A temporal saliency map for modeling auditory attention

Emine Merve Kaya and Mounya Elhilali
Department of Electrical and Computer Engineering
The Johns Hopkins University
Baltimore, MD 21218, USA
Email: {merve, mounya}@jhu.edu

***Abstract*—The auditory system is flooded with information throughout our daily lives. Rather than processing all of this information, we selectively shift our attention to various auditory events - either events of interest (top-down attention) or events that capture our attention exogenously (bottom-up). In this work, we are concerned with aspects of human attention that are bottom-up stimulus-driven. Saliency of an auditory event is measured by how much the event differs from the surrounding sounds that precede it in time. To calculate this, we propose a novel auditory saliency map that is defined only over time. The proposed model is contrasted against previously published auditory saliency maps which treat the two-dimensional auditory time-frequency spectrogram as an image that can be analyzed using visual saliency models. Instead, our proposed model capitalizes on the rich high-dimensional feature space that defines auditory events; where each acoustic dimension is processed across multiple scales. These normalized feature maps are then combined over time into a single temporal saliency map. The peaks of the temporal saliency map indicate the locations of the salient events in the auditory scene. We validate the accuracy of the proposed model in simulated test scenarios of simple and complex sound clips. By exploiting the unique aspects of auditory processing that cannot be readily captured by visual processes, we are able to outperform other auditory saliency models; all while highlighting the commonalities and differences between the two modalities in processing salient events in everyday scenes.**

I. Introduction

How do our brains cope with the myriad of information our sensory systems are faced with on a daily basis? One factor that alleviates the problem is that we have the remarkable ability to quickly recognize an object in the environment that does not blend in. We can detect a red bird in a bush, a baby crying in a classical concert, or a foul smell in the house easily. We do not process every face in a crowd at once; instead our attention quickly shifts to the few faces that are either the most different, or most interesting to us. This helps us filter the incoming sensory information in a way that allows us to make the best decision about how to act. It has been found that neural mechanisms exist to carry out this process of filtering relevant information from the scene to be analyzed more intensively. [1] This fight for the spotlight of attention is based on both exogenous properties of the object (bottom-up saliency) as well as cognitive processes (top-down). [2] Top-down attention comes into play mainly when there is a task to be completed, whereas bottom-up saliency is fixed throughout changes in task or situation. [2], [3]

The first computational model of bottom-up saliency was proposed by Koch and Ullman in 1985 [4] and further developed by Itti et al. [5] for the visual system. The basis of the model was related to the "feature integration theory" [6], where various features are extracted from the spatial input. Saliency information that is extracted from each feature is then combined to produce a spatial map of saliency. The saliency map encodes the bottom-up conspicuity of each location in the visual scene. This framework was shown to replicate human overt and covert attention [7]. It is a fast parallel-processed mechanism; and top-down influences are not difficult to be combined with the bottom-up system. Since its proposal, the saliency map framework for the visual system has been greatly extended upon.

In the auditory modality, only a few such systems have so far been proposed. The first model, by Kayser et al. [8] treats the auditory input spatially by considering the time-frequency representation of an auditory signal ("the auditory image") as the input of the saliency model. Thereafter, a saliency mechanism closely following the framework of Itti is applied. This mechanism is able to match experimental results of simple salient stimuli, such as finding a salient natural tone among noise. However, its performance is inherently limited in that it only extracts visual features from the auditory image, which is not enough to capture a lot of significant information from an auditory signal. The second model, by Kalinli et al. [9] builds on Kayser's model by adding two more features that are extracted from the auditory image, and using a different normalization scheme. While this definitely improves the type of saliency of auditory stimuli that can be captured by the model, it is still bounded by the same problem as Kayser's. The last model, by Duangudom et al. [10] focuses on spectro-temporal receptive field (STRF) output computed from the auditory image as the feature base of the system and slightly expands the normalization procedure. However, the model is oversaturated by STRF output in various scales and rates, and does not represent a complete account of sound perception in the brain.

