Rodríguez-Hidalgo, Antonio, Peláez-Moreno, Carmen, Gallardo-Antolín, Ascensión. (2018). The robustness of echoic log-surprise auditory saliency detection. *IEEE ACCESS*, v. 6, pp. 72083-72093.

DOI: https://doi.org/10.1109/ACCESS.2018.2882055

The robustness of Echoic log-surprise auditory saliency detection

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

***Abstract*—The concept of saliency describes how relevant a stimulus is for humans. This phenomenon has been studied under different perspectives and modalities, such as audio, visual or both. It has been employed in intelligent systems to interact with their environment in an attempt to emulate or even outperform human behavior in tasks such as surveillance and alarm systems or even robotics.**

In this work, we focus on the aural modality and our goal consists in measuring the robustness of Echoic log-surprise in comparison with a set of auditory saliency techniques when tested on noisy environments for the task of saliency detection. The acoustic saliency methods that we have analyzed include Kalinli's saliency model, Bayesian log-surprise, and our proposed algorithm, Echoic log-surprise. This last method combines an unsupervised approach based on the Bayesian log-surprise and the biological concept of echoic or Auditory Sensory Memory by means of a statistical fusion scheme where the use of different distance metrics or statistical divergences, such as Renyi's or Jensen-Shannon's among others, are considered. Additionally, for comparison purposes, we have also compared some classical onset detection techniques, such as those based on Voice Activity Detection (VAD) or Energy thresholding.

Results show that Echoic log-surprise outperforms the detection capabilities of the rest of the techniques analyzed in this work under a great variety of noises and signal-to-noise ratios, corroborating its robustness in noisy environments. In particular, our algorithm with the Jensen-Shannon fusion scheme produces the best F-scores. With the aim of better understanding the behavior of Echoic log-surprise, we have also studied the influence of its control parameters, depth and memory, and their influence at different noise levels.

***Index Terms*—acoustic saliency, echoic memory, multi-scale, statistical divergence, Jensen-Shannon, acoustic event detection**

I. Introduction

In this paper, we address the problem of auditory saliency detection, a task that requires an understanding of human perception and signal processing. Auditory saliency can be defined as a property of particular sounds to stand out perceptually. Several efforts have been made to model the aspects that make a signal salient or relevant, using experiments that combine high cognitive visual or acoustic loads with the detection of subtle changes in audio or studying how human response depends on the availability of attentional resources [1], [2], [3]. It is worth distinguishing this bottom-up phenomenon only related with the intrinsic characteristics of the input sound from that of *attention* where the saliency of a sound is influenced by the task the listener is performing, that is, a top-down phenomenon.

All the authors are with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain.

Email: arodigue@pa.uc3m.es, carmen@tsc.uc3m.es, gallardo@tsc.uc3m.es

In contrast with the field of visual saliency modeling where eye-tracking devices provide empirical ground-truth labels, it is difficult and costly to obtain labelled data for auditory saliency. For this reason, unsupervised methods are usually preferred for this task. Examples of these are the models of Kayser et al. [4], Kalinli et al. [5], Schauerte et al. [6] or [7]. The first two proposals are partially inspired on the visual saliency model proposed by Itti et al. [8] and adapted to the particular properties of audio signals. In these works, the input acoustic representation is the spectrogram, an image-like representation that includes both time and frequency information in a single bi-dimensional structure. These saliency algorithms extract from it characteristics related with temporal and frequency contrasts, among others, which are finally processed considering several scale resolutions. The resultant multi-scale scheme obtains several across-scale combinations by means of a center-surround operation that after a normalization stage and a summation operation produces the final saliency map.

On the other hand, Schauerte et al. [6] and [7] adopt an statistical approach where the Acoustic Bayesian Surprise proposed by the former was later refined by the latter basically by the inclusion of a logarithmic transformation, a perceptually motivated non-linearity. As previously mentioned models, Bayesian Surprise also uses a spectrogram as input representation. It processes each frequency band independently, and it models parametrically the statistical distribution of a particular time frame of the chosen frequency band. This distribution is compared with the previous one employing the Kullback-Leibler (KL) divergence. If there is no novelty or surprise in the acoustic signal, both distributions exhibit similar properties. However, if there is a sudden and meaningful change in the acoustic signal the distribution of the last temporal instant becomes distinct, and the KL divergence produces a high value representing this change in the analyzed frequency band.

In [9], we further elaborated on the aforementioned Bayesian Surprise and Log-surprise introducing the concept of echoic memory or Auditory Sensory Memory (ASM). This concept explains the amount of time that humans need to forget an acoustic signal they have recently perceived [10], [11], [12], [13], [14], [15], [16]. According to different authors, this temporal span goes from 10 to 20 seconds, and depends on certain parameters such as the age of each individual. In order to capture this temporal behavior, we proposed a multi-scale approach and introduced two concepts in our mathematical model: depth and memory. With the definition of depth, we quantify the number of log-surprise saliency signals we consider to make the final decision, and the memory represents how many time frames we use for the computation of those

saliency signals. All the resultant signals computed for these memory values were combined making use of distance metrics or statistical divergences, namely we compare Kullback-Leibler [17], Jensen-Shannon [18] and Renyi divergences [19] and Cramer [20] and Bhattacharyya [21] distances. We named this scheme Echoic Log-surprise [9].

On the other hand, one of the most challenging problems in signal processing is to ameliorate the effects of noise in real world scenarios. Robust algorithms and techniques have been devised to improve the performance of many applications in adverse conditions. For example, in Automatic Speech Recognition (ASR) [22], it is customary to test new developments under mismatched conditions, this is, when training and testing sets are collected under different environmental conditions. This is useful to measure the robustness of new algorithms to unexpected changes in the type and amount of noise. There is a plethora of noise-robust methods for ASR, which deal with robust feature extraction [23], [24], noise removal [25], [26] or even deep learning based techniques [22], [27], [28], [29].

In this paper, we analyse the robustness of the Echoic Log-surprise algorithms and we compare it with other auditory saliency detection mechanisms. To the best of our knowledge there are no previous contributions that consider the robustness of acoustic saliency algorithms in several noisy conditions, although it is clear that acoustic conditions in real-life scenarios are never optimal and the performance of any system might decrease dramatically. Here, we study in depth the robustness of our Echoic Log-surprise algorithm by adding six noise signals obtained from the DEMAND dataset [30], and also white Gaussian noise. In our experiments we consider SNR values ranging from $SNR = -5dB$ to $SNR = 20dB$, providing also the results of the noiseless configuration for the sake of comparison. Even if Echoic log-surprise apparently performed adequately in clean environmental conditions we expect to determine what are the limits of this technique, and to finally find out if it withstands the noise in comparison with the rest of the saliency techniques under analysis.

