Journal on Multimodal User Interfaces manuscript No.
(will be inserted by the editor)

Prioritizing foreground selection of natural chirp sounds by tempo and spectral centroid

Francesco Tordini · Albert S. Bregman · Jeremy R. Cooperstock

Received: date / Accepted: date

Abstract Salience shapes the involuntary perception of a sound scene into foreground and background. Auditory interfaces, such as those used in continuous process monitoring, rely on the prominence of those sounds that are perceived as foreground. We propose to distinguish between the salience of sound events and that of streams, and introduce a paradigm to study the latter using repetitive patterns of natural chirps. Since streams are the sound objects populating the auditory scene, we suggest the use of global descriptors of perceptual dimensions to predict their salience, and hence, the organization of the objects into foreground and background. However, there are many possible independent features that can be used to describe sounds. Based on the results of two experiments, we suggest a parsimonious interpretation of the rules guiding foreground formation: after loudness, tempo and brightness are the dimensions that have higher priority. Our data show that, under equal-loudness conditions, patterns with fast tempo and lower brightness tend to emerge and that the interaction between tempo and brightness in foreground selection seems to increase with task difficulty. We propose to use the relations we uncovered as the underpinnings for a computational model of foreground selection, and also, as design guidelines for stream-based sonification applications.

This research has been generously supported by the Networks of Centres of Excellence: Graphics, Animation, and New Media (GRAND), and by the Natural Sciences and Engineering Research Council (NSERC) of Canada, grant #203568-06.

F. Tordini (✉) · J. R. Cooperstock
Centre for Intelligent Machines, McGill University, 3480 University Street, Montreal, QC H3A 0E9, Canada
E-mail: tord@cim.mcgill.ca

A. S. Bregman
Department of Psychology, McGill University, 1205 Docteur Penfield Avenue, Montreal, QC H3A 1B1, Canada

Keywords Auditory scene analysis · salience · feature extraction · sonification · foreground selection · natural sounds.

1 Introduction

The design of auditory displays, such as warning systems and mobile assistive technologies, must deal with sonic information design, management of attention, and salience. Our long-term objective is to create a tool that assists in sound scene design by predicting the perceived auditory foreground. The salience of a sound can be defined as its prominence relative to other sounds or, more generally, with respect to a background. Even though 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. Bottom-up mechanisms, including salience, shape the listener's involuntary organization of the sounds generating the scene [1,2]. Therefore, salience likely plays an important role in the design of effective sonification strategies, that is, the use of non-speech audio to present and represent information [3,4]. To guide such sonification strategies, it may be valuable to employ a computational model that maps a set of acoustical features to the perceived salience of a sound in a scene.

There are two important challenges to achieve such a model. First, the lack of an adequate operational definition of salience hinders the collection of perceptual data as ground truth. Second, despite a possibly infinite set of acoustic and perceptual features from which we might choose for use in salience prediction, the literature does not offer concrete guidance as to their relevance, apart from the obvious feature of loudness.

In the present article, we make an initial effort to address both of these challenges. We first introduce a distinc-

tion between sensory and perceptual salience. The former is a measure of the novelty of a sound "event", whereas the latter represents the quality driving the formation of the auditory foreground and background. We will focus on perceptual salience, complementing the literature on salience models that exclusively addresses sensory salience [5–11]. We equate perceptual salience with faster and more accurate stream selection. We use this definition to design an oddball task based on auditory streaming that uses two concurrent, spatialized, repetitive patterns of natural chirp sounds, and present the results of two experiments based on this paradigm. Second, our analysis suggests that when streams have comparable loudness, our collected data can be best explained by the high-level acoustic features of brightness and tempo.

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 [3,12,13]. The complexity and importance of taking into account the perceptual and cognitive dimensions when designing sonification systems are well documented [14,15].

The use of natural, environmental sounds is an interesting complement to traditional iconic, or metaphoric ones, especially 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 [16,17]. Iconic, symbolic sounds are often perceived as artificial. Their acceptability under prolonged listening conditions can only be achieved by careful sound design. Instead, natural sounds tend to be better accepted: the prediction of their perceived salience would streamline the design cycle of many sonification problems [18]. In this context, the attempts to translate Bregman's principles of auditory scene analysis (ASA) [19] into sonification design rules have been frequent in recent years, although lacking consistency. The stream-based sonification approach proposed by Barrass and Best [20] is a relevant example.

Salience prediction is also important for applications in other fields, for example in mobile assistive technologies [21] and warning signal design [22–24].

2.2 Computational models of salience

Kayser et al. [5] were the first to propose a feature-driven computational model of auditory salience and to compare 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. Their experiments dealt with monaural, lateralized sounds treated in isolation on a stereo background presented at a lower level, and were designed around a detection task with intensity being the only independent factor. Their work inspired our research described here. However, since sounds rarely occur in isolation, especially in most natural environments, our approach differs in that we present pairs of sound patterns, with equal loudness, in a binaural scenario.

Extensions of the monaural algorithm proposed by Kayser et al. [5] add cochlear [7] and loudness models [10] as a preprocessing stage, and pitch as an additional feature. Kalinli [8] uses pitch both for speech tracking purposes and as an added feature to her salience model of syllable onsets, and to improve speech segmentation in automatic speech recognition (ASR) applications. Slaney et al. [9] addressed auditory salience in a spatial scenario in the context of speech separation and ASR. Other authors, including Kaya et al. [11], propose a statistical approach that avoids the explicit use of perceptual features.

All current computational models implement the concept of novelty and can be regarded as detectors of salient boundaries, i.e., onsets. They all share the same "memory span" 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. More generally, current models of auditory salience focus on local (short term) rather than global (long term, summarized) features. Furthermore, these salient-onset detectors are conceived as segmentation tools for automatic sound analysis rather than for the prediction of the foreground/background representation of the sounds populating a scene. The latter approach would be more useful in an interactive sonification context [12,25].

3 Redefining auditory salience

Imagine yourself as a medical school resident on your first day in an intensive care unit (ICU). Among the other tasks, your responsibility is to monitor the status of ventilation and blood pressure levels of the patient by listening to the sounds produced by two instruments. These produce periodic, regular patterns of equally loud sounds, each with its own pace and characteristic timbre, resulting in a crowded auditory environment [26,27]. Your goal is to promply detect any deviation from normality for the two variables, both critical to the patient. However, it is unlikely that you can sustain your attention to monitor them both, continuously. In such a scenario, can we predict which instrument you will by attending to the most, and with less effort, and which will be your perceived foreground, if any? In the remainder of this section, we suggest that we should distinguish between the salience

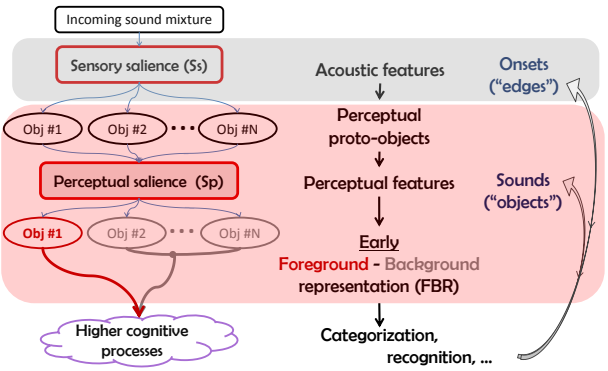

Fig. 1: Types of auditory salience. Sensory salience (S_s) works on a local time scale to identify salient onsets and contribute to the formation of the proto-objects (the sound streams). Perceptual salience (S_p) shapes the "early" foreground-background representation of the scene by selecting one object (e.g., Obj #1) to be used by higher, cognitive processes. These, in turn, can manipulate the object definition and selection, as represented by the ascending arrows [1,2].

of events and that of sound streams, i.e., the auditory objects [2] in a scene. Then, starting from a set of features motivated by research on timbre perception [28,29], we select those that we hypothesize to be sufficient to explain the behavioral data collected in our experiments.

