TOWARD AN IMPROVED MODEL OF AUDITORY SALIENCY

Francesco Tordini*, Albert S. Bregman, Jeremy R. Cooperstock
McGill University, Montréal, QC, Canada
tord@cim.mcgill.ca

Anupryia Ankolekar, Thomas Sandholm
Hewlett-Packard Laboratories
Palo Alto, CA, United States

ABSTRACT

While visual saliency models are approaching maturity, their auditory counterparts remain in their infancy. This is mainly due to the difficulties of gathering basic data, and oversimplifications such as an assumption of monaural signals. Moreover, conventional testing approaches for evaluating auditory saliency models tend to be overly simplistic.

To address these shortcomings, we developed an experimental procedure for testing auditory saliency along with more formalized stimulus-selection criteria to support more versatile and ecologically relevant saliency models. This work is described, along with an analysis of some relevant acoustical correlates that emerge from the experiments. The results motivate the formulation of a measure of sound complexity and appear to favor time-domain, rather than frequency-domain analysis to describe saliency. Finally, some conclusions are drawn regarding the definition of an expanded feature set to be used for auditory saliency modeling and prediction in the context of natural, everyday sounds.

1. INTRODUCTION

The saliency of a sound can be defined in natural language as its prominence relative to other sounds or, more generally, with respect to a *background*. If we then consider a prototype sound scene consisting of M natural sounds and a background, having a lower perceived level than the sounds, the typical expected outcome of what we may call saliency analysis is a ranking of the M sounds in terms of their relative saliency, and a correlated set of shared acoustical and perceptual features that we could use to predict that ranking. In what follows we assume that the sound sources are distinguishable, that is, the source separation problem is not an issue. This condition can be reinforced by using spectrally separated sounds or by spatial separation (i.e., spatial release from masking).

The so-called *bottom-up approach* to saliency prediction attempts to find a mapping between sound features and perceived saliency based on models and rules derived from the biological structure and the perceptual organization of the human auditory system.

A first important challenge for the bottom-up models comes from the difficulty of gathering *perceptual ground truth*, that is, sounds labeled and ranked in terms of their perceived saliency.

(*) corresponding author. This research was funded through the HP Labs Innovation Research Program 2001 (HPL-IRP2011), the Graphics, Animation, and New Media (GRAND) Networks of Centres of Excellence, and a student award from the Centre for Interdisciplinary Research in Music, Media and Technology (CIRMMT)

The second main challenge is the selection of the features to be used for the signal analysis and saliency prediction. An exhaustive search is clearly not practical, given the size of the possible acoustical and perceptual feature set. Some guidelines must then be drawn to help select a starting point. A taxonomy of the sounds must also be chosen in order to clearly define the class of sounds addressed by the analysis (see, for example, Adiloğlu et al. [1] or Gaver [2]).

With this work we meant to address both problems. We propose an experimental schema that allows collection of behavioral data from subjects listening to a pair of sounds, presented on a background, in a binaural scenario. We also test some descriptors and attempt to find the ones that best describe the saliency ranking inferred from the behavioral data.

Finding statistically significant results was not our primary goal, rather it was testing the new conceptual design and its performance with subjects listening to real natural sounds. Here we present the preliminary results coming out of this framework.

We started from the simplest scenario, with one sound only, presented in a fixed spatial location and with a certain pattern. This represents the baseline for the second experiment where we define a simple approximation of a natural scene by using two spatialized sounds (M=2) presented over a background. We also decided to limit the scope of our study to everyday sounds that are neither speech nor man-made music. More specifically, we used bird songs for this work.

1.1. Background and Related Work

Saliency and attention are intimately related, but is the latter that has attracted most of the research effort in the field of cognitive psychology, where the task-driven (or "top-down") is the leading approach (see Wright and Ward [3] for a modern review with a special focus on vision, and Spence and Santangelo [4] for a review in the multimodal scenario.). As a result, signal driven (or "bottom-up") models typically come from psychophysics and psychoacoustics [5] but hardly deal with the concept of saliency which points to a perceptual, rather than sensory, concept. Saliency, with its most natural definition, is actually *in between* "bottom-up" and "top-down" and perhaps this is the reason why it does not find an easy placement in the research agenda from a psychological perspective and most psychological works only adopt it as a qualitative concept. This may also explain why very few perceptual saliency models are available. More specifically, a closed loop between modeling, perceptual ground truth, and applications is far from being robust for audition, although a noticeable attempt was made by Kayser et al. [6] who proposed a feature-driven computational model and compared its predictions to the results of two behavioral experiments. Their monaural auditory saliency model

was based on three feature maps: intensity, frequency and temporal contrast. Even if temporal contrast allows continuity constraints to be put over the temporal envelope, this model builds on monaural intensity maps and therefore cannot capture nor explain effects due to the phase relationship between signal waveforms that permit localization and spatial release from masking. Also, the experiments run by Kayser et al. [6] 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 was inspired by that of Kayser et al. [6], but is different in that we present sounds in pairs over a background, and in a binaural scenario. We also attempted to formalize some criteria useful for the design of the sound corpus to be used. We therefore aimed to capture perceptual data that are ecologically more valid (the importance of spatial perception beyond sound localization is well developed [7, 8, 9, 10]).

On the other hand, saliency is a "handy", powerful concept from the application point of view, therefore making it interesting to other research communities.

To our knowledge, the first work to address auditory saliency in a spatial scenario was done by Slaney et al. [11] in the context of speech separation and ASR. They introduced the concept of binaural saliency as captured by binaural onsets obtained from the differential cross-correlation of the cochlear filter-bank output spikes computed using ITDs only. This work represents a notable evolution with respect to the monaural saliency models available so far. Extensions of the monaural algorithm proposed by Kayser et al. [6] add cochlear and loudness models (e.g., references [12, 13, 14]) at the preprocessing stage, and pitch as an additional feature. Kalinli [15, 16] uses pitch both for speech tracking purposes and as an added feature to her *auditory gist* preprocessing stage. With the latter she introduces a pre-attentive model where she attempts to bridge the gap between top-down and bottom-up models.

None of the works above addressed the problem of gathering the perceptual ground truth data, relying, instead, on performance measures defined in terms of automatic (i.e., machine based) speech recognition rates [11, 15], to evaluate their systems.

1.2. Saliency and Sonification

Research on the design of warning signals [17, 18, 19] and mobile assistive technologies (see references [20, 21] for audio-only mobile application examples) implicitly deal with saliency and the management of attention in their sound design principles and guidelines (see references [22, 23] for useful reviews).

The main themes of the research agenda present in the Sonification Report [24], generated five years after the first ICAD conference, show very little need for modification after almost two decades of research. An auditory display uses sound to communicate information. Sonification is defined as a subtype of auditory display that uses non-speech audio to present and represent information. Kramer and colleagues [24] 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 a good effort from the research community. In fact, the complexity and the importance of taking into account the perceptual and cognitive dimensions while designing sonification systems are well documented in a recent comprehensive collective work on sonification (see, in particular, chapters 2 and 4 [25, 26]).

Modern sonification calls for the use of natural sounds instead of metaphoric, iconic ones and designs with "sourcy" environments where real, dynamic sounds are not presented in isolation.

The umbrella term *ecological psychoacoustics* was used by Walker and Kramer [27] 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) [28] into sonification design rules has been more frequent, although still tepid.

The stream-based sonification by Barrass and Best [29] is a good example in this direction. They tested and extended the so-called van Noorden diagrams [30] to dimensions other than F0 of simple tones such as brightness, intensity and panning (ILD) of noise bursts. They aimed to design sonifications that could control streaming and take listening attention into account by studying galloping sequences. .

The testing paradigm we propose here is rather simple yet it incorporates many of the items discussed so far, namely (almost) galloping patterns, dynamic sounds derived from natural ones, and spatialization. It aims to serve as a tool for the robust inference of the perceptual saliency of sounds for naive subjects. It also allows us to validate saliency parametrization via novel features extracted from the sounds. For these reasons it seems like a good companion for sonification design tasks from an *ecological psychoacoustics* standpoint.

2. CAPTURING SALIENCY GROUND TRUTH

What we present here is a preliminary design and the first results obtained from a small sample of subjects. We carried out an assessment experiment (Exp1) followed by a main experiment (Exp2). Both experiments used spatialized sounds rendered on the horizontal plane (no elevation added) and played on a background (possible backgrounds also included silence). More specifically we used one sound pattern for Exp1 and two sound patterns for Exp2. Each pattern was built using an elementary bird chirp, or a *simple sound* repeated at constant rate (see section 2.3.2 for details on the stimuli). The bird chirps had different durations, which produced a natural asynchrony between the two streams in Exp2, as shown in Fig. 2 .

The experimental paradigm presented here relies on the assumption that, after segregation and streaming have occurred, stream selection is a competitive process that makes the "most salient" of two concurrent auditory streams more likely to be attended (see Bregman [28] for an introduction to simultaneous and sequential segregation, and streaming, and Jones [31] for a review of the perception of temporal patterns).

We call this schema *Segregation of Asynchronous Patterns* (*SOAP*).

Throughout this work we assume that source separation is not an issue. Segregation is guaranteed by the structural differences and by the spatial separation characterizing the sounds in the playback schema.

The subject listens to the sounds via supra-aural headphones in a quiet environment and is presented with a neutral visual field.

The subjects' task is to detect the occurrence of a single shortened interstimulus interval (ISI) in an otherwise isochronous sequence of 12 repetitions of a sound.

In Exp1, each trial consists of a single sequence. The presentation side of the sound sequences is fully balanced and their order is randomized for each participant. This is a simple detection task since no competitor streams are present.

