THE LOUD BIRD DOESN'T (ALWAYS) GET THE WORM: WHY COMPUTATIONAL SALIENCE ALSO NEEDS BRIGHTNESS AND TEMPO

Francesco Tordini, Albert S. Bregman, and Jeremy R. Cooperstock*

McGill University
Montréal, QC, Canada

ABSTRACT

Salience shapes the involuntary perception of a sound scene into foreground and background. A computational model of salience would provide a strong perceptual baseline for the sonification designer. However, there is a lack of ground truth to evaluate the proposed models and to measure their performance with respect to human perception. This paper describes three contributions. First, we introduce a behavioral definition of salience. We describe an experiment based on our definition that tests a corpus of natural communication sounds. Our results suggest that salience is well described by three perceptual dimensions: not only loudness, but also, tempo and brightness. Second, we extract the most significant acoustical features and analyze their relation with salience, as measured by our ground truth. The context effects emerging from our analysis confirm the difference between salience and novelty. Finally, we suggest some necessary characteristics of the computational salience model based on the analyzed features.

1. INTRODUCTION

The design of auditory displays, such as warning systems and mobile assistive technologies, must deal with information delivery using sound, management of attention, and salience. Our long-term objective is to create a tool that assists in sound scene design by predicting salience.

The salience of a sound can be defined as its prominence relative to other sounds or, more generally, with respect to a background. Although the distinction between salience and attention is debated, it is well accepted that salience represents "bottom up" processes while attention deals with "top down", task-driven ones.

Sonification is a subtype of auditory display that uses non-speech audio to present and represent information [1, 2]. For an effective sonification, it is necessary to predict the salience of the sounds that will be used. This is because bottom-up mechanisms, including salience, shape the listener's involuntary organization of the sounds generating the scene [3].

* corresponding author: tord@cim.mcgill.ca. FT is with the Department of Electrical and Computer Engineering and the Centre for Interdisciplinary Research in Music, Media and Technology (CIRMMT) at McGill University, www.cirmmt.org.

To understand the effects of salience on scene perception, we need a computational model that maps a set of acoustical features to the perceived salience of a sound. There are two important challenges to do so: the difficulty of gathering perceptual data (our ground truth), and the selection of the features to be used for salience prediction.

The ground truth has to be collected using behavioral experiments that allow labeling and ranking of a set of sounds based on their perceived salience.

With respect to the second challenge, there is a possibly infinite set of acoustic and perceptual features from which we might choose. Therefore, the ability to predict the salience of a sound using a reduced set of such features is highly desirable. This work addresses both issues, first through an experimental paradigm that extracts ground truth, and second using those experimental results to select features. These features represent the building blocks for our computational model of salience.

2. RELATED WORK

2.1. Salience and sonification

Sonification implicitly deals with salience and the management of attention in its sound design principles and guidelines (see, for example, Hunt et al. [4] or Bakker et al. [5]).

The main themes of the research agenda present in the Sonification Report [1] show very little need for modification after almost two decades of research. Kramer et al. [1] specified sonification as the "transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication or interpretation". The challenges behind the words "relation" and "perceived" used therein still deserve attention from the research community. In fact, the complexity and the importance of taking into account the perceptual and cognitive dimensions when designing sonification systems are well documented [6, 7].

Modern sonification calls for the exploration of the use of natural sounds as a complement, or alternative to metaphoric, iconic ones and the use of designs with "sourcy" environments where real, dynamic sounds are not presented in isolation. The use of natural, environmental sounds is especially interesting when generating immersive, continuous soundscapes. The sonification of continuous data needs an auditory display that can be easily distinguished from the background when necessary, but can also be allowed to fade out of attention, and not be annoying or intrusive when not desired [8, 9]. Iconic, symbolic sounds are often perceived as artificial and their acceptability under prolonged listening conditions is the result of a very careful sound design. Natural

sounds tend to be better accepted: the prediction of their perception would streamline the design cycle of most sonification problems. However, as recently highlighted by Dubus and Bresin [10], there is a lack of perceptual evaluation studies on sonification. This is particularly true if considering complex, natural sounds.

Walker and Kramer [11] used the umbrella term *ecological psychoacoustics* to summarize the extensions to traditional psychoacoustics that would have been crucial for a successful design of auditory displays beyond loudness, masking effects, pitch, etc. Since then the attempts to translate Bregman's principles of auditory scene analysis (ASA) [12] into sonification design rules has been more frequent, although lacking consistency.

The stream-based sonification by Barrass and Best [13] is a good example in this direction. The authors tested and extended the so-called van Noorden diagrams [14] to dimensions other than the fundamental frequency (F0) of simple tones such as brightness, intensity and panning, i.e., interaural level difference (ILD), of noise bursts. They aimed to design sonifications that could control streaming and take listening attention into account by studying galloping sequences. Gossman [15] gives a high level discussion on the limits of simultaneity in sonification.

Salience prediction is also important for applications beyond the field of information display, for example in mobile assistive technologies [16, 17] and warning signals design [18, 19, 20].

2.2. Computational models of salience

Salience and attention are intimately related. Attention has attracted most of the research efforts in cognitive psychology where the leading approach is task-driven, or "top-down" [21, 22]. On the other hand signal driven, or "bottom-up", models come from psychophysics and psychoacoustics [23]. However, these fields fail to deal with the concept of salience, which is shaped by a perceptual rather than a sensory approach. This is the reason why salience does not find an easy placement in the research agenda from a psychological perspective and it is mostly used as a qualitative concept. This may also explain why few perceptual auditory salience models are available. More specifically, a closed loop between modeling, perceptual ground truth and applications is far from being robust for audition, even though a noticeable attempt was made by Kayser et al. [24] who proposed a feature-driven computational model and compared its predictions to the results of two behavioral experiments. Their monaural auditory salience model was based on three feature maps: intensity, frequency and temporal contrast. Even if temporal contrast allows one to put continuity constraints over the temporal envelope, this model builds on monaural intensity maps and therefore can neither capture nor explain effects due to the phase relationship between signal waveforms that permit localization and spatial release from masking. Furthermore, the experiments run by Kayser et al. [24] dealt with monaural, lateralized sounds treated in isolation on a stereo background, and were designed around a detection task with intensity being the only independent factor. However, sounds rarely occur in isolation. In fact, in most natural environments it is unusual to hear a single sound in isolation.

The present work is inspired by that of Kayser et al. [24], but it presents sounds in pairs and in a binaural scenario. We attempted to formalize some criteria to inform the design of a sound corpus that uses natural recordings. We therefore aimed to capture perceptual data that are ecologically more valid.

On the other hand, salience is a "handy", powerful concept from the application point of view, therefore making it interesting to other research communities. To our knowledge, only Slaney et al. [25] addressed auditory salience in a spatial scenario in the context of speech separation and automatic speech recognition (ASR). They introduced the concept of binaural salience as captured by binaural onsets obtained from the differential cross-correlation of the cochlear filter-bank output spikes computed using interaural time differences (ITDs) only. Their work represents a notable evolution with respect to the monaural salience models that were available at that time. Extensions of the monaural algorithm proposed by Kayser et al. [24] add cochlear [26, 27] and loudness models [28] as a preprocessing stage, and pitch as an additional feature. Kalinli [29, 30] uses pitch both for speech tracking purposes and as an added feature to her "auditory gist". With the gist she attempted to introduce a pre-attentive model to be used as preprocessor for ASR applications.

None of the above-mentioned works addressed the problem of gathering the perceptual ground truth data to evaluate their models. They rely, instead, on performance measures defined in terms of automatic (i.e., machine based) speech recognition rates [25, 29].

All current computational models can be regarded as detectors of salient boundaries, or onsets. They implement the concept of "novelty" using principled designs motivated by perceptual studies, as described above, or more general statistical approaches [31]. They all share the same "memory" in that novelty is evaluated using a short time window, typically in the half-second range. They therefore exclude, for example, the possibility of capturing those aspects of salience related to tempo changes. Moreover, these salient-onset detectors are conceived as analysis tools for sound mixtures rather than for the prediction of the foreground/background representation of the sounds populating a scene.

A salience model that is capable of predicting the relative salience of multiple sounds before they are added to an existing auditory scene would be more appropriate within the sonification context, where synthesis-oriented design tools are needed.

The paradigm we present in the next section is simple yet it incorporates many of the items discussed so far, namely almost galloping patterns, dynamic sounds derived from natural ones and spatialization. Our experiments aim to collect data about the perceptual salience of natural sounds used to create synthetic scenes. They therefore represent a good tool for research on sonification from an ecological psychoacoustics standpoint.

3. CAPTURING PERCEPTUAL SALIENCE

We introduce a test pipeline that allows the collection of perceived salience and loudness data from listeners, presented with a pair of sound streams in a binaural scenario. The organization of the framework is illustrated in Fig. 1. *Loudness* is obviously an important component of salience and can overshadow other features. Therefore, we controlled for large differences between sounds by equalizing their level using a loudness-matching test run in a preliminary session. As indicated in Fig. 1, loudness judgments are also verified at the end of the salience battery to evaluate perceptual consistency across different subjects and the impact of residual loudness differences on salience (i.e., foreground/background selection).

The salience battery, in particular the SOAP stage (see Fig. 2), relies on the assumption that after segregation and streaming have occurred, stream selection is a competitive process that makes the

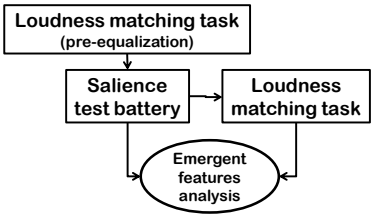

Figure 1: Tests pipeline: perceived loudness is initially used to equalize sounds prior to the salience battery (see Fig. 2) and then (post-battery) to complement detection performance data.

most salient of two concurrent auditory streams more likely to be in the foreground. The salience battery gives us more insight on the relations between response times (RTs) and task complexity, leading to our proposed definition of salience in terms of behavioral response:

> A sound is salient, i.e., belongs to the foreground, when its selection in a complex scene is "as easy" as its detection in isolation, i.e., over silence.

This definition has the advantage of being purely behavioral and independent of the features implied by a particular computational model. We use RT to measure the ease of detection and also to control memory effects on performance.

3.1. Salience battery

Our salience battery (Fig. 2) consists of three consecutive tests. This is done in order to separate the effects of cognitive load from salience effects. We start from the simplest scenario with one sound at a time, presented in a fixed spatial location. In the second step we introduce the spatialization, and in the third step the second sound stream. Test 3 represents a simple approximation of a natural scene and is based on the "streaming of asynchronous sounds patterns" (SOAP) task [32].

