This is a postprint version of the following published document:

Rodríguez-Hidalgo, A., Peláez-Moreno, C., &
Gallardo-Antolín, A. (2018). Echoic log-surprise: A
multi-scale scheme for acoustic saliency detection.
Expert Systems with Applications, 114, 255-266.

doi:<https://doi.org/10.1016/j.eswa.2018.07.018>

© Elsevier, 2018

This work is licensed under a [Creative Commons Attribution-NonCommercialNoDerivatives 4.0 International License](https://creativecommons.org/licenses/by-nc-nd/4.0/).

Echoic log-surprise: a multi-scale scheme for acoustic saliency detection

Antonio Rodríguez-Hidalgo, Carmen Peláez-Moreno,
Ascensión Gallardo-Antolín

*Department of Signal Theory and Communications
Universidad Carlos III de Madrid (Spain)*

Abstract

Saliency detection is a key cognitive mechanism that prioritizes particular stimuli over others to allow the brain to take decisions about the relevance of each incoming stimuli in the process of exploring the world. In spite of the similarities in the way the human brain processes auditory and visual cues, a remarkable difference exists in terms of the time span of sensory memories retained: given the more ephemeral nature of auditory stimuli in comparison with their visual counterparts, echoic memories last longer than iconic ones. In this paper, we introduce a bottom-up auditory saliency detection algorithm that profits from this observation. Specifically, we put forward a new technique, that we termed *echoic log-surprise* that combines an unsupervised statistical approach based on Bayesian log-surprise and the biological concept of echoic or Auditory Sensory Memory.

Our algorithm computes several independent log-surprise cues in parallel considering a wide range of memory values, with the aim of leveraging saliency information from different temporal scales. Then, we explore several statistical metrics to combine these multi-scale signals in a single temporal saliency signal including Renyi entropy, Jensen-Shannon divergence, Cramer or Bhattacharyya distances.

Again in contrast with visual saliency empirical frameworks there is no consensus about the way acoustic saliency detection should be evaluated

Email addresses: `arodigue@pa.uc3m.es` (Antonio Rodríguez-Hidalgo),
`carmen@tsc.uc3m.es` (Carmen Peláez-Moreno), `gallardo@tsc.uc3m.es`
(Ascensión Gallardo-Antolín)

mainly due to the lack of devices that could take the role of eye-trackers in providing ground truth measurements. We have adopted Acoustic Event Detection tasks as adequate proxies for our purposes. Results suggest that our echoic log-surprise method outperforms classical techniques such as energy thresholding or voice activity detection and previous state-of-the-art saliency algorithms, such as Kalinli’s and conventional log-surprise.

Keywords: acoustic saliency, echoic memory, multi-scale, statistical divergence, Jensen-Shannon, acoustic event detection

1. Introduction

Humans live immerse in a complex environment, where our sensors receive a vast amount of stimuli. Since our brain cannot process all these inputs simultaneously, it prioritizes some of them through the mechanism of *attention*. The property of a sensory cue that determines its relevance is named *saliency*, whether acoustical, visual or any other sensory type. Irrespective of the modality, two main precedence mechanisms can be distinguished: *top-down*, where there is an specific task or aim guiding the exploration of the environment, or *bottom-up*, where the exploration is only driven by the inherent characteristics of the stimuli.

Visual saliency has received a lot of attention with a wide variety of algorithms available such as (Itti et al., 1998; Itti & Baldi, 2005; Zhang et al., 2008; Pan et al., 2016; Kümmerer et al., 2016; Pan et al., 2017), for example. Following (Zhao & Koch, 2013), visual saliency procedures can be split into three different stages. First, a feature extraction phase aiming at producing relevant and compact representations of the input. Second, further processing of the individual characteristics, usually relying on human visual mechanisms’ imitation to produce a set of support images. And third, combination or fusion of the previous images through different techniques, most recently based on supervised machine learning schemes. The progress on visual saliency systems has been made possible due to gaze capturing devices or eye-trackers. Moreover, many authors have made the datasets publicly available (Zhao & Koch, 2013) contributing to the development of this discipline. This pretty general scheme has been applied to different tasks, such as image retargeting (Chen et al., 2015), human action recognition (Yi & Lin, 2013) and improving mass detection of breast cancer (Agrawal et al., 2014), among others.

In this work, however, we focus on acoustic saliency, where our goal is to determine if an acoustic event stands out from its acoustic background or environment. In contrast with its visual saliency counterpart, there is a lack of audio databases labelled with this characteristic (Kaya & Elhilali, 2017) and therefore, most of the solutions adopt an unsupervised modelling strategy. One successful approach consists of reproducing the first stages of the visual saliency algorithms stated above by using spectrogram-like representations of the audio signal as the input image, followed by some adaptation of the parameters of the model for the particular case of audio. A foundational model is due to (Kayser et al., 2005), where those modifications include the usage of different auditory features such as frequency and temporal contrasts. From those bidimensional features they generate a pyramid with different scales that are related using a center-surround mechanism inherited from its visual saliency counterpart (Itti et al., 1998) that emulates the behavior of the visual receptive field. Then, the resultant maps are normalized and finally added to conform an acoustic saliency map. We notice that this work was one of the first approaches to acoustic saliency and in spite of considering several pyramid scales did not introduce any specific acoustic memory mechanisms.

Later on (Kalinli & Narayanan, 2009) kept the original structure developed by Kayser and introduced two new features, namely *orientations* and *pitch*. Nevertheless, some authors argued that acoustic saliency should be computed considering exclusively past information a feature that proposals such as Kayser violate since they employ the spectro-temporal equivalent of spatial filters leading to a non-causal analysis. Taking this into account, (Kaya & Elhilali, 2012) proposed a model that strictly uses causal features in a scheme inspired by that of Kayser. The acoustic characteristics that they put forward are the amplitude envelope of the temporal waveform of the acoustic signal, the spectrogram, bandwidth, rate information and pitch.

Instead of looking for a biologically inspired model for saliency, an alternative approach consists in the development of statistical models of saliency based on the idea that rare feature values indicate the presence of salient events. In (Tsuchida & Cottrell, 2012) a Bayesian approach where a cochleogram computed using Gammatone filters is employed as the audio spectro-temporal representation. After reducing the its dimensionality using Principal Component Analysis (PCA) a Gaussian mixture model is fit to get the distribution that models the prior knowledge about the features. Finally, saliency at each time instant is computed as the self-information of the feature vector observed at this specific time, which can be defined as its probability according

to the aforementioned model, in such a way that the smaller the probability the more salient the feature is considered.