In the experimental part of this work, we use Acoustic Event Detection (AED) and Classification (AEC) datasets to assess our proposal. These are considered suitable proxies for the saliency detection problem, more difficult to grab and annotate. AEC annotations include onset time, offset time and an event class label from which we only employed the onset as our target, more closely related to the saliency phenomenon. We compare the robustness of our system against the algorithm proposed by *Kalinli* [5], *Bayesian Log-surprise* [7], a simple *Energy thresholding* system and a Voice Activity Detector (VAD) [31]. The datasets we have considered for this task are DCASE-2016 (Task 2) [32] and CLEAR06 UPC-TALP [33]. In comparison with our previous work [9] we have dropped DARES-G1 dataset [34], since all the algorithms that we tested on it seemed to perform poorly. After analyzing the available annotations we concluded that they were poorly annotated for the task that we are considering.

The remainder of this paper is organized according to the following scheme: we start outlining the contributions of our work, followed by Section III where we explain the theory behind *Echoic log-surprise* and statistical fusion. In Section IV we describe the experimental setup, including information about the noises used in our experiments as well as the datasets and metrics used. Finally, in Sections V and VI we explain the results gathered from our experiments and the conclusions that can be obtained from them together with some future lines of work.

II. Contributions

The main contribution of this paper is the analysis of robustness against background noise of the *Echoic log-surprise* saliency algorithm. We have compared this robustness with that of the main auditory saliency detection algorithms in the literature. In the comparison we have also included some classical detection mechanisms such as voice activity detection. Since we consider that real world scenarios comprise different noise sources, both stationary and non-stationary, we have used a noise dataset that includes acoustic data from both indoor and outdoor environments [30], in addition to the classical white Gaussian noise. Finally, similarly to our previous work [9] we have tested the performance of all the systems using two different and non-related AED/C datasets, which proved to be a useful approach in order to avoid overfitting the specifications of our saliency systems to a particular dataset.

III. Methods

Figure 1 shows the block diagram of the multi-scale saliency system used in this work, whose technical and simulation details were thoroughly explained in [9]. It is divided in three different stages: feature extraction, multi-scale saliency determination and statistical fusion.

A. Feature extraction

An adequate representation of the input signal is key for obtaining a good saliency detection. A very common approach employs the *spectrogram*, which conveys information from both the temporal and frequency domains. In fact, we can consider the spectrogram as an image-like representation of the acoustic information and then use techniques imported from visual saliency detection, an inspiration that some authors [4], [5], [6] have used in their acoustic saliency proposals.

A spectrogram shows the evolution of the signal along time for different uniformly distributed frequency bands by means of the Fourier Transform. However, for our systems we take into consideration other aspects of Human Auditory System (HAS). It is well established [35], [36] that human hearing does not consider all the frequency bands to be equally important. In fact, humans are more sensitive to the information that is concentrated in the lower frequency bands of the spectra. Speech is, non surprisingly, located in the frequency bands that the HAS privileges. Consequently, some works have developed alternative ways to model the spectrogram, taking into consideration this perceptual behavior. Those spectro-temporal representations are usually termed *cochleograms* and, in this case, are implemented by applying a critical-band analysis over the audio signal, considering a filter-bank whose frequency ranges are based on the behavior of the HAS. A

very well-known choice for this filter-bank is the Mel-scaled filter-bank that is in the core of the most popular feature extraction procedure for speech and audio-related tasks [35]. In particular, the spacing of these filters in the frequency domain is determined by the Mel-scale [35], [36] according to the following equation:

$$B(f) = 1125 \cdot \ln(1 + f/700) \tag{1}$$

where $B(f)$ represents the Mel-scale transformation of the frequency f measured in Hz, that it is approximately linear for frequencies below 1 KHz and logarithmic for frequencies above 1 KHz. A common simplification uses triangular shapes with their bandwidths increasing as the central frequency grows higher mimicking the critical bands [36].

In summary, in order to obtain the cochleogram, the spectrogram of the raw signal $x(t)$ is computed by using overlapped Hamming windows, and passed through a triangular mel-scaled filter-bank, as the one depicted in the schematic in Figure 1. Then, for each one of the bands, the energy is calculated, yielding the cochleogram $X(k, n)$, where k and n represent, respectively, the k_{th} sub-band of the Mel-scale filter bank and the frame index. This stage is common for all the saliency and detection techniques implemented in this paper.

B. Multi-scale saliency computation

Our algorithm is based on the concept of *log-surprise* where the logarithm of the Kullback-Leibler (KL) divergence is used to determine the level of dissimilarity between the audio cochleogram $X(k, n)$ at two different temporal instants (or frames) n and $n-1$. High values of *log-surprise* indicate that there is a change in the acoustic signal and, therefore, the occurrence of an acoustically salient event.

The *log-surprise* for each Mel-frequency band k is computed according to the following equation:

$$d_{log-surp}(k, n) = \ln\Big(D_{KL}(\mathcal{X}_{k,n} \| \mathcal{X}_{k,n-1})\Big) = \ln\left(\frac{(\mu_{k,n} - \mu_{k,n-1})^2}{2\sigma^2_{k,n-1}} + \frac{1}{2}\Big(\frac{\sigma^2_{k,n}}{\sigma^2_{k,n-1}} - 1 - \ln\frac{\sigma^2_{k,n}}{\sigma^2_{k,n-1}}\Big)\right), \tag{2}$$

where $\mathcal{X}_{k,n}$ is the probability density function of the cochleogram $X(k, n)$ estimated for the band k at fram n. Assuming that $\mathcal{X}_{k,n}$ is normally distributed $\mathcal{X}_{k,n} \sim \mathcal{N}(\mu_{k,n}, \sigma^2_{k,n})$, we need to compute the values of $\mu_{k,n}$ and $\sigma^2_{k,n}$ for each band k using a buffer with N frames. This represents the memory of the *log-surprise* and is a crucial parameter of our algorithm. Since $d_{log-surp}(k, n)$ is calculated independently for each band, the global *log-surprise* signal $s(n)$ is obtained as follows:

$$s(n) = \frac{1}{N_{mel}} \sum_{k=1}^{N_{mel}} d_{log-surp}(k, n). \tag{3}$$

where N_{mel} is the number of frequency bands of the Mel-scaled filter bank.

The multi-scale stage of our *Echoic log-surprise* algorithm in an extension of this method where several *log-surprise* saliency signals with different memory values are computed providing information from the same audio signal at different temporal resolutions.

There are two control parameters, namely the *depth* of the system dth and the initial memory N_1. The value of dth determines the number of levels, i.e. the number of saliency signals that are going to be computed $s_i(n), i \in \{1, 2, ..., dth\}$. The parameter N_1 indicates the memory used for the calculation of the first level saliency $s_1(n)$. The remainder saliency signals $s_i(n), i \in \{2, ..., dth\}$ are obtained by using increasing values of memory N_i verifying that $N_i = N_1 \cdot N_{i-1}, i \in \{2, ..., dth\}$ [1] Consequently, setting up the system with a depth $dth = 3$ and an initial memory $N_1 = 2$ implies that we obtain three saliency signals $s_1(n)$, $s_2(n)$ and $s_3(n)$ with memory values $N_1 = 2$, $N_2 = 4$ and $N_3 = 8$ frames, which correspond to buffer sizes of $40ms$, $80ms$ and $160ms$ respectively, when using an analysis window size of $20ms$. In this work, we consider models up to a depth of $dth = 10$ and a fixed initial memory $N_1 = 2$ frames, which provides a maximum memory value of $N_{10} = 1024$ frames corresponding to a buffer size of 20.48 seconds. This amount is closer to the ASM temporal values proposed in [10], [11], that lay between 10 and 20 seconds.