3.1 Sensory and perceptual salience

We propose two definitions, sensory salience (S_s) and perceptual salience (S_p), to differentiate between the contributions of local (short term) features and global (long term) ones, as illustrated in Figure 1. Sensory salience is the property that measures how noticeable an event is with respect to its temporal neighborhood. Perceptual salience is the property that measures the likelihood that a stream, for example a pattern of similar sounds, belongs to the auditory foreground when competing with other streams.

In Figure 1 we use the term proto-object ("Obj #..") to identify the building block of a sound scene, that is, a stream. In our scheme an event, such as an onset, contributes to the definition of a proto-object and then belongs to it. We assume that stream formation is a primitive process that can be explained with a bottom-up model [19]: this is supported by the evidence for feature-based separation at the level of the auditory cortex [1].

In order to understand whether an event belongs to the foreground of an auditory scene, we suggest that it is necessary, but not sufficient, that the event be salient in the sensory sense. In other words, the mechanisms regulating perceptual salience follow those determining sensory salience. Higher order (top-down) processes will only deal with the proto-objects (the streams) [2] that are organized in an early auditory scene. It is S_p that drives this initial foreground-background organization of the streams. Interactions with higher cognitive functions certainly exist, but emerge only after the initial perception of the sound scene, which is our present focus. However, it is beyond the scope of this article to delve into discussion and analysis of such connections and their role in the stability of scene perception after the initial phase [30].

Returning to our medical scenario, the sound patterns produced by the two intruments have different timbre and characteristic pace. Naturally, they are not synchronized, so the streams will merge into an incoherent babble. However, the listener can, at will, focus on either of the two patterns, effectively steering away from the other.

From the computational standpoint, a sensory salience model is based on short-term features and fixed time windows [6,5]. It should highlight any event corresponding to a large discontinuity in the spectrotemporal space defined by the mix, generating a map of salient onsets. It is unlikely, however, that a listener's attention will actively jump from one sound to the other, alternating between the two streams, over an extended period. Presumably, the listener will attend to what is perceived as the foreground stream, eventually switching to the other if it becomes more salient.

A model based on sensory salience (S_s) cannot account for the existence of such a "default" foreground stream. Our hypothesis is that by modeling S_p using long term (global) features associated with the proto-objects, we can predict the default foreground, as observed with our experiments. The duration and the position of the time windows that we use to define the features in Section 3.2 are not fixed, since they depend on the sub-units of the streams defined by the salient onsets (in the S_s sense).

Our experimental paradigm, described in Section 4.1, aims to capture S_p using a detection task based on a simple streaming scenario. Our assumption is that detection performance is a measure of the prevalence of one stream over the other: if there exists a default foreground stream, more of its anomalies will likely be detected than those of the background stream, even if the anomalies are equally distributed across streams and have comparable S_s values.

3.2 Acoustical and perceptual features

We considered a set of relatively independent audio features, in the sense introduced by Peeters et al., using correlational analysis of dissimilarity ratings, followed by hierarchical clustering [29]: spectral centroid, spectral spread, spectral flatness, effective duration, amplitude modulation, and frequency modulation [29]. Based on previous literature on salience prediction [5,10,31] we also added computed loudness [32] and effective duration of the autocorrelation function (τ_e) [31]. The latter is a measure of harmonicity and spectral spread. We suggest that a further reduction of the

feature set is possible if the aim is not to model dissimilarity but, instead, stream selection–our research goal. In the interests of space, we show in Table 1 only the three perceptual features that were retained and studied in depth (Column 1), the corresponding acoustical features (Column 2), and the terms used to describe them (Column 3). In the following subsections, we introduce the techniques used to compute the descriptors for loudness, tempo (derived from duration), and brightness. These features are the most general and powerful descriptors for natural sounds [28]. A model capable of predicting the default foreground based on these global features would be a convenient design tool for sonification applications. In fact, parametric mappings based on these features are often used for continuous process sonification [16], but have not been generalized to complex sounds such as those used in our studies.

We will use calligraphic symbols to distinguish the duration of the perceptual features ($\mathscr{D}$), tempo ($\mathscr{T}$), and brightness ($\mathscr{B}$), from their acoustical counterparts. Departing from the literature reviewed in Section 2.2 we compute our global descriptors using adaptive, rather than fixed, time windows defined by two consecutive onsets belonging to a sound stream

perceptual feature	**acoustical feature**	**global descriptor**
loudness (pLOUD)	computational loudness (cLOUD)	$cLOUD_{75\%}$
duration ($\mathscr{D}$), tempo ($\mathscr{T}$)	duration (effective, physical)	D_{eff}, D_{phys}
brightness ($\mathscr{B}$)	weighted spectral centroid (sC_{ISO})	$sC_{\text{ISO }50\%}$

Table 1: Features selected for the data analysis. The global descriptors are used as acoustical "labels" for each sound. Like Peeters et al. [29] we use robust statistics: 3^{rd} quartile (75%) for cLOUD and median value (50%) for sC_{ISO}. Tempo is calculated from D_{phys}.

3.2.1 Perceptual and computational loudness

Loudness is obviously an important component of salience and can overshadow other features. We measure it to prepare equally loud sounds for our experiments, so as to permit the effects of other features to emerge from the data.

Here, we distinguish between two measures of loudness. Perceptual loudness (pLOUD) is a measure, reported in dB, based on a loudness-matching task, where the sound level is adjusted until it is judged equal in loudness to a fixed reference [33]. pLOUD is used in most perceptual studies using natural, dynamic sounds.

Computational loudness (cLOUD) is a measure, reported in sones, assigned by a loudness model to a sound over a determined time window. We introduce a summarized cLOUD measure ($cLOUD_{75\%}$), defined as the 75$^{\text{th}}$ percentile in the set of short term loudness (STL) values [32] for each sound [34].

3.2.2 Duration, effective duration, and tempo

We define two measures of the duration ($\mathscr{D}$) of a sound. The physical duration (D_{phys}) is the actual duration of the soundfile $D_{\text{phys}} = M/FS$, where M is the number of samples and FS is the sampling frequency. The effective duration (D_{eff}) is a measure of the time the signal is perceptually meaningful. In this work we used the definition proposed by Peeters et al. [29] applied to the envelope of the short-term loudness (STL) of the bird chirps. We use D_{eff} in Section 3.2.3 to define the time window for the estimation of the global brightness of a sound. Our experiments use repetitive patterns of bird chirps. The tempo ($\mathscr{T}$) of the patterns is given by $\mathscr{T} = 60 \cdot 10^3/(D_{\text{phys}} + \Delta t)$, measured in beats per minute (bpm); Δt is the duration in milliseconds of the inter-stimulus interval (ISI) between two consecutive chirps. More details about the patterns are given in Section 4.1.

3.2.3 Spectral centroid and brightness

Brightness ($\mathscr{B}$) is considered one of the independent perceptual dimensions for most sound categories and is particularly relevant for environmental sounds [28]. For this purpose, we use the spectral centroid [29], more specifically, the complex spectral centroid (sC) suggested by Misdariis et al. [28]. We used 2048 points for the FFT (FS = 44.1 kHz) 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 and minimize the correlation between sC and perceived loudness, we proposed the use of a spectrally weighted centroid (sC_{ISO}) [31]. The weighting is applied to the signal prior to the calculation of the sC and uses the profile proposed by the international standard ISO 226:2003 [35]. In this work, we take the median value of the spectral centroid ($sC_{\text{ISO }50\%}$) of each sound as its global spectral centroid, or "summarized brightness", evaluated in accordance with its effective duration D_{eff}.