In Exp2 there are two such sequences running 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 with stream competition and higher perceptual load. It is illustrated in Fig. 1.

In both experiments the subject has to indicate the location (L/R) of the detected change by pressing one of two buttons on the keyboard. Subjects are free to choose their preferred method of pressing the keys, using either fingers of the same or different hands, whichever they found most comfortable.

Response time (RT, in ms) and detection accuracy (either True or False) are recorded.

Subjects may attempt to monitor both patterns presented in Exp2 as a single stream relying on an "overall rhythm" that may characterize the auditory scene. This would make the selection process superfluous and reduce the given task to a simpler side-detection one. The asynchrony of the two patterns shown in Fig. 2 is an important element of our design as it discourages such attention strategy by breaking the temporal relations between the two patterns, making the scene more difficult to follow as a whole.

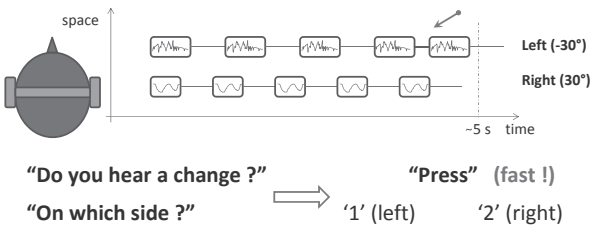

Figure 1: The SOAP paradigm in a spatial scenario. The red arrow in the top right corner highlights the change (shortened ISI) presented to the subject's left ear.

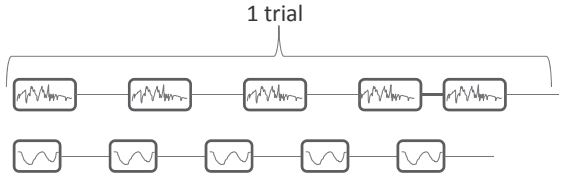

Figure 2: Structure of a trial. Chirps are separated by $\Delta t = 250\ ms$. The red segment shows where the ISI reduces to $\Delta t' = 80\ ms$. A trial is a group of K chirps, with K=12 for this experiment (the position of the red segment is randomized in a range between between a lower limit K=6 and an upper limit K=11). Consecutive trials are separated by 2.5 s of silence and followed by a short noise burst located in front of the subject, acting as "fixation point" and preparing the subject to the next trial.

2.1. Participants

A total of seven (N=7) subjects (average age = 28.2; 5 male) volunteered for the experiments. They all reported normal hearing.

2.2. Design

A within-subjects full factorial design was utilized with sound type, presentation side and ISI value being the independent factors. The same sample of subjects participated in both experiments. This was needed to assess the typical response time (RT) for each participant during Exp1.

2.3. Materials

Both experiments used the same hardware and software setup.

2.3.1. Apparatus

The tests were implemented using the Pure Data (PD) language (v0.42.5-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 response times were measured by a custom PD sub-patch. Sound preprocessing and behavioral data analysis was done using GNU Octave (v.3.6.1, 64bit) custom scripts. 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.

2.3.2. Stimuli

Seven bird chirps, two simple sounds (beep-like bursts, used for Exp1 only) were used as elements of the foreground patterns presented to the subjects during the assessment (Exp1) and principal experiment (Exp2). Bird chirp duration varied from 215 to 530 ms; the two simple sounds were respectively 110 and 325 ms long. See Fig. 3 and Fig. 4 for spectral details. Three backgrounds (silence, pink noise, or natural noise (defined here as a mix of human babble noise and pink noise)) were used across sessions, as in Kayser et al. [6]. Backgrounds were played at -10 dB with respect to the foreground.

All sounds were peak normalized (see reference [32] for a discussion of the perceptual effects of Peak and RMS normalization) and loudness equalized according to subjective listening tests and final adjustments by ear. It should be noted, however, that methods for the assessment of the loudness of time varying signals are highly debated.

A pink noise burst (duration = 38ms, intensity = 50% with respect to the foreground sounds) was played before each trial with an inter stimulus interval (ISI) of 850 ms with respect to the first sound of the trial to exclude forward masking. This noise burst was not lateralized (i.e., it was played centered, in front of the subject) and acted as an acoustic fixation point (a *centering sound*) to minimize the risk of lazy listening attitudes that is, paying attention to one ear only. Although its usefulness is intuitive, the effectiveness of such a centering sound is harder to measure that of its visual counterpart (the effect of the fixation cross displayed on the screen during visual tests can be monitored via eye-tracking) and its design needs further investigation and the use of an appropriate symbolic sound in place of the noise burst.

Original bird sounds were mono, with 16 bit coding and Fs=44.1 kHz.