C. Statistical fusion

Finally, all the saliency signals computed for the different scales are combined in a fusion stage, by comparing the information they carry by means of statistical divergences and distances. For doing this, for each of the saliency signals $s_i(n), i \in \{1, 2, ..., dth\}$ and each frame n, a running histogram $h_i(n), i \in \{1, 2, ..., dth\}$ is obtained considering the previous consecutive M frames. Then, these dth histograms are combined according to the chosen fusion scheme, producing the final *echoic log-surprise* saliency signal $s_{echoic}(n)$. Our fusion mechanism is inspired in the concept of echoic memory, which states that an unexpected sound is usually remembered from 10 to 20 seconds. Our fusion mechanism combines data with memory values covering the previous timespan, which means that we keep acoustic information from several temporal values at the same time, and use it to compute a saliency signal.

For a generic statistical divergence or distance d_{fusion}, we can define the *echoic log-surprise* as follows:

$$s_{echoic}(n) = \begin{cases} d_{JSD}\{h_1(n), \cdots, h_{dth}(n)\}, & \text{for } JSD \\ \sum_{i=1}^{dth-1} d_{fusion}\{h_i(n), h_{i+1}(n)\}, & \text{otherwise} \end{cases} \tag{4}$$

[1] In a preliminary experimentation, other rules for setting the memory values for the different levels were tried, as for example, a linear relationship such as $N_i = A(i-1) + N_1, i \in \{2, ..., dth\}$, with different values of the slope A an $N_1 = 2$. Nevertheless, best results were obtained with the exponential rule used in this paper, suggesting that our method benefits from the use of a large range of memory values in the computation of the different saliency signals $s_i(n)$.

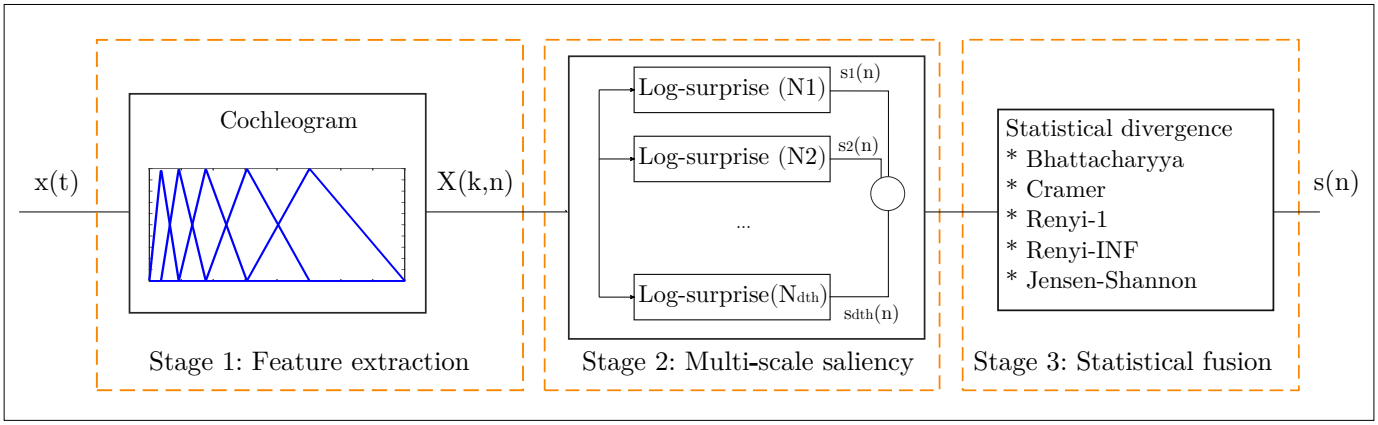


Fig. 1: General schematic of multi-scale saliency system.

	Num. of files	Num. of events	Num. of classes	$f_s(Hz)$
D16_T2	72	2000 (aprox)	11	44100
UPC_TALP	30	1030 (aprox)	14	44100

TABLE I: Technical parameters of the datasets.

In this paper, we have employed several fusion schemes d_{fusion} based on the distances of Cramer [20] and Bhattacharyya [21], Renyi-1 and Renyi-INF entropies [19] and Jensen-Shannon divergence (JSD) [18], which are notated as d_{Cramer}, d_{Bhatta}, $d_{Renyi-1}$, $d_{Renyi-INF}$ and d_{JSD} respectively. More details about how they are used for fusion in this context can be found at the aforementioned work [9].

IV. Experimental setup

A. Datasets

Similarly to [9], we have worked with two different datasets that were developed for AED/AEC tasks. A summarized version of the technical specifications of each database is illustrated in Table I.

1) D16_T2: This particular dataset was developed for an AED/AEC challenge celebrated during 2016, named "Detection and Classification of Acoustic Scenes and Events 2016" (DCASE2016) [32], [37], from which we have chosen the Task 2. This task consists on the detection and classification of certain acoustic synthetic events. Data was recorded using a sampling frequency of $f_s = 44100Hz$, with a resolution of 24 bits. There are 72 audio clips with different Event-to-Background Ratio (EBR), divided into validation and test subsets. We use the validation subset, formed by 18 files, for the configuration of certain global parameters of our system. On the other hand, the test subset, composed of 54 audio files, is used to obtain the final results.

2) UPC-TALP database of isolated meeting-room acoustic events: This database was produced for CLEAR06 [33], a workshop focused on AED, among other tasks. This database was recorded with an array of microphones. We chose to use the one defined as number three for our experiments since this was the one employed for labeling the acoustic events. There are 30 audio clips, with a resolution of 16 bits and $f_s = 44100Hz$.

B. Noise

For the contamination of the previously mentioned audio datasets, we have used the DEMAND collection of noises [30] that comprises different real-world noise files acquired using an array of microphones at $f_s = 48000Hz$, from which we have chosen the channel two of the array. The noise collection is divided into six categories, four of them captured indoor and the two remaining in open air scenarios. From of total of 18 noise files, we have selected six different ones for our analysis, one per category:

- DKITCHEN: belongs to the 'Domestic' category, and contains audio recorded in a kitchen during the preparation of a meal.
- NFIELD: was captured from a sport field where there were a number of people. It belongs to the 'Nature' category.
- OHALLWAY: contains the sounds of groups of people passing, which were captured on a hallway. It belongs to the 'Office' category.
- PCAFETER: category is 'Public'. As its name describes, it was captured on a cafeteria inside of an office.
- SCAFE: was also acquired on a cafeteria, but placed on a public square instead. It is labeled into the 'Street' category.
- TBUS: contains sounds captured inside a public bus. Its category is 'Transportation'.

There are mainly two reasons to choose this dataset. Firstly, the sounds that it contains were captured considering a wide variety of real-life scenarios, allowing to test the behavior of all the analyzed systems in a diversity of acoustic environments. Secondly, other noise datasets developed for speech-related tasks such as Noisex-92 [38] and Chime-4 [39] have a sampling frequency of $f_s = 16kHz$, or $f_s = 8kHz$ in the case of Aurora-2 [40]. However, since we are working with a higher maximum frequencies we considered that DEMAND is more appropriate, being its sampling frequency $f_s = 48kHz$.