4 Experimental framework

4.1 Perceptual salience: operational definition and streaming paradigm

We collect the S_p data from the listeners using our salience battery [31], shown in Figure 2. The battery consists of three consecutive tests in order to separate the effects of cognitive

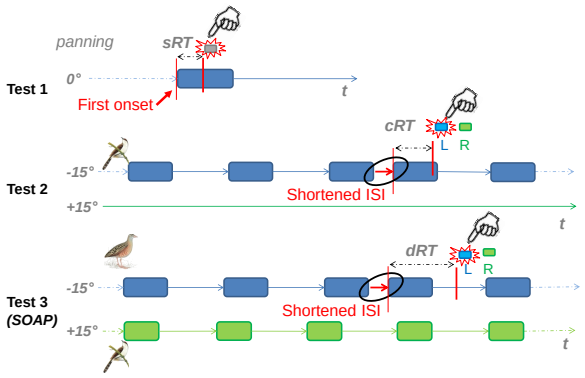

Fig. 2: The salience battery starts with the *simple detection* test, which presents a single sound. The *spatial detection* test adds spatialization, and the *spatial discrimination* test adds a second stream. The red arrows represent the events that the subjects must detect. In the examples shown here,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

load from salience effects, and to define personalized baselines for the response time (RT) of each participant. The simple detection test involved the simplest scenario: one sound at a time, presented monaurally. The spatial detection test introduces spatialization. Finally, the spatial discrimination test presented two asynchronous and concurrent sequences in a binaural scenario (streaming of asynchronous patterns, SOAP), in which one of these sequences contained a shortened interval. In the SOAP test, at each trial, the listeners were asked to detect the occurrence and identify which stream contains the shortened inter-stimulus interval (ISI), as illustrated in Figure 2 (blue and green rectangles of Test 3). Trials were 5 to 7 s long and separated by 2 s of silence. A pattern consisted of a sequence of identical sounds, equally spaced by a $\Delta t = 250$ *ms* (normal ISI). If listeners detected the anomaly (shortened ISI, e.g., 80 ms), they used their dominant hand to press the appropriate key on a keyboard as quickly as possible. Their responses to the anomalies were placed into one of three categories: left, right, or none (no key pressed within the acceptable time window). The occurrence of the shortened ISI was balanced between the two sides and its position within a pattern was randomly generated at each trial. We collected response times (RTs) and detection accuracy for each participant. RT was used as a measure of performance and to filter false positives from the detection accuracy data.

The SOAP test was based on the assumption that after segregation and streaming have occurred, stream selection is a competitive process in which the more salient of two concurrent auditory streams is more likely to be in the foreground. Thus, there is an increased probability of detecting anomalies in the more salient stream. The ease of stream selection can be measured in two ways, RT or accuracy: in terms of RT, the faster the response, the easier the pattern is to follow. Similarly, for accuracy, once false positives are removed from the data, the higher the detection rate, the easier it is to isolate a pattern from the mixture, effectively suppressing the background.

In Sections 5 and 6 we describe our two experiments in which this salience battery was used to collect the S_p data from the listeners.

4.2 Salience measures and statistical analysis

A response during the SOAP test was considered perfect when the participant detected the "shortened ISI" event and the side on which it occurred. Therefore, each trial had three possible outcomes: detection (D), detection with side error (SidERR), and miss (MISS). False positives, that is, key strokes recorded before the shortened ISI, or on the wrong side (SidERR), were eliminated from the data. The analysis we present in this article focuses on the D dataset. To improve sampling, each condition was repeated during the test, producing D counts for each condition for each participant. Traditional ANOVA is inadequate to analyze such counts, since our repeated measures results in correlation between data points. Similarly, ANOVA is unsuitable for analysis of the proportions (percentages) because of the biases due to uneven sampling. Averaging data points corresponding to the same condition was also not an option, since this would have caused a loss of information and masked the discrete nature of the original D dataset. Instead, we employ the method of generalized estimating equations (GEE) [36], a non-parametric extension of generalized linear models (GzLM), which is often used to analyze longitudinal and other correlated response data, particularly if responses are binary or in the form of counts [37]. As we were analyzing count data, we used a negative binomial regression model with a log link function.

4.3 Corpora of natural sounds: motivation and design

There are several motivations to collect perceptual data using natural sounds. Theunissen and Elie [38] propose the use of natural sounds to probe the auditory system and introduce an analytic framework for the characterization of the auditory stimulus-response using animal vocalizations. Moreover, natural sounds are better accepted than synthetic sounds, especially when the application requires prolonged listening [18], as with most sonification designs for continuous process monitoring [16]. However, since it is unfeasible to cover the entire field of natural sounds, we chose to use a very limited selection of non-human communication sounds, namely bird chirps. These offer a large choice of temporal and spectral textures, while being relatively homogeneous in terms of familiarity to typical listeners, which is less likely the case for a broader selection that includes

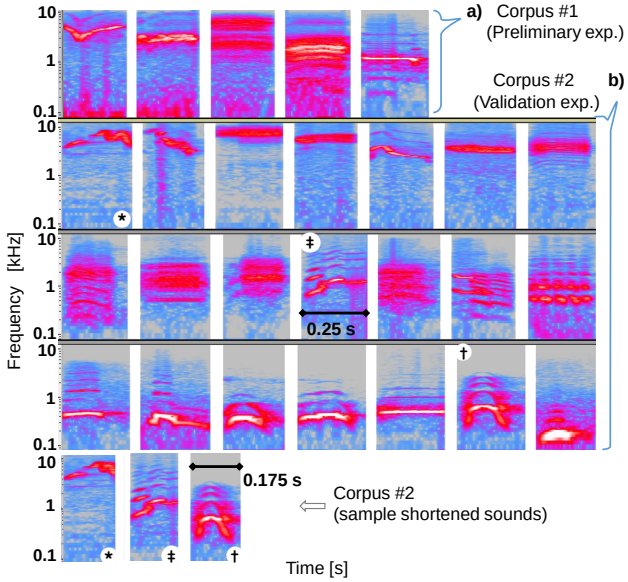

Fig. 3: Spectrograms of the original (long) chirps for the two corpora. In the interests of space, the corresponding sets of shortened sounds are omitted and only three examples from the second corpus are shown in the last row. Same-symbol pairs identify long/short versions for three sample sounds from Corpus #2.

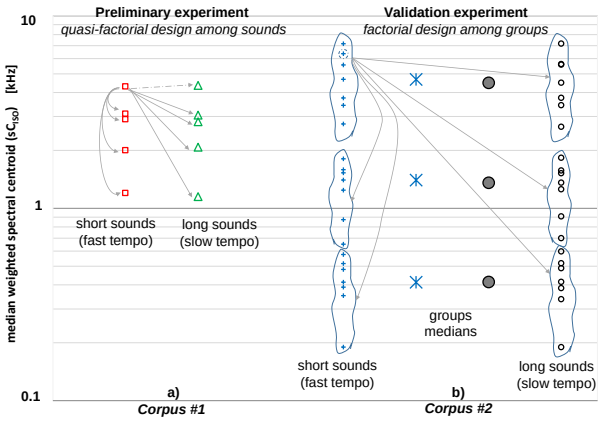

Fig. 4: Corpora designs. Sounds are represented by their global brightness value ($sC_{\mathrm{ISO\ 50\%}}$). **a)** Corpus#1 (preliminary experiment), quasi-factorial design between *sounds*. The dashed line represents the omitted combination (example for the first element). **b)** Corpus#2 (validation experiment), full factorial design between *groups* of sounds. A sound from each group was randomly selected at each trial. Larger symbols represent the median spectral centroids for each group.

other animals such as cats and dogs. This choice helps avoid the risk that other higher level attributes of the sounds could dominate the measured effects. Consistent with Theunissen and Elie [38], our hypothesis is that the features with statistically significant effects on listeners' event-detection performance are also more perceptually salient within the set of all possible features: they emerge naturally from the use of uncontrolled, yet biologically relevant, stimuli.

Particular care is needed when compiling corpora of natural sounds. In this respect, we followed two rules: sounds must be countable, i.e., distinguishable, or separable, when presented in a mixture with more than one chirp, and they must be equally loud since large loudness differences likely overshadow the effects of other features on salience.