In this line of thought, (Schauerte & Stiefelhagen, 2013) employed the so-called Bayesian surprise adapted for auditory cues. Bayesian surprise is computed using Kullback-Leibler divergence for each frequency band of a spectro-temporal representation, considering a particular amount of past temporal information. Schauerte et al. proposed two parametric models based on Gamma and Gaussian distributions. Since the source code was made available from the authors, this will be one of our baselines. In (Rodríguez-Hidalgo et al., 2016), we demonstrated that the logarithmic transformation of the Bayesian surprise is advantageous to avoid scaling and normalization problems, since the original Bayesian surprise algorithm detects extremely salient elements that usually mask the rest of possible candidates. We termed this modification *log-surprise*.

Here, we go a step forward by considering the concept of Auditory Sensory Memory (ASM), which is defined as the capacity of our brain to store certain acoustic information unexpectedly perceived (Cooper et al., 2006) and take into consideration the differences that exist in terms of the time span of auditory and visual sensory memories retained: given the more ephemeral nature of auditory stimuli in comparison with their visual counterparts, echoic memories last longer than iconic ones. Several authors (Cowan, 1984; Schröger, 2007) suggest that acoustic information can be stored for approximately 10 – 20 seconds. However, posterior works (Bottcher-Gandor & Ullsperger, 1992; Sams et al., 1993; Gomes et al., 1999; Glass et al., 2008; Grossheinrich et al., 2010) have empirically determined that this storage time depends on the maturational development of subjects, being close to 10 seconds for adults. Our proposal computes Bayesian log-surprise in parallel at several time spans, considering memory lengths from 140 milliseconds to 14 seconds. Then, the output saliency signals are introduced into a fusion stage, where we compute pairwise statistical divergences and distances that are finally added to obtain a global temporal saliency signal. The spectro-temporal representation that we consider in this work are the the log-Mel frequency bands.

To evaluate our proposal we have applied our saliency detectors to the task of Acoustic Event Detection (AED) since the human mechanisms to solve an AED task are, in principle, those of saliency and attention. As already mentioned, there are no databases labelled with the acoustic saliency (Kaya & Elhilali, 2017) and therefore there is no consensus about how to as-

sess these algorithms. There are, however, AED challenges such as DCASE (Mesaros et al., 2016b) with different target cases such as acoustic scene recognition (Schröder et al., 2017) or audio tagging (Xu et al., 2017) that are useful for our purposes. We have also employed DARES-G1 (Grootel et al., 2009) and CLEAR06 UPC-TALP (Temko et al., 2007), which include the onset and offset labels of certain acoustic events. Given the availability of tags, usually the focus of these challenges is also on Acoustic Event Classification (AEC) and, for this reason, most of the adopted solutions employ supervised techniques for classification and feature extraction, among many others (Schröder et al., 2017; Xu et al., 2017; Schröder et al., 2015; Virtanen et al., 2016). It is worth noting that our goal is not to solve the AED/AEC task but to compare unsupervised saliency detection algorithms. Therefore, we test the performance of our system against two other saliency proposals, *Kalinli* (Kalinli & Narayanan, 2009) and *log-surprise* (Rodríguez-Hidalgo et al., 2016) but also against a statistical model-based *Voice Activity Detector* (VAD) (Sohn et al., 1999) and a simple *Energy thresholding* of the spectro-temporal representation.

The rest of the paper is organized as follows: first, we present an outline of our contributions followed by an explanation of how Bayesian surprise is applied to saliency detection that sets the scene for the description of our approach in Section 4. Then, the experimental set up is presented in Section 5 followed the experiments in Section 6 with some conclusions and further work closing the paper in Section 7.

2. Contributions

Our first contribution is the so-called multi-scale saliency, which essentially considers several acoustic saliency signals computed at various time spans and fuses them using statistical divergences. With this novel method, we try to emulate the Auditory Sensory Memory (ASM) mechanism, which determines how long humans can remember acoustic information before our brain forgets it. In order to do so, we use Bayesian *log-surprise* considering several memory values, and finally combine all the aforementioned signals, which would eventually lead to a more reliable acoustic saliency signal.

Our second contribution is related with the analysis of the fusion stage, where we consider several statistical divergences and distances to combine the individual saliency signals, as we want to determine how critical the fusion operation is by itself. The best results are obtained using the *Jensen-*

Shannon divergence, that depends directly of Shannon entropy, a remarkably relevant concept of Information Theory.

Our last contribution is the comparison and analysis of different acoustic saliency techniques (a simple energy thresholding, a statistical model-based voice activity detector, and *Kalinli* and *log-surprise* algorithms) against our proposal for the task of acoustic event detection. In particular, we have experimented the different methods on three databases that cover several realistic scenarios and recording conditions. As results show, our multi-scale saliency algorithm outperforms the baseline methods for all the databases considered.

3. Bayesian Surprise

The essence of *Bayesian surprise* (Itti & Baldi, 2005) is the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) defined as:

$$D_{KL}(X_t||X_{t-1}) = \frac{(\mu_t - \mu_{t-1})^2}{2\sigma_{t-1}^2} + \frac{1}{2} \left(\frac{\sigma_t^2}{\sigma_{t-1}^2} - 1 - \log \frac{\sigma_t^2}{\sigma_{t-1}^2} \right) \quad (1)$$

in the case of two Gaussian random variables $X_t \sim \mathcal{N}(\mu_t, \sigma_t^2)$ and $X_{t-1} \sim \mathcal{N}(\mu_{t-1}, \sigma_{t-1}^2)$ where μ_t and μ_{t-1} represent the means of each random variable and σ_t and σ_{t-1} , their standard deviations. The variable X_{t-1} models the past distribution of the signal whilst X_t represents the current signal distribution. Consequently, for Bayesian surprise the KL divergence is used to measure the similarity between two consecutive time segments.

Now, to compute the Bayesian surprise of the present audio signal compared to the past, or in other words, the new information that the signal carries at present and was absent in the past, we obtain a spectro-temporal representation $X(\omega, t)$ from the audio signal $x(t)$, for example, a spectrogram as the one depicted in Figure 1. Then, we update μ and σ^2 for every temporal instant t , and compute the KL divergence for the analyzed temporal segment according to Eq. (1) as a function of (ω, t) , since the analysis is performed independently for each frequency band of the chosen spectro-temporal representation. This way, a surprise map is produced as the ones shown in Figure 1. Finally, in order to get a temporal saliency signal we average the content of all the frequency bands producing a unidimensional signal $s(t)$ representing acoustically salient events along the time axis.