In this paper, we propose an alternative biologically plausible method of capturing bottom-up auditory attention. We stay within the boundaries of the original saliency framework by Itti; however, we take a slightly different approach to treating incoming auditory information than was executed in previous

978-1-4673-3140-1/12/$31.00 ©2012 IEEE

auditory saliency models. Our key contribution is that we treat an auditory scene as a single dimensional temporal input at all times, rather than treating it as an image. This is not to say we do not use the frequency-time representation: We do. However, it is only one feature component of the system, and even then, we treat every frequency channel as a temporal signal and do not use contrasts between adjacent frequency channels. With our newly proposed features for use in auditory saliency extraction, we follow the original framework of finding salient points, which are in the end combined to yield a single temporal auditory saliency map.

II. Previous Auditory Saliency Maps

The structure of the original saliency map framework in [5] is as follows. The original image is filtered at various levels to provide multiscale features, usually from scale 1 to scale 8 where the image at each (i+1)th scale is half of the size of the image at (i)th scale. All of these features are subjected to center surround differences by cross-scale subtraction of a subset of pairs of features. This results in “feature maps”. The feature maps are then each normalized so that they accurately suppress the background in a scene, at the same time boosting the salient information. The normalized feature maps are added across scales to produce one “conspicuity map” for each feature. These conspicuity maps are averaged to form the final saliency map. Thus, the size of the resulting saliency map is a scaled version of the size of the original image. Therefore it can be directly mapped to the visual space to find the location of the salient event in the scene.

Kayser’s model first forms the spectrogram of the auditory signal. The spectrogram is treated as the auditory image and the rest of the processing will closely follow the framework explained above. Three features are extracted from the spectrogram: Intensity, frequency contrast, and temporal constrast. Since all of these features are features of the spectrogram rather than directly on the auditory signal, every feature in scale 1 is of the same size as the original spectrogram. Center surround differences are calculated in the same manner as Itti’s model. The normalization process is the similar in theory as the one used in [5]. Following, the conspicuity maps are averaged to form the final saliency map. The most salient event can be found as the time instance where the maximum of the saliency map occurs.

Kalinli’s model builds on top of the Kayser model by adding orientation and pitch information. Orientation information is extracted from the spectrogram in 45 and 135 degree angles. Pitch is calculated following the temporal hypothesis of pitch extraction and then mapped to the frequency axis of the spectrogram to provide a feature map the same size as the other maps. The rest of the processing is the same as Kayser’s map, except for the normalization which is done as the iterative normalization described in [11], which we also use here for the time dimension.

Duangudom’s model uses time-frequency energy, temporal modulation, spectral modulation, and spectro-temporal modulation. The center-surround stage is removed, and only normalization remains after feature extraction. The normalization is done in the same method as [5] (Model 1). A variation is also examined, in which normalization is done on local patches of maps instead of globally (Model 2), which results in slightly higher correlation to human reports of saliency.