3.1.1. Simple detection

Each trial consists of a short sequence of two sounds (K=2) that are presented at random points over time and centered in space. The subject has to detect the first onset of each trial by pressing a key, which determines the simple response time (sRT).

3.1.2. Spatial detection

A symmetrical spatialization is introduced ($\pm$ 15 degrees, on the horizontal plane). The sound sequences are played one at a time, on either side of the head (Test 2 in Fig. 2). The presentation side is fully balanced and the order is randomized for each participant. This is a simple detection task since no competitor streams are present. A trial is defined by an isochronous sequence of sounds. Consecutive trials (groups of k sounds, with $k \in$ 12–14) are separated by 2.5 s of silence and followed by a short noise burst located in front of the subject, acting as "auditory fixation point" and preparing the subject for the next trial. Sounds within each sequence are separated by an interstimulus interval (ISI) of 250 ms. This value was chosen in order to minimize forward and backward loudness masking effects and also to produce patterns with normal tempo values, considering the duration of the stimuli described in Sec. 3.2. The subjects' task is to detect the occurrence of a single shortened ISI in an otherwise isochronous sequence. This interval, represented by a red arrow in Fig. 2, can be as short as 80 ms. Its position within the trial is randomized over time but constrained to take place after 1.6 s from the start of each trial to ensure proper streaming onset. The subject has to indicate the location (L/R) of the detected change by pressing one of two keys. The time to press a key determines the choice response time (cRT).

3.1.3. Spatial discrimination

Two sequences run concurrently, one on each side of the head, in which only one of the two sequences contains the shortened interval. This is a more complex task, based on the SOAP paradigm, with stream competition and higher perceptual load. The time to press a key determines the discrimination response time (dRT). Accuracy scores and RT values are logged for each participant.

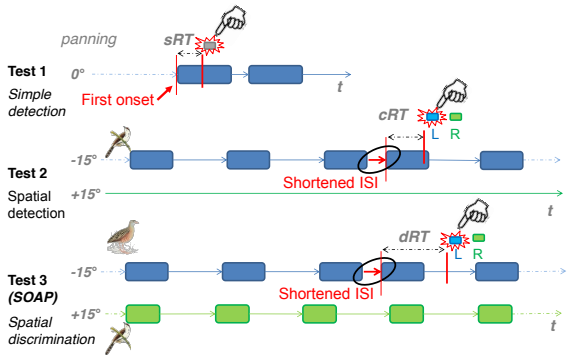

Figure 2: The salience battery starts with the *simple detection* test, which presents a single stream. The *spatial detection* test adds spatialization, and the *spatial discrimination* test a second stream. The red arrows represent the events that the subjects must detect. In the examples shown here, the shortened ISI is presented to the participant's left ear.

The three types of RT are used to measure the participants' responses to tasks with different complexities and then create personalized RT baselines.

With a task such as SOAP, a subject may detect the shortened ISI event equally well for two sounds (i.e., with equal accuracy, or proportion correct, PC), but with different RT values suggesting that, for the slower response, he probably reviewed the recent scene in his working memory to reach a better performance. We used the different RTs defined in Fig. 2 to test the effects of different time windows on the detection dataset (PC). We probed the validity of our operational definition of salience by comparing the unconstrained responses with the trimmed, fast ones. In terms of the RTs defined in Fig. 2 this means that for the trimmed dataset we look for the detections where dRT is close to cRT. The effects of this trimming are presented in Section 4.1.

3.2. Stimuli

Natural sounds may be classified using complex taxonomies according to the meaning to the subject and to their nature (see,

for example, the seminal paper by Gaver [33] or the more recent survey offered by Temko in the context of acoustic event detection [34]). In this work we wanted to avoid mechanical and impact sounds because of their very peculiar temporal structure. We also wanted to use sounds with little semantic content to an average human listener. The class of non-human communication sounds and, more specifically, bird chirps seemed a good candidate corpus within the broader animal sounds class. Bird chirps offer a large choice of temporal and spectral textures while being relatively homogeneous in terms of familiarity (as opposed to a broader selection including other animals like cats and dogs).

Five recordings of bird chirps were taken as starting point (sounds (1,3,5,7,9), average duration 190 ms). Five replicas with longer duration were generated preserving the spectral properties of the original sounds (sounds (2,4,6,8,10), average duration 230 ms). The resulting ten sounds were used for the salience and the perceptual loudness tests of Figure 1. Two "beep-like" laboratory-generated sounds with different duration and spectral centroids (respectively, 100 ms/950 Hz and 300 ms/1400 Hz) were added to the sound corpus and used together with the bird chirps for the first two stages of the salience battery to probe effects of sound category on RTs. The reference sounds used for the loudness-matching tests were not used for the salience battery. All sounds can be downloaded from `http://srl.mcgill.ca/~tord/SOAPsounds/`.

The average tempo of the patterns used for the salience battery (Fig. 2) was 129 bpm: 136 bpm for the five short sounds, and 122 bpm for the corresponding long replicas. The perturbation introduced by the shorter ISI corresponded, on average, to a local glitch of +70 bpm.

3.3. Participants

A pilot group of N=7 volunteers (age = 28 $\pm$ 3, 2 females) was used for the preliminary loudness equalization phase. A separate group of thirty one (N=31) participants (age = 21.7 $\pm$ 2.6, 19 females) participated in the salience experiments. Out of these, 12 were paid and recruited through the McGill classifieds listing while the remaining 19 were McGill undergraduate students compensated with course extra credit. They all reported normal hearing.

3.4. Design, materials and apparatus

A within-subjects full factorial design was utilized with sound type, presentation side and ISI value being the independent factors. Each participant ran the two preliminary test blocks in Fig. 2 to assess his RT baseline, followed by the third block, split in three sessions and combining ten bird sounds (pairwise comparisons) as well as change on either side. Catch trials with no change were included (5% of the good trials). The preliminary blocks used two type of sounds, a simple one, derived from a sinusoidal burst, and the same bird chirps used by the SOAP test.

All sounds were preliminary peak normalized and loudness equalized by using the median adjustments (in dB) applied by the seven pilot participants (Fig. 1). Sounds were mono, with 16 bit coding and Fs=44,100 Hz. The average listening level was 78 dB SPL.

All tests were performed in a quiet room (average noise floor 70 dBA). We used a pair of JVC HANC250 supra-aural headphones that provided acceptable noise insulation and high comfort levels to minimize fatigue effects. All experiments used the same hardware and software setup. The tests were implemented using the Pure Data (PD) language (v0.43.4-extended) running on a Hewlett-Packard laptop with Intel Core Duo P7450 2.13 GHz, with Win7-64bit operating system. An ESI GIGAPORT-HD ASIO USB interface was used to minimize latency. Subjects' RTs were measured and logged by a custom PD sub-patch. Sound preprocessing, feature extraction and data analysis were done using *GNU Octave* v3.8.2 custom scripts and IBM SPSS v20.0.

4. DATA ANALYSIS

A response was considered perfect when the participant detected the "shortened ISI" event *and* the side on which it occurred. The overall average performance across participants was high (84% perfect detections, 7% imperfect detections with side errors, 8% missed events). Fig. 3 shows the PC values for each sound under four conditions that will be explained in the next subsection.

Participants correctly classified 96% of the catch trials confirming that the "no-event" condition could be easily discriminated. The data analysis in the following subsections is relative to the dataset with the perfect detections. The counts for each of the sounds in the dataset of the missed events is strongly correlated with those in the PC dataset ($\rho = 0.9, N = 10, p < .001$). Since the empirical distributions of the responses are negatively skewed, we used nonparametric tests for the statistical analysis. However, we also verified that similar conclusions could be reached using parametric models.

No effect of age, sex, handedness, or compensation method (monetary or course credit) was observed on detection performance (PC data).

One-way ANOVA analysis confirmed that there are no main effects of duration, sound type, trial pattern, or loudness on the response to sounds presented in isolation during the first two stages of the salience battery (Fig.2).

4.1. Time course of salience

A detailed analysis of the RT dataset is beyond the scope of this paper. However, we report here some results that are useful to evaluate the operational definition of salience that we proposed in Section 3. The RT values across all participants showed a positively skewed, long-tailed distribution, typical of RT measurements. Our analysis of the RT dataset revealed patterns similar to the ones observed for the PC dataset. We found a strong monotonic negative correlation between the median response time for each sound (dRT) and its median detection rate (PC) ($\rho = -0.78, N = 10, p < .001$). This confirms our hypothesis that a sound with high detection rate is associated with a faster response, while lower detection rates correspond to longer response times. This in turn supports our idea that "salience is fast".

In test 3 we defined an acceptance window to discriminate, for each participant, the late detections from the early ones. Such a window is defined using the RTs from test 1 (sRT) and test 2 (cRT). We used the minimum *sRT* from the pooled data ($\mathrm{sRT}_{\min} = 230$ ms) to define the lower limit of the acceptance window and filter accidental key pressings. We compared the median PC values of each sound under four conditions: unconstrained RT and three different acceptance windows that filtered "slow" detections. We defined the upper bound for each of the acceptance windows using different statistics of the response time to the sounds presented

in isolation (cRT). For example, the use of $\text{cRT}_{95\%}$ (RTs corresponding to the 95% percentile of the cRT distribution for each participant, $\widetilde{cRT}_{95\%} = 900$ ms) means that we discarded all the responses that arrived later than this value, corresponding to 22% of the PC dataset. As illustrated in Fig.3 the relationships between the PC of the sounds is not perturbed by the acceptance time windows. Therefore, it seems reasonable to conclude that, for the perfect detections in the test 3 of Fig.2, the sounds were detected as if they were presented in isolation, with just a longer RT due to the higher task complexity ($\widetilde{dRT} = 610$ ms, $\widetilde{cRT} = 470$ ms). Since we consider PC a measure of salience of a particular sound, we suggest that the observed relationship between the PCs of the sounds indicates intrinsic properties that bring them to the foreground when competing with each other.

