i on entire feature maps	0.7224
2 Local inhibition	0.8043

Table 4: Correlation between subject and model responses for scene pairs where more than 50% of subjects agreed on salience.

Based on the results in Tables 3 and 4, for these sounds, the model does well in predicting what scenes humans would consider salient. Additional experiments are being performed to further evaluate the model for different types of stimuli.

4.4 Conclusions and Future Work

The auditory saliency model presented in this paper predicts what in an auditory scene stands out perceptually to observers. It can be used to sort through the elements of a complex scene and determine what is most important and captures an listener's attention. The model is physiologically motivated and uses auditory receptive field models and adaptive inhibition to form the saliency map. The model was validated experimentally, and there was a fairly strong correlation between auditory scenes chosen as salient by the model and scenes chosen as salient by human subjects. Additionally, some simple examples of the saliency map were used to demonstrate that it can predict known experimental results, but the saliency map can also be used for more complex auditory scenes. Currently, the model is a bottom-up processing model, but we know that top-down influences can also affect perception. We are working on adding top-down input to the model and more experiments are underway to further evaluate the model's performance and its numerous applications in auditory scene analysis and other areas.

REFERENCES

[1] L. Itti, C. Koch, and E. Nieber, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] A. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, 1990.

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

[4] K. Wang and S. Shamma, "Spectral shape analysis in the central auditroy system," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 382–395, Sept 1995.

[5] B.M. Calhoun and C.E. Schreiner, "Spatial frequency filters in cat auditory cortex," in *Proceedings of the 23rd Annual Meeting Society of Neuroscience*, 1993.

[6] D.M. Green, "Frequency and the detection of spectral shape change," in *Auditory Frequency Selectivity*, B.J.C. Moore and R.D. Patterson, Eds., pp. 351–360. NATO ASI Series, 1986.

[7] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *Eurasip Journal on Applied Signal Processing*, vol. 2003, no. 7, pp. 668–675, June 2003.