We enforced the former requirement by means of a preliminary informal listening session using a group of three volunteers. For the latter, we pre-equalized the sounds to minimize their differences in loudness using the measures described in Section 3.2.1. We compiled two corpora for the design of the stimuli for the experiments. The first corpus (10 sounds: 5 chirps and 5 shortened replicas) were used for the preliminary experiment (Section 5.1), and the second corpus (42 sounds: 21 chirps and 21 shortened replicas) for the validation experiment (Section 6.1). The replicas with shorter duration were generated using the SoundStretch utility [39], preserving the spectral content of the original chirps. The spectrograms of the original (long) sounds for both corpora are shown in Figure 3, together with three examples of shortened replicas from the second corpus.

The sounds of the second corpus were pre-classified and conceptually grouped according to their global brightness ($sC_{\mathrm{ISO\ 50\%}}$ values, Figure 4b) to test the robustness of the effect this feature has on foreground selection, i.e., salience. Although the first experiment uses a factorial design between sounds, in the second one we run all the combinations between groups, with sounds within groups randomly selected at each trial. The second experiment therefore tests the effect of *average* brightness differences between sounds on stream selection, a much more critical requirement than that tested in the preliminary experiment (Section 5). These design differences are summarized in Figure 4. All sounds are available for download. [1]

4.4 Materials and apparatus

Original sounds were monaural, with 16 bit coding and FS = 44.1 kHz. All tests were performed in a controlled laboratory room (average noise floor 26 dBA), and involved the use of a pair of Sennheiser HD650 circumaural headphones. The average listening level was 78 dB SPL (measured using a B&K artificial ear simulator).

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 an Intel Core Duo P7450 2.13 GHz CPU, running the Windows7-64bit operating system. An ESI GIGAPORT-HD ASIO USB interface was used to minimize latency. Subjects' RTs were measured with 10 ms accuracy and logged by a custom PD sub-patch. Sound preprocessing,

[1] http://srl.mcgill.ca/~tord/SOAPsounds/

feature extraction and data analysis were performed with GNU Octave v3.8.2 custom scripts and IBM SPSS® v20.0.

5 Preliminary Experiment

We tested the correlation of the descriptors presented in Section 3.2 with respect to the participant's detection data. The rank correlation was computed across all conditions, independent of the choice of stream not containing the audio event (the green one in Figure 2). Then, we focused on the emergent feature from the correlation analysis to look for context effects as a proxy for salience.

5.1 Stimuli

Five recordings of bird chirps were taken as a starting point (Figure 3a, Corpus#1, average $D_{phys} = 243$ ms, SD= 7 ms), from which five replicas with shorter duration (average D_{phys}= 190 ms, SD= 6 ms) were generated using the SoundStretch utility [39]. All sounds underwent preliminary equalization using a loudness-matching task performed by a small group (N= 7) of pilot listeners [40,34]. The average tempo of the patterns used for the salience battery (Figure 2) was 129 bpm: 121 bpm (SD= 2) for the five long sounds and 136 bpm (SD= 2) for the corresponding short replicas. The standard ISI value was $\Delta t = 250$ ms. The average trial duration was 5.5 s. The shortened ISI event position within the trial was randomized in the second half of each trial to allow enough time for the buildup of the streams [41]. The perturbation introduced by the shorter ISI ($\Delta t' = 80$ ms, $-68\%\Delta t$) corresponded, on average, to a local tempo "glitch" of +70 bpm. An almost constant asynchrony would result when comparing two patterns using chirps with similar D_{phys} (and therefore tempo). We countered this problem by adding a zero mean random jitter ($\Delta t \pm \delta$, $\delta_{max} = 30$ ms) to each period of every sequence.

5.2 Design

Participants ran the two preliminary tests in Figure 2 to assess their RT baseline for each chirp sound, followed by the third (SOAP) test where a within-subjects design was utilized with chirp (5 sounds), chirp duration (long, short), presentation side (left, right), and ISI value (normal, shortened) being the independent factors. All tests used the same sounds. Combinations with the same chirp on both ears (independently of its duration) were excluded to ensure stream separability. Each condition was repeated twice. Trial sequence was randomized for each participant. Catch trials with no ISI change were included (5% of the number of good trials). When present, the shortened ISI value was 80 ms (32% of the standard ISI of 250 ms). The average duration of a trial was 5.6 s.

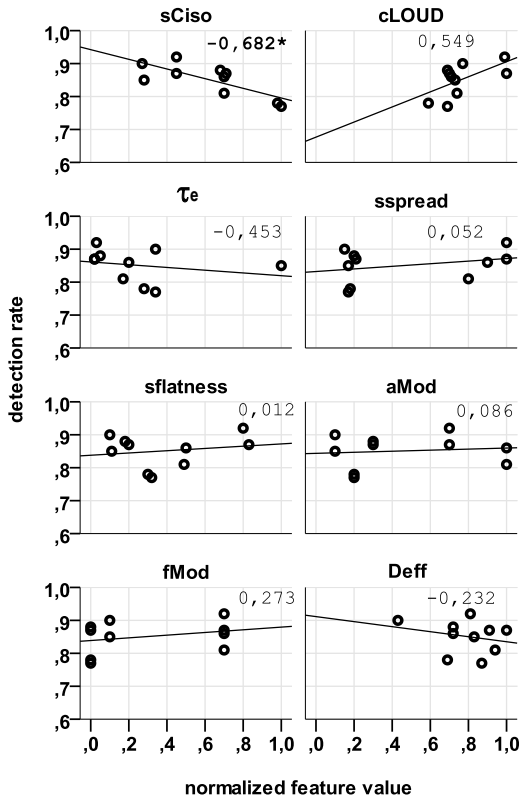

Fig. 5: Median feature values for each sound and the associated detection rate, averaged across conditions and participants. The Spearman rank correlation (ρ) value is shown for each one of the 8 features initially considered for this experiment. Brightness, represented by sC_{ISO}, has a strong, negative, correlation ($\rho = -0.682$) with the perceptual data. * indicates statistical significance at p< .05 level, with Bonferroni correction.

5.3 Participants

A group of thirty one (N= 31) participants (age = 21.7 ± 2.6; 19 females) participated in the experiment. 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.

5.4 Brightness as emergent feature

We considered the detection rates for the 10 sounds of the corpus, averaged across conditions and participants. Then, we computed the Spearman rank correlation (ρ) with respect to each of the eight descriptors introduced in Section 3.2. We summarize the main results in Figure 5, plotting the median values of the features for each sound, normalized across

sounds. Brightness, represented by the weighted spectral centroid (sC_{ISO}), emerged as the feature with the strongest, negative, rank correlation with the mean detection rate ($\rho = -0.682$, p$< .05$), followed by the computed loudness (cLOUD, $\rho = 0.549$, $p = .1$), and by the effective duration of the autocorrelation envelope (τ_e, $\rho = -0.432$, $p = .18$). Even if not significant, the moderate negative correlation observed for τ_e is consistent with the results of our previous study that used the same paradigm and a different corpus [31]. τ_e is a measure of harmonicity and spectral spread of a sound; in our dataset it shows a strong, negative correlation with spectral spread ($\rho_{\tau_e/\texttt{sspread}} = -0.633$, p$< .05$) and amplitude modulation ($\rho_{\tau_e/\texttt{aMod}} = -0.644$, p$< .05$). The effective temporal duration (D_{eff}) and the physical duration (D_{phys}, not shown) of the chirps both have weak to moderate, negative, and not significant correlation values with the detection data. So, for example, a short D_{phys} value (i.e., fast tempo), considered alone, is not a strong predictor for high detection rate.

In the analysis that follows, we study the role played on detection rate by the perceived brightness of the stream containing the event. We look for context effects considering the brightness difference (ΔsC_{ISO}) between the pattern with the event and the one with no ISI change. The analysis also includes the tempo values of the patterns used in each condition, since D_{phys} was explicitly manipulated in the experimental design, and also, to enable comparison of our results with previous research on tempo perception, especially in the field of warning sound design [22–24].