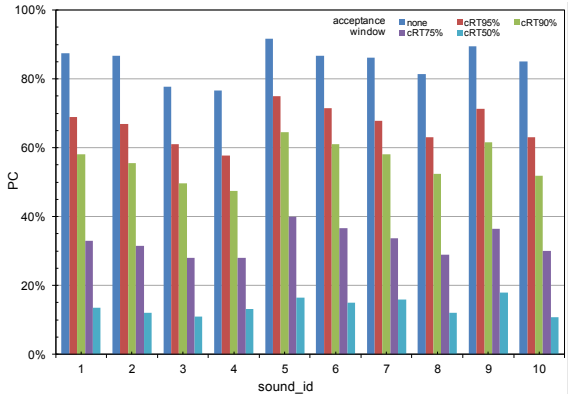

Figure 3: Effects of the acceptance time window on the detection data (PC). The effects on PC of three acceptance windows defined using sRT and cRT are compared with the unconstrained PC values. There is a strong monotonic correlation between all conditions except for the shortest time window ($\text{cRT}_{50\%}$)

4.2. Effects of loudness

We distinguish between two types of loudness. The perceptual loudness (pLOUD) is the quality of a sound that can be observed using a loudness-matching task, as we did in our experiments. It is reported in dB,representing the level adjustment that the subject applies to bring a sound to the same loudness as a reference sound. The computational loudness (cLOUD) is the measure, usually reported in sones, that a computational model of loudness associates to a sound.

In our experiments (Fig. 1) we evaluated pLOUD of the 10 bird chirps using two alternative reference sounds. We tried to minimize perceptual loudness differences with the preliminary loudness equalization step shown in Fig. 1. Subsequently, we analyzed the dB adjustments from the loudness-matching test that followed the salience battery and we did not find statistically significant differences between the sounds. Nevertheless, we could observe differences as large as 3 dB between the median level adjustments of some sounds. In particular, sounds (1,2) and (5,6) were perceived to be 2-3 dB "louder" than the other bird chirps. The just noticeable difference (JND) between two pure tones is in the range of 1 dB, but larger values are typically reported for natural sounds. Therefore, we expected to observe small residual effects of the sounds' level on PC.

For cLOUD, we implemented the model proposed by Glasberg and Moore [35] and compared it with the pLOUD and PC data. Further details concerning the extraction of cLOUD are given in Sec. 5.1 and results are summarized in Fig. 4. We found that cLOUD has a strong monotonic correlation with pLOUD when considering the median dB adjustment across participants. This was verified using the datasets obtained using the two reference sounds, i.e., R1 ($\rho = 0.80, N = 10, p = 0.005$) and R2 ($\rho = 0.92, N = 10, p = 0.001$).

4.3. Effects of sound duration and tempo

We evaluated the effects of the different duration between the group of the short sounds and that of the longer ones. A Wilcoxon signed-rank test showed that the average tempo rates (136 and 122 bpm) induced by the different average duration elicited a statistically significant change in the detection performance PC ($Z = -3.83, p < .001$). The variance observed on the short sounds was smaller than that of the long sounds ($\chi^2 = 13.6, p < .001$) with the median PC_{short} 3% higher than the median PC_{long}. The same effect was observed on the dRT distributions. The limited effect size is due to the fact that the test was easy for most participants. Nevertheless this result suggests that fast patterns tend to be dominant when competing with their slower counterparts.

4.4. Effects of brightness

Brightness is considered one of the independent perceptual dimensions for most sound categories and it is particularly relevant for environmental sounds [36]. We considered the complex spectral centroid (sC) as a good acoustical descriptor for brightness [36]. We provide more details about the extraction of this feature in Sec. 5.3. We wanted to investigate the effects of the distribution of the sC on the PC data. As illustrated in Fig. 5 we tie together sounds 1 and 7 to obtain four statistically independent classes of sounds. A Kruskal-Wallis H test showed that there was a statistically significant difference in PC score between the four different sC classes, $\chi^2(3) = 32.5, p < .001$, with a significant difference between the sounds with higher sC (sounds 3 and 4) and all the others (Mann-Whitney U post-hoc tests, $p < .002$). This suggests that in a simple scene with rhythmic patterns as the one represented by SOAP, an anomaly in the pattern of the sounds with lower spectral centroids is more easily detected. In other words, the sounds with lower spectral centroid are more likely to be in the foreground, provided that there is a sufficient spectral distance from the competing sounds. This is confirmed by the strong monotonic negative correlation between PC and sC ($\rho = -0.68, N = 10, p < .001$).

5. ACOUSTIC FEATURES

5.1. Computational loudness

We extracted the computational loudness (see Sec. 4.2) using the model proposed by Glasberg and Moore [35] for time-varying sounds. Our implementation used 128 ERB bands and the ANSI-S3.4 outer-ear model [35]. We considered the short-term loudness (STL) finding that the 75$^{\text{th}}$ percentile of the STL of each sound matches the corresponding pLOUD value better than the median STL, or maximum STL. This seems reasonable since it is similar to considering the RMS value of the STL. The boxplots in Fig. 4

summarize the distributions of cLOUD for the sounds used in our experiments.

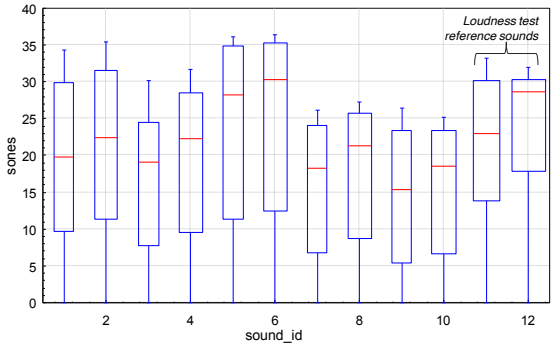

Figure 4: Boxplots of the short term loudness [35] used for our computational loudness (cLOUD). The last two boxplots are relative to the two reference sounds (R1 and R2) used for the loudness-matching task. Sounds 1,3,5,7 and 9 are the original, short, bird chirps. The longer versions have higher median loudness values, but their distributions overlap with those of the corresponding short versions.

5.2. Effective duration

The values of duration and tempo reported in Sec. 3.2 and the analysis presented in Sec. 4.3 are obtained using the physical durations (D_{phys}) of the sounds. In order to evaluate the effective duration (D_{eff}) of the bird chirps we used the definition proposed by Peeters et al. [37]. However, we did not expect to observe large effects of D_{eff} on PC since the interactions reported in Sec.4.3 are in fact related to the tempo of the pattern and, therefore, to the inter-onset-interval (IOI). There is a low, negative rank correlation between D_{eff} and PC ($\rho = -0.3, p = 0.4$).

5.3. Spectral centroid

We first computed the spectral centroid (sC) following Misdariis et al. [36] to have an estimate of the perceived brightness of each sound. We used 2048 points for the FFT and a time window of 46 ms, with a step size of 5 ms for all computations. To include the effects of the uneven sensitivity to different frequencies of the human hearing system, we introduced a spectral weighting, prior to the calculation of the sC. This weighting used the profile proposed by the international standard ISO 226:2003 [38]. The evolution of the weighted spectral centroids (sC_{ISO}) for each of the original bird chirps is illustrated in Fig.5.

The relationship of sC_{ISO} with PC was analyzed in Sec. 4.4. We also tested the correlation between PC and sC without the spectral weighting obtaining a non-significant, lower value ($\rho = 0.6, N = 10, p = 0.65$), as summarized in Table 1.

6. DISCUSSION

Our analysis of the ground truth collected using the salience battery (Fig. 2) confirmed that the preliminary loudness equalization

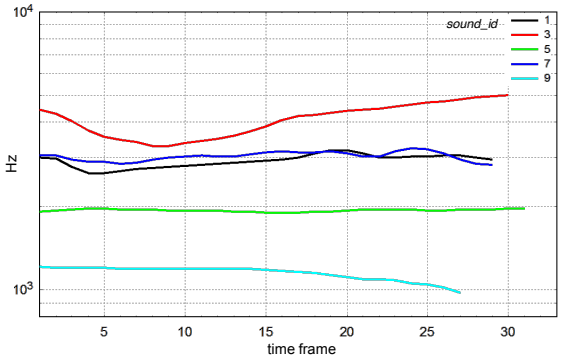

Figure 5: Evolution of the weighted spectral centroids (sC_{ISO}) for the five original bird chirps (2048 pt FFT, 46 ms time window, stepsize 5 ms). The longer versions of the chirps (sounds 2,4,6,8 and 10) have equivalent profiles. The FFT of each sound was compensated using the inverse of the ISO 226:2003 [38] equal loudness curve relative to 78dB SPL.

was successful. In fact, we did not observe main effects of perceptual (pLOUD) or computational (cLOUD) loudness on the detection data (PC dataset). Nevertheless, it is important to report the high correlation between cLOUD and pLOUD.

After equalizing for loudness, the two perceptual dimensions of tempo and brightness emerge as significant to salience measurements. The validity of these perceptual dimensions for similarity judgments is well supported by perceptual literature [36, 39], especially for environmental sounds, which are the focus of our study.

We considered a simple parametrization for these dimensions by using one acoustical feature for each, namely, duration and spectral centroid. Using the SOAP paradigm, we observed an interaction between these two features as predictors of salience: the sounds with higher salience, i.e., higher median detection rate, $\widetilde{PC}$, are those with faster tempo *and* lower spectral centroid. This is consistent with the findings of Hove et al. [40] who observed an advantage for temporal perception of musical tones at lower frequencies: it is easier to detect a glitch in the tempo produced by a kick-drum than in the melody of an electric guitar when the two instruments are playing together. The superior performance we found associated with lower frequencies and faster tempos leads to the important distinction between salience and novelty. If the performance of our subjects was driven by novelty only, i.e., the shortened ISI event by itself, then all sounds would have shown the same performance, especially since the test was quite easy. Instead, the relationship between the subjects' detection data, sC and tempo is suggestive of context effects, that is, salience. In other words, while novelty is agnostic with respect to the direction of a change, salience is not.

The effectiveness of the spectral centroid and the contribution of the residual loudness, measured by cLOUD, as salience predictors, are illustrated in Table 1. The advantage of including a perceptually motivated spectral weighting in the computation of the spectral centroid is supported by the higher rank correlation value associated with sC_{ISO}.

There is a gap between a computational model of salience and the set of acoustical descriptors that emerge from the behavioral data. We need to define the rules to combine such emergent acous-

	cLOUD	$-sC$	$-sC_{\mathrm{ISO}}$	**cLOUD**$-sC_{\mathrm{ISO}}$
ρ	0.322	0.608	0.669*	0.778*

$(^*)p < .05$, N=10.

