Pattern Recognition Letters 38 (2014) 78–85

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Automatic detection of auditory salience with optimized linear filters derived from human annotation ☆

Kyungtae Kim [a,*], Kai-Hsiang Lin [b], Dirk B. Walther [c], Mark A. Hasegawa-Johnson [d], Tomas S. Huang [e]

[a] Mobile Communications Division, Samsung Electronics Maetan 3-dong, Yeongtong-gu, Suwon-si, Gyeonggi-do 443-742, Republic of Korea
[b] Department of Electrical and Computer Engineering, University of Illinois, 2253 Beckman Institute, 405 N. Mathews Urbana, Illinois 61801, United States
[c] Department of Psychology, Ohio State University, 1825 Neil Avenue Columbus, Ohio 43210, United States
[d] Department of Electrical and Computer Engineering, University of Illinois, 2011 Beckman Institute, 405 N. Mathews Urbana, Illinois 61801, United States
[e] Department of Electrical and Computer Engineering, University of Illinois, 2039 Beckman Institute, 405 N. Mathews Urbana, Illinois 61801, United States

ARTICLE INFO

Article history:
Received 9 May 2013
Available online 28 November 2013

Keywords:
Auditory salience
Conference room
Detection
Nonlinear programming

ABSTRACT

Auditory salience describes how much a particular auditory event attracts human attention. Previous attempts at automatic detection of salient audio events have been hampered by the challenge of defining ground truth. In this paper ground truth for auditory salience is built up from annotations by human subjects of a large corpus of meeting room recordings. Following statistical purification of the data, an optimal auditory salience filter with linear discrimination is derived from the purified data. An automatic auditory salience detector based on optimal filtering of the Bark-frequency loudness performs with 32% equal error rate. Expanding the feature vector to include other common feature sets does not improve performance. Consistent with intuition, the optimal filter looks like an onset detector in the time domain.

1. Introduction

In our daily lives we are often confronted with an overwhelming amount of sensory information, far exceeding the processing capabilities of our brains. How can we still make sense of the world around us, for instance, in a busy traffic situation? We need mechanisms to select the relevant or important information out of the data deluge accosting our sensory systems. Our brains achieve this with selective attention – a process of preferentially processing some stimuli over others.

Attention can be driven from the top down by intent and volition, or it can be triggered from the bottom up by intrinsic properties of the stimulus that make the stimulus highly noticeable, or salient (Itti and Koch, 2001; Connor et al., 2004). In a traffic situation, for example, we may decide to pay attention to street signs or to the traffic report on the radio, but the siren and flashing lights of an approaching ambulance will nevertheless immediately grab our attention.

As a mechanism of deliberate, goal-directed orienting of our senses top-down attention reflects our longer-term cognitive strategy. For instance, in preparing a lane change we will pay special attention to traffic from behind in the rear view mirrors and even orient our head to the side to look over our shoulder before initiating the lane change. Bottom-up attention, on the other hand, allows us to react to salient or surprising stimuli (Itti and Baldi, 2006), whether they are an attacking predator or a pedestrian jumping in front of our car.

In fact, many signals in our environment are designed in such a way that they trigger our bottom-up attention system. For instance, flashing lights are used to attract our attention to a waiting message on the answering machine or to another driver's intention to make a turn. Salient sounds trigger our bottom-up attention when we forget to take the cash from the ATM or when a fire alarm is wailing at a volume that is impossible to ignore. In many cases, sounds are better suited to attract our attention than visual stimuli, because we do not need to be oriented toward them in order to perceive them, and, unlike our eyes, our ears are never shut.

In the visual domain, bottom-up salience is believed to be driven by a number of low-level features, such as local color and luminance contrasts and oriented edges (Koch and Ullman, 1985). Contributions to stimulus salience from these features are combined into a saliency map (Itti et al., 1998), which is then used to guide sequential scanning of a scene, in order to serialize perception of individual objects (Walther and Koch, 2006). Attempts have been made to apply a similar concept to the auditory domain, e.g., by computing a visual saliency map of the spectrogram of an auditory stimulus with slightly adapted features (Kayser et al., 2005; Kalinli and Narayanan, 2009; Kalinli et al., 2009; Segbroeck and Hamme, 2010).

☆ This paper has been recommended for acceptance by S. Sarkar.

* Corresponding author. Tel.: +82 31 301 3838.
E-mail address: kktahn.kim@samsung.net (K. Kim).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.11.010

Salience in both the visual and the auditory domains can be loosely described as something being different from its immediate neighborhood (Kayser et al., 2005; Kayahara, 2005; Coensel et al., 2009). Here, neighborhood can be understood in the sense of space, time, frequency, or any other feature space. However, beyond these superficial similarities, there are important differences between visual and auditory salience. For instance, auditory events often overlap in time. Segregation of overlapping sounds is much harder than segmenting visual objects from an image. Furthermore, acoustic signals are processed continuously in real time. This has implications for the speed of processing as well as the shape of filters that are used. Filters in the time domain need to be asymmetric, because they can only use current and past but not future parts of the signal. In the visual domain, on the other hand, image space is typically assumed to be isotropic, leading to symmetric filters. This serves to illustrate that the detection of auditory salience is more complicated than applying visual salience detection to a graphical representation (e.g., a spectrogram) of an audio signal.

To our knowledge no systematic effort has been made to identify which features are essential for auditory salience. In this paper we use a data-driven approach to this issue. Based on the annotations of salient audio events by human participants we derive the optimal filter for auditory salience. To this end we have to solve several issues: (i) we have to devise a protocol for the annotation of a large corpus of audio data; (ii) we have to acquire a sufficient number of annotations to allow inference to the salience of audio events; (iii) we have to separate the effects of stimulus-driven bottom-up attention (salience) from task-driven top-down attention (expectation); and (iv) we have to develop a detection algorithm, which can consider time–frequency variations of acoustically sensed signals in an efficient way. We report solutions to all four problems in the following sections.

2. Establishing ground truth for auditory salience

One of the major reasons holding back research on auditory salience is the difficulty of acquiring and interpreting ground truth data from human observers. Kayser et al. attempted to measure audio salience by asking human subjects to choose the more salient out of two sounds (Kayser et al., 2005). They used natural sounds such as animal sounds with additional noise to eliminate any top-down semantic associations.

Unlike Kayser's study, we here consider signals in which both top-down and bottom-up attention allocation processes may be active. This leaves us with the conundrum of separating top-down from bottom-up contributions. In our approach to this problem we measure inter-observer agreement. If transcribers are not told to listen for any specific class of audio events, then their cognitive models of the task should vary somewhat from transcriber to transcriber, and therefore their expectation-driven (top-down) attention allocation should also vary. Audio events that are noticeable due to top-down attention should vary across individuals much more than events that are salient due to bottom-up factors, which should be more uniform among people. In other words, any sound may catch the attention of someone who is listening for it, but a more salient sound should catch the attention of more transcribers than a less salient sound.

By its design this approach cannot distinguish between situation-driven attention that is shared among most individuals and purely stimulus-driven salience. However, the distinction between top-down and bottom-up attention defined in this manner has been used successfully in the investigation of visual attention (Einhäuser et al., 2007). We therefore adopt the discrimination of detected events into observer-general and observer-specific components as an operational definition of bottom-up and top-down attention for this work.

2.1. Salience annotation

We used the AMI Meeting Corpus (AMI project, date last viewed 7/15/2010,) to investigate audio salience. The AMI corpus was designed by a 15 member multi-disciplinary consortium dedicated to the research and development of technology that will help groups interact better. It consists of 100 h of recordings of meetings and includes close-talk and far-field microphones, individual and room-view video cameras, and output from a slide projector and an electronic whiteboard. The meetings were recorded in English using three different rooms with different acoustic properties and include mostly non-native speakers of English. The dialogues in the corpus are usually designed to capture completely natural and uncontrolled conversations.

Various unpredictable acoustic events and background noise in the corpus provide us with a diverse acoustic scene, which is important to cover the range of potential auditory events as widely as possible. Naturally, the AMI corpus does not cover all possible acoustic scenes, but it shows more variations in human interactions in a real environment than any other available database.

We arbitrarily selected 12 h of recordings from the AMI corpus. We mixed the recordings from the microphone arrays in the selected sessions into one recording. We then asked 12 annotators to listen to the recordings and annotate salient passages per mouse click in a custom interface. Participants were given the following instructions:

"Imagine that you were in the conference room you are listening to. You might focus on the conversation between members in the conference room or not. During listening you should mark the moment when you hear any sound which you unintentionally pay attention to or which attracts your attention. The sound might be any sound, including speech."

We intentionally gave as little guidance as possible about the nature of the acoustic events that should be annotated in order to minimize top-down influences. The annotation resulted in a binary signal, where 1 denoted "noticeable" and 0 "not-noticeable" sounds. All 12 participants annotated all 12 recordings. Annotations of the same recording were combined by summing the binary signals, resulting in annotation scores in $\{0, 1, \ldots, 12\}$. To account for variations among subjects in the precise start and end times of annotated events we re-aligned the annotations before summation. We set starting points to the earliest among the salience annotations (marked as 1) and ending points to the latest for each annotation event, which ensures that the annotation event contains the acoustic signal that captured annotator attention.

Observing the summed annotation scores, acoustic events selected as salient include pulling up chairs, slamming a door, and footsteps. Vocal sounds like coughing and laughing are sometimes annotated with high scores. In spite of their low sound pressures, some quiet sounds such as tapping a mouse on a desk get high score, whereas annotations for some loud sounds such as loud speech are less consistent across participants. Acoustic events with medium scores tend to be semantically similar to those with high scores (e.g., laughter sometimes receives high scores, but sometimes receives only medium scores). The semantic overlap between the high-score region and medium-score region suggests that not all sounds in the medium-score region are the objects of top-down attention allocation; instead, it may be the case that sounds that are perceptually salient, but with a lower degree of saliency, might receive salience labels from only a subset of the annotators, and might therefore wind up in the medium-score region. This reasoning suggests that, while high-scoring sounds are salient, not all salient sounds are high-scoring.

2.2. Classification of salient audio events

We classify annotated audio events into mostly top-down and mostly bottom-up according to varying amounts of inter-observer consistency. Events that are annotated by many observers are, presumably, salient. Events that are annotated by only one or very few participants are presumably not salient, but rather, have attracted the attention of a few annotators because of some attribute of the top-down attention allocation schemes employed by those individual annotators. Finally, there are parts of the audio signal that draw nobody's attention.

In order to formalize this distinction, we define three types of audio events as follows:

A_0 : not noticeable
A_1 : noticed by those who listen for it (top-down attention allocation)
A_2 : noticed by almost everyone(salient)

It is reasonable to assume that the three event types, A_0, A_1, and A_2, occur exclusively, and that they cover all possible cases, that is $\Pr(A_0) + \Pr(A_1) + \Pr(A_2) = 1$.

The distribution of annotation scores can be expressed as:

$$\Pr(k) = \Pr(k|A_0)\Pr(A_0) + \Pr(k|A_1)\Pr(A_1) + \Pr(k|A_2)\Pr(A_2), \quad (1)$$

where k represents the number of votes, $k \in \{0, \ldots, 12\}$.

If we assume for simplicity that individuals vote independently of each other, then the distribution of the annotation scores is a mixture of binomial distributions:

$$\Pr(k) = \sum_{j=0}^{2} \pi_j B(k; N, p_j), \quad \{\pi_j \equiv p(A_j)\}, \quad (2)$$

where $B(k; N, p)$ represents a binomial distribution with the parameters N and p. Provided a sufficient number of votes to derive a reliable histogram, the best distribution parameters π_j^* and p_j^* can be obtained by fitting (2) to the histogram. We then use the estimated distribution to derive salience evaluations of the audio signal.

2.3. Applying the statistical model to salience annotations

The parameters of Eq. (2), π_j^* and p_j^*, can be estimated from annotation scores using Expectation–Maximization (EM) (Dempster et al., 1977). The details of the derivation of the distribution parameters are described in Appendix A. The distributions with the parameters derived through the EM algorithm together with the histogram of the annotation scores are shown in Fig. 1.

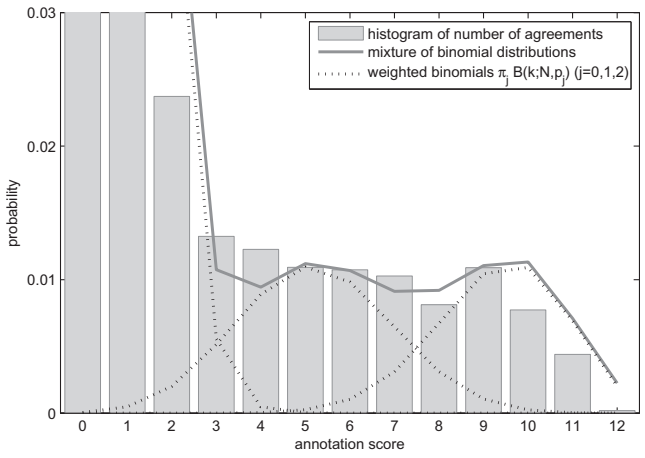

Fig. 1. Histogram and binomial mixture model of the annotation scores of audio events (annotation score = number of annotators who reported attending to the event).

As shown in Fig. 1, the mixture of binomial distributions matches the histogram well. The formal boundary between salience and non-salience can be obtained from the estimated mixture of binomial distributions. Here we define a salient event as $\omega_0 = \{A_2\}$ and a non-salient event as $\omega_1 = \{A_0 \bigcup A_1\}$. The minimum probability at the optimal error threshold k^* is defined by

$$\Pr(\omega_1|k) \begin{cases} \geqslant & \Pr(\omega_0|k) \quad k \geqslant k^* \\ < & \Pr(\omega_0|k) \quad k < k^* \end{cases}, \quad (3)$$

where $\Pr(\omega_0|k) \propto \pi_2 B(k; N, p_2)$, and $\Pr(\omega_1|k) \propto \pi_0 B(k; N, p_0) + \pi_1 B(k; N, p_1)$. As shown in Fig. 1, the resulting optimal threshold is $k^* = 8$.

3. Automatic detection of acoustic salience

Based on the collected annotations, we have developed a method to automatically detect auditory salience. We adopt a linear discriminant approach, which requires less training data than other more complex detection algorithms. Although this approach is fairly simple, it can nevertheless demonstrate viability of auditory salience detection.

As a first processing step we apply a time-frequency (Bark) invariant linear filter to the feature space (Zwicker and Fastl, 1998). Bottom-up attention is commonly believed to be triggered by rapid changes in the sound signal, which means that humans catch a salient piece of sound by noticing variations in the prevalent sound-scape over a short time. A linear time-invariant frequency-invariant filter is sufficient to emphasize consistent variation in the short-time auditory feature sequence, hence it is sufficient to represent the known attributes of salience detection. Because it has fewer model parameters than most other classifiers, a time-frequency invariant linear filter is also useful to prevent over-training. More details are described below.

3.1. Linear discrimination for salience filtering

In this work, we consider a simple linear discriminant for detecting auditory salience, which can be described as:

$$\begin{aligned} \omega_0 &: \mathbf{w}^T \mathbf{y}_t > b, \\ \omega_1 &: \mathbf{w}^T \mathbf{y}_t < b, \end{aligned} \quad (4)$$

where ω_0 and ω_1 represent salient and non-salient events, and b is a threshold. The vectors $\mathbf{w}$ and $\mathbf{y}_t$ are a weight vector for the linear discriminant and a feature vector at time t, respectively. The feature vector y_t is generated from cochlear output $x(i,t)$, where i is frequency (in Bark) and t is time. The mapping from $x(i,t)$ to y_t is computed in a data-driven fashion, using as few assumptions as possible. In order to limit the number of trainable parameters (and thereby avoid overtraining), the mapping is assumed to be a linear time-invariant filter, which we will call the salience filter. The salience filter is applied to cochlear output to transform it into a salience signal. It is reasonable to assume that the salience filter affects a finite duration of cochlear output, because bottom-up attention is triggered within 60–80 ms of an event (Schneider and Shiffrin, 1977). Therefore, we express the salience feature $\mathbf{y}_t$ as:

$$\mathbf{y}_t = \mathbf{Y}_t \mathbf{h}, \quad (5)$$

such that

$$\mathbf{Y}_t(i, (2V+1)j + k + 1) = x(i - U + j, t - V + k), \quad (6)$$

$$\left\{\begin{array}{l} 1 \leqslant i \leqslant D \\ 0 \leqslant j \leqslant 2U \\ 0 \leqslant k \leqslant 2V \end{array}\right\},$$

where each row of $\mathbf{Y}_t$ contains a neighborhood of $(2U+1)\times(2V+1)$ elements centered at $x(i,t)$ and the vector $\mathbf{h}$ contains the coefficients of the salience filter whose dimension is $(2U+1)\times(2V+1)$. In this work, the perceptual loudness used in Perceptual Evaluation of Speech Quality (PESQ) is adopted to express the cochlear output signal, $x(i,t)$ which is the loudness in the ith half-Bark frequency bin in the tth frame (ITU-T Recommendation P.862, 2001). Eq. (5) allows our model of salience to consider the loudness in D different frequency bands and in $(2V+1)$ consecutive frames, without requiring us to train a $D\times(2V+1)$ – dimensional linear discriminant model. High-dimensional linear discriminant models often fail to generalize from training data to novel testing data. The two-step salience filtering model described in Eqs. (4) and (5) reduces the dimension of the learning problem from $D\times(2V+1)$ to $D+(2U+1)\times(2V+1)$, by making the very strong assumption that the salience filter is frequency-invariant in the Bark frequency domain. The assumption of a frequency-invariant salience filter is neither supported nor refuted by any available psychological evidence, but it is at least consistent with the assumptions made in Kayser et al. (2005). Here we design the salience filter h to cover ±1 Barks ($U=2$). For AMI recordings sampled at 16 kHz, the PESQ perceptual loudness vector (ITU-T Recommendation P.862, 2001) has a dimension of $D=49$, and is extracted every 4 ms. In order to cover a short-term memory window of 60 ms, we want the salience filter to have the length $(2V+1)=15$ frames, resulting in a total linear discriminant dimension of $D+(2U+1)\times(2V+1)=124$ dimensions while the number of acoustic feature elements used for a single decision unit is $D\times(2V+1)=735$.