Ideally, this scheme should detect those acoustic events that are particularly salient. However, in (Rodríguez-Hidalgo et al., 2016) we demonstrated

that the extremely spiky nature of $s(t)$ made it difficult to detect some weaker events that were masked by others with much larger amplitudes. A natural way to solve this problem was to apply a logarithm transformation to the saliency map obtaining what was called Bayesian *log-surprise*, the hypothesis being that the distributions of the log-energies are better modelled with the Gaussian distributions employed. An example is depicted in Figure 2, where yellow regions and red lines represent, respectively, the ground truth (actual positions of the acoustic events) and the salience signals $s(t)$.

4. Our model

As we explained previously, in order to compute Bayesian surprise we need to estimate μ and σ for every t and w for a given spectro-temporal representation. However, their estimation depends on the amount of data processed. Therefore, we need to define the memory of the system, i.e., the number of previous samples N that will be considered. The example depicted in Figure 1 shows three surprise maps that were computed using different N values. The images show that the performance of the surprise is strongly-bounded to the memory of the system producing smoother maps as N grows.

Our proposal here uses the Bayesian log-surprise, as depicted in the schematic of Figure 3 but there is a need to find a balanced value of the memory: on the one hand, using a small value of N could be helpful to detect every possible event of an acoustic signal, but it would also produce false-alarms. On the other, a big value of N would detect only those events whose spectral properties are particularly salient, rejecting events that are relevant (a low recall). To overcome these issues we propose a multi-scale memory system that is divided in two different stages as depicted in Figure 3: multi-scale saliency determination and fusion.

4.1. Multi-scale saliency

The first stage computes saliency features considering Gaussian log-surprise, and depends on two different values: dth and N_0 , that represent the depth or number of temporal scales (levels) used in the system and the initial memory value, respectively. Using the initial value $N = N_0$, we establish that the deeper the system is, the bigger the memory used to compute Bayesian log-surprise should be.

Our scheme tries to emulate the concept of Auditory Sensory Memory, which according to (Cooper et al., 2006) can be defined as a passive storage of coded acoustic information processes in two different stages. During the first stage acoustic features are analyzed and coded, a task that could take 200–300 ms according to (Cowan, 1984). In the second stage, the information would be retained from approximately 10 – 20 seconds according to (Cowan, 1984; Schröger, 2007). Consequently, a listener would perceive and remember an acoustic event for the aforementioned period of time.

In order to emulate the coding stage of ASM we consider dth values of memory samples N for producing a set of saliency signals $\{s_i(t), i = 1, 2, \dots, dth\}$ that will be subsequently fused in the next stage of our scheme. By doing so we expect to keep an acoustic record of the information that is perceived by a person along time, and how significant the appearance of new information is when it is compared with the already stored one.

Finally, when all the log-surprise signals are computed we normalize them to be in the range $[0, 1]$.

4.2. Fusion

Once the results of the multi-scale memory saliency stage are got they need to be fused to obtain the final $s(t)$ function. As all the previously computed saliency signals depend on the temporal axis, it seems adequate to combine them using a measure of their amount of information. Consequently, we analyze several alternatives in this Section.

A more detailed representation of this stage is depicted in Figure 4, where the fusion module is divided into three different steps. In the first one we divide each surprise signal into audio chunks of M samples, considering always that selecting a chunk for the time instant t_s implies that we process all the surprise elements of the i_{th} saliency signal $s_i(t)$ that belong to the range $(t_s - M) < t \leq t_s$.

Once we have obtained the segments for each time instant t_s of every surprise signal available, $s_i(t)$, we compute their histograms with a number of bins, n_{bins} , on the second step of the stage. The central square box of Figure 4 shows three graphical representations of the histograms computed for each time segment, individually for each saliency signal $s_i(t)$.

Finally, in the histogram fusion step we need to compute the divergence or distance operation for each time segment considering the corresponding histograms, which determines the differences among the available log-surprise

signals. In the example depicted in Figure 4 we show three different log-surprise signals, and how fusion works for two different time segments, $T1$ and $T2$. After the histogram fusion step we obtain a temporal cue that represents our multi-scale memory saliency signal.

For this purpose, we consider four distances or divergences for distributions as we explain in the next paragraphs. The comparison of more than two distributions can be either be solved by performing pairwise comparisons or by using a metric with this ability as the Jensen-Shannon one.

As we consider always more than two saliency signals and their respective histograms, we compute their divergences and distances considering pairs of ascending memory values. For example, if we have a depth of $dth = 3$, we would combine the histograms from $s_1(t)$ with $s_2(t)$ yielding $s_{1:2}(t)$, then the histograms of $s_2(t)$ with the ones of $s_3(t)$ would give us $s_{2:3}(t)$. Finally, we would add together the outcome signals obtaining the final saliency signal $s(t) = s_{1:2}(t) + s_{2:3}(t)$. Consequently, we compare each saliency signal only with the one that has the immediately higher memory value, in order to favor some redundancy in the resultant saliency signal $s(t)$. Finally, $s(t)$ is thresholded along the temporal axis with a given threshold, which in this case corresponds to its average value, producing the final binary signal which indicates the presence or absence of a relevant acoustic event. Note that in the example depicted in Figure 4 we have represented the particular case of Jensen-Shannon divergence in which the summation operation is not necessary, as it will be explained in next paragraphs, where the divergences and distances we have considered in the fusion stage are detailed.

4.2.1. Rényi divergence

Rényi divergence (van Erven & Harremoës, 2012) represents a generalization of the Kullback-Leibler divergence, and its mathematical equation is defined as follows:

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \left(\sum_{i=1}^{n_{bins}} \frac{p_i^\alpha}{q_i^{\alpha-1}} \right), \quad (2)$$

where P and Q represent the two distributions whose divergence we are computing. Moreover, several divergences can be obtained depending on the value that the parameter α takes considering always that $\alpha \geq 0$. In this work we focus on two divergences:

- $D_1(P||Q) = \sum_{i=1}^{n_{bins}} p_i \log(p_i/q_i)$, which represents the Kullback-Leibler divergence for two discrete random variables.
- $D_\infty(P||Q) = \log \sup_i (p_i/q_i)$, that is computed as the logarithm of the supremum of all the ratios of probabilities.

Independently of the value that α takes $D_\alpha \geq 0$ as it is demonstrated in (Cover & Thomas, 2006), considering that the bigger D_α is the bigger will be the differences among P and Q .

4.2.2. Cramer distance

This distance (Székely, 2002), also known as Energy distance, is defined as the L2-norm of two distributions P and Q as it is depicted below:

$$D_{Cramer}^2(P, Q) = \sum_{i=1}^{n_{bins}} (p_i - q_i)^2, \quad (3)$$

where P and Q values must be real-valued. It is worth noting that, $D_{Cramer}^2 \geq 0$ for every possible value of P and Q .

4.2.3. Bhattacharyya distance

If we consider two distributions P and Q , it is defined in (Bhattacharyya, 1943) as

$$D_{Bhatta}(P, Q) = -\log \left(\sum_{i=1}^{n_{bins}} \sqrt{p_i q_i} \right). \quad (4)$$

Bhattacharyya distance also verifies that $D_{Bhatta} \geq 0$.

4.2.4. Jensen-Shannon divergence

Jensen-Shannon divergence can be obtain for an arbitrary number of distributions P_i for $i : \{1, dth\}$ where dth is the total number of distributions in our case, i.e. the number of log-surprise signals $s_i(t)$. Jensen-Shannon divergence (Endres & Schindelin, 2003) can be defined as

$$D_{JSD}(P_1, \dots, P_{dth}) = H\left(\sum_{i=1}^{n_{bins}} \pi_i p_i\right) - \sum_{i=1}^{n_{bins}} \pi_i H(p_i), \quad (5)$$

where P_i represents the probability of the i -th distribution, $H(P_i)$ is the Shannon entropy for distribution P_i and π_i the weight that we want

to assign to it. Since we do not have a priori information about which of the distributions is more likely, we set this weight uniformly for every distribution considered. Finally, Jensen-Shannon divergence bounds are $0 \leq D_{JSD}(P_1, \dots, P_{dth}) \leq \log(dth)$.

Note that Jensen-Shannon divergence $D_{JSD}(P_1, \dots, P_{dth})$ is able to measure divergences between a number dth of signals at the same time, and, therefore, it is not necessary to combine partial salience signals by using the addition operation as in the previous cases.

5. Experimental setup

5.1. Baseline algorithms

Since we are testing our auditory saliency algorithms on the task of acoustic event detection, we compare our system with the energy computed directly from the spectro-temporal representation, a voice activity detector, and two different auditory saliency techniques: *Kalinli* model and *Bayesian log-surprise*.

5.1.1. Energy thresholding

This scheme computes the energy from the spectrogram considering Parseval's theorem (Huang et al., 2001), which states that the energy can be obtained both from time and frequency domains.

Then, we set a threshold equal to the average value of the energy of the audio signal. Values above this threshold are considered to be salient, a simple proposal that should work reasonably well when the Signal-to-Noise Ratio (SNR) is high. This approach can be considered as a very basic VAD algorithm, introduced later, as it only uses the energy magnitude to distinguish among relevant signal and background.

5.1.2. Voice Activity Detector (VAD)

A voice activity detector is a technique specifically designed for the task of human speech detection on an acoustic signal (Ramirez et al., 2007; Sangwan et al., 2007; Chang et al., 2006; Yoo et al., 2015). It consists of two basic elements, a feature extractor and a decision module that actually classifies an audio clip into speech or non-speech classes. Sometimes, spectral subtraction techniques are applied to the input auditory cue in an attempt to reduce the noise of the signal and ease the detection.

The VAD model that we use is the one proposed by (Sohn et al., 1999) and included in Voicebox Toolbox (Brookes, 1997), which actually uses the Likelihood Ratio Test (LRT) as the decision rule to discriminate between signal and noise. The features introduced into the LRT are the Discrete Fourier Transform (DFT) coefficients of the signal, noise and noisy signal which are independently modeled using Gaussian distributions. Finally, a Hidden-Markov-based hang-over scheme is employed to smooth the output signal/noise decision, which reduces misdetections.

5.1.3. *Kalinli saliency models*

This model is inspired in *Kayser* acoustic saliency model (Kayser et al., 2005), which computes three conspicuity maps from the spectrogram: intensity, frequency and temporal contrast. Then, those maps are resized into several scales to form three different dyadic pyramids, one for each feature. Then, a center-surround difference is computed, which resembles certain operations developed by human eye. Finally, the processed maps are normalized and fused in order to get the final saliency map.

Kalinli model (Kalinli & Narayanan, 2009) differs from the previous one in which two additional features are introduced, namely pitch and orientations, and in the use of a more sophisticated normalization process. The Matlab implementation that we used can be found at (Macaluso, 2010).

5.2. *Spectro-temporal representations*

As auditory saliency algorithms are usually inspired in their visual counterparts, it is more practical to model acoustic information using an image-like representation. A very common approach is the spectrogram as an image, as it represents the magnitude of different frequency bands along the temporal axis, producing a two-dimensional representation of the acoustic cue.

However, alternative spectro-temporal representations can also be considered. Cochleograms are biologically-inspired spectro-temporal representations that mimic the behaviour of the human cochlea, in particular, the non-linear logarithmic-like transformation of the frequency domain taking into account the critical bands described by (Fletcher & Munson, 1933), i.e. assuming that the cochlea resembles a bank of filters whose central frequencies are non-linearly distributed according a logarithmic law.

Triangular shaped filters are typically employed to model these bands. Figure 5 is a depiction of them following

$$B(f) = 1125 \ln(1 + f/700), \quad (6)$$

where $B(f)$ represents the Mel-scale (Huang et al., 2001) transformation of the frequency f measured in Hz.

Once the input signal $x(t)$ is filtered using the bank of filters, the log-energies per band are computed, obtaining the cochleogram.

Both spectro-temporal representations are represented as $X(\omega, t)$ in this work. However, a preliminary experimentation determined that log-Mel energy bands produce better results than the spectrogram for all the saliency modalities analyzed. Consequently, for the sake of brevity, we use the first one as the standard spectro-temporal representation for the experiments and analysis in the rest of the paper.

5.3. Databases

In this work, we decided to make use of several datasets that were actually collected for the task of AED/AEC tasks. These usually involve two different problems: the detection of acoustic events and their classification according to a specific set of class labels. We are considering auditory saliency as an unsupervised task and therefore we do not make use of any label whatsoever to train the systems. Consequently, our model should be compared with unsupervised event detection algorithms, without any kind of classifier at the output. Three datasets are used in this work:

5.3.1. DARES-G1

This dataset (Grootel et al., 2009) consists of approximately 800 different classes, and almost 3000 labeled events. Each audio clip has a duration of 60 seconds, recorded with a sampling frequency $f_s = 44100Hz$ and a resolution of 16 bits. The recordings were captured on different scenarios, such as streets, public buildings, etc.

5.3.2. UPC-TALP database of isolated meeting-room acoustic events

This dataset was recorded for the AED task CLEAR06 (Temko et al., 2007), where several experiments were developed. We focus on the subset *UPC-TALP*, particularly in the microphone number 3, which comprises 30 files and approximately 1030 tags. Some of the available classes are: *door knock*, *steps*, *chair moving*, *key jingle*, etc. Audio clips were sampled using $f_s = 44100Hz$ and 16 bits of resolution.

5.3.3. Detection and Classification of Acoustic Scenes and Events 2016 (DCASE2016)

This database (Mesaros et al., 2016b) was developed for a challenge celebrated during 2016. It comprises four different sets, which focus on four independent tasks. We particularly selected the ones used in Tasks 2 and 3, whose annotations include three elements: onset, offset and tag of the event. Tasks 1 and 4 are related with Scene Classification and Audio Tagging, respectively, and do not include any temporal label that can be used for our evaluation. All the audio clips are recorded using a sampling frequency of $f_s = 44100Hz$, and 24 bits of resolution. The duration of each clip depends on the analyzed subset.

Task 2 subset was synthetically developed, and contains data for three different Event-to-Background Ratio (EBR) values (-6 , 0 and $6dB$), which determines the relationship between event and background energies. This subset contains 72 audio clips including more than 2000 labeled events and is divided into two different groups: validation or development set (18 audio clips) and test (the remaining 54 clips). The acoustic events belong to 11 different categories: *clearing throat*, *coughing*, *door knock*, *door slam*, etc. In addition, there are two different scenarios, *monophonic* and *polyphonic*, where monophony implies that there exists a single event for a certain time window, whilst there are several overlapping acoustic events occurring at the same time for polyphony.

Task 3 contains audio clips from two acoustic scenarios, namely *residential area* and *home*, with their own specific labels (almost 1500).

5.4. Metrics

Measuring acoustic saliency is a difficult task due to the lack of labelled data. Therefore, we will follow the guidelines and algorithms developed for the DCASE2016 challenge (Mesaros et al., 2016b) for event detection. Their evaluation code includes two different metrics: segment and event based with labeled onsets, offsets and class. As we are not concerned with the classification stage, we will focus on the onset-offset evaluation exclusively without considering class labels. Furthermore, we choose the event-based evaluation, which is more restrictive than the segment-based modality as it only considers temporal windows with a tolerance of hundreds of milliseconds, instead of a chunk of one second length as it happens with segment-based evaluation.

Distance operation	dth	N	F-score
Cramer	4	42	0.6905
JSD	5	34	0.7690
Renyi-1	4	22	0.7203
Renyi-Inf	4	16	0.7560
Bhattacharyya	5	34	0.7365

Table 1: Multi-scale saliency parameters for each divergence/distance operation.

The metric used in this work is therefore, the F-score,

$$F = 2 \frac{P \cdot R}{P + R}, \quad (7)$$

where P and R are the precision and recall scores, respectively.

We considered the onset exclusively with a tolerance margin of 200 ms, as defined for the challenge, and disposed of the classification labels. More details about metrics for AED can be found at (Mesaros et al., 2016a).

In spite of this, with the purpose of having a better understanding of the performance of the different algorithms, we have analyzed the detection score for each of the classes individually.

5.5. Parameter settings

Depending on the stage of our system, some technical parameters need to be discussed. First of all, each audio signal is resampled from its original sampling frequency to $F_s = 22000Hz$. Then, the spectrogram is computed considering 1024 frequency components, a Hamming window of 20 ms and an overlapping factor of the 50%.

If the audio clip is stereophonic, both channels are averaged in order to get a single *monophonic* input. Then, to obtain the cochleogram we use a mel-scale filterbank with 150 filters. This choice guarantees the necessary number of filters to apply the desired depth in the saliency algorithms. We have studied several values of depth dth and initial memory N_0 ranging $dth \in \{1...6\}$ and $N_0 \in \{14...44\}$. Optimal values in combination with the divergence metric chosen have been obtained using the held-out validation set from DCASE2016 Task 2.

For all the compared systems the fusion stage histograms are computed using $M = 50$ temporal samples and $n_{bins} = 20$ bins. Note that each sample corresponds to 20 ms.

6. Experiments

6.1. Experiment 1: performance analysis under several monophonic datasets

First we will analyze how the different detection algorithms perform under two *monophonic* datasets, namely: DARES-G1 and UPC-TALP.

Figure 6 shows the results, where we can observe how all the log-surprise based algorithms outperform the baselines. In fact, *Kalinli* performs poorly in comparison with the rest. Also, *Energy* and *VAD* produce better results than *Kalinli* for both databases.

Though the performance of all the log-surprise based algorithms is quite similar, some differences between both databases can be appreciated: in both cases, the *Jensen-Shannon* divergence obtains the best results followed by the basic *Log-surprise* in the case of DARES-G1 and *Bhattacharyya* divergence in the case of UPC-TALP.

6.2. Experiment 2: DCASE2016, a study of monophony and polyphony

In this section we focus on the DCASE2016 dataset. As we mentioned in section 5.3, this dataset comprises two different tasks. Task 2 contains synthetically generated files with three Event-to-Background rates (EBR). Also acoustic events may overlap in time, which represents the *polyphonic* scenario. Task 3 considers two different scenarios: *home* and *residential-area*. The content clips of this subset were recorded in real environments, where events overlap naturally and therefore EBR was not measured.

Table 2 shows the results obtained for Task 2, where the top performing algorithm is *Jensen-Shannon* (JSD) immediately followed by *Renyi-Inf* and the rest of our fusion saliency proposals. We observe that neither *Energy* nor *VAD* produce good results in comparison with the rest. In addition, the F-scores achieved by *VAD* are worse than those obtained by the simple *Energy* algorithm. One possible explanation for this behavior is that VADs are usually developed for speech detection, which implies that their design characteristics are optimized for this particular task. The other baseline algorithm, namely *Kalinli*, performs poorly again as in the previous section. On the other hand, the basic *log-surprise* without the multi-scale fusion is able to produce better results than all the rest baseline algorithms though worse than all the multi-scale versions.