5.5 Detection data analysis: brightness and tempo effects

The average detection rate with Test 2 of the battery (Figure 2) was 99% for all sounds, confirming that the detection of the shortened ISI was easy and that the differences in the SOAP test were due to the selection of the stream, rather than to the detection of the event itself. An omnibus analysis confirmed that there were no main effects of duration, sound (bird chirp), or trial pattern on the response to sounds in a single-stream presentation (Test 2, Figure 2).

The SOAP task (Test 3) was found easy by most listeners, confirmed by the high overall average performance across participants (84% perfect detections, 7% imperfect detections with side errors, 8% missed events). Participants correctly withheld 96% of the catch trials confirming that the "no-event" condition could be easily discriminated. Personalized RT baselines, based on the RTs in Test 1 and 2, were used to filter false positives. No effect of age, sex, handedness, compensation method (monetary or course credit), and presentation side of the event was observed on detection performance.

We used a negative binomial regression model with log link-function fitted with the GEE method [36] to analyze the effects of tempo ($\mathscr{T}$, derived from D_{phys}) and brightness ($\mathscr{B}$, represented by $sC_{\text{ISO }50\%}$). The χ^2 values were computed using the Wald statistic.

We initially considered the original factors used in the design ("chirp" and "tempo"), finding strong main effects effects of both on detection performance, but no clear interpretation of the relationship between chirps was possible. However, focusing on brightness, we noticed that some chirps had very similar median sC_{ISO} values, leading us to consider reducing the number of discrete levels (bins) into which brightness was divided. We tied together the top three sounds in Figure 4a, defininining three clusters of chirps, corresponding to three frequency bands: high, mid-high, and mid-low. This clustering is optimal in the sense that the centers of the three bands (clusters) are equally spaced on a logarithmic scale by 0.8 octaves. In this light, our analysis confirmed the strong main effects of tempo ($\chi^2_{\mathscr{T}}(2) = 35$, p $< .001$) and brightness ($\chi^2_{\mathscr{B}}(4) = 58$, p $< .001$), and a strong interaction between the two factors ($\chi^2_{\mathscr{B}x\mathscr{T}}(8) = 80$, p $< .001$).

For the remainder of this article, we use uppercase letters to identify the pattern with the shortened ISI (the event to be detected) and lower case letters for the pattern without ISI change. For example, the label "`Fs`" describes the case were a pattern with fast tempo ($\mathscr{T}$), containing the event, is competing with a pattern with slow $\mathscr{T}$, without the event. For brightness ($\mathscr{B}$), we use the difference between the sC_{ISO} values, subtracting the median sC_{ISO} of the pattern with no ISI change from that of the pattern with the event. Therefore, a negative ΔsC_{ISO} describes a condition where the pattern with the event has lower $\mathscr{B}$ than the pattern with no ISI change.

5.6 Discussion

We first consider the main effects of the two factors, tempo and brightness, on the average detection performance across participants.

The asymmetric profile shown for tempo (main effects) in Figure 6 suggests a context effect within this factor. While listeners have no problems in detecting the shortened ISI when the stream carrying containing it has a fast $\mathscr{T}$ (`Fs` condition), performance drops when the event occurs within a slow pattern (`Sf` condition). Unsurprisingly, this suggests that fast patterns are more likely to be in the foreground, consistent with literature on tempo perception [42,43] and warning sounds [22,24]. Participants performed 6% better when the event occurred within a fast pattern, irrespective of brightness. Thus, fast patterns are more salient. Interestingly though, the highest average detection rate (88%) was measured when the patterns had equal tempo (`Ff` or `Ss`), suggesting that other factors are responsible for the performance in this condition.

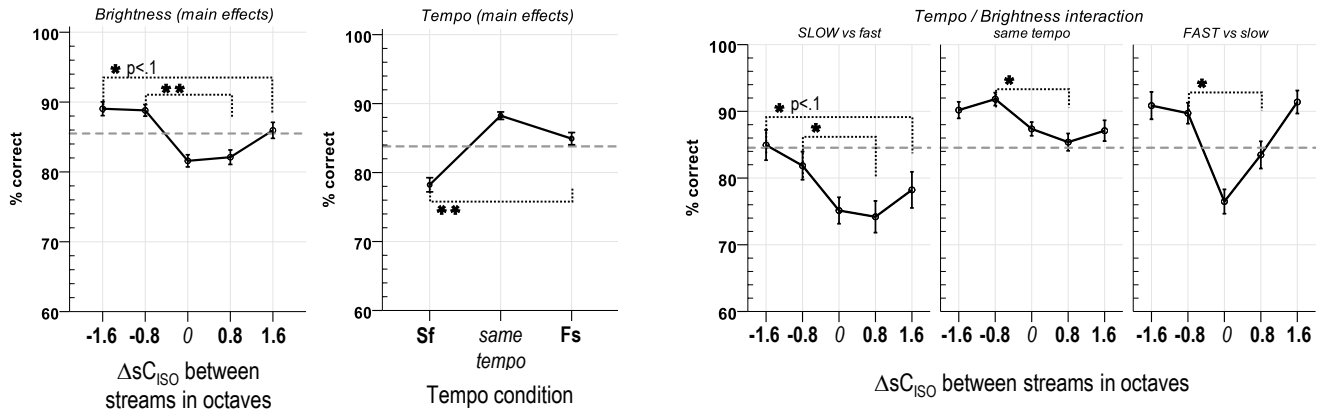

Fig. 6: Brightness and tempo effects (mean detection ± SEM): main effects and interactions between features. Mean detection performance is higher when the event is presented in patterns with lower $\mathscr{B}$ (conditions with negative ΔsC$_{\mathrm{ISO}}$) and faster $\mathscr{T}$. The rightmost panel shows the strong interactions between $\mathscr{B}$ and $\mathscr{T}$. Dashed lines correspond to average % correct, and asterisks indicate statistically significant asymmetries discussed in the text (* p< .05, ** p< .001, Bonferroni).

We can interpret the brightness (main effects) panel of Figure 6 in the same manner. Here, the abscissae indicate the distance, in octaves, between the clustered chirp sounds (Figure 4a) that are used for the patterns. When the shortened ISI is present in a pattern with lower $\mathscr{B}$, the average detection rate is greater than when the event is in a stream with higher $\mathscr{B}$. In this experiment, when the event was present in a pattern with low $\mathscr{B}$ the performance was 4 to 6% better than the case with high $\mathscr{B}$, irrespective of the $\mathscr{T}$ values. Thus, patterns using dark sounds were more salient. The advantage for sounds with lower $\mathscr{B}$ is also confirmed in the rightmost panel of Figure 6 where the interactions between $\mathscr{B}$ and $\mathscr{T}$ conditions are shown. Our findings concerning $\mathscr{B}$ are consistent with those of Hove et al. [44], who observed an advantage for the temporal perception of musical tones at lower frequencies, but seem to counter the intuitive fact that high frequency sounds are more noticeable than low frequency ones [45]. However, most studies use stimuli with similar spectral occupation and equalize them to have equal RMS levels [30,45]. Instead, we used dynamic stimuli with different spectrotemporal characteristics, equalized according to their perceived loudness. The equalization introduced uneven gain structure among the sounds and may be a reason for the different responses we collected. The superior performance we found associated with lower frequencies and faster tempi 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 exhibited the same performance, especially since the test was quite easy. Instead, the relationship between the subjects' detection data, $\mathscr{B}$ and $\mathscr{T}$ 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.

6 Validation Experiment

Our second experiment aimed to replicate the results of the preliminary one, but using a larger corpus of natural sounds, therefore increasing the possibility for other features to emerge. Moreover, we investigated the relation between task difficulty and salience by using three difficulty levels.