In addition, for the sake of comparison with other robustness studies, we have also used white Gaussian noise in our tests. We named this modality *WHITE*.

In summary, we have seven different noise configurations, which were added to the audio signals using Voicebox Toolbox [41] considering SNR values from $-5dB$ to $20dB$ in $5dB$ steps. The noise addition algorithm computed the signal level using the P.56 [42] ITU-T recommendation. Finally, we have also obtained the results for the noiseless condition.

C. Parameter setting and evaluation

All the audio clips were downsampled to $f_s = 22000Hz$. Cochleograms were computed performing first a Fast Fourier Transform (FFT) with 1024 frequency bins and subsequently transformed into the Mel-scale using a triangular filter bank with 150 filters. For the FFT we used a Hamming window of $20ms$ and an overlapping of 50%.

The fusion algorithm uses an initial memory of $N_1 = 2$, and a maximum depth of 10. Regarding the histograms computed to fuse all the saliency signals when $dth > 1$, we consider a temporal length of $M = 50$ frames per signal and 20 bins per histogram.

For the evaluation, we use the event-based metric proposed for the DCASE2016 challenge [43], where the F-score is computed for the onset of each audio event as follows:

$$F = 2\frac{P \cdot R}{P + R}, \tag{5}$$

where R represents the recall while the precision is represented as P. We considered a tolerance of $\pm 200ms$ (as in the DCASE2016 challenge) and a minimum duration of $60ms$ for each acoustic event. Events lasting less that this value were removed. Since this work focuses in the detection problem and the classification task is out of its scope, we did not evaluate the systems in terms of classification accuracy.

V. Results and Discussion

In this Section we report the results achieved by the different variants of the *echoic log-surprise* saliency detection algorithm and compare them with the ones obtained by the following baseline techniques: a simple method consisting on an *energy thresholding* of the audio signal, a *VAD* based on the work of Sohn et al. [31] using the implementation available in Voicebox toolbox [41], the *Kalinli* saliency algorithm [5] inspired on the works of Itti [8] and Kayser [4] and downloaded from [44], and the *log-surprise* based saliency detector [6], [7] using the code provided by the authors. We included Table II as a reminder of the techniques and the abbreviations used in this work.

Notice that the figures in most of the cases depict the results as a function of the SNR, after averaging the results of the two datasets and all their noisy versions. The exception is shown in Section V-F, where we present a more detailed analysis of the performance of the considered techniques under each one of the noisy scenarios.

Technique name	Shortname	Ref
Energy thresholding	Energy	[35]
Kalinli saliency	Kalinli	[5], [44]
Log-surprise saliency	Log-surprise	[6], [7]
Voice Activity Detector	VAD Sohn	[31]
Echoic Log-surprise considering different fusion techniques	Bhattacharyya Cramer Jensen-Shannon, JSD Renyi-1 Renyi-INF	[9]

TABLE II: Summary of the techniques used in this work.

A. Pre-analysis 1: choosing the memory value 'N' for 'Log-surprise'

In this analysis we have considered a noiseless scenario, since our goal is to provide guidelines for the adequate choice of the memory value N of *log-surprise*. Our initial hypothesis, based on a visual examination of the *log-surprise* saliency signal, was that a memory of $N = 50$ should produce a reasonable detection performance. A bigger value smoothed output signals excessively, and a smaller one was not able to cope with false positives, which would affect and reduce drastically the F-score of the system, confirming our preliminary hypothesis.

B. Pre-analysis 2: the robustness of baseline saliency detection algorithms

To assess the robustness of our *echoic log-surprise* saliency detection algorithm we need to establish how baseline algorithms perform in the same noisy conditions.

Figure 2 depicts the F-scores obtained for these algorithms as a function of the SNR. Our first observation is that, as we expected, noise affects the detection performance of all the methods. However, we observe that *Kalinli* presents the worst performance although it is almost invariant to SNR, since its F-score for $SNR = -5dB$ only changes slightly in comparison with the noiseless configuration. Finally, the results achieved by *log-surprise* suggest that this algorithm is the most robust of the classical techniques considered in this work, since it produces the higher F-scores for every proposed SNR value.

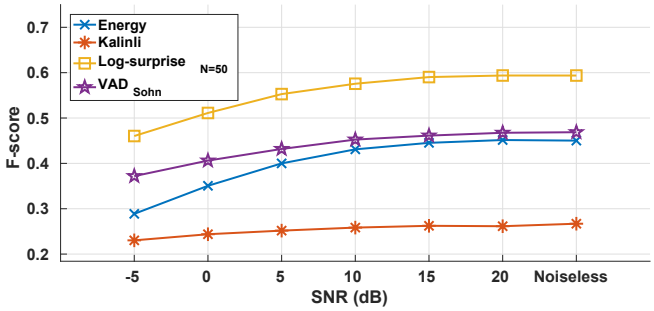


Fig. 2: F-score against SNR for the classical detection techniques analyzed in this work.

C. Analysis 1: review of the depth under several SNR scenarios

During this analysis of the performance of the proposed *echoic log-surprise* algorithm, we aim at demonstrating that an adequate choice of the depth parameter, dth, produces advantageous results when compared against the classical detection techniques that we mentioned in this work. We average the results obtained for the two datasets and all the noisy conditions which allows us to analyse the global performance of our fusion proposals. We study the influence of the SNR and different values of dth. The obtained graphs for each fusion scheme are depicted in the Figures 3a to 3e for Bhattacharyya, Cramer, Jensen-Shannon, Renyi-1 and Renyi-INF, respectively.

If we focus on the shape of the aforementioned figures, we observe that the deeper our system works, the higher the F-score becomes, a behavior that is repeated for all the proposed SNR values. In fact, when we consider a value $dth > 5$ the F-score of the system increases at a slower pace, which means that from that depth the system keeps improving but the obtained values are quite similar until it reaches what we consider to be the general optimal working point, at $dth = 8$. In the Figure 3d we can observe that using a superior value of dth deteriorates the performance of the system for the particular case of Renyi-1, a behaviour that is not shared by any of the other fusion proposals. The best F-score is obtained using Jensen-Shannon, approximately 0.62 for the noiseless condition and $dth = 10$ but the other fusion algorithms provide close results at their optimal dth values.

D. Analysis 2: the robustness of the echoic log-surprise as compared to classical saliency detection techniques

The goal of this analysis is the comparison of the classical approaches studied in subsection V-B against the multi-scale proposal that we introduced.

Figure 4 depicts two individual graphs that show the results for all the saliency detection techniques we consider at different values of depth dth, a parameter that, despite not being critical for the classical techniques, it is an essential part of our proposed fusion models. The first depth value is $dth = 1$, which, for the fusion schemes, represents the basic configuration where a single *log-surprise* signal is computed with initial memory $N_1 = 2$. The second depth value that we considered is $dth = 8$ since according to the analysis performed in Section V-C is the optimal point of work for all the fusion techniques. The last value of dth in this analysis is $dth = 10$, the limit that we set for our saliency systems. Thus, our current analysis comprises the maximum and minimum values of dth, as well as the optimal point of work.