3.2. Cost function for the linear discriminant

As used in general automatic detection, the linear discriminant is optimized in a minimum mean square error sense:

$$F(\mathbf{w},\mathbf{h},b)=\frac{1}{T}\sum_{t=0}^{T-1}(a_t-\hat{a}_t)^2 \tag{7}$$

$$\hat{a}_t=f(\mathbf{w}^T\mathbf{y}_t-b) \tag{8}$$

$$=f(\mathbf{w}^T\mathbf{Y}_t\mathbf{h}-b), \tag{9}$$

where T is the total number of the training samples and $a_t=1$ or 0 means salience or non-salience, respectively. The binary decision function f can be approximated by a sigmoidal function. Since the scales of $\mathbf{h}$ are redundant, we also assume that the salience filter $\mathbf{h}$ has unit energy, without loss of generality. The salience filter $\mathbf{h}$ and the linear discrimination weighs $\mathbf{w}$ are obtained by solving the following nonlinear optimization problem:

$$\min_{\mathbf{w},\mathbf{h}} F(\mathbf{w},\mathbf{h},b)\ s.t.\ \|\mathbf{h}\|=1. \tag{10}$$

Various nonlinear programming methods can solve the problem (10). Here we use the projection method introduced by Gill and Murray (1974). Details of the derivation are given in Appendix B. Applying the projection method to problem (10) gives the following iterative equations:

$$\left\{\begin{array}{ll} \mathbf{w}_{(k+1)} & =\mathbf{w}_{(k)}-\frac{1}{\mu}\nabla_{\mathbf{w}}F(\mathbf{w}_{(k)}) \\ \mathbf{h}_{(k+1)} & =\mathbf{h}_{(k)}-\frac{1}{\mu}\nabla_{\mathbf{h}}F(\mathbf{h}_{(k)}) \\ \mathbf{h}_{(k+1)} & \Leftarrow\frac{\mathbf{h}_{(k+1)}}{\sqrt{\mathbf{h}_{(k+1)}^T\mathbf{h}_{(k+1)}}} \end{array}\right\}. \tag{11}$$

The iterative optimization consists of gradient descents for the two variables, $\mathbf{h}$ and $\mathbf{w}$ and an energy normalization for $\mathbf{h}$, which appears to be a reasonable solution considering the unit energy constraint.

4. Experimental results and discussion

We experimentally test our linear discriminant solution with the salience annotation data from Section 2. We compare its performance with a salience map based on computer vision techniques (Kayser et al., 2005), which, to our knowledge, is the only published solution for automatic acoustic salience detection so far (and has been used to measure auditory salience in other studies (Kalinli and Narayanan, 2009; Kalinli et al., 2009; Segbroeck and Hamme, 2010)). It is also instructive to inspect the resulting salience filter, because it may help to interpret how humans process auditory salience.

4.1. Salience detection with Kayser's saliency map

Kayser et al. proposed an auditory salience map (Kayser et al., 2005) adopting an existing visual salience detection method (Itti et al., 1998). Their computer-vision based auditory (CVBA) salience map extracts several acoustically salient features in parallel, representing various levels of sound feature analysis by auditory neurons. In their work, sound intensity, frequency contrast, and temporal contrast are used as acoustic salience features. Evidence for the salience of each feature is compared across scales with a center-surround mechanism. To obtain a feature-independent scale, these time-frequency maps for the features are normalized using an asymmetric sliding window, which reflects psychoacoustic temporal masking effects. Finally, the salience maps from individual features are combined in analogy to the idea of feature integration (Treisman and Gelade, 1980).

The CVBA saliency map shows salience in the time-frequency domain. For our task, discrimination happens only in the time domain. A frequency vector in the saliency map at each time slot needs to be summarized to a single value. In this experiment, a linear discrimination rule is applied to the salience vector in the frequency domain to get a discrimination criterion.

$$\begin{array}{l} \omega_0:\mathbf{v}^T\mathbf{m}_t>c \\ \omega_1:\mathbf{v}^T\mathbf{m}_t<c, \end{array} \tag{12}$$

where ω_0 and ω_1 are salient and non-salient events, respectively. The parameters $\mathbf{v}$ and c are a weight vector and a threshold for the linear discrimination, respectively and $\mathbf{m}_t$ is a feature vector which consists of frequency elements at time t in Kayser's saliency map. The linear discriminant for Kayser's saliency map is trained in a squared error sense just like the proposed method; the two methods differ only in the computation of $\mathbf{y}_t$ and $\mathbf{m}_t$. The CVBA map as used here is down-sampled by a factor of 2 to achieve a temporal resolution of 4 ms, and it has 128 dimensions in the frequency domain.

4.2. Performance evaluation

As explained in Section 2, we used the AMI corpus for automatic auditory salience detection. We collected annotation data for approximately twelve hours of recordings, which were arbitrarily selected from the AMI corpus. (Selected sessions for the experiments: EN2001a, EN2001b, EN2001d, EN2001e, EN2002a, EN2002b, EN2002c, EN2002d, EN2003a, EN2009b, ES2002c, ES2002d, ES2003a, ES2003d, ES2004b, ES2007a, ES2007c, ES2007d, ES2008a, ES2008b, ES2008c, ES2008d, ES2009a, ES2009c, ES2009d, ES2013c, ES2015a, IB4002, IN1002, IN1007,

IS1002b, IS1002d, IS1003a, IS1004a, IS1004c, IS1004d, IS1006a, IS1008b, IS1008d, IS1009b, IS1009c.) First, the annotation data are divided into six segments of about 2 h duration. Five out of the six segments are used for training to estimate the salience filter and the discrimination weight, and the remaining one is used for the detection test with the discrimination parameters obtained during the training. The same procedure is repeated six times for cross validation. The results are combined over the 6 combinations. To measure how the proposed algorithm performs, Detection Error Tradeoff (DET) curves are drawn in Fig. 2. The DET results for the six cross validation folds are combined using the method proposed in Adler and Schuckers (2005). The black solid line represents the method proposed in this paper, and the dotted line the linear discrimination with the computer-vision based auditory salience map (CVBA). At equal error rate, the proposed method outperforms CVBA by approximately 8.7% under the linear discrimination assumption. The performance enhancement is statistically reliable at a 99% confidence level (McNemar's test with Yates' correction) as shown in Table 1(a) (Gillick and Cox, 1989). The gray solid line represents the performance of a linear discrimination on the perceptual loudness features without the salience filter (NOSF), which can tell whether the salience filter contributes to the overall performance above mere loudness detection. As shown in Fig. 2, the salience filter enhances discrimination performance by 2.5% at equal error rate. The performance enhancement is significant at the 99% confidence level as shown in Table 1(b).