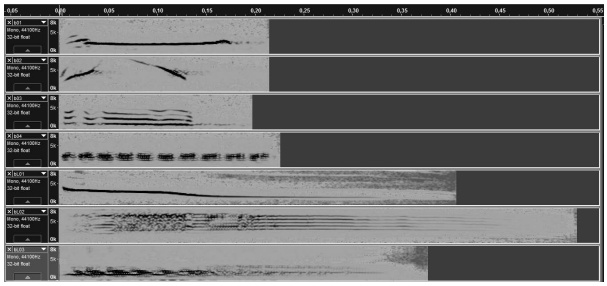

Figure 3: Spectrograms of the bird chirps sound-set used in our SOAP experiment. Top to bottom: b01 (Mustached Warbler), b02 (Marsh Warbler), b03 (Long-tailed tit), b04 (Little Auk), bL01 (Northern Cardinal), bL02 (Gray Catbird), bL03 (Carrion Crow).

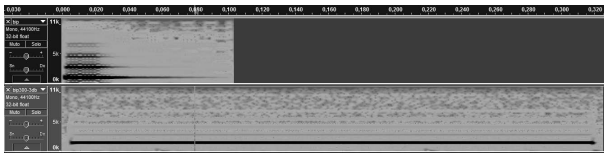

Figure 4: Spectrograms of the "beep-like" sounds used to test the RT *simple case*. Top: short "beep" (110 ms) with exponential decay. Bottom: long "beep" (325 ms) with sharp onset and decay.

2.4. Procedure

All subjects ran the preliminary test (Exp1) to assess their median RT where the sound used for the each trial could be either a "beep-like" one, derived from a complex tone, or one of the bird chirps used in Exp2, which we refer to as the "mixed" condition. Three subjects also repeated the same test using a trial sequence with patterns made of beep-like sounds only. These repeated sessions allowed the experimenter to have more insight into the effect of the task and complexity of the task and the sounds on the RT values which were to be used as baselines for the main experiment (Exp2). The comparison of the gathered RT values for these three subjects are shown in Fig. 5. As described in Section 2, the sound patterns were presented in isolation over the spatialized backgrounds, similar to Kayser et al. [6]. Exp1 took approximately 5 minutes to complete.

Participants were allowed to rest for a few minutes after Exp1 before starting the main experiment (Exp2). The latter was broken into three consecutive sessions to allow the participants to rest during the long sequence of trials needed by the full factorial design that uses the patterns of seven bird sounds, their presentation side, and the side of the ISI change as factors. The three sessions differed only in the background that was used which was held constant for the session duration. Catch trials with no change were included (in 5% of the trials). The subject was allowed to rest after each session. Each of the three sessions of Exp2 took approximately 4 minutes to complete. A summary of the main factors used for the design of the two experiments is provided in Table 1.

3. RESULTS AND DISCUSSION

3.1. Data preprocessing

We first processed the RT data to detect outliers and then transformed the accuracy results coming from Exp2 from categorical (binary) to a normalized, continuous accuracy score. To this end we pooled per bird/per subject data. The normalized, continuous accuracy values were computed for each participant as the detection frequency of each bird.

The distribution of RT values was observed to have a lognormal tendency but, to address outlier detection over a small sample, we used median and mean absolute deviation (MAD) which are location/scale robust statistics ([33]) rather than log–transformed RT values and standard deviation. A summary of the preprocessing steps taken for each of the two variables is provided in Table 2

Table 1: design summary of the experiments

Experiment	Variant	Sounds	Design notes
Exp1	"mixed"	all sounds in Figs. 3–4	Full factorial using sound (9 values) and side (L/R). One sound (sequence) per trial. Shortened-ISI position in the sequence is randomized. Trial order is randomized for each participant.
	"beep"	top sound in Fig. 4	as above, but using one sound only ("beep")
Exp2	-	all sounds in Fig. 3	Pairs of sounds (two sequences per trial, as shown in Fig. 1). Full factorial design with variables sound (7 values), side (L/R) and shortened-ISI presentation side (L/R). Shortened-ISI position in the sequence is randomized. Trial order is randomized for each participant.

Table 2: Preprocessing applied to detection accuracy and response time (RT) data

Variable	Data Type (raw)	Data Type (final)	Preprocessing
RT	continuous	continuous	Median RT used within subjects. Outlier: negative values and data points exceeding $3IQR$
Accuracy	binary (1/0)	continuous, normalized	Pooled binary values per subject, per bird. Continuous accuracy computed as accuracy rate. Missed- and catch–trials are not counted in.

3.2. Simple task (Exp1)

As expected a simpler task allows for shorter RT values. This was observed in Exp1, with respect to RT values from Exp2. We also observed, as shown in Fig. 5, that within the same task, the complexity of the presented sounds does play a role in the subject response.

This could be explained in terms of expectation since in the "mixed" presentation the "possible" scene was less predictable than in the "beep"-only presentation, where there was only one possible sound. However, the position of the event to be detected is randomly distributed within the last quarter of the trial, meaning that the listener already has become familiar with at least 3.5 s of the same pattern and suggesting that factors other than violation

of expectation may be more relevant here. In fact, the variation of the median RT between the two variants of Exp1 can be seen as an effect of the complexity of the scene along time (as patterns are presented in isolation in Exp1). We think this is different from early recognition since the subjects had enough exposure to all seven bird sounds to familiarize with them.