6.1 Stimuli

Twenty-one natural recordings of bird chirps (average D_{phys} = 248 ms, SD= 12 ms), covering a frequency range from 200 Hz to 7 kHz, were taken as a starting point (Figure 3b). As with the first corpus, the only selection criterion was their perceptual separability when presented in a mixture. Twenty-one replicas with shorter duration (average D_{phys} = 177 ms, SD= 9 ms) were generated using the SoundStretch utility [39]. The average tempo of the patterns used for the salience battery (Figure 2) was 131 bpm: 120 bpm (SD= 3) for the 21 long sounds, and 141 bpm (SD= 3) for the corresponding short replicas. The loudness levels of the 42 sounds were estimated using $cLOUD_{75\%}$ and then adjusted to a common reference value corresponding to value of 78 dB SPL. The sounds were ranked using their global brightness $sC_{ISO\ 50\%}$ and grouped into three clusters (high, mid, and low) each containing seven long sounds, and three clusters for the corresponding short sounds (see Figure 4b). The sounds were chosen so that the centers of mass of the clusters were equally spaced on a logarithmic scale by 1.6 octaves.

6.2 Design

Participants ran the two preliminary tests in Figure 2 to assess their RT baseline, followed by the third (SOAP) test,

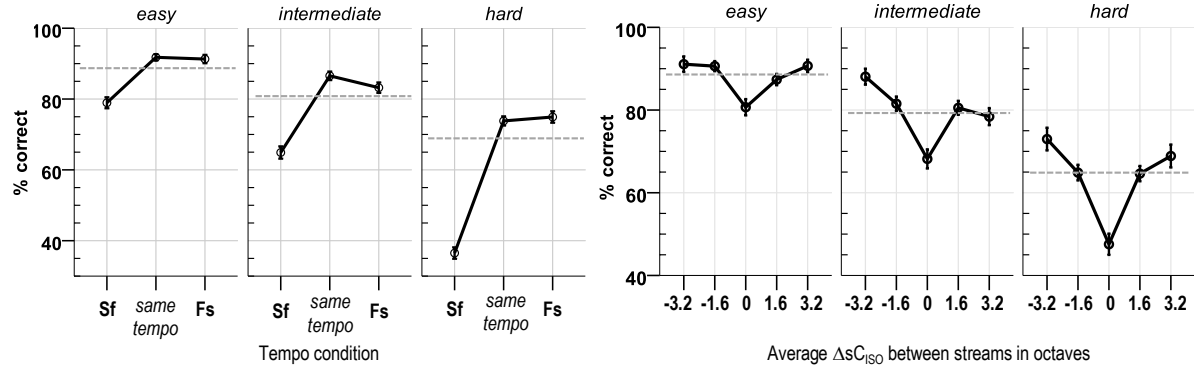

Fig. 7: Brightness and Tempo effects (mean detection ± SEM): effects of task difficulty on the main effects of $\mathscr{B}$ and $\mathscr{T}$. Task difficulty changes with the duration of the shortened ISI ($\Delta t'$): `easy` (80 ms), `intermediate` (120 ms), and `hard` (160 ms). The normal ISI is always 250 ms. Mean detection performance is higher when the event belongs to patterns with lower $\mathscr{B}$ (conditions with negative $\Delta \text{sC}_{\text{ISO}}$) and faster $\mathscr{T}$. Dashed lines correspond to the average detection rates. Interactions between $\mathscr{B}$ and $\mathscr{T}$ are presented in Figure 8.

combining the 42 bird sounds, as well as change on either side. Groups of seven chirps were defined considering three brightness bands (Figure 4b). For the SOAP test we used a within-subjects design with brightness group (3 groups: high, mid, low), chirp duration (long, short), presentation side (left, right), and ISI value (normal, shortened) being the independent factors. Each condition was presented twice. At each trial a sound was randomly selected from the corresponding group and trial sequence was randomized for each participant. The average duration of a trial was 5.6 s. Catch trials with no ISI change were included (20% of the trials). Each participant ran three blocks of the SOAP test. Each block used a different difficulty level, determined by the duration of the shorter ISI ($\Delta t'$) compared with the standard one (Δt) in Figure 2: easy ($\Delta t' = 80$ ms, $-68\%\Delta t$), intermediate ($\Delta t' = 120$ ms, $-52\%\Delta t$), hard ($\Delta t' = 160$ ms, $-36\%\Delta t$).

The factorial design between groups (Figure 4b), instead of sounds (Figure 4a), is a way of minimizing test duration while using a large corpus of sounds. With respect to the preliminary experiment, this dilutes the effect of brightness on event detection rate since we only analyze the effects of the average brightness difference between patterns. Therefore, this design defines a more challenging environment for the emergence of brightness in terms of its effects on stream selection. Finally, working with clusters, we can reduce the risk of familiarization by randomly sampling a sound from the corresponding group at each trial.

6.3 Participants

A group of N= 31 (median age = 21, range 18–42; 23 females) McGill undergraduate students participated in this experiment and were compensated with course extra credit. They all reported normal hearing.

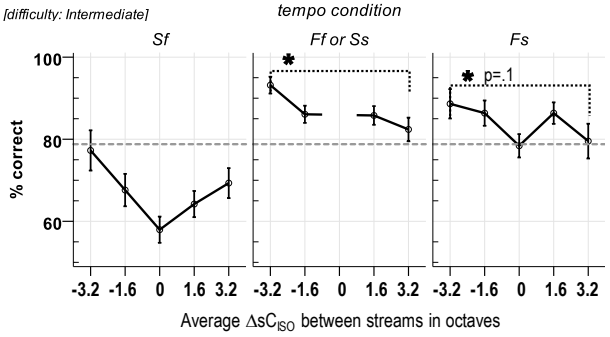

Fig. 8: Across features interactions for task with intermediate difficulty ($\Delta t' = 120$ ms). Mean detection rates (± SEM) are shown as a function of the brightness separation between patterns (in octaves), under different $\mathscr{T}$ conditions. The effect size of $\mathscr{B}$ is greatest when the patterns have the same tempo (`Ff` or `Ss`). Asterisks indicate statistically significant asymmetries (* p< .05, with Bonferroni correction)

6.4 Data analysis

Similarly to our analysis in the preliminary experiment, the RT baselines were collected and used to filter false positives. Our analysis focused on the effects of $\mathscr{T}$, $\mathscr{B}$, and task difficulty ($\Delta t'$ value) on the detection performance averaged across participants.

We fitted a negative binomial regression model to the D dataset using the GEE method [37]. Significant main effects were observed for all factors $\mathscr{B}$, $\mathscr{T}$, and difficulty ($\Delta t'$), with $\mathscr{T}$ playing a larger role ($\chi^2_{\mathscr{T}}(2) = 232$, $\chi^2_{\Delta t'}(2) = 140$, $\chi^2_{\mathscr{B}}(4) = 105$, p< .001). The interactions between tempo, brightness, and difficulty were also significant and with comparable effect sizes ($\chi^2_{\mathscr{T}\times\Delta t'}(4) = 90$, $\chi^2_{\mathscr{B}\times\Delta t'}(8) = 55$, p< .001), as shown in Figure 7.

The effect of the $\mathscr{B}\times\mathscr{T}$ interaction term increases with the difficulty of the task: $\chi^2_{\mathscr{B}\times\mathscr{T}(easy)} < \chi^2_{\mathscr{B}\times\mathscr{T}(intermediate)} < \chi^2_{\mathscr{B}\times\mathscr{T}(hard)}$. For an easy task ($\Delta t' = 80$ *ms*, much shorter than Δt, therefore easy to detect) the differences between

the patterns along $\mathscr{T}$ or $\mathscr{B}$ considered alone seem to provide sufficient cues to reach a high detection performance.