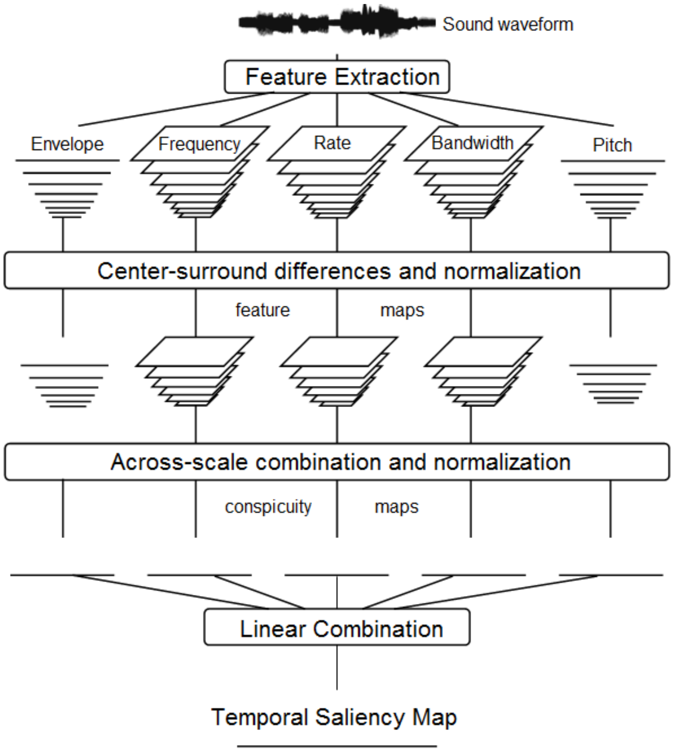

Fig. 1. Architecture of the temporal saliency model. Features shown as a line are one-dimensional, features shown as a block are two-dimensional. The dimensions of different features and feature maps may be different.

Kayser has tested his method on simple sounds among noise, and show that his detection results match psychoacoustic experiment results of perceived human saliency. Kalinli has used the model to detect prominent syllables in speech, for which detection results of 60-80% are obtained, where performance is calculated based on how well it matches the manually labeled data. Duangudom has tested on same type of stimulus as Kayser, with reports of lower performance (Correlation mean across subjects for Model 1=0.48, Model 2=0.53) than Kayser has reported ($r=0.56$, $p<0.01$).

III. A Temporal Saliency Map

The model we propose in this paper uses 5 features: Waveform envelope, spectrogram, rate, bandwidth, and pitch. The features envelope and pitch are always kept one dimensional throughout processing. The other features are first computed in two dimensions: They are still treated as an entity that varies primarily among time, however the feature is computed for multiple frequencies/octaves to obtain the highest amount of information.

The waveform envelope is obtained by the Hilbert transform of the original waveform. The main advantage of including the envelope as a feature is two-fold: It is easier to detect loud

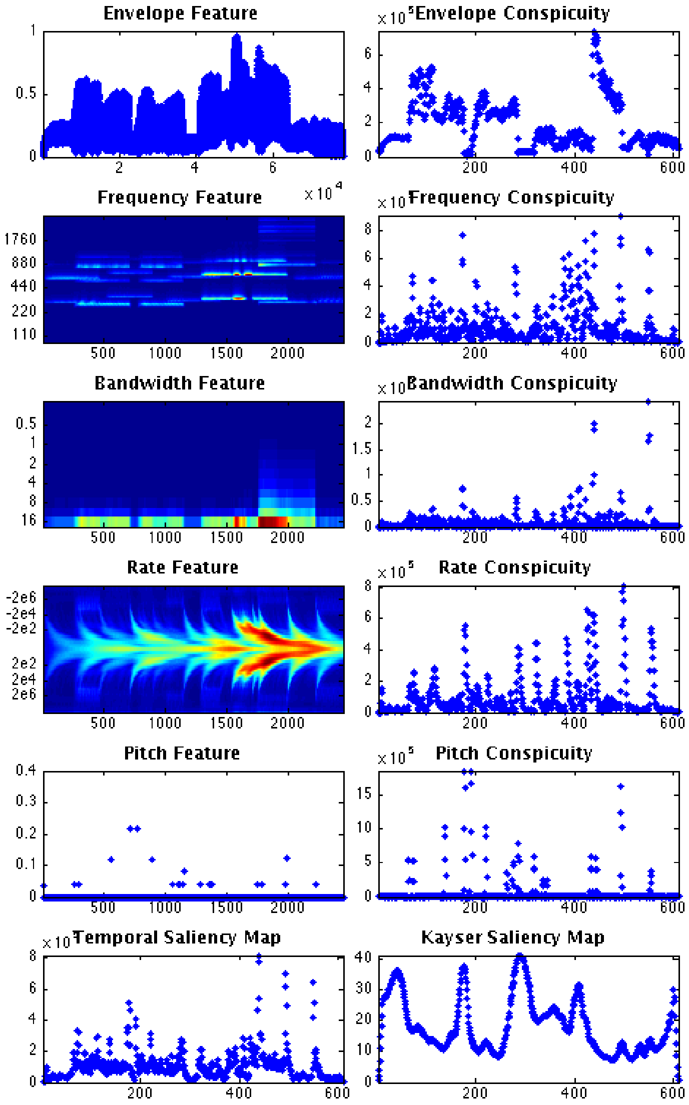

Fig. 2. Saliency results for a timbre-varying target. The background instrument is a violin and the foreground instrument is a flute. The left figures are, from top: Envelope, frequency, bandwidth, rate, pitch features; followed by saliency results of our model. The right figures are the five conspicuity maps belonging to the feature on the left; followed by saliency results of Kayser's model. The top three peaks of our model correspond to the beginning, middle and end of the target note. The top peaks of Kayser's model all correspond to background notes.

events with the waveform, and also the shape of a tone helps to characterize its timbre. The envelope is computed by first taking the magnitude of the Hilbert transform of the data, and then running a Butterworth filter of cutoff 60Hz and order 6 through it. The waveform at scale 1 is convolved with a 5-sample length Gaussian and decimated to half length. This process is repeated until all 8 scales are obtained.

The auditory spectrum of sound used mimics the information processing of the early auditory system. [12] The center frequencies of the bandpass filters convolved with the input are evenly distributed on a logarithmic scale. The filtered signal is put through a mechanism including high-pass filtering, non-linear compression, and low-pass filtering, in simulation of inner hair cells. The final output is the integration of a lateral inhibitory network. [13] Our spectrogram is computed with time windows of length 2ms with no overlap and 128 channels over 5.4 octaves. In our experiments, we use data with a sampling rate of 16kHz, which gives us center frequencies ranging between approximately 100Hz and 4kHz. The high windowing rate is chosen to give us better resolution for further rate processing. The spectrogram is convolved with a 2-D Gaussian of 5 samples in either direction, and downsampled to obtain all scales.

Bandwidth and rate information are computed from the spectrogram by filtering it with cortical bandpass filters in time and frequency channels. The computation of these features mimic the response of neurons at the mammal auditory cortex, which are tuned to a range of spectral resolution and temporal modulation. [14] It has been found that the auditory system uses these spectrotemporal modulations, which are shown to capture properties of speech intelligibility for humans [15]. The characteristic ripple frequencies to compute the bandwidth feature are selected uniformly between 2^{-2} and 2^{4} cycles per octave. The frequencies of the filters for the rate feature are selected uniformly between 2^{0} and 2^{8}, each at up and down directions. The spectrogram at every scale is filtered by the appropriate filters for that scale. The rate filter frequencies are adjusted at every level due to downsampling, the high frequencies are not computed at that level. This results in different lengths for the second dimension in the rate feature.

Pitch is obtained from template-matching. For each time window we select the pitch as the maximum of the cross-correlation lag. This also gives us a saliency score, which corresponds to the level of the correlation function. To reduce random noise effects, we discard the pitch information of the time values which have a saliency that is lower than the difference between the mean and standard deviation of all saliency scores along time. We take the logarithm of the remaining pitch values and take their derivative so that resulting high peaks correspond to changes in pitch. The scales of the pitch feature are obtained identical to the envelope feature.

After obtaining these features in 8 scales, center-surround differences are found, mimicking the properties of local cortical inhibition. The process of calculation is across-scale subtraction between a center (fine) scale and a surround (coarse) scale, with the result being rectified. The fine scales are selected as $c \in \{2, 3, 4\}$ and the coarse scales are $s = c + d$ where $d \in \{3, 4\}$; giving us 6 feature maps for each feature. All of these feature maps are normalized so that the minimum value across the map is 0, and the maximum value, summed for all frequencies or bandwidths, at each time instance is 1. For envelope and pitch, this just means that the feature map is scaled between 0 and 1. However, the two dimensional features will be scaled between 0 and a number between 0 and 1. The previous models all used scaling between 0 and 1 for this point, however, the features they use are all of the same dimensions, whereas all of our features have different dimensions. Since the final salient score will be found by taking the maximum of the sum across the dimension that is not time, if we were to scale all maps between 0 and 1, when summed across the second dimension, the two dimensional maps would naturally produce a higher number, which does not necessarily correspond to a higher saliency. This problem is compensated for when we do this adaptive normalization.

