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.

© 2013 Elsevier B.V. All rights reserved.

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.

Download English Version:

https://daneshyari.com/en/article/533894

Download Persian Version:

https://daneshyari.com/article/533894

Daneshyari.com