The interactions between the levels of tempo and brightness are presented in Figure 8, in the case of intermediate task difficulty ($\Delta t' = 120\ ms$). We found that listeners performed 9 to 10% better when the event was presented in patterns having low sC_{ISO} than when the shortened ISI was on streams with high sC_{ISO}. This is the same result observed in the preliminary experiment, but with statistical significance limited to the cases with larger spectral separation and with neutral (`Ff`, `Ss`) or favourable (`Fs`) tempo conditions. The mitigated effect of $\mathscr{B}$ with respect to the preliminary experiment was an anticipated consequence of the factorial design among groups of sounds rather than individual chirps (see Figure 4b). This likely caused the greater variance around the average performance of each of the $\mathscr{B}$ conditions in Figure 7 and 8.

6.5 Discussion

The patterns shown for the main effects of tempo and brightness in Figure 7 are consistent with the ones we found with the preliminary experiment and shown in Figure 6. The effect of tempo differences on detection performance, considered alone (leftmost panel in Figure 7), confirm that fast patterns are more salient than slow ones. Considering the effects of brightness differences between the streams, we found, not surprisingly, that the larger the spectral separation, the better the impact of brightness was on detection performance. However, our validation experiment also confirmed what we observed in the first experiment: under equal loudness conditions, darker patterns seem to have higher salience than brightner ones.

Moreover, the relative importance of brightness with respect to tempo increased with the difficulty of the task. This can be interpreted in terms of resources allocated to solve the foreground-background problem by the listener: $\mathscr{B}$ is "recruited" to compensate for the lack of power provided by $\mathscr{T}$. The harder the task, the more $\mathscr{B}$ is used. However, when $\mathscr{T}$ is a strong enough cue (e.g., `Fs` condition) the effect of $\mathscr{B}$ is smaller: the listener does not need to recruit it. This is also "computationally" plausible: if a cue is sufficient to bring a stream to the foreground, there is no reason to over-analyze the scene by resolving brightness differences.

Finally, although the evaluation of the $cLOUD_{75\%}$ global descriptor is out of the scope of this paper, we report that according to an informal questionnaire given at the end of the experiment, the listeners agreed on the fact that the sounds of the second corpus were "overall equally loud".

7 General discussion

Our two experiments demonstrated that average detection performance is uneven across the different conditions of $\mathscr{T}$ and $\mathscr{B}$ used in our simple auditory scenes. This was true even when the task was easy, as in the preliminary experiment and in the first block of the validation experiment. We found that once loudness differences between sounds are minimized, the values of $\mathscr{T}$ and $\mathscr{B}$ of the two streams have significant main effects on listeners' ability to detect the shortened ISI event. This was largely expected since spectral separation and/or tempo differences help parse a scene into separate concurrent streams [19] and, hence, assign the event to the correct pattern. The larger those differences, the higher the improvement in detection performance (the main effects).

For $\mathscr{B}$, we considered differences between the $sC_{\mathrm{ISO}\ 50\%}$ of the two streams, ranging from 0 to 3 octaves. However, spectral separation does not explain the statistically different values observed for conditions with the same absolute brightness difference between patterns. Therefore, we suggested that those differences in the detection performance (as large as 10%, p. < 05) are a proxy for context effects within the levels of a feature, in this case, brightness.

In our two studies, under equal loudness conditions, dark sounds seem more salient than bright ones. It is of course unsurprising to observe unequal sensitivity along a particular dimension. Indeed, it is natural to expect fast patterns to be more salient than slow ones, as we found. However, the result that dark sounds are more salient than bright ones is surprising in light of both common sense and much of the literature on pitch perception, which would suggest the opposite. This result motivates further investigation of the important differences between context effects in pitch and brightness, especially for non-laboratory stimuli. With natural sounds, such as bird chirps, there is no reason why perception should follow the same rules for brightness as it does for pitch. However, it is critically important to carry out such investigation under conditions of equal loudness.

Context effects also exist between independent features such as tempo and brightness. We found that these two dimensions interact and that their interplay depends on the difficulty of the task. We interpreted this finding in terms of resources needed to parse an auditory scene. When tempo differences are not sufficient to bring a pattern to the foreground (as in the `Sf` condition) the listener makes more use of the brightness differences between streams. This seems biologically plausible: there is no need to use extra computational power (i.e., resolving brightness differences) if the tempo difference between the streams is sufficient.

In our studies we focused on the detection accuracy data. However, it is important to report that the discrimination response time (dRT) values collected with the SOAP tests on

both experiments were negatively correlated with the detection rate of each sound ($\rho < -0.75$ on all tests, p$<$ .001). This confirms our hypothesis that a sound with a high detection rate is associated with a faster response, while lower detection rates correspond to longer response times.

Finally, our discussion on salience is adequate for repetitive patterns using chirp sounds. At present, we cannot generalize our results about the feature set and their relative importance to scenarios with episodic sounds (those that happen only once), or with very slow patterns, when the tempo perception breaks down. Moreover, replicating these results using classes of natural sounds other than chirps and different tempi would be beneficial [38].

8 Conclusions and future work

We introduced two definitions to distinguish *sensory* (S_s) from *perceptual* (S_p) salience. We proposed a behavioral definition for S_p and used it to inform the design of our tests (Figure 2). From a sonic information standpoint, S_p is the salience of the stream that carries the information. It measures how likely it will pop out from a mixture and become the foreground. S_s, instead, measures the informative content of an event, such as an onset, within a stream.

Using our streaming paradigm and a correlation analysis we selected three global descriptors from a larger set of features: brightness, tempo and loudness. We argue that, beyond loudness, tempo and brightness are the next important dimensions to predict stream selection and the formation of the auditory foreground. These perceptual features are also the most important ones in similarity judgments [28], especially for environmental sounds, which are the focus of our study.

We used the loudness descriptor to equalize the sounds and better observe the effects of the other two factors. We argue that an equal-loudness equalization procedure is more appropriate than the traditional equal-intensity setting when studying brightness effects on scene perception, especially when considering natural sounds, which typically exhibit non-uniform spectral content distribution. This avoids a loudness bias towards high frequencies, which would result from an equal-intensity setting, often used in studies on pitch.

Under the equal-loudness condition tempo differences seem, overall, more salient than those differences in brightness; faster patterns are more salient than slower ones and, with sufficient spectral separation, darker sounds are more salient than brighter ones. Also, the negative correlation between response times and detection accuracy, i.e., the fact that correct detections are fast, is consistent with our operational definition and with the assumption that salience is an early perceptual process.

Our results with respect to tempo are in agreement with those of the large literature addressing the perception of tempo in audition [42]. Numerous such examples exist in the design of auditory warning sounds and the field of music perception, and these all support our results, i.e., fast is more salient than slow. However, to our knowledge, the same cannot be said for the brightness of a sound, especially with respect to the study of context effects. Although several studies address the quantitative measure of sound brightness, using features such as the spectral centroid, we have not come across any prior studies explicitly addressing brightness and context effects. We believe the main source of confusion here results from the abundance of studies addressing perception and context effects for pitch. However, pitch and brightness represent different qualities, especially for natural, i.e., complex, sounds. In this regard, our study represents a first step towards the understanding of context effects for brightness and how this feature can be used to control scene perception.

The relationships between and within $\mathscr{T}$ and $\mathscr{B}$ are the underpinings of a computational model of perceptual salience that, starting from a model of loudness, aims to predict foreground selection in a complex sound scenes.

Furthermore, the relations we observed among the features and within their levels can be used effectively as guidelines for continuous process sonification, since $\mathscr{T}$ and $\mathscr{B}$ can be understood easily and parametrized by the designer.

9 Acknowledgments

FT thanks J. Blum for the long discussions during this research, F. Grond, the colleagues at the Shared Reality Lab, the JMUI editors and reviewers for their valuable input.