These feature maps are now subjected to iterative nonlinear normalization [11] to boost salient parts and suppress smooth parts. We do not use frequency contrast because we are only concerned about the information among time, rather than the adjacency of the values in the second dimension. All feature maps (M) are convolved with a one-dimensional Difference-of-Gaussian filter (DoG) of length 50ms for the excitatory part with equally long inhibitory parts. The result of the convolution is added to the map, and the 2%th value of the map at the beginning of the normalization (C_{inh}) is subtracted. Negative values of the result are mapped to 0. Concisely, the following transformation is applied at each iteration, for 10 iterations:

$$M \leftarrow |M + M * DoG - C_{inh}|_{\geq 0}$$

The normalized feature maps are combined across scale into a single conspicuity map for each feature. The across scale addition is made by interpolating every map into a single scale and adding them. The two dimensional maps are averaged in their second dimension to provide a map that varies only in time, so that it will be compatible with the other temporal features. These conspicuity maps, which are now all the same one dimensional size, are averaged to provide the final temporal saliency map.

IV. Results

We tested our model on a set of stimuli made up of single tones of various musical instruments being played. The instruments used are violin, flute and harmonica. The length of the stimulus is set at 5 seconds, each containing 10 notes of length 1 second, overlapping with each other every 0.5 seconds. In each scene, only one aspect of musical instrument tone is varied: Timbre, pitch, or loudness. Our goal is to show that the features chosen in this study are able to capture these perceptual properties of sound, laying a foundation for selecting saliency among these dimensions which are most recognized by humans. In this section, we will investigate a few stimuli to enlighten the working of the model.

First, let us look at a timbre difference in the target note. In this example, the background notes are violin, and the foreground note is harmonica, which is the 8th note out of 10 notes in the scene. The loudness of each note is constant and the notes vary only slightly in a range of 5 semitones. The temporal saliency map correctly finds the location of the target note in this case, as seen in Figure 2. We can also see that Kayser's model is unable to find the location of the target note, the saliency is relatively low during the duration of that note. We can look at our 5 conspicuity maps to see where the saliency is coming from. The loudness map has found the most salient location as the 7th note. The frequency map has selected the target note as most salient. The bandwidth map has found the beginning and ending locations of the target note as most salient. The rate map has also found the target note as salient. The pitch feature has not actually found any significant difference, we can see that the differences are no bigger than 3 semitones, but the normalization procedure has boosted these small differences as if they are salient. We will discuss this issue in the next section. For now, we can simply note that frequency and rate are the main contributors for the final saliency score. For this case of timbre difference, frequency, bandwidth and rate features have all contributed to finding the location of the salient note, even if it is not perfectly reflected in their respective conspicuity maps.