Finally, from a listening and qualitative standpoint, the "beep" sounds seem less complex than the bird chirps used for the "mixed" presentation. The faster detection of events in the former agrees with the shorter processing times of direct (visual) cues over symbolic cues, typically reported from attention research using cue/target designs. For a much broader discussion on the time course of attention, see Wright and Ward [3], pp. 23–29.

However, these observations may only be taken as preliminary given the limited number of subjects used here. More data are currently being collected.

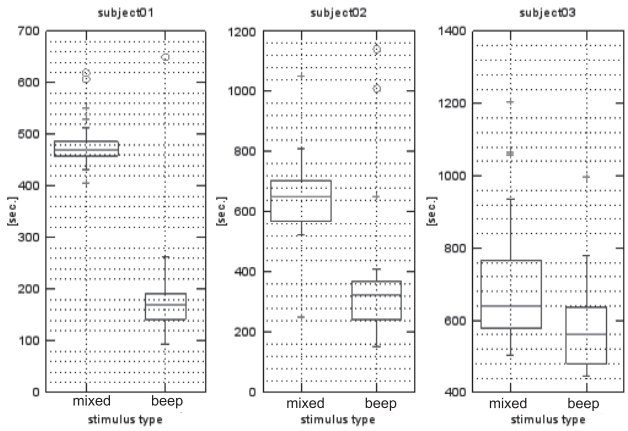

Figure 5: Within-subject (N=3) comparison of Response Time (given in ms) for different sessions of experiment Exp1 using stimuli of different complexity. See first row of Table 1 for design details.

3.3. Complex Streaming Task - SOAP (Exp2)

Our main conjecture about saliency and streaming is that high accuracy scores are tied to high saliency values. Using the pooled accuracy values (per bird, per subject) we can therefore attempt to define a saliency scale of the sounds. RT values are usually in agreement with the accuracy values in the pooled dataset, supporting the idea that if changes occur in the "active" stream, their detection is faster. This correlation is shown in Fig. 6

However, as observed in section 3.2, RT values are highly sensitive to the task and to the perceived complexity of the sound scene and of the sounds that are used, while accuracy scores seem to be less sensitive with respect to these factors. Accordingly we decided to use the pooled accuracy as "leading" behavioral data for the evaluation of the saliency of the sounds. This does not mean, however, that information provided by RT is not used. On the contrary, RT values gathered from Exp1 helped identify the acceptance time window to be used for the determination of accuracy. Background type (silence, pink noise, or natural noise) seemed to have no impact on the RT variance or on Accuracy scores and it was therefore ignored in the pooling process. Most of the subjects reported that it was easier to attend to both sound patterns when their rhythms were synchronized. This is in line with our statement about the importance of asynchrony for the reinforcement of streaming.

Finally, It was noted and reported by the subjects that the centering sound is often assimilated to the background as it becomes highly predictable after a few trials, therefore reducing its effectiveness. A solution to be adopted for future designs is to change the centering sound at every repetition, or use spearcons [34].

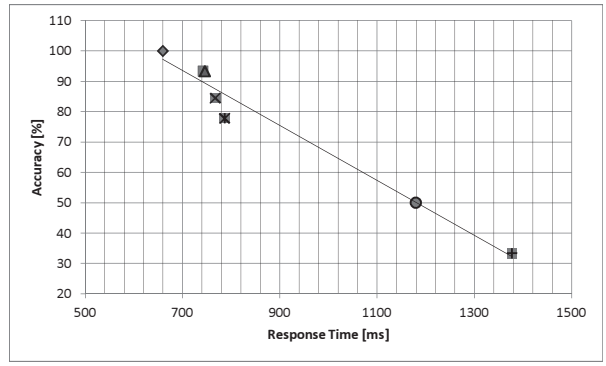

Figure 6: Accuracy % and RT pooled values across the subjects. Each point corresponds to a sound. The resulting saliency ranking, in descending order, is [b03,b01,bL03,b04,b02,bL01,bL02].

4. PRELIMINARY FEATURE ANALYSIS

Our initial selection of features with which to describe the bird chirps came from the broad set analyzed by Peeters et al. [35]. As the main focus of the present work was not a complete feature search, we only attempted to find initial correlations between the saliency ranking shown in Fig. 9 and some of the features that appeared in the global, local, temporal and spectral classes defined in reference [35], namely temporal centroid (defined as the center of gravity of the energy envelope), spectral centroid, harmonicity, and effective duration. Centroid definitions, frame size and hop-size for the windows used to calculate the time-varying descriptors are consistent with those specified in Peeters et al. [35].

However, for harmonicity (W_1 /W_0) and effective duration (τ_e) we decided to use the definitions outlined by Ando starting from the autocorrelation function (ACF) of the signals [36]. The effective duration is a measure of the decay time of the man lobe of the autocorrelation function. For more detailed information on the computation techniques for the(τ_e) we refer to the work of D'Orazio et al. [37]. As shown in Fig. 8, τ_e, the effective duration of the autocorrelation function, seems to have better correlation [$\rho(7) = 0.87$, $p < .01$, and $r(7) = 0.76$, $p < .05$] with the accuracy scores than other features do (see also Fig. 7) and therefore serves as a good predictor for the saliency ranking of the sound corpus we used. In Fig. 9 we show the relations between the behavioral and the saliency ranks predicted by τ_e. As already discussed, the ranking derived from the response time (RT) is more noisy leading to weaker correlation values with τ_e [$\rho(7) = -0.43$, $p < .337$, and $r(7) = -0.65$, $p < .115$] and is provided here only for comparison purposes.

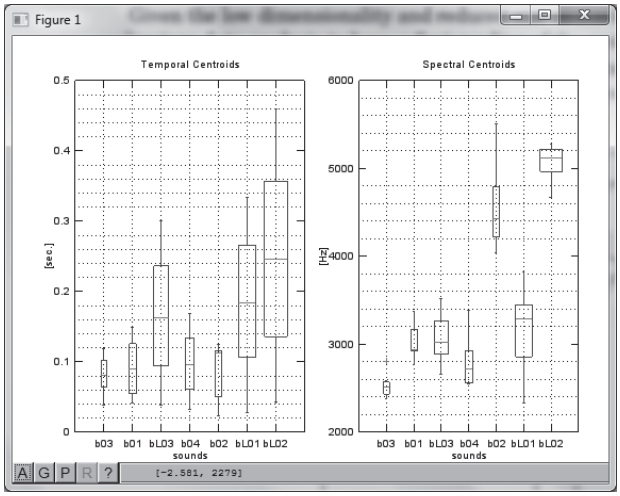

Figure 7: Statistics of the running temporal and spectral centroids of the sounds used for the SOAP experiment. Window-size=60 ms, hop-size=20 ms, FFT size=4096.

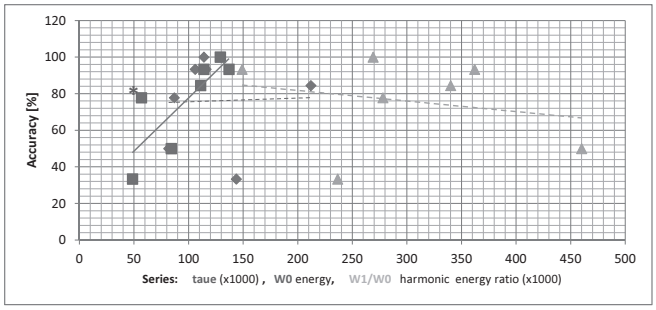

Figure 8: Relations between Accuracy% and the three features: τ_e *or "taue"* (ACF effective duration), W_0 (ACF value at lag=0), W_1/W_0 (W_1 is the value of the first peak of the ACF for $lag > 0$: the main harmonic). The red asterisk (*) indicates the data point "b04" in the τ_e series; τ_e and W_1/W_0 data are multiplied by 1000 for ease of display. Data points of τ_e are displayed in the same order as in Fig. 6.

Extreme saliency values find agreement between the perceptual and the feature driven dimensions, while the area with average saliency shows a less clear picture. In fact, this can be an effect of the small sample size.

5. CLOSING THE LOOP BETWEEN FEATURES AND SALIENCY: A CRITICAL VIEW

Since a formal definition of auditory saliency of natural sounds remains far from maturity, the method proposed here is intended simply to overcome the limitation of current approaches to auditory saliency that only use monaural schemes and avoid the presence of competing stimuli. Moreover, the use of sound patterns provides a suitable context for the analysis of the behavioral data over different timescales. The importance of time characteristics for saliency modeling is emerging from recent studies ([38, 39]).

Our approach builds on the idea that the speed and accuracy of detection of an event within a time pattern serves as an indirect measure of auditory saliency. We assume that when monitoring

	FEATURE BASED	BEHAVIORAL	
	τe	pooled Accuracy	pooled RT
Max	bL03	b03	b03
	b03	b01	b04
	b01	bL03	b02
	b04	b04	b01
	bL01	b02	bL03
	b02	bL01	bL01
min	bL02	bL02	bL02

SALIENCY

Figure 9: Saliency rankings from behavioral data (pooled accuracy and RT values) and from the effective duration (τ_e) values extracted from the running autocorrelation function of each bird sound.

a mixture of two sound patterns, the greater ease of spotting an anomaly in the stream with higher saliency will result in faster detection with fewer errors. However, although it may seem intuitive, we should bear in mind that the pattern and the sound we use as building block span different time scales. Thus, our measurements may well reflect local properties of the sound, the global pattern, or both, all of which represent other possible dimensions of saliency. Our work therefore represents an effort to narrow the definition of saliency, starting from a behavioral perspective, and using natural sounds in a binaural, ecologically valid, framework.

Ultimately, we are attempting to discover acoustic features that are correlated with performance on the experimental task, and propose these as causes of auditory saliency. Initially, we have conducted some local feature search, computed on the individual bird chirps, to determine whether these serve as predictors for detection accuracy. Our preliminary results suggest that some temporal descriptors are indeed well correlated with accuracy scores. However, the connection between features and accuracy needs to be addressed using a broader feature set and verified by experiments with more subjects, a task that is currently ongoing. Potentially relevant features will then be tested with other sounds and experimental framework to see whether their role in saliency can be generalized.

6. CONCLUSIONS AND FUTURE WORK

Despite the scarcity of auditory saliency computational models, several application areas are recognizing their potentials and the added value that they could bring in: ASR [16, 11, 40], HCI [41], sonification and sound design [23] among the others.

We developed a new experimental design to gather perceived saliency data taking advantage of the primitive processes (in the sense of [28]) regulating segregation and streaming of spatialized asynchronous patterns (SOAP). This schema is an evolution with respect to the current monaural, intensity based, saliency detection tests [6], in that it generalizes to situations that have more than one sound, separated in space, therefore allowing a better view over the nature of (primitive) saliency. It also introduces an ecologically valid testing framework for sonification applications by using "more natural" scenes.

Finding statistically significant results was not the goal of this preliminary work.. Rather it was testing this proof-of-concept design and its performance with real natural sounds. We explored

relationships among perceptual dimensions and acoustic feature candidates.

Even if highly correlated, response time (RT) values and accuracy of detection scores showed different sensitivity to the design variables and therefore we prefer to take accuracy as lead behavioral indicator for saliency. However, it is worth performing a more in-depth analysis of the behavior of RT with respect to the structure of the sounds. Our future research will try to clarify this relation using an evolution of the current assessment test (Exp1).

We tested the SOAP design with a set of bird chirps which is a subset of the everyday sounds, our intended scope (we exclude speech and man-made music from our definition of everyday sounds). The analysis of the preliminary data supports the robustness of the design and allowed the experimenters to explore some features that attempt to predict the observed data. The effect of different, low-intensity, continuous backgrounds, lacking any event structure, proved to be negligible when they were held constant over a block of trials. This confirms the intuitive idea that, apart from loudness masking effects, saliency is a local property associated with the sound objects that are actually "different" and changing with respect to the rest of the scene. This poses several questions about the definition of "the rest of the scene" since, ultimately, we will want to deal with relative saliency instead of an absolute one. Certainly, an absolute saliency, such as the one identified with this preliminary study, is a needed starting point for explorations in this direction.

Finally, our initial feature analysis over the set of sounds that we used favors the use of temporal descriptors such as the Effective Duration (τ_e) of the autocorrelation function (ACF) to predict the saliency ranking induced by our behavioral data.

Future work will be devoted to validate the proposed schema and initial findings on a larger subject pool. We also need to generalize to different sets of sounds. This will also serve to observe the behavior of the signal descriptors tested so far, to validate the goodness of (τ_e) as a saliency predictor and to expand the set of useful features, with a special focus on temporal descriptors as suggested in references [35, 36].

7. ACKNOWLEDGMENTS

F.T. thanks Dario D'Orazio and Simona de Cesaris for the discussions on the extraction methods for τ_e.

8. REFERENCES

[1] K. Adiloğlu, R. Anniés, E. Wahlen, H. Purwins, and K. Obermayer, "A Graphical Representation and Dissimilarity Measure for Basic Everyday Sound Events," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 5, pp. 1542–1552, Jan 2012. [Online]. Available: http://hal.inria.fr/hal-00684620

[2] 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.

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

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

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

[6] 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. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960982205011103

[7] C. Darwin, "Spatial hearing and perceiving sources," in *Auditory Perception of Sound Sources*, ser. Springer Handbook of Auditory Research, W. A. Yost, A. N. Popper, and R. R. Fay, Eds. Springer US, 2007, vol. 29, pp. 215–232.

[8] R. Carlyon and H. Gockel, "Effects of harmonicity and regularity on the perception of sound sources," in *Auditory Perception of Sound Sources*, ser. Springer Handbook of Auditory Research, W. A. Yost, A. N. Popper, and R. R. Fay, Eds. Springer US, 2007, vol. 29, pp. 191–213.

[9] B. C. Moore and H. Gockel, "Factors influencing sequential stream segregation," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 320–333, 2002. [Online]. Available: http://www.ingentaconnect.com/content/dav/aaua/2002/00000088/00000003/art00004

[10] C. J. Darwin and R. W. Hukin, "Auditory objects of attention: the role of interaural time differences." *Journal of experimental psychology. Human perception and performance*, vol. 25, no. 3, pp. 617–629, June 1999. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/10385981

[11] 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'12, 2012*, May 2012.