Table 1: Spearman correlation between PC and some of the tested features. cLOUD is the 75^{th} percentile of the short term loudness [35]. sC is the complex spectral centroid [36]. $\mathrm{sC}_{\mathrm{ISO}}$ is the weighted centroid with ISO226:2003[38] compensation. sCs have negative correlation with PC.

tical features into a model. The higher correlation of the naïve composite feature $F = cLOUD - sC_{\mathrm{ISO}}$ suggests that computational loudness and spectral centroid are complementary, although not purely additive, and lead to a higher rank correlation when they are considered together rather than independently. This supports the assumption that these features capture two quasi-independent perceptual dimensions, i.e., pLOUD and brightness. Once we have a set of features we need a criterion to combine them within our computational model. To achieve this, we introduced a preliminary normalization step to add together heterogeneous features. We normalized sC and cLOUD with respect to the values of the other sounds of the corpus. In doing so, we assumed that the working memory works to generate a wider "context" for each sound, including all the other sounds in our small corpus. We suggest that a normalization strategy across all the sounds that are used is necessary for a computational model that predicts the relative salience of a sound with respect to its "competitors".

The SOAP paradigm is not suitable to test for interactions between working memory and the temporal characteristics of the isolated bird chirps (e.g., $\mathrm{D}_{\mathrm{eff}}$). This is because the tempos of the patterns are constant during a trial and the event to be detected comes after streaming build up, therefore masking the effects of different $\mathrm{D}_{\mathrm{eff}}$ values.

7. CONCLUSION AND FUTURE WORK

We proposed a behavioral definition of salience and used it to inform the design of our battery of experiments (Fig. 2). The results of the interactions between response times and detection performance support the validity of our operational definition and confirm that salience is an early perceptual process.

We looked at the problem of foreground/background selection using a pair of isochronous patterns presented in a spatial scenario and a corpus of natural bird recordings. Within these boundaries, we measured perceptual salience using the detection performance of an anomalous event. We demonstrated that our experimental paradigm is sensitive to context effects, which represent the main difference between salience and novelty. This provides further support for our behavioral definition of salience. We also showed that our ground truth supports the use of three perceptual dimensions: tempo, brightness and loudness.

In order to establish the foundations of a computational model of salience, we studied the relation between salience and the main acoustical features representing the three perceptual dimensions, namely duration, spectral centroid and computational loudness. We reviewed the techniques to extract these acoustical features and proposed a perceptually weighted spectral centroid to have a higher correlation with our ground truth. We presented a strategy to combine the features using a normalization across sounds and we showed that this simple additive approach leads to a better correlation with our ground truth. This is an important step towards the computational model of salience based on the acoustical descriptors of the three perceptual dimensions that we considered.

As future work, we will use the salience battery to cross-validate the emergent features presented here on a different corpus of natural sounds, including human communication sounds such as unconnected speech syllables or digits.

More research needs to be done to understand the role played by the effective duration of a sound in determining its salience. In order to do so we will consider "memorability" as a proxy for salience. Therefore, we will avoid repetitive patterns and use, instead, a sequences of different sounds. This, in turn, will allow us to explore salience while considering more than two sounds.

8. ACKNOWLEDGMENTS

FT thanks Jeff Blum and Ilja Frissen for their useful comments on earlier versions of the paper. Financial support for travel was provided by McGill University's Faculty of Engineering and CIRMMT.

9. REFERENCES

[1] G. Kramer, B. N. Walker, T. Bonebright, P. Cook, J. Flowers, and N. Miner, "The sonification report: Status of the field and research agenda," Report prepared for the National Science Foundation by members of the International Community for Auditory Display. Santa Fe, NM: International Community for Auditory Display (ICAD1999), 1999.

[2] T. Hermann, "Taxonomy and definitions for sonification and auditory display," in *Proc. 14th Int. Conf. Auditory Display (ICAD2008)*, ICAD. Paris, France: ICAD, 06 2008.

[3] A. Gutschalk and A. R. Dykstra, "Functional imaging of auditory scene analysis," *Hearing Research*, vol. 307, pp. 98–110, 2014.

[4] A. Hunt, T. Hermann, and S. Pauletto, "Interacting with sonification systems: Closing the loop," *Int. Conf. on Information Visualisation*, pp. 879–884, 2004.

[5] S. Bakker, E. van den Hoven, and B. Eggen, "Knowing by ear: leveraging human attention abilities in interaction design," *Journal on Multimodal User Interfaces*, pp. 1–13, 2011.

[6] B. N. Walker and M. A. Nees, "Theory of sonification," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Berlin: Logos Publishing House, 2011, pp. 9–39.

[7] J. G. Neuhoff, "Perception, cognition and action in auditory display," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Berlin: Logos Publishing House, 2011, pp. 63–85.

[8] P. Vickers, "Sonification for process monitoring," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Berlin: Logos Publishing House, 2011, pp. 455–491.

[9] S. Barrass and G. Kramer, "Using sonification," *Multimedia Syst.*, vol. 7, no. 1, pp. 23–31, Jan 1999.

[10] G. Dubus and R. Bresin, "A systematic review of mapping strategies for the sonification of physical quantities," *PLoS One*, vol. 8, 2013.

[11] B. N. Walker and G. Kramer, "Ecological psychoacoustics and auditory displays: Hearing, grouping, and meaning making," *Ecological psychoacoustics*, pp. 150–175, 2004.

[12] A. S. Bregman, *Auditory Scene Analysis - the perceptual organization of sound*. MIT Press, 1990.

[13] S. Barrass and V. Best, "Stream-based sonification diagrams," in *Proc. 14th Int. Conf. Auditory Display (ICAD2008)*, IRCAM. Paris, France: IRCAM, 2008.

[14] L. P. A. S. Van Noorden, "Temporal coherence in the perception of tone sequences," Ph.D. dissertation, unpublished. Eindhoven: Institute for Perception Research. Eindhoven University of Technology, Eindhoven, NL, 1975.