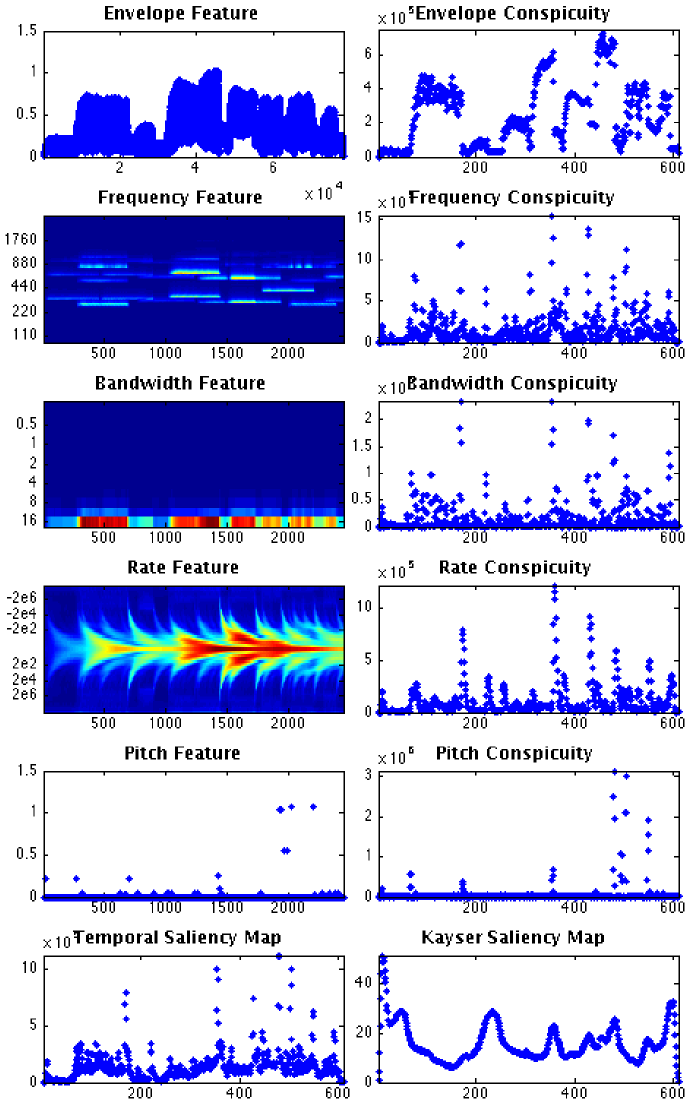

Fig. 3. Saliency results for a pitch-varying target. The target note has a pitch that is 5 semitones higher than the highest pitch of the background notes. The figures are the same order as Figure 2. The top peak of our model correspond to the time shortly after the target note began playing. Kayser's model has not found any significantly salient note.

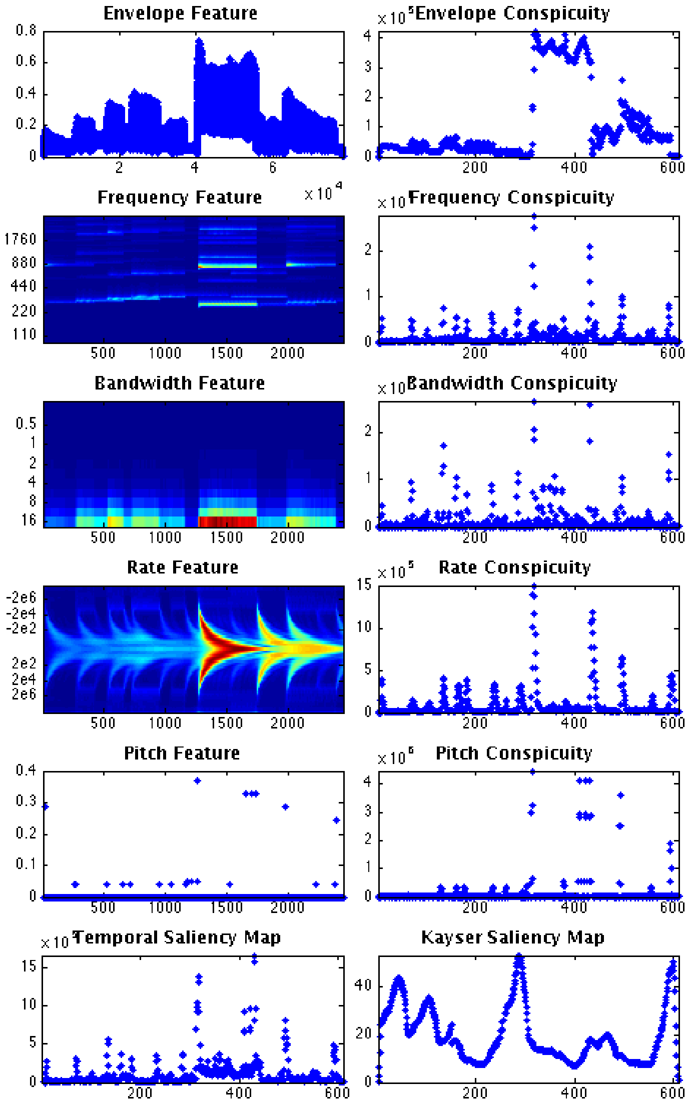

Fig. 4. Saliency results for a loudness-varying target. The ratio of the target to the background is 10db. The figures are the same order as Figure 2. The top two peaks of our model correspond to the beginning and ending times of the target note. Kayser's model has selected the less loud sections of the scene as more salient.