References

1. A. Gutschalk and A. R. Dykstra, “Functional imaging of auditory scene analysis,” *Hearing Research*, vol. 307, pp. 98–110, 2014.
2. S. Denham and I. Winkler, “Auditory perceptual organization,” in *The Oxford Handbook of Perceptual Organization* (J. Wagemans, ed.), pp. 1–31, Oxford University Press, March 2014.
3. 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 (ICAD). Santa Fe, NM., 1999.
4. T. Hermann, “Taxonomy and definitions for sonification and auditory display,” in *Proceedings of the 14th International Conference Auditory Display (ICAD2008)*, (Paris, France), pp. 1–8, IRCAM, Jun. 2008.
5. 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.
6. O. Kalinli and S. Narayanan, “A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech,” *Proceedings of the Interspeech, Antwerp, Belgium*, pp. 1–4, 2007.
7. B. De Coensel, D. Botteldooren, B. Berglund, and M. E. Nilsson, “A computational model for auditory saliency of environmental

sound," in *Proceedings of the of the* 157^{th} *meeting of the Acoustical Society of America (ASA)*, vol. 125, (Portland, OR, USA), pp. 2528–2528, 2009. Poster 1pPP36.
8. O. Kalinli, *Biologically Inspired Auditory Attention Models With Applications In Speech And Audio Processing*. PhD thesis, University of Southern California, CA, USA, Dec. 2009.
9. M. Slaney, T. Agus, S. Liu, M. Kaya, and M. Elhilali, "A model of attention-driven scene analysis," in *Proceedings of the of the IEEE International Conference on Acoustics, Speech, and Signal processing (ICASSP)*, (Kyoto, Japan), pp. 145–148, May 2012.
10. V. Duangudom, *Computational auditory saliency*. PhD thesis, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, Dec. 2012.
11. E. M. Kaya and M. Elhilali, "Investigating bottom-up auditory attention," *Frontiers in Human Neuroscience*, vol. 8, no. 327, pp. 1–12, 2014.
12. A. Hunt, T. Hermann, and S. Pauletto, "Interacting with sonification systems: Closing the loop," *International Conference on Information Visualisation*, pp. 879–884, 2004.
13. S. Bakker, E. van den Hoven, and B. Eggen, "Knowing by ear: leveraging human attention abilities in interaction design," *Journal on Multimodal User Interfaces*, vol. 5, no. 3-4, pp. 197–209, 2012.
14. B. N. Walker and M. A. Nees, "Theory of sonification," in *The Sonification Handbook* (T. Hermann, A. Hunt, and J. G. Neuhoff, eds.), pp. 9–39, Berlin: Logos Publishing House, 2011.
15. J. G. Neuhoff, "Perception, cognition and action in auditory display," in *The Sonification Handbook* (T. Hermann, A. Hunt, and J. G. Neuhoff, eds.), pp. 63–85, Berlin: Logos Publishing House, 2011.
16. P. Vickers, "Sonification for process monitoring," in *The Sonification Handbook* (T. Hermann, A. Hunt, and J. G. Neuhoff, eds.), pp. 455–491, Berlin: Logos Publishing House, 2011.
17. S. Barrass and G. Kramer, "Using sonification," *Multimedia Systems*, vol. 7, pp. 23–31, Jan 1999.
18. T. Hildebrandt, T. Hermann, and S. Rinderle-Ma, "A sonification system for process monitoring as secondary task," in 5^{th} *IEEE Conference on Cognitive Infocommunications (CogInfoCom)*, pp. 191–196, Nov 2014.
19. A. S. Bregman, *Auditory Scene Analysis - the perceptual organization of sound*. MIT Press, 1990.
20. S. Barrass and V. Best, "Stream-based sonification diagrams," in *Proceedings of the 14th International Conference Auditory Display (ICAD2008)*, (Paris, France), pp. 1–6, IRCAM, 2008.
21. Á. Csapó, G. Wersényi, H. Nagy, and T. Stockman, "A survey of assistive technologies and applications for blind users on mobile platforms: a review and foundation for research," *Journal on Multimodal User Interfaces*, vol. 9, no. 4, pp. 275–286, 2015.
22. 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.
23. 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.
24. 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.
25. A. Hunt and T. Hermann, "Interactive sonification," in *The Sonification Handbook* (T. Hermann, A. Hunt, and J. G. Neuhoff, eds.), pp. 273–298, Berlin: Logos Publishing House, 2011.
26. C. Meredith and J. Edworthy, "Are there too many alarms in the intensive care unit? an overview of the problems," *Journal of Advanced Nursing*, vol. 21, no. 1, pp. 15–20, 1995.
27. R. A. Stevenson, J. J. Schlesinger, and M. T. Wallace, "Effects of divided attention and operating room noise on perception of pulse oximeter pitch changesa laboratory study," *Anesthesiology*, vol. 118, no. 2, pp. 376–381, 2013.
28. 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 Journal of Audio Speech Music Processing*, vol. 2010, pp. 6:1–6:26, Jan 2010.
29. G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The timbre toolbox: Extracting acoustic descriptors from musical signals," *Journal of the Acoustical Society of America*, vol. 130, pp. 2902–2916, 2011.
30. D. Pressnitzer and J.-M. Hupé, "Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization," *Current Biology*, vol. 16, no. 13, pp. 1351 – 1357, 2006.
31. F. Tordini, A. Bregman, A. Ankolekar, T. E. Sandholm, and J. R. Cooperstock, "Toward an improved model of auditory saliency," in *Proceedings of the* 19^{th} *International Conference Auditory Display (ICAD2013)*, (Łódź, Poland), pp. 189–196, Jul. 2013.
32. B. R. Glasberg and B. C. J. Moore, "A model of loudness applicable to time-varying sounds," *Journal of the Audio Engineering Society*, vol. 50, no. 5, pp. 331–342, 2002.
33. M. Florentine, A. N. Popper, and R. R. Fay, eds., *Loudness*, vol. 37 of *Springer Handbook of Auditory Research*. New York, NY, USA: Springer US, 2011.
34. F. Tordini, A. Bregman, and J. R. Cooperstock, "The loud bird doesn't (always) get the worm: Why computational salience also needs brightness and tempo," in *Proceedings of the* 21^{st} *International Conference Auditory Display (ICAD2015)*, (Graz, Austria), pp. 236–243, Jul. 2015.
35. International Organization for Standardization (ISO), "BS-ISO-226:2003(E) Acoustics. Normal equal-loudness-level," standard, Geneva, CH, Sep. 2003.
36. A. Agresti, *Categorical Data Analysis*. Wiley Series in Probability and Statistics, Wiley, 3 ed., 2014.
37. J. A. Hanley, A. Negassa, J. E. Forrester, *et al.*, "Statistical analysis of correlated data using generalized estimating equations: an orientation," *American Journal of Epidemiology*, vol. 157, no. 4, pp. 364–375, 2003.
38. F. E. Theunissen and J. E. Elie, "Neural processing of natural sounds," *Nature Reviews Neuroscience*, vol. 15, pp. 355–366, Jun. 2014.
39. O. Parviainen, "Soundstretch utility." Version 1.9.0 (2015-05-18), available at `http://www.surina.net/soundtouch/soundstretch.html`.
40. F. Tordini, "Is there more to saliency than loudness?," in 6^{th} *Workshop on Speech in Noise (SPiN): Intelligibility and Quality*, (Marseille, France), Jan. 2014.
41. A. S. Bregman, "Auditory streaming is cumulative," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 4, pp. 380–387, Aug. 1978.
42. J. D. McAuley, "Tempo and rhythm," in *Music Perception, Springer Handbook of Auditory Research 36* (M. Riess Jones, R. R. Fay, and A. Popper, eds.), Springer Science+Business Media, 2010.
43. C. A. Gonzalez and C. L. Baldwin, "Effects of pulse rate, fundamental frequency and burst density on auditory similarity," *Theoretical Issues in Ergonomics Science*, vol. 16, no. 2, pp. 1–13, 2014.
44. 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. 10383–10388, 2014.
45. T. Bouchara, C. Jacquemin, and B. F. G. Katz, "Cueing multimedia search with audiovisual blur," *ACM Transactions of Applied Perception*, vol. 10, pp. 7:1–7:21, June 2013.