[15] J. Gossmann, "A perspective on the limited potential for simultaneity in auditory display," in *Proc. 18th Int. Conf. Auditory Display (ICAD2012)*, M. A. Nees, B. N. Walker, and J. Freeman, Eds. Atlanta (GA), USA: Georgia Institute of Technology, 2012.

[16] J. Blum, M. Bouchard, and J. R. Cooperstock, "What's around me? spatialized audio augmented reality for blind users with a smartphone," in *8th Annual Int. Conf. on Mobile and Ubiquitous Systems:Computing, Networking and Services (Mobiquitous)*, Copenhagen, Denmark, Dec. 2011, best Papers Session.

[17] J. M. Loomis, R. G. Golledge, and R. L. Klatzky, "Navigation system for the blind: Auditory display modes and guidance," *Presence: Teleoperators and Virtual Environments*, vol. 7, no. 2, pp. 193–203, 1998.

[18] R. D. Patterson, "Guide lines for auditory warning systems on civil aircraft," Instituut voor Perceptie Onderzoek, RP/ne 82/01, Manuscript no. 413/II, Feb. 1982.

[19] E. E. Wiese and J. D. Lee, "Auditory alerts for in-vehicle information systems: The effects of temporal conflict and sound parameters on driver attitudes and performance," *Ergonomics*, vol. 47, no. 9, pp. 965–986, 2004.

[20] C. Suied, P. Susini, and S. McAdams, "Evaluating warning sound urgency with reaction times," *Journal of experimental psychology: applied*, vol. 14, no. 3, pp. 201–212, 2008.

[21] R. D. Wright and L. M. Ward, *Orienting of attention*. Oxford, Oxford University Press, 2008.

[22] C. Spence and V. Santangelo, "Capturing spatial attention with multisensory cues: A review," *Hearing Research*, vol. 258, no. 1-2, pp. 134–142, 2009.

[23] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, 2nd ed. Springer, Apr. 1999.

[24] 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.

[25] M. Slaney, T. Agus, S. Liu, M. Kaya, and M. Elhilali, "A model of attention-driven scene analysis," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal processing (ICASSP), 2012*, Kyoto, Japan, May 2012.

[26] B. De Coensel, D. Botteldooren, B. Berglund, and M. E. Nilsson, "A computational model for auditory saliency of environmental sound," in *Proc. of the 157th meeting of the Acoustical Society of America (ASA)*, vol. 125, no. 4, part 2, Portland, OR, USA, 2009, pp. 2528–2528, poster 1pPP36.

[27] B. De Coensel and D. Botteldooren, "A model of saliency-based auditory attention to environmental sound," in *20th International Congress on Acoustics (ICA)*, Sydney, Australia, Aug. 2010, pp. 1–8.

[28] V. Duangudom, "Computational auditory saliency," Ph.D. dissertation, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, Dec. 2012.

[29] O. Kalinli, "Biologically inspired auditory attention models with applications in speech and audio processing," Ph.D. dissertation, University of Southern California, CA, USA, Dec. 2009.

[30] O. Kalinli and S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," *Proc. Interspeech, Antwerp, Belgium*, pp. 1–4, 2007.

[31] E. M. Kaya and M. Elhilali, "Investigating bottom-up auditory attention," *Frontiers in Human Neuroscience*, vol. 8, no. 327, 2014.

[32] F. Tordini, A. Bregman, A. Ankolekar, T. E. Sandholm, and J. R. Cooperstock, "Toward an improved model of auditory saliency," in *Proc. 19th Int. Conf. Auditory Displays (ICAD2013)*, Łódź, Poland, Jul. 2013.

[33] W. W. Gaver, "What in the world do we hear? an ecological approach to auditory event perception," *Ecological Psychology*, vol. 5, pp. 1–29, 1993.

[34] A. Temko, "Acoustic event detection and classification," Ph.D. dissertation, Speech Processing Group, Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona, Spain, Dec. 2007.

[35] B. R. Glasberg and B. C. J. Moore, "A model of loudness applicable to time-varying sounds," *J. Audio Eng. Soc*, vol. 50, no. 5, pp. 331–342, 2002.

[36] N. Misdariis, A. Minard, P. Susini, G. Lemaitre, S. McAdams, and E. Parizet, "Environmental sound perception: Metadescription and modeling based on independent primary studies," *EURASIP J. Audio Speech Music Process.*, vol. 2010, pp. 6:1–6:26, Jan 2010.

[37] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The timbre toolbox: Extracting acoustic descriptors from musical signals," *J. Acoust. Soc. Am. (JASA)*, vol. 130, pp. 2902–2916, 2011.

[38] International Organization for Standardization (ISO), "BS-ISO-226:2003(E) Acoustics. Normal equal-loudness-level," Geneva, CH," Standard, Sep. 2003.

[39] S. McAdams and E. Bigand, *Thinking in Sound: The Cognitive Psychology of Human Audition*, ser. Oxford science publications. Clarendon Press, 1993.

[40] M. J. Hove, C. Marie, I. C. Bruce, and L. J. Trainor, "Superior time perception for lower musical pitch explains why bass-ranged instruments lay down musical rhythms," *Proceedings of the National Academy of Sciences*, vol. 111, no. 28, pp. 10 383–10 388, 2014.