Next, let us look at what happens when there is a pitch difference. When the target note has a lower pitch, both our system and Kayser's system do well. Kayser can do well even without pitch information simply due to frequency contrast because the fundamental frequency of that note will be lower than any harmonics of other notes so there will always be a clear difference at that time. However, it is not so straightforward when there is a higher pitched note mixed in among notes with lower pitch. In this example, seen in Figure 3, we are using flute notes. The stimulus setup is the same as the previous example, and now our 8th note has a pitch that is 5 semitones higher than the masking notes. From the conspicuity maps, we see the main contributors to the saliency at the correct point are pitch and envelope. The other features have found a previous note with higher energy as more salient. Although this was successfully suppressed by the normalization, it has still produced a very high saliency, close to the maximum. Kayser's model was unable to find the correct note, with many peaks throughout time around the same saliency level, as would be expected from the image properties of this spectrogram.

	Our model	**Kayser's model**	
Hit at 1st peak	70%	15%	
Hit at 1-3 peaks	100%	40%	
	1st peak	**1st peak**	**1-3 peaks**
Hit for timbre	33.3%	0%	0%
Hit for pitch	87.5%	37.5%	75%
Hit for loudness	83.3%	0%	33%

Fig. 5. Detection rates of the target musical notes. Background notes vary only slightly in pitch, while the foreground note can be differing in instrument (timbre), pitch, or loudness. A hit occurs when a peak of the saliency map corresponds to the time of the target note being played.

Finally, we look at an example of what happens when there is a loudness difference. This scene is made up of harmonica notes, with the target to mask ratio (TMR) being 10dB. As can be seen from the plots in Figure 4, the envelope expectedly gives the highest saliency for the duration of the target note, with the other features that consider intensity also giving high values at the boundaries of the target note. The pitch coincidentally had a 4 semitone difference at the time of the high TMR note, and this small difference is again highly boosted in the system. Interestingly, Kayser's map could not find this difference, which should be straightforward to find when the spectrogram is treated as an image. Its time and frequency contrast features have overruled the intensity feature, so it found the relatively more silent parts of the scene more salient. However, this does not correspond to what is salient for a human in this case, and indeed TMR is one of the most easily detected dimensions of saliency for humans.

We ran a test on 20 variations of the stimulus described above: Timbre is varied 6 times (violin, flute, harmonica pairs), pitch 8 times (5st and 10st differences, low and high) and TMR 6 times (7dB and 10dB). A hit is defined as a peak in the saliency map that corresponds to the location of the target note at any instance while the note is playing. We calculated at which peak the target note is found at when peaks are ordered by magnitude. Detection results are presented in Figure 5.

V. Discussion

Out of the tested dimensions of human perception of saliency, we can see from the example results how our system is able to perform better for auditory saliency detection. Clearly, not all examples are able to yield the same result; mostly due to complications in calculating features. However, we see that the most important part of the mechanism to automatically detect saliency is the normalization and combination part. After the feature extraction stage, the biological system is able to easily detect the components that stand out and should be salient. Computationally, this task is not trivial.

Even from the three examples we have pointed out in the previous section, we can see various problems in normalizing features. For example, if we are to look at only one feature instead of the general picture, we can see that the pitch feature is always detecting some point as salient even if the relative difference of pitch is not high. Our background varies within 5 semitones, so it is natural to find small differences in pitch within the background. However, the system does not know that these small differences should not be boosted, unless we were to treat the pitch dimension separately. This is a limitation of the normalization procedure used; it will always boost outliers regardless of scale.