As an example, if we compare *Jensen-Shannon* against *log-surprise* for the *monophonic* scenario considering $EBR = -6dB$ we observe that our proposal improves the results of the latter in approximately a 44.1%, and

a 43.9% for $EBR = 6dB$, a meaningful result considering that the *Jensen-Shannon* multi-scale method is based on *log-surprise*. Similarly, from the comparison of *Jensen-Shannon* with the best non-saliency algorithm, which is *Energy* we can observe that it produces an increasement in the F-score of 158.1% for $EBR = -6dB$ and 69.7% for $EBR = 6dB$.

We observe that for every tested algorithm, the F-score increases directly with the EBR value, which seems logical since a low value of the EBR implies that background interferences are more harmful for the foreground event. Even for lower EBR configurations *Jensen-Shannon* produced the best results, followed by the rest of our fusion proposals.

As we want to analyze more in depth the behaviour of our proposal in the polyphony case, we also test the performance of our algorithms using the Task 3 dataset. As we explained previously in this Section, this task is divided into two subsets with totally different acoustic events and environments, *residential area* and *home*, which were not synthetically generated but captured in real environments. Consequently, the environmental background plays a critical role for this task, as it may affect negatively the deployment of our algorithms.

Results can be observed in Figure 7 where again all of our *log-surprise* based saliency algorithms are top-performers with the fusion mechanism of *Jensen-Shannon* producing the best results. It should be noticed that the performance for *residential-area* is much lower compared to the *home* subset. In fact, both results are far below the ones obtained for Task 2 in this section. This behaviour can be explained if we take into account the difficulties that real environments introduce: first of all, the acoustic characteristics of the environment itself that interfere severely with the acoustic events to be detected. Second, the fact that we are using unsupervised algorithms and the AED labeling scheme employs a holistic definition of *event* that sometimes contains several unlabeled *sub-events*, for example, we can find events with the tag *bird-singing* of almost 50 seconds long, where silences among the bird tweets would allow us to define *sub-events* that have not been tagged. Our algorithms detect each of these *sub-events* as separate units and therefore are considered as false alarms in the evaluation. This behavior can also be observed in other classes, such as *wind-blowing* and *people-speaking*. It could be argued that after some time, repeated *sub-events* of the same class would stop to be salient but the modelling of this is an issue we leave for further investigation.

Table 2: F-scores for DCASE2016 Task 2.

	EBR (dB)	Energy	VAD	Kalinli	Log-surp.	Cramer	Bhatta	Renyi-1	Renyi-Inf	Jensen-Shannon
Task 2: Monophonic	-6	0.2828	0.3039	0.1913	0.5057	0.6305	0.6824	0.6578	0.6910	0.7289
	0	0.4247	0.3248	0.2290	0.6211	0.7665	0.8301	0.7451	0.8206	0.8432
	6	0.5321	0.3526	0.2220	0.6270	0.8152	0.8754	0.8435	0.8720	0.9024
Task 2: Polyphonic	-6	0.4873	0.3280	0.2860	0.5184	0.5524	0.5842	0.5831	0.6397	0.6691
	0	0.5452	0.3463	0.3202	0.5629	0.6129	0.7064	0.6865	0.7245	0.7366
	6	0.5576	0.3684	0.3219	0.6179	0.6470	0.7691	0.7136	0.7664	0.7927
Average Std. Deviation		0.4716	0.3373	0.2617	0.5755	0.6708	0.7413	0.7049	0.7524	0.7788
		0.1302	0.0415	0.0797	0.0997	0.1191	0.1177	0.1127	0.1054	0.0935

6.3. Experiment 3: class performance

With this experiment we try to determine which classes or categories are more easily detected, and what properties describe them.

In this section, we aim at providing a more in-depth analysis of the *polyphonic* scenario of DCASE2016 Task 2 with the three values of the EBR provided. F-scores are depicted in Figure 8 as a heatmap for the available events. We can split the labelled events into two separate groups: natural vs. non-natural. The former define those events produced by humans (*speech*, *clearthroat*, *cough* and *laughter*) and the rest are considered non-natural.

As we expected, the overall behaviour of the *log-surprise* based schemes outperform the baseline algorithms for each of the events and the results improve with higher EBRs.

For natural events, it can be observed that the most difficult natural class to detect is *speech*, whose main difference against the other three natural classes is that it is not as impulsive as the rest like coughing or laughing. Specifically, it is worth highlighting the marked worsening of the results for the event *speech* as the EBR decreases, where we observe for *Jensen-Shannon* a reduction in the performance of 25.3% when comparing the $EBR = -6dB$ and $6dB$ conditions, in contrast with the second worst natural class (*clearthroat*) with a reduction of 13.55%. We can summarize the results of the natural category considering that in average, decreasing the EBR from $6dB$ to $-6dB$ imposes a reduction of 11.6% in the F-score for *Jensen-Shannon*, of 13.7% for *log-surprise* and of 11.8% for *Energy*. Consequently, results suggest that *Jensen-Shannon* is both the best performer and the most robust algorithm, since it suffers the smallest degradation for this particular category. *Log-surprise* performs better than *Energy* in absolute terms, but the latter suffers a smaller degradation when EBR decreases. Any case, F-scores achieved by *Energy* are considerably lower than those obtained by *log-surprise* and, especially, our multi-scale proposals.

The analysis of the non-natural events shows that all the tested systems fail to detect the onset of the class *drawer*. A listening inspection of that class audio clips reveals that the recorded audio consists of the sliding sound of a drawer, followed by its closing slam. In this situation, our system detects that the most salient event is clearly the final impact of the drawer, leading to an imprecise delayed output signal with respect to the annotation. On the other hand, from the seven classes that belong to the non-natural category there are four which depend strongly on the value of the EBR, namely *pageturn*, *drawer*, *phone*, and *doorslam*. For the particular case of *Jensen-Shannon* we

observe that the reduction of the EBR for the non-natural category decreases the F-score in approximately 22.6% from $EBR = 6dB$ to $-6dB$. For *log-surprise*, we observe a reduction of 37.4% and a 13.1% for *Energy*. Again, results suggest that our fusion proposal presents the most robust behavior, taking into account that *Jensen-Shannon* achieves the best results in absolute terms.

7. Conclusions

In this work, we have presented a new approach for bottom-up acoustic saliency, called *echoic log-surprise*, that is based on the combination of several acoustic saliency signals computed at various time spans. Each individual signal is obtained by using the *log-surprise* algorithm with different memory values. The fusion is performed by comparing the information carried out by these individual saliency signals by means of statistical divergences and distances. In particular, *Cramer* and *Bhattacharyya* distances and *Renyi-1*, *Renyi-Inf* and *Jensen-Shannon* divergences have been tried.

We have measured the performance of our multi-scale saliency system for the acoustic event detection task by employing three different datasets that cover a great variety of scenarios and noisy conditions. For comparison purposes, we have considered two basic techniques: an energy thresholding of the spectro-temporal representation and a voice activity detector, and two state-of-the-art acoustic saliency methods: Kalinli and log-surprise.

As results show, *log-surprise* and our multi-scale proposals achieve considerably better results than the classical algorithms *Energy thresholding*, *VAD* and *Kalinli*. Comparing *log-surprise* with the different *echoic log-surprise* versions, it can be observed that, in general, all the latter ones improves the F-scores obtained by the former in the databases considered. Regarding the different divergences proposed for the fusion stage, in general terms, the best is *Jensen-Shannon*, followed by *Renyi*. It is worth mentioning that both divergences are obtained from the entropy defined by Information Theory (Cover & Thomas, 2006), which justifies their similar performance. The fusion scheme based on the *Bhattacharyya* distance also produces good results, although worse than *Jensen-Shannon*. The *Cramer* distance is the worst multi-scale system, although it clearly outperforms the classic approaches.

As summary, we can conclude that the combination of several *log-surprise* saliency signals considering various scales of memory is useful for the AED task and we can theorize that it could be used as a part of a more complex

acoustic event classification system for determining the time instants where a relevant acoustic event occurs prior to the classification process itself.

From the interpretation point of view, we have intended to develop a biologically-inspired acoustic saliency method, emphasizing the temporal nature of audio signals and the use of several memory scales that try to resemble the ASM mechanism of the human aural perception (echoic memory) which determines how acoustic information is stored and remembered by our brain.

Further works should take into consideration three aspects: first of all, we consider that a more depth analysis of the robustness of the aforementioned techniques could be useful under a greater variety of scenarios and noisy conditions. Secondly, in this work we have used statistical distances and divergences to fuse the statistical information carried out by the *log-surprise* signals. However, we believe that more advanced techniques could be developed to combine this information. Finally, future works should try to integrate our saliency algorithms into a more complex acoustic event and sound scene classification system and combine them with advanced machine learning methodologies such as Deep Learning techniques.

Acknowledgment

This work is partially supported by the Spanish Government-MinECo project TEC2014-53390-P.

Agrawal, P., Vatsa, M., & Singh, R. (2014). Saliency based mass detection from screening mammograms. *Signal Processing*, 99, 29 – 47. doi:<http://dx.doi.org/10.1016/j.sigpro.2013.12.010>.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35, 99–109.

Bottcher-Gandor, C., & Ullsperger, P. (1992). Mismatch negativity in event-related potentials to auditory stimuli as a function of varying interstimulus interval. *Psychophysiology*, 29, 546–550.

Brookes, M. (1997). VOICEBOX: Speech processing toolbox for MATLAB. <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>.

- Chang, J.-H., Kim, N. S., & Mitra, S. K. (2006). Voice activity detection based on multiple statistical models. *IEEE Transactions on Signal Processing*, *54*, 1965–1976. doi:10.1109/TSP.2006.874403.
- Chen, Y., Pan, Y., Song, M., & Wang, M. (2015). Image retargeting with a 3d saliency model. *Signal Processing*, *112*, 53 – 63. doi:http://dx.doi.org/10.1016/j.sigpro.2014.11.001. Signal Processing and Learning Methods for 3D Semantic Analysis.
- Cooper, R. J., Todd, J., McGill, K., & Michie, P. T. (2006). Auditory sensory memory and the aging brain: A mismatch negativity study. *Neurobiology of Aging*, *27*, 752 – 762. doi:http://dx.doi.org/10.1016/j.neurobiolaging.2005.03.012.
- Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. Wiley.
- Cowan, N. (1984). On short and long auditory stores. *Psychological Bulletin*, *96*(2), 341–70.
- Endres, D. M., & Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, *49*, 1858–1860. doi:10.1109/TIT.2003.813506.
- van Erven, T., & Harremoës, P. (2012). Rényi divergence and kullback-leibler divergence. *CoRR*, abs/1206.2459.
- Fletcher, H., & Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *The Bell System Technical Journal*, *12*, 377–430. doi:10.1002/j.1538-7305.1933.tb00403.x.
- Glass, E., Sachse, S., & von Suchodoletz, W. (2008). Development of auditory sensory memory from 2 to 6 years: an mmn study. *Journal of Neural Transmission*, *115*, 1221–1229.
- Gomes, H., Sussman, E., Ritter, W., Kurtzberg, D., Cowan, N., & Vaughan, H. J. (1999). Electrophysiological evidence of developmental changes in the duration of auditory sensory memory. *Developmental Psychology*, *35*(1), 294–302.

- Grootel, M., Andringa, T., & Krijnders, J. (2009). DARES-G1: Database of Annotated Real-world Everyday Sounds. In *Proceedings of the NAG/DAGA Meeting 2009*. Rotterdam.
- Grossheinrich, N., Kademmann, S., Bruder, J., Bartling, J., & Von Suchodoletz, W. (2010). Auditory sensory memory and language abilities in former late talkers: A mismatch negativity study. *Psychophysiology*, *47*, 822–830.
- Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. (1st ed.). Upper Saddle River, NJ, USA: Prentice Hall PTR.
- Itti, L., & Baldi, P. (2005). A principled approach to detecting surprising events in video. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (pp. 631–637 vol. 1). volume 1. doi:10.1109/CVPR.2005.40.
- Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 1254–1259. doi:10.1109/34.730558.
- Kalinli, O., & Narayanan, S. (2009). Prominence detection using auditory attention cues and task-dependent high level information. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*, 1009–1024. doi:10.1109/TASL.2009.2014795.
- Kaya, E. M., & Elhilali, M. (2012). A temporal saliency map for modeling auditory attention. In *2012 46th Annual Conference on Information Sciences and Systems (CISS)* (pp. 1–6). doi:10.1109/CISS.2012.6310945.
- Kaya, E. M., & Elhilali, M. (2017). Modelling auditory attention. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*, 20160101. doi:10.1098/rstb.2016.0101.
- Kayser, C., Petkov, C. I., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology*, *15*, 1943–1947. doi:10.1016/j.cub.2005.09.040.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, *22*, 79–86. doi:10.1214/aoms/1177729694.

- Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). Deepgaze II: reading fixations from deep features trained on object recognition. *CoRR*, *abs/1610.01563*. URL: <http://arxiv.org/abs/1610.01563>.
- Macaluso, E. (2010). MT_TOOLS : Computation of saliency and feature-specific maps. <http://www.slneuroimaginglab.com/mt-tools>.
- Mesaros, A., Heittola, T., & Virtanen, T. (2016a). Metrics for polyphonic sound event detection. *Applied Sciences*, *6*, 162. doi:10.3390/app6060162.
- Mesaros, A., Heittola, T., Virtanen, T., Benetos, E., Foster, P., Lagrange, M., Lafay, G., & Plumbley, M. D. (2016b). IEEE AASP challenge: Detection and classification of acoustic scenes and events 2016. <http://www.cs.tut.fi/sgn/arg/dcase2016/>.
- Pan, J., Canton-Ferrer, C., McGuinness, K., O'Connor, N. E., Torres, J., Sayrol, E., & Giró i Nieto, X. (2017). Salgan: Visual saliency prediction with generative adversarial networks. *CoRR*, *abs/1701.01081*. URL: <http://arxiv.org/abs/1701.01081>.
- Pan, J., McGuinness, K., Sayrol, E., O'Connor, N. E., & Giró i Nieto, X. (2016). Shallow and deep convolutional networks for saliency prediction. *CoRR*, *abs/1603.00845*. URL: <http://arxiv.org/abs/1603.00845>.
- Ramirez, J., Segura, J., & Gorriz, J. (2007). *Voice Activity Detection. Fundamentals and Speech Recognition System Robustness*. INTECH Open Access Publisher.
- Rodríguez-Hidalgo, A., Gallardo-Antolín, A., & Peláez-Moreno, C. (2016). Towards aural saliency detection with logarithmic Bayesian Surprise under different spectro-temporal representations. *Proceedings of Iberspeech 2016*, (pp. 99–108).
- Sams, M., Hari, R., Rif, J., & Knuutila, J. (1993). The human auditory sensory memory trace persists about 10 sec: Neuromagnetic evidence. *Journal of Cognitive Neuroscience*, *5*, 363–370.
- Sangwan, A., Zhu, W. P., & Ahmad, M. O. (2007). Design and performance analysis of bayesian, neyman-pearson, and competitive neyman-pearson voice activity detectors. *IEEE Transactions on Signal Processing*, *55*, 4341–4353. doi:10.1109/TSP.2007.896118.

- Schauerte, B., & Stiefelhagen, R. (2013). "Wow!" Bayesian surprise for salient acoustic event detection. *IEEE International Conference on Acoustics, Speech and Signal Processing*, .
- Schröder, J., Goetze, S., & Anemüller, J. (2015). Spectro-temporal gabor filterbank features for acoustic event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*, 2198–2208. doi:10.1109/TASLP.2015.2467964.
- Schröder, J., Moritz, N., Anemüller, J., Goetze, S., & Kollmeier, B. (2017). Classifier architectures for acoustic scenes and events: Implications for dnnns, tdnnns, and perceptual features from dcase 2016. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*, 1304–1314. doi:10.1109/TASLP.2017.2690569.
- Schröger, E. (2007). Mismatch negativity: A microphone into auditory memory. *Journal of Psychophysiology*, *21*, 138–146.
- Sohn, J., Kim, N. S., & Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, *6*, 1–3. doi:10.1109/97.736233.
- Székel, G. J. (2002). E-statistics: The energy of statistical samples, .
- Temko, A., Malkin, R., Zieger, C., Macho, D., Nadeu, C., & Omologo, M. (2007). Clear evaluation of acoustic event detection and classification systems. *LNC3 4122*, (pp. 311–322).
- Tsuchida, T., & Cottrell, G. (2012). Auditory saliency using natural statistics. *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, (pp. 1048–1053).
- Virtanen, T., Mesaros, A., Heittola, T., Plumbley, M., Foster, P., Benetos, E., & Lagrange, M. (2016). *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*. Tampere University of Technology. Department of Signal Processing.
- Xu, Y., Huang, Q., Wang, W., Foster, P., Sigtia, S., Jackson, P. J. B., & Plumbley, M. D. (2017). Unsupervised feature learning based on

- deep models for environmental audio tagging. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25, 1230–1241. doi:10.1109/TASLP.2017.2690563.
- Yi, Y., & Lin, Y. (2013). Human action recognition with salient trajectories. *Signal Processing*, 93, 2932 – 2941. doi:http://dx.doi.org/10.1016/j.sigpro.2013.05.002.
- Yoo, I. C., Lim, H., & Yook, D. (2015). Formant-based robust voice activity detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23, 2238–2245. doi:10.1109/TASLP.2015.2476762.
- Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8, 32. doi:10.1167/8.7.32.
- Zhao, Q., & Koch, C. (2013). Learning saliency-based visual attention: A review. *Signal Processing*, 93, 1401 – 1407. doi:http://dx.doi.org/10.1016/j.sigpro.2012.06.014. Special issue on Machine Learning in Intelligent Image Processing.

Figure 1: Comparison of the spectrogram of an auditory signal against several surprise maps with different memory values. In this case, a sample corresponds to 20 ms. Logarithm and different magnitude ranges were applied to ease visualization.

Figure 2: Output saliency signals $s(t)$ computed both using surprise and log-surprise, illustrating the spiky nature of the Bayesian surprise signal.

Figure 3: Schematic of the proposed algorithm using a depth of four levels. The orange boxes highlight the contribution of this work: the multi-scale log-surprise scheme, where several memory-valued components are deployed, and a statistical fusion stage, which is in charge of the combination of the previously computed saliency cues.

Figure 4: Detailed schematic of the Fusion Stage, where three saliency signals are decomposed into several histograms and finally fused using *Jensen-Shannon Divergence (JSD)*, producing our multi-scale saliency proposal. In this example, we pay special attention to two particular acoustic segments, that we name $T1$ and $T2$ respectively.

Figure 5: Example of a filterbank whose 20 elements are triangular filters, spaced according to Mel-scale. Notice that the higher is the central frequency, the bigger is the separation between different filters and their critical bandwidth.

Figure 6: Results obtained for Experiment 1, where we compared the performance of several detection algorithms against our proposal, considering two different datasets.

Figure 7: Results obtained for DCASE2016 Task 3, where we considered the two subsets available independently.

Figure 8: F-scores obtained for every available class of DCASE2016 Task 2 for the three EBR values considered. The first four classes represent the so-called natural category, whereas the remainder ones belong to the non-natural category.