In the graphs to the left of Figure 4 we find several bar diagrams that depict F-score results for each saliency technique considering all the SNR values. To the right, there are three diagrams with red bars that represent the range of performance improvement of each system from $SNR = -5dB$ to the noiseless condition. Hence, the lowest value of each bar represents the value obtained for $SNR = -5dB$, and the top shows the noiseless one. The right hand side graphs of Figure 4 eases the analysis providing means to quickly and visually compare the results at both extremes of the SNR range.

From both types of diagrams, it can be observed that, as expected from the results obtained in subsection V-A, the fusion systems perform poorly and significantly worse than the classical techniques at $dth = 1$. In that case, the dominant technique is *log-surprise* with a memory value of $N = 50$ followed by *VAD*, *Energy* and *Kalinli*. These four techniques do not depend on depth.

For $dth = 8$ both graphs suggest that the classical saliency techniques produce worse results than our fusion proposals for every considered SNR. As a matter of fact, when we set $SNR = -5dB$ almost all of the fusion systems produce similar F-scores, that increase at a similar pace with the value of the SNR. The exceptional case is JSD that performs slightly better than the rest of the fusion techniques for the same noise configurations.

Finally, when we set $dth = 10$ we observe that, as we mentioned in Section V-C, Renyi-1 produces the worst results among the fusion techniques and its increase of performance with the improvement of the SNR is more moderate. It should be mentioned that its value is also similar to the one obtained for $dth = 8$.

After this analysis, we consider that an adequate point of operation for all the fusion techniques could be $dth = 8$, since it would work equally well for all the proposed techniques. Nevertheless, the recommended technique would be Jensen-Shannon no matter what dth value we use, since it is not as sensitive as the rest of the techniques to this parameter when it is big enough.

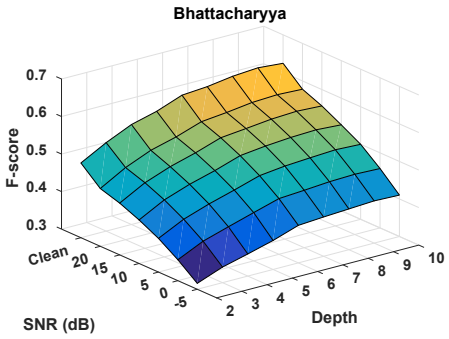
E. Analysis 3: Precision and Recall scores

To obtain more insight about the behaviour of the different algorithms we have depicted the Precision-Recall (P-R) graph of Figure 5. For each detection technique, it shows three ellipses corresponding to $SNR = -5dB, 10dB$ and clean conditions, whose horizontal and vertical axes are proportional to the standard deviation value of *precision* and *recall* obtained for the files of *D16_T2* validation subset respectively and their centers are situated in the average P-R position. The black arrows indicate the directions in which these average P-Rs move as SNR increases.

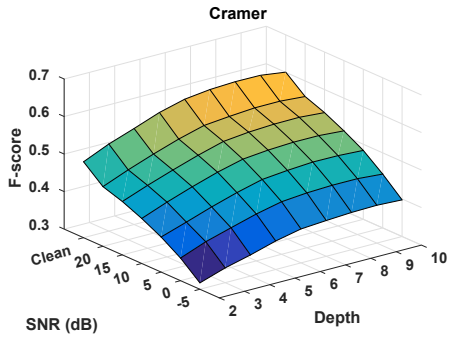
First of all, we observe that *energy*, *VAD* and *Kalinli* tend to have high *recall* and small *precision*, no matter what SNR we consider. A low *precision* score implies that these three systems produce large number of false positive values. A high *recall* score indicates, however, that some of these detected onsets actually are well positioned and produce true positives.

Log-surprise is different from the three previous ones, since it produces higher *precision* than *recall*. This means that the onsets detected by this algorithm are most of the times properly placed in time, but some of the ground-truth onsets are missed.

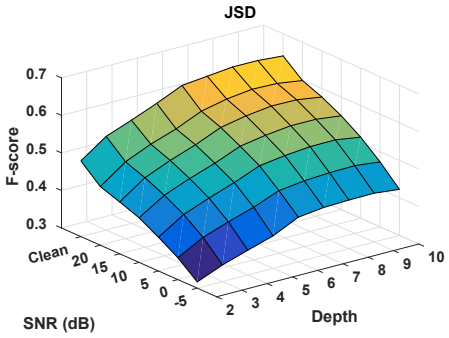
When analyzing the variation along the SNR we observe that *Kalinli* is the technique whose ellipses appear closer to the origin of the P-R graph, and therefore, it performs worse than the other techniques for the reasons that we have already



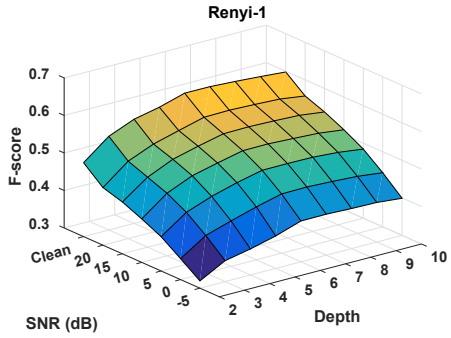
(a) F-score for Bhattacharyya fusion.



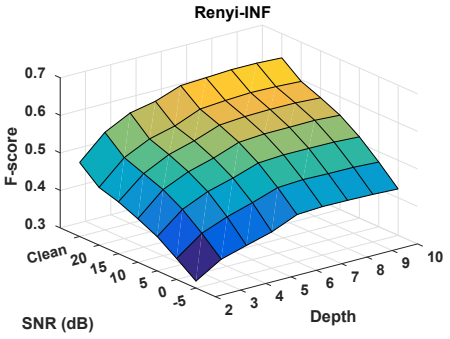
(b) F-score for Cramer fusion.



(c) F-score for Jensen-Shannon fusion.



(d) F-score for Renyi-1 fusion.



(e) F-score for Renyi-INF fusion.

Fig. 3: Variation of the F-score for our fusion saliency proposals under different values of SNR and dth.

explained. *Energy* and *VAD* are placed in similar positions of the graph, although the eccentricity of their corresponding ellipses is quite different, indicating a bigger dispersion among the *recall* scores than in the *precision* axis for *VAD*. *Energy* increases its *precision* with the SNR, whilst *VAD* increases its *recall*. In both cases, these increments occur mainly towards one of the axis, which would explain why both of them keep producing similar F-scores as observed in Figure 2. *Log-surprise* mainly increases its *recall* with the SNR, which in conjunction with its high *precision* would explain why it outperforms the rest of the techniques of this analysis.

The superiority the proposed *Echoic log-surprise* method in comparison with the non multi-scale algorithms can be observed for *Jensen-Shannon* with $dth = 10$ in Figure 5 (green ellipses). Interestingly, its starting point (at $SNR = -5dB$) is situated in the same P-R position than *log-surprise* for the noiseless condition (yellow ellipses) with slightly better *precision* than *recall*. The bigger the SNR the bigger both the *precision* and the *recall* obtained for *Jensen-Shannon* and consequently the F-score ending on the equal *precision* and *recall* dotted line.

Finally, as a general comment referred to the eccentricity of the ellipses, we observe that most of the times there is a bigger dispersion in the *recall* dimension. This means that, for a certain detection technique, a similar amount of false positive values along the validation files is found, but there is a big variation in the number of false negatives.

F. Analysis 4: comparison of the saliency techniques for every independent noise configuration

Finally, we analyse, the behaviour of all the detection saliency systems considering all the noisy signals independently, and averaging the results for both audio datasets. The results are illustrated in Figure 6. These graphs show a comparison of all the classical techniques against Jensen-Shannon with $dth = 8$ where each subfigure represents a different type of noise, namely *DKITCHEN*, *NFIELD*, *OHALLWAY*, *PCAFETER*, *SCAFE*, *TBUS* and *WHITE*.

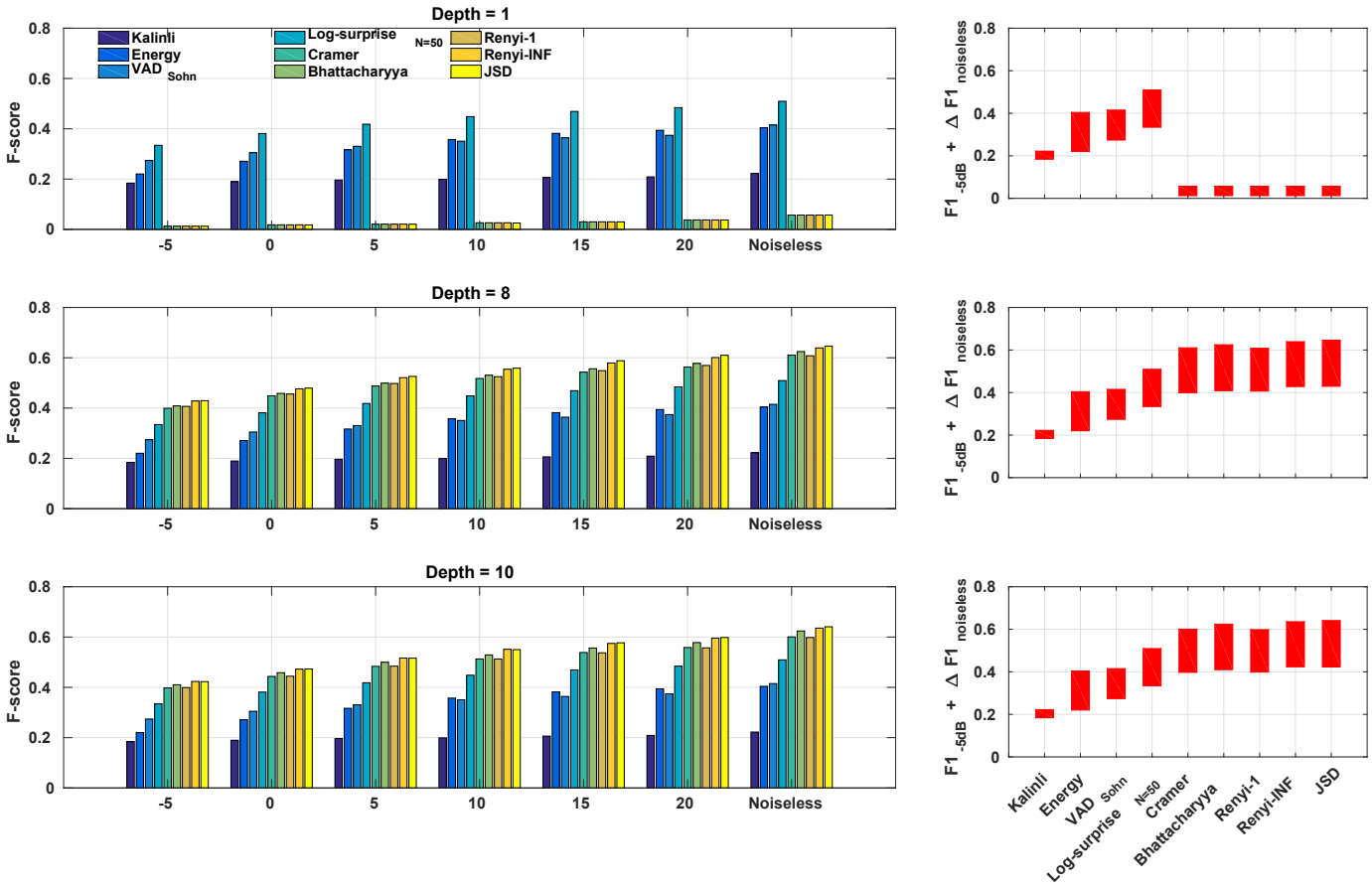


Fig. 4: Results in terms of F-score for all the fusion saliency techniques at different values of the depth parameter. Results arison purposes.

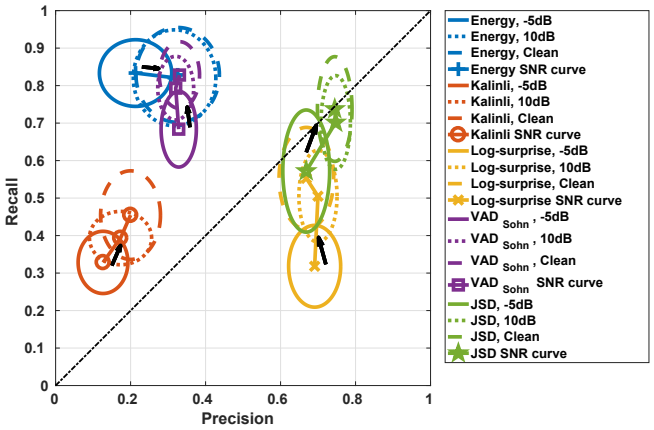


Fig. 5: Precision-Recall (P-R) graphic for the proposed classical detection techniques, considering different SNR configurations. The axes of each ellipse are proportional to the standard deviation value of Precision and Recall of *DCASE_T2* validation subset and its centers represents the corresponding average P-R pair. The black arrows indicate the directions in which the average P-Rs move as SNR increases.

If we analyze the results obtained for the classical techniques, we observe that *Kalinli* performs similarly for each noise signal. We observe certain variations for *VAD* that produces similar results for every SNR value when the noise configurations are DKITCHEN, NFIELD or OHALLWAY. For the rest of them, it is clearly affected when $SNR = -5dB$, and the higher the SNR the better it performs. *Energy* presents a similar trend to *VAD* although its performance in very noisy conditions ($SNR \leq 0dB$) degrades with respect to *VAD* in most of the noises. Two special cases are NFIELD and TBUS which, with the exception of *energy*, strongly affect the performance of all the systems, showing curves that remain low and almost flat. *Log-surprise* and *Jensen-Shannon* show a similar behaviour for all the considered noises and SNR values, although none of the classical techniques outperforms *Jensen-Shannon* for any of the noises.