The normalization problem is also apparent in the timbre example, where we clearly see a large difference in the bandwidth feature. However, this is not accurately reflected in the conspicuity map due to the smooth fade-out and low energy being suppressed by the Difference-of-Gaussian filter. The result is that even though bandwidth successfully found the target, it has a lower peak than other features have, so its final influence is not very high.

Even though we have not done a comparison with Kalinli's model, we can claim the results will not be very different from Kayser's. This is due to the simple fact that the second model has only added pitch and orientation information, and uses the same normalization we have used. In our case of musical notes, which tend to have a flat "image", orientation will not give any significant results. Additionally, even without the extra pitch information, Kayser's model does generally well in the case of lower pitch. The addition of the pitch feature will cause an improvement when the target has a higher pitch, however it will not help with other types of saliency. The different normalization scheme may be an improvement over Kayser's, but from our examples and results we can see that our features capture information Kayser's features could not find; the normalization will not help when there is no significant information in the features. We have not done a comparison with the Duangudom model since its performance, as reported, is not greater than Kayser's.

The main implication of these results is that, while our newly proposed feature set is sufficiently rich, the normalization scheme of the visual saliency map may not be the best solution for the auditory saliency problem. As humans, we are able to see from individual features where the saliency lies; however, the current normalization method is not able to capture the information we see, suppressing and boosting incorrect parts instead. This framework depends on selection of parameters for its performance. We have only slightly altered the parameters of the original visual model, but have confirmed experimentally that varying parameters does not increase overall performance. This tells us that the issue is more fundamental, and that a method more fitting for an auditory paradigm might be necessary.

VI. Conclusions

We have presented a novel saliency extraction mechanism from an auditory scene. Following the visual saliency map frameworks, which the previous auditory saliency maps have also not strayed from, we extract multiscale features in parallel from the auditory signal. Center surround differences are subjected to an iterative normalization, which allows us to recombine the feature streams into a single saliency map. We have proposed for the first time to extract the saliency among the time dimension only, allowing us to have one dimensional features, along with features whose spatial dimension does not have to be the same. Results on complex auditory scenes were presented and contrasted with performance from a previous auditory saliency model, allowing us to demonstrate why our model is superior to previous models. Our results show that although we use a rich feature space that is able to capture significant properties of sound, the normalization method that works well on visual scenes may not be the optimal choice for the auditory scene. Future work on this model should take this into consideration and build a normalization mechanism fitting for the auditory system to complete a biologically inspired auditory saliency model.

Acknowledgments

This research was supported by ONR grant N000141010278, NSF grant IIS-0846112, AFOSR grant FA9550-09-1-0234 and NIH grant 1R01AG036424.

References

[1] J. Coull, "Neural correlates of attention and arousal: insights from electrophysiology, functional neuroimaging and psychopharmacology," *Progress in Neurobiology*, vol. 55, no. 4, pp. 343 – 361, 1998.

[2] L. Itti and C. Koch, "Computational modelling of visual attention," *Nat Rev Neurosci*, vol. 2, no. 3, pp. 194–203, 03 2001.

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

[4] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry." *Hum Neurobiol*.

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

[6] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97 – 136, 1980.

[7] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, pp. 1489–1506, 2000.

[8] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: An auditory saliency map," *Current Biology*, vol. 15, no. 21, pp. 1943 – 1947, 2005.

[9] O. Kalinli and S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *INTERSPEECH-2007*, 2007.

[10] V. Duangudom and D. V. Anderson, "Using auditory saliency to understand complex auditory scenes," in *15th European Signal Processing Conference (EUSIPCO 2007)*, 2007.

[11] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *J. Electronic Imaging*.

[12] X. Yang, K. Wang, and S. Shamma, "Auditory representations of acoustic signals," *Information Theory, IEEE Transactions on*, vol. 38, no. 2, pp. 824 –839, mar 1992.

[13] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[14] S. Shamma, H. Versnel, and N. Kowalski, "Ripple analysis in ferret primary auditory cortex. i. response characteristics of single units to sinusoidally rippled spectra," 1994.

[15] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (stmi) for assessment of speech intelligibility," *Speech Communication*, vol. 41, pp. 331 – 348, 2003.