These results show that temporal variations in the auditory signal as detected by the proposed salience filter method can affect acoustic salience. Note that the CVBA performs poorly, even compared to mere loudness detection (NOSF). The difference between these two methods is determined by the feature vector used for linear discrimination. Despite having more features for linear discrimination, the CVBA approach of applying visual salience to the time-spectral map is more likely to obscure acoustic salience characteristics than to build up an acoustic salience map. Even though cognitive mechanisms for auditory salience might resemble those for visual salience, we have to conclude that a linear discriminant compression of the CVBA salience map does not match our annotation data.

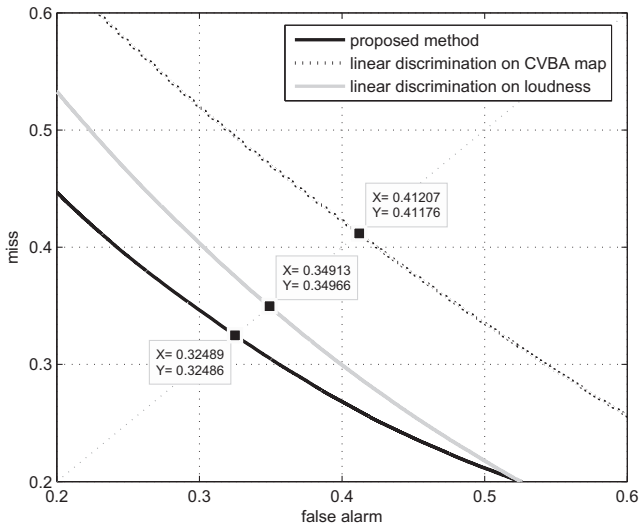

Fig. 2. Detection Error Tradeoff curves for the proposed method and for detection with the CVBA map.

4.3. Observation of the salience filter

It is also instructive to inspect the salience filter that we obtained through the training process, since it may give us insights into how humans detect salient acoustic stimuli. Fig. 3 shows the impulse response of the salience filter in the time-frequency (Bark) domain. From the shape of the impulse response, both temporal and frequency contrast detectors can be observed. The temporal contrast detector enhances onsets. The frequency contrast detector enhances high-frequency vs. low-frequency energy, though the utilization of frequency contrast may be altered or countered in each particular frequency bin because the linear discriminant vector $\mathbf{w}$ differentially weights the outputs of the function $\mathbf{y}_t = \mathbf{Y}_t\mathbf{h}$. The temporal shape of the impulse response is almost the same across different Bark frequency values with little variation.

The consistency of temporal shapes of the impulse response across Bark frequencies in Fig. 3 suggests independent processing of each of the half-Bark frequency bands. We run another test to compare spectral-context-dependent vs. spectral-context-independent salience filters. The size of the context-independent salience filter is set to 1 by 15 in frequency by time units. The result is shown in Fig. 4. The context-independent salience filter (dotted line, 1×15) performs slightly better than the context-dependent one (solid line, 5×15). Except at band edges, the 5×15 filter should be capable of duplicating the computations of the 1×15 filter, therefore it is worthwhile to list possible reasons for its slightly inferior performance. (1) The context-dependent case requires the estimation of more parameters than the context-independent case. Therefore, using the same amount of training data, the context-independent model may generalize better to the test data, while the context-dependent model may be more prone to overfitting to the training data. (2) The context-dependent filter has some degradation at the upper and lower edges of the Bark frequency spectrum. At the ends of the Bark spectrum, the saliency filter is fed mirrored values beyond the Bark range, which is apparently suboptimal.

4.4. Discrimination with conventional acoustic features

We have also tested the possibility that salience detection may depend on other acoustic features. For this purpose, we concatenated

Table 1
Contingency table of the acoustic salience detection results for McNemar's test statistic with Yates' correction.

	CVBA: correct	CVBA: incorrect
(a) Proposed method vs. linear discriminant on the computer-vision based auditory salience map (CVBA) (Kayser et al., 2005)		
Proposed method: correct	613,166	314,357
Proposed method: incorrect	199,485	240,556
		$\chi^2 = 2.6 \times 10^4$
	NOSF: correct	NOSF: incorrect
(b) The proposed method vs. loudness without salience filter (NOSF)		
Proposed method: correct	762,169	165,354
Proposed method: incorrect	156,525	283,516
		$\chi^2 = 2.4 \times 10^2$

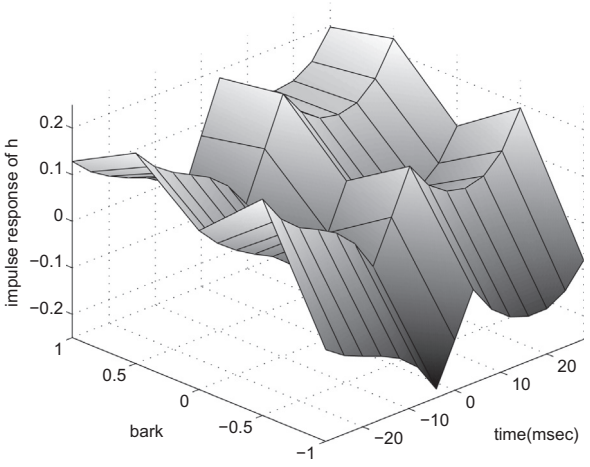

Fig. 3. Estimated impulse response (**h**; Time-Bark plot).

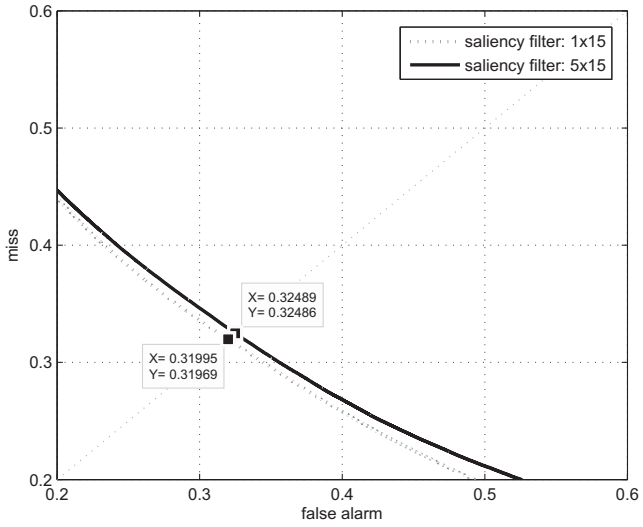

Fig. 4. Detection Error Tradeoff curves for the proposed method with 1 by 15 and 5 by 15 saliency filters.

four non-linear features: (1) zero-crossing rate, (2) spectral flatness measure ($\int log(S(f,t))df - log(\int S(f,t)df)$, where $S(f,t)$ is spectral energy at frequency f and time t) (Niyogi et al., 1999), (3) pitch (T_0), and (4) pitch prediction coefficient ($R(T_0)/R(0)$, where $R(\cdot)$ is the autocorrelation of the acoustic signal). These features cannot be linearly calculated from the loudness, but they are commonly used as acoustic features in speech signal processing; pitch and pitch prediction, in particular, are probably computed by the auditory brainstem based on phase locking on the auditory nerve (Licklider, 1951; Cariani, 1999). Care should be taken when combining these features with loudness features because the salience filter reflects spectral relation. For instance, a zero-crossing rate (ZCR) attached near the first Bark band of the loudness would result in a different error rate from a ZCR attached near to the last Bark band. To eliminate dependency on the location of the second feature, we applied only the spectral-context-independent saliency filter. The temporal length of the saliency filter used in this comparison is set to 60 ms (1×15). The EERs (Equal Error Rates) for the combinations of loudness features with non-linear features (1)–(4) are shown in Table 2. Note that the discrimination performance can get worse even though more features are added since all the training and test sets are different. As shown in Table 2, none of the feature combinations outperform discrimination based only on loudness. The fact that the same saliency filter is applied for all frequency elements in our approach means that our approach could cause poorer performance when the additional features have different temporal characteristics. Even if the additional features are related to acoustic salience, the combination of the feature and the loudness could perform worse than using just one of them. This fact tells us that the results in Table 2 do not mean that the additional features, zero-crossing rate, spectral flatness, pitch and pitch prediction coefficient, have no connection to auditory salience. The combination of the loudness and zero-crossing rate outperforms the other combinations in Table 2. This should be interpreted to mean that the linear relation between loudness and auditory salience is closer to that between the zero-crossing rate and auditory salience than the other features.

Table 2
Equal Error Rate for linear discriminations with the feature combinations.

Features	Dimension	Equal error rate
Proposed method (loudness)	49	0.3198
Loudness + zero-crossing rate	50	0.3271
Loudness + spectral flatness	50	0.3958
Loudness + pitch (T_0)	50	0.4345
Loudness + $R(T_0)/R(0)$	50	0.4313
Loudness + all the features	53	0.3922
MFCC	13	0.3446

We also tested Mel Frequency Cepstral Coefficients (MFCC) as an input to the linear discrimination. The discrimination result is shown in the bottom of Table 2, and is poorer than that produced using loudness. There are at least two possible explanations for this result. First, the MFCC may have less relation with auditory salience than loudness. Second, the consistency of the salience filter in the Bark domain may be better than in the cepstral domain.

4.5. Selection of threshold for statistical purification

One of the crucial steps in this work is to build up ground truth for auditory salience by thresholding polling data. The question arises how the choice of the threshold affects the discrimination results. So far we have used a threshold of 8 out of 12 annotations for salience. What happens if we choose other values? We experimentally tested other threshold values, and the results are shown in Fig. 5. For this experiment, we did not segment the annotated data into two parts but rather used all the data for training. Note

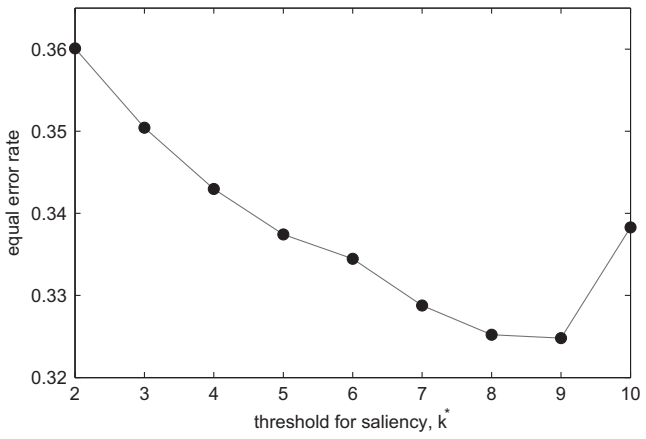

Fig. 5. Equal error rate varying thresholds for the number of salience annotations.

that therefore the performances differ somewhat from those in Fig. 2. For simple comparison, equal error rates are compared across the thresholds of 2–10. As the threshold increases, the equal error rate decreases all the way to 9 and then increased again for 10. The minimum error rate is hit at 9, but the difference between 8 and 9 is small.

5. Conclusions

This paper proposes a method for collecting annotations of salience in auditory data. A mechanism is derived for the automatic detection of auditory salience based on linearly filtering a loudness chunk of 60 ms duration. The salience filter is extracted from the annotation data using linear discrimination. The filter is shaped like an onset detector in time, which agrees with intuition about the way humans process auditory signals. The proposed salience detection algorithm performs at 68.0% accuracy. These results show that it is possible to detect salient acoustic events automatically. More sophisticated detection mechanisms than linear discriminant analysis may give even better results in the future.

The ability to detect salient auditory events may find applications in auditory surveillance. In many situations, microphones are less expensive to deploy than cameras, allowing for more comprehensive coverage. Furthermore, since detection of auditory salience is faster to compute than image-based visual salience, it may serve as a first warning system, triggering subsequent detailed analysis of an auditory and/or visual feed by more specialized algorithms or by a human operator.

Appendix A. Expectation–Maximization (EM) algorithm for the sum of multiple binomial distributions

This appendix develops EM equations for the sum of three binomial distributions, which represents the distribution of polling scores by human subjects during the salience annotation. The general equations of the EM algorithm are given as

$$\text{Expectation step}: \quad Q(\theta|\theta^{(j)}) = \mathbb{E}_{\mathbf{Z}|\mathbf{k},\theta^{(j)}}[\log \mathcal{L}(\theta;\mathbf{k},\mathbf{Z})],$$
$$\text{Maximization step}: \quad \theta^{(j+1)} = \arg\max_{\theta} Q(\theta|\theta^{(j)}),$$

where $\theta^{(j)} = \{\pi^{(j)}, p^{(j)}\}$ at the jth iteration and $\mathbf{Z}$ is the latent variable that determines the component from which the observation $\mathbf{k}$ originates. In our problem, the likelihood $\mathcal{L}(\theta;\mathbf{k},\mathbf{Z})$ is expressed with the binomial mixture distribution:

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}|\mathbf{k},\theta^{(j)}}[\log \mathcal{L}(\theta;\mathbf{k},\mathbf{Z})] &= \mathbb{E}_{\mathbf{Z}|\mathbf{k},\theta^{(j)}}\left[\log\left[\prod_{t=0}^{T-1}\sum_{i=0}^{M-1}\mathbb{I}(Z_t=i)\cdot\pi_i B(k_t;p_i)\right]\right]\\ &= \sum_{t=0}^{T-1}\sum_{i=0}^{M-1}\eta_{i,t}^{(j)}\cdot\log\pi_i B(k_t;p_i)\\ &= \sum_{t=0}^{T-1}\sum_{i=0}^{M-1}\eta_{i,t}^{(j)}\cdot\left[\log(\pi_i)+\log\left(\frac{N!}{k_t!(N-k_t)!}\right)\right.\\ &\quad \left.+k_t\log p_i+(N-k_t)\log(1-p_i)\right], \end{aligned} \tag{A.1}$$

where

$$\eta_{i,t}^{(j)} = \mathbb{E}_{\mathbf{Z}|\mathbf{k},\theta^{(j)}}[\mathbb{I}(Z_t=i)] = \Pr(Z_t=i|k_t;\theta^{(j)}) = \frac{\pi_i^{(j)}B(k_t;p_i^{(j)})}{\sum_{l=0}^{M-1}\pi_l^{(j)}B(k_t;p_l^{(j)})}$$

and $\mathbb{I}$ is an inidicator function. The total number of votes, N is dropped off from the notation of the binomial distribution $B(k;N,p)$ for convenience. The linear form of $Q(\theta|\theta^{(j)})$ in (A.1) means that the values of π and p can be maximized independently of each other.

$$\boldsymbol{\pi}^{(j+1)} = \arg\max_{\boldsymbol{\pi}}\sum_{l=0}^{M-1}\left[\log\pi_l\sum_{t=0}^{T-1}\eta_{l,t}^{(j)}\right]$$
$$p_i^{(j+1)} = \arg\max_{p_i}\left[\sum_{t=0}^{T-1}\eta_{i,t}^{(j)}k_t\right]\log p_i+\left[\sum_{t=0}^{T-1}\eta_{i,t}^{(j)}(N-k_t)\right]\log(1-p_i).$$

Solving for π_i under the constraint that $\sum_i\pi_i=1$, we get:

$$\pi_i^{(j+1)} = \frac{\sum_{t=0}^{T-1}\eta_{i,t}^{(j)}}{\sum_{l=0}^{M-1}\sum_{t=0}^{T-1}\eta_{l,t}^{(j)}}. \tag{A.2}$$

In an analogous manner, the second parameter p_i becomes:

$$p_i^{(j+1)} = \frac{\sum_{t=0}^{T-1}\eta_{i,t}^{(j)}k_t}{\sum_{t=0}^{T-1}\eta_{i,t}^{(j)}k_t+\eta_{i,t}^{(j)}(N-k_t)} = \frac{\sum_{t=0}^{T-1}\eta_{i,t}^{(j)}k_t}{\sum_{t=0}^{T-1}\eta_{i,t}^{(j)}N}. \tag{A.3}$$

Appendix B. Numerical optimization of a nonlinear function with an equal power constraint

In this appendix we derive iterative equations to solve the optimization problem for the salience filter and the discriminant weight. The solution derived here is a special case of the projection method for nonlinear programming, which is described in Gill and Murray (1974). First our nonlinear optimization problem is rephrased as

$$\min_{\mathbf{w},\mathbf{h}} F(\mathbf{w},\mathbf{h}) \ s.t. \ \|\mathbf{h}\| = 1.$$

Letting $\mathbf{z} \equiv [\mathbf{w};\mathbf{h}]$ for convenience a solution $\hat{\mathbf{z}}$ of the problem corresponds to a stationary point of its Lagrangian function,

$$L(\mathbf{z},\lambda) = F(\mathbf{z}) - \lambda(\mathbf{h}^T\mathbf{h}-1).$$

Thus we have

$$\left\{\begin{array}{rl}\nabla L(\hat{\mathbf{z}},\lambda) & = 0\\ \mathbf{h}^T\mathbf{h} & = 1\end{array}\right\}.$$

We wish to make local approximations to the functions about a current point $\mathbf{z}$, so we consider the problem of finding the step $\Delta\mathbf{z}$ from this point, which attains the feasible point with the lowest value of the objective function within a distance δ from $\mathbf{z}$:

$$\min_{\Delta\mathbf{z}}\{F(\mathbf{z}+\Delta\mathbf{z}) : (\mathbf{h}+\Delta\mathbf{h})^T(\mathbf{h}+\Delta\mathbf{h}) = 1, \quad \Delta\mathbf{z}^T\Delta\mathbf{z} \leqslant \delta^2\}$$

The solution must satisfy Karush–Kuhn–Tucker conditions:

$$\left\{\begin{array}{l}\nabla F(\mathbf{z}+\hat{\mathbf{z}}) - \lambda'(\mathbf{h}+\Delta\mathbf{h}) + \mu\Delta\mathbf{z} = 0\\ (\mathbf{h}+\Delta\mathbf{h})^T(\mathbf{h}+\Delta\mathbf{h}) = 1\end{array}\right\}, \tag{B.1}$$

where μ is the Kuhn–Tucker multiplier. We choose δ small enough to justify approximating the optimization target function by its first order Taylor expansion, $F(\mathbf{z}+\delta\mathbf{z}) \approx F(\mathbf{z}) + \Delta\mathbf{z}^T\nabla F(\mathbf{z})$. Substituting the approximated target function in (B.1) and splitting (B.1) into $\mathbf{w}$ and $\mathbf{h}$ parts gives

$$\left\{\begin{array}{l}\nabla_{\mathbf{w}}F(\mathbf{w}) + \mu\mathbf{w} = 0\\ \nabla_{\mathbf{h}}F(\mathbf{h}) - \lambda'(\mathbf{h}+\Delta\mathbf{h}) + \mu\Delta\mathbf{h} = 0\\ (\mathbf{h}+\Delta\mathbf{h})^T(\mathbf{h}+\Delta\mathbf{h}) = 1\end{array}\right\}. \tag{B.2}$$