VI. Conclusions

In this work we have presented the robustness analysis of the *echoic log-surprise* saliency technique in comparison with state of the art methods using an AED task. After performing several tests using two different datasets and seven SNR conditions, we have observed that an adequate depth parameter dth clearly helps improving the detection performance of our system. A first analysis showed that increasing the depth up to $dth = 8$ was advantageous. However, for $dth = 10$ the performance of *Renyi-1* began to deteriorate while that of the rest of the multi-scale algorithms remained almost

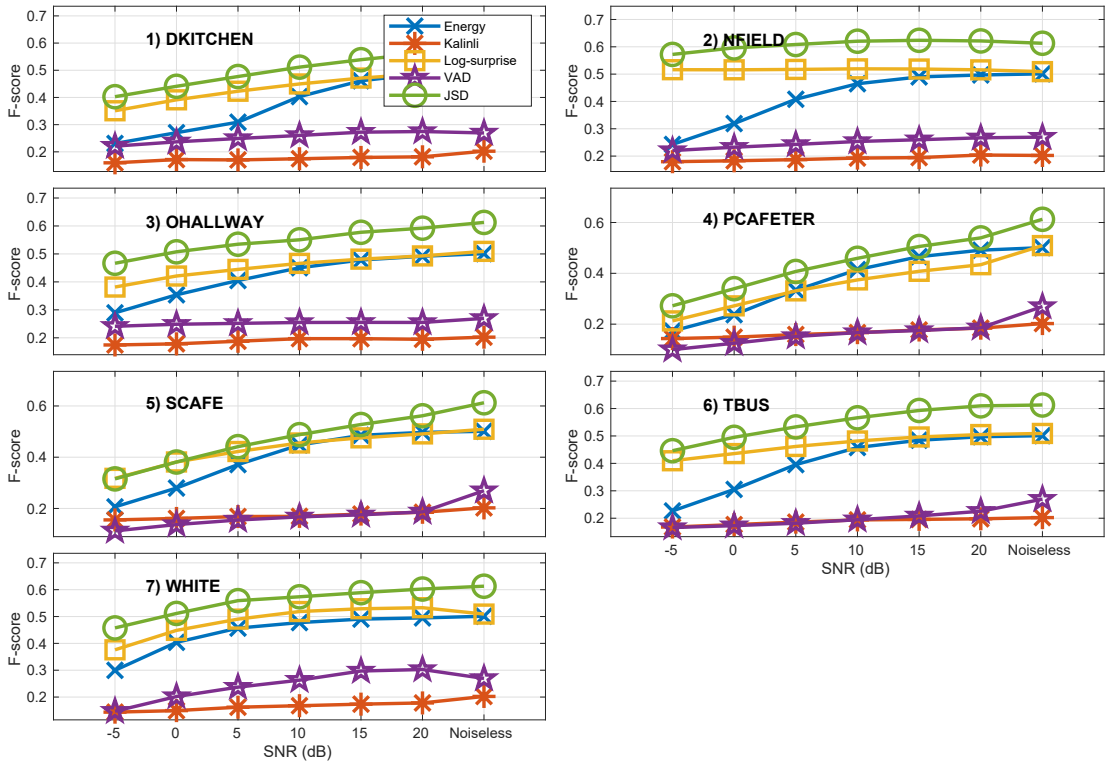


Fig. 6: Results in terms of F-score as a function of SNR for all the classical saliency detection techniques considering several noise signals, which are compared against Jensen-Shannon *Echoic log-surprise*.

unchanged. We set the initial memory value $N_1 = 2$ since, for the maximum dth we considered, amounts as using a buffer memory of 20.48 seconds, a reasonable size if we establish a correspondence with ASM values [10], [11]. However, we leave for future investigations, the empirical determination of this value.

Our second analysis allowed us to compare the performance of *echoic log-surprise* with some classical saliency detection techniques, considering averaged results for all the datasets and noise signals. We discovered that for small values of dth. the classical algorithms outperformed *echoic log-surprise*. In particular, for $dth = 1$, the best F-score was obtained for *log-surprise*, which performed clearly better than *VAD* and *Kalinli* for all the SNR configurations. This result is particularly interesting, since *log-surprise* has a memory value of $N = 50$ and it is equivalent to constrain the *echoic log-suprise* multi-scale algorithm to a single scale with $N_1 = 50$. Note, however, that the single-scale configuration for *echoic log-surprise* that we show in our analyses corresponds to $dth = 1$ and $N = 2$, a memory value that is clearly insufficient to model the salient nature of the acoustic events and clearly worse than the results we show for *log-surprise*. Increasing the value dth, however, makes the multi-scale approaches considerably better than the classical ones. For example, Jensen-Shannon with $dth = 10$ outperforms *log-surprise* for almost a 17%. The results for the second analysis also showed that Jensen-Shannon produced the best F-scores through all the ranges of SNR and noiseless configurations.

Finally, we made a detailed analysis of the performances for each of the noises independently. Results showed that though some types of noise are more detrimental than others, *echoic log-surprise* produced the best results for all the considered SNR values showing a high degree of robustness as well.

Future work will focus on the integration of new statistical fusion techniques, which might lead to even better results. Also a more in depth study of different alternatives for the determination of the optimal memory values for each level may lead to improvements in the detection mechanism. Using our algorithms to solve other tasks such as on-set detection for Music Information Retrieval (MIR) or to aid the training of attentive mechanisms in Long-Short Term Memory architectures that even combine audio and visual cues as in [45] have been also identified as promising research directions.

ACKNOWLEDGMENT

This work is partially supported by the Spanish Government-MinECo projects TEC2014-53390-P and TEC2017-84395-P.

REFERENCES

[1] R. Southwell, A. Baumann, C. Gal, N. Barascud, K. Friston, and M. Chait, "Is predictability salient? A study of attentional capture by auditory patterns," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372(1714), p. 20160105, 2017. [Online]. Available: http://dx.doi.org/10.1098/rstb.2016.0105

[2] T. Petsas, J. Harrison, M. Kashino, S. Furukawa, and M. Chait, "The effect of distraction on change detection in crowded acoustic scenes," *Hearing Research*, vol. 341, pp. 179 – 189, 2016.

[3] J. H. McDermott, D. Wrobleski, and A. J. Oxenham, "Recovering sound sources from embedded repetition," *Proceedings of the National Academy of Sciences*, vol. 108, no. 3, pp. 1188–1193, 2011. [Online]. Available: http://www.pnas.org/content/108/3/1188.abstract

[4] 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.

[5] O. Kalinli and S. Narayanan, "Prominence detection using auditory attention cues and task-dependent high level information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 1009–1024, July 2009.

[6] B. Schauerte and R. Stiefelhagen, ""Wow!" Bayesian surprise for salient acoustic event detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[7] A. Rodríguez-Hidalgo, A. Gallardo-Antolín, and C. Peláez-Moreno, "Towards aural saliency detection with logarithmic Bayesian Surprise under different spectro-temporal representations," *Proceedings of Iberspeech 2016*, pp. 99–108, 2016.

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, Nov 1998.

[9] A. Rodríguez-Hidalgo, C. Peláez-Moreno, and A. Gallardo-Antolín, "Echoic log-surprise: A multi-scale scheme for acoustic saliency detection," *Expert Systems with Applications*, vol. 114, pp. 255 – 266, 2018.