[12] V. Duangudom and D. V. Anderson, "Using auditory saliency to understand complex auditory scenes," *European Signal Processing Conference EURASIP*, no. Eusipco, pp. 1206–1210, 2007. [Online]. Available: http://eurasip.org/Proceedings/Eusipco/Eusipco2007/Papers/C1L-D01.pdf

[13] 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, JASA*, vol. 125, no. 4, part 2, 2009, pp. 2528–2528, poster 1pPP36.

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

[15] O. Kalinli, "Biologically inspired auditory attention models with applications in speech and audio processing," Ph.D. Thesis. Submitted to Univ. of Southern California, December 2009.

[16] 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.

[17] R. D. Patterson, "Guide lines for auditory warning systems on civil aircraft," Instituut voor Perceptie Onderzoek, RP/ne 82/01, Manuscript no. 413/II, February 1982. [Online]. Available: https://dl.dropbox.com/u/37237083/CNBHpapers/AuditoryWarningSystems_CAA82017.pdf

[18] R. Patterson, "Auditory warning sounds in the work environment," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 327, no. 1241, pp. 485–492, 1990. [Online]. Available: http://www.biomedsearch.com/nih/Auditory-warning-sounds-in-work/1970894.html

[19] 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. [Online]. Available: http://dx.doi.org/10.1037/1076-898X.14.3.201

[20] 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 International Conference on Mobile and Ubiquitous Systems:Computing, Networking and Services (Mobiquitous)*, Copenhagen, Denmark, December 2011, best Papers Session.

[21] 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. [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/105474698565677

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

[23] 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. [Online]. Available: http://dx.doi.org/10.1007/s12193-011-0062-8

[24] 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 (ICAD), 1999.

[25] 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.

[26] 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.

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

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

[29] S. Barrass and V. Best, "Stream-based sonification diagrams," in *Proceedings of the 14th International Conference on Auditory Display (ICAD2008)*, S. Patrick and O. Warusfel, Eds., IRCAM. Paris, France: IRCAM, 2008.

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

[31] M. R. Jones, "Temporal expectancies, capture, and timing in auditory sequences," in *Attraction, Distraction and Action Multiple Perspectives on Attentional Capture*, ser. Advances in Psychology, C. L. Folk and B. S. Gibson, Eds. North-Holland, 2001, vol. 133, pp. 191–229. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016641150180011X

[32] E. Bigand, C. Delbé, Y. Gérard, and B. Tillmann, "Categorization of extremely brief auditory stimuli: Domain-specific or domain-general processes?" *PLoS ONE*, vol. 6, no. 10, p. e27024, 10 2011.

[33] P. J. Rousseeuw and S. Verboven, "Robust estimation in very small samples," *Computational Statistics & Data Analysis*, vol. 40, no. 4, pp. 741 – 758, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167947302000786

[34] T. Dingler, J. Lindsay, and B. N. Walker, "Learnabiltiy of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech," in *Proceedings of the 14th International Conference on Auditory Display*, Paris, France, 2008, inproceedings. [Online]. Available: Proceedings/2008/DinglerLindsay2008.pdf

[35] 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.

[36] Y. Ando, T. Okano, and Y. Takezoe, "The running autocorrelation function of different music signals relating to preferred temporal parameters of sound fields," *The Journal of the Acoustical Society of America*, vol. 86, no. 2, pp. 644–649, 1989. [Online]. Available: http://link.aip.org/link/?JAS/86/644/1

[37] D. D'Orazio, S. De Cesaris, and M. Garai, "A comparison of methods to compute the 'effective duration' of the autocorrelation function and an alternative proposal." *J Acoust Soc Am*, vol. 130, no. 4, p. 1954, 2011. [Online]. Available: http://www.biomedsearch.com/nih/comparison-methods-to-compute-effective/21973350.html

[38] S. H. Chon and S. McAdams, "Timbre saliency vs. timbre dissimilarity ? what is the relationship?" vol. 19, no. 1. ASA, 2013, p. 035054. [Online]. Available: http://link.aip.org/link/?PMA/19/035054/1

[39] E. Kaya and M. Elhilali, "A temporal saliency map for modeling auditory attention," in *Information Sciences and Systems (CISS), 2012 46th Annual Conference on*, 2012, pp. 1–6.

[40] M. Slaney, V. Mahadevan, A. Thomas, and K. Horiguchi, "Salient distractors in speech: A recognition task for measuring acoustic salience," in *Proc. 13th Annual Conference of the International Speech Communication Association (Interspeech'12)*, 2012.

[41] C. Roda, Ed., *Human Attention in Digital Environments*. New York, NY, USA: Cambridge University Press, 2011.