The step for $\mathbf{w}$ is easily obtained from the first equation of (B.2). By substituting $\Delta\mathbf{h}$ from the second equation of (B.2) into the third equation, we get the Lagrange multiplier λ':

$$\lambda' = \mu - \sqrt{(\mu\mathbf{h}-\nabla_{\mathbf{h}}F(\mathbf{h}))^T(\mu\mathbf{h}-\nabla_{\mathbf{h}}F(\mathbf{h}))}. \tag{B.3}$$

Substituting (B.3) into the second equation of (B.2) gives the final adaptation steps:

$$\left\{\begin{array}{l}\Delta\mathbf{w}=\frac{1}{\mu}\nabla_{\mathbf{w}}F(\mathbf{w})\\ \Delta\mathbf{h}=\frac{\mu\mathbf{h}-\nabla_{\mathbf{h}}F(\mathbf{h})}{\sqrt{(\mu\mathbf{h}-\nabla_{\mathbf{h}}F(\mathbf{h}))^{T}(\mu\mathbf{h}-\nabla_{\mathbf{h}}F(\mathbf{h}))}}-\mathbf{h}\end{array}\right\}. \tag{B.4}$$

Rewriting (B.4) as iterative equations for $\mathbf{w}_{(k+1)}$ and $\mathbf{h}_{(k+1)}$ with $\mathbf{w}_{(k)}$ and $\mathbf{h}_{(k)}$ at the previous step consequently gives the gradient descent optimization solution followed by unit energy normalization of $\mathbf{h}$:

$$\left\{\begin{array}{l}\mathbf{w}_{(k+1)}=\mathbf{w}_{(k)}-\frac{1}{\mu}\nabla_{\mathbf{w}}F(\mathbf{w}_{(k)})\\ \mathbf{h}_{(k+1)}=\mathbf{h}_{(k)}-\frac{1}{\mu}\nabla_{\mathbf{h}}F(\mathbf{h}_{(k)})\\ \mathbf{h}_{(k+1)}\Leftarrow\frac{\mathbf{h}_{(k+1)}}{\sqrt{\mathbf{h}_{(k+1)}^{T}\mathbf{h}_{(k+1)}}}\end{array}\right\}. \tag{B.5}$$

References

Adler, A., Schuckers, M.E., 2005. Calculation of a composite det curve. In: Proceedings of Audio- and Video-based Biometric Person Authentication 2005, pp. 279–288.

AMI project, date last viewed 7/15/2010. Idiap dataset distribution portal: AMI meeting corpus. <http://corpus.amiproject.org>.

Cariani, P., 1999. Neural timing nets for auditory computation. In: Comput. Models Auditory Funct.. IOS Press, Burke, pp. 233–247.

Coensel, B.D., Botteldooren, D., Berglund, B., Nilsson, M.E., 2009. A computational model for auditory saliency of environmental sound. J. Acoust. Soc. Am. 125 (4), 2528.

Connor, C.E., Egeth, H.E., Yantis, S., 2004. Visual attention: bottom-up versus top-down. Current Biol. 14, R850–R852.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. J. R. Stat. Soc. 39 (1), 1–38.

Einhäuser, W., Mundhenk, T., Baldi, P., Koch, C., Itti, L., 2007. A bottom up model of spatial attention predicts human error patterns in rapid scene recognition. J. Vis. 7 (10), 1–13.

Gill, P.E., Murray, W., 1974. Numerical Methods for Constrained Optimization. Academic Press, New York, pp. 29–66.

Gillick, L., Cox, S., 1989. Some statistical issues in the comparison of speech recognition algorithms. In: International Conference on Acoustics, Speech, and Signal Processing, pp. 532–535.

Itti, L., Baldi, P., 2006. Bayesian surprise attracts human attention. Adv. Neural Inf. Process. Syst. 19, 547–554.

Itti, L., Koch, C., 2001. Computational modelling of visual attention. Nat. Rev. Neurosci. 2 (3), 194–203.

Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. 20 (11), 1254–1259.

ITU-T Recommendation P.862, 2001. In: Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, pp. 1–28.

Kalinli, O., Narayanan, S., 2009. Prominence detection using auditory attention cues and task-dependent high level information. IEEE Transactions on Audio, Speech, and Language Processing 17 (5), 1009–1024.

Kalinli, O., Sundaram, S., Narayanan, S., 2009. Saliency-driven unstructured acoustic scene classification using latent perceptual indexing. In: IEEE International Workshop on Multimedia Signal Processing, MMSP'09, pp. 1–6.

Kayahara, T., 2005. Auditory salience defined by frequency difference captures visual timing. In: Perception. Suppl. p. 217.

Kayser, C., Petkov, C.I., Lippert, M., Logothetis, N.K., 2005. Mechanisms for allocating auditory attention: an auditory saliency map. Current Biology 15 (21), 1943–1947.

Koch, C., Ullman, S., 1985. Shifts in selective visual attention: towards the underlying neuronal circuitry. Human Neurobiology 4 (4), 219–227.

Licklider, J.C.R., 1951. A duplex theory of pitch perception. Experientia VII 4, 128–134.

Niyogi, P., Burges, C., Ramesh, P., 1999. Distictive feature detection using support vector machines. In: International Conference on Acoustics, Speech, and Signal Processing, pp. 425–428.

Schneider, W., Shiffrin, R.M., 1977. Controlled and automatic human information processing: 1. detection, search, and attention. Psychological Review 84 (1), 1–66.

Segbroeck, M.V., Hamme, H.V., 2010. Advances in missing feature techniques for robust large vocabulary continuous speech recognition. IEEE Transactions on Audio, Speech, and Language Processing 19 (99), 123–137.

Treisman, A., Gelade, G., 1980. A feature-integration theory of attention. Cognitive Psychology 12 (1), 97–136.

Walther, D., Koch, C., 2006. Modeling attention to salient proto-objects. Neural Networks 19 (9), 1295–1407.

Zwicker, E., Fastl, H., 1998, . second updated ed. Psychoacoustics, Facts and Models second updated ed. Springer, New York, pp. 149–172.