[10] N. Cowan, "On short and long auditory stores," *Psychological Bulletin*, vol. 96(2), pp. 341–70, 1984.

[11] E. Schröger, "Mismatch negativity: A microphone into auditory memory," *Journal of Psychophysiology*, vol. 21, pp. 138–146, 2007.

[12] E. Glass, S. Sachse, and W. von Suchodoletz, "Development of auditory sensory memory from 2 to 6Â years: an mmn study," *Journal of Neural Transmission*, vol. 115, no. 8, pp. 1221–1229, 2008.

[13] H. Gomes, E. Sussman, W. Ritter, D. Kurtzberg, N. Cowan, and H. J. Vaughan, "Electrophysiological evidence of developmental changes in the duration of auditory sensory memory," *Developmental Psychology*, vol. 35(1), pp. 294–302, 1999.

[14] C. Bottcher-Gandor and P. Ullsperger, "Mismatch negativity in event-related potentials to auditory stimuli as a function of varying interstimulus interval," *Psychophysiology*, vol. 29, no. 5, pp. 546–550, 1992.

[15] M. Sams, R. Hari, J. Rif, and J. Knuutila, "The human auditory sensory memory trace persists about 10 sec: Neuromagnetic evidence," *Journal of Cognitive Neuroscience*, vol. 5, no. 3, pp. 363–370, 1993.

[16] N. Grossheinrich, S. Kademann, J. Bruder, J. Bartling, and W. Von Suchodoletz, "Auditory sensory memory and language abilities in former late talkers: A mismatch negativity study," *Psychophysiology*, vol. 47, no. 5, pp. 822–830, 2010.

[17] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 03 1951.

[18] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1858–1860, 2003.

[19] T. van Erven and P. Harremos, "Rényi Divergence and Kullback-Leibler Divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, July 2014.

[20] G. J. Székely, "E-statistics: The energy of statistical samples," Bowling Green State University, Tech. Rep. 02-16, 2002.

[21] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99–109, 1943.

[22] J. Li, L. Deng, Y. Gong, S. Member, R. Haeb-umbach, and S. Member, "An Overview of Noise-Robust Automatic Speech Recognition," vol. 22, no. 4, pp. 745–777, 2014.

[23] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct 1994.

[24] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 4574–4577.

[25] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr 1979.

[26] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec 1979.

[27] D. Yu and J. Li, "Recent progresses in deep learning based acoustic models," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 3, pp. 396–409, 2017.

[28] Y. Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition," in *Interspeech 2016*, 2016, pp. 2369–2372. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-879

[29] D. Serdyuk, K. Audhkhasi, P. Brakel, B. Ramabhadran, S. Thomas, and Y. Bengio, "Invariant representations for noisy speech recognition," 2016. [Online]. Available: http://arxiv.org/abs/1612.01928

[30] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," 2013.

[31] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan 1999.

[32] A. Mesaros, T. Heittola, T. Virtanen, E. Benetos, P. Foster, M. Lagrange, G. Lafay, and M. D. Plumbley, "IEEE AASP challenge: Detection and classification of acoustic scenes and events 2016," http://www.cs.tut.fi/sgn/arg/dcase2016/, 2016.

[33] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR Evaluation of Acoustic Event Detection and Classification Systems," *LNCS 4122*, pp. 311–322, 2007.

[34] M. Grootel, T. Andringa, and J. Krijnders, "DARES-G1: Database of Annotated Real-world Everyday Sounds," in *Proceedings of the NAG/DAGA Meeting 2009*, Rotterdam, 2009.

[35] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.

[36] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," *The Bell System Technical Journal*, vol. 12, no. 4, pp. 377–430, 1933.

[37] T. Virtanen, A. Mesaros, T. Heittola, M. Plumbley, P. Foster, E. Benetos, and M. Lagrange, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*. Tampere University of Technology. Department of Signal Processing, 2016.

[38] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247 – 251, 1993.

[39] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535 – 557, 2017.

[40] D. Pearce, H.-G. Hirsch, and E. E. D. Gmbh, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *in ISCA ITRW ASR2000*, 2000, pp. 29–32.

[41] M. Brookes, "VOICEBOX: Speech processing toolbox for MATLAB," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1997.

[42] ITU-T, "Objective measurement of active speech level," International Telecommunication Union, Tech. Rep. p.56 (12/11), 1994.

[43] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[44] E. Macaluso, "MT_TOOLS : Computation of saliency and feature-specific maps," http://www.brainreality.eu/mt_tools/, 2010.

[45] H. Zhang, X. Cao, and R. Wang, "Audio visual attribute discovery for fine-grained object recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 7542–7549.

Antonio Rodríguez-Hidalgo received his B.Sc. in Telecommunication Engineering in 2014 from Universidad de Granada (Spain). Afterwards, he completed his Telecommunication Engineering M.Sc. and Multimedia and Communications M.Sc. from Universidad Carlos III de Madrid (Spain) in 2016. He has been a research visitor in Chait-Lab at University College London (Ear Institute, UCL, UK) in 2017. He is currently pursuing his Ph.D. in Universidad Carlos III de Madrid, where he has gained experience in processing a wide variety of signals: MRI, audiovisual content, EEG/MEG, ultrasounds, etc. His main research interests include salience algorithms, signal processing and machine learning, as well as feature engineering.

Carmen Peláez-Moreno received her Telecommunication Eng. degree from the Public University of Navarre in 1997 and Ph.D. from the University Carlos III of Madrid in 2002. Her Ph.D. thesis has been awarded a 2002 Best Doctoral Thesis Prize from the Spanish Official Telecom. Eng. Association (COIT-AEIT). From March to Dec. 2004, she participated in the International Computer Science Institute's (ICSI, Berkeley (CA)) Fellowship Program. Since Nov. 2009, she is an Associate Professor in the Department of Signal Theory and Communications at the University Carlos III of Madrid. Her research interests include speech recognition and perception, multimedia processing, machine learning and data analysis. She has co-authored over 60 papers in prestigious international journals, books and peer-reviewed conferences.

Ascensión Gallardo-Antolín received her Ph. D. in Telecommunication Engineering from the Polytechnic University of Madrid, Spain, in 2002. She has been a visiting scientist at the International Computer Science Institute (ICSI, Berkeley, USA) in 2005, the German Research Center for Artificial Intelligence (DFKI, Saarbrucken, Germany) in 2006 and the Centre for Speech Technology Research (CSTR, University of Edinburgh, UK) in 2013. Currently, she is an Associate Professor at the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain. She has coauthored more than 70 peer-reviewed papers in international journals and national and international conferences. She has participated in several research projects including some of the Spanish Council on Science and Technology and the UE. She has received the Best Ph. D. Thesis Award from the Professional Association of Telecommunication Engineers of Spain (COIT) and the Ph. D. Excellence Award from the Polytechnic University of Madrid. Her main research interests include automatic speech recognition, audio classification and segmentation, multimedia information retrieval, auditory and visual salience models and signal processing for multimedia human-machine interaction.