"WOW!" BAYESIAN SURPRISE FOR SALIENT ACOUSTIC EVENT DETECTION

B. Schauerte *R. Stiefelhagen*

Institute for Anthropomatics
Karlsruhe Institute of Technology
Vincenz-Prießnitz-Str. 3
76131 Karlsruhe, Germany
http://cvhci.anthropomatik.kit.edu/~bschauer/

ABSTRACT

We extend our previous work and present how Bayesian surprise can be applied to detect salient acoustic events. Therefore, we use the Gamma distribution to model each frequencies spectrogram distribution. Then, we use the Kullback-Leibler divergence of the posterior and prior distribution to calculate how "unexpected" and thus surprising newly observed audio samples are. This way, we are able to efficiently detect arbitrary, unexpected and thus surprising acoustic events. Complementing our qualitative system evaluations for (humanoid) robots, we demonstrate the effectiveness and practical applicability of the approach on the CLEAR 2007 acoustic event detection data.

Index Terms— Acoustic event detection, Acoustic saliency, Cognition, Probability, Algorithms

1. INTRODUCTION

Attention is the cognitive process that has to identify subsets within sensory inputs that contain important information to focus subsequent complex and slow processing operations on the potentially relevant information. This is a key capability in biological and artificial systems that enables real-time processing despite limited computational capacities. However, as a consequence, attention has to process all incoming sensory information (e.g., from the millions of human sensory receptors), making it computationally challenging. But, since attention serves as a gateway to later processing steps, efficient, reliable, and fast attentional allocation is key to efficient processing of complex natural scenes.

Unfortunately, auditory and acoustic attention has not yet found its way into as many practical applications as its visual counterpart (see, e.g., [1–3]). One of the possible reasons for this could be that only few run-time efficient, reliable, and robust models for auditory attention exist that have been proposed and tested for practical applications. In this paper, we propose the use of acoustic surprise for the detection of arbitrary salient acoustic events. In principle, surprise defines important stimuli as statistical outliers given the previous observations of a signal, which naturally also integrates the concept of novelty detection. We primarily developed acoustic surprise to focus the computational resources and control the overt attention of (humanoid) robots, see Fig. 1, and smart environments (see [1, 3–5]). In such applications, acoustic surprise can serve two purposes: First, we can focus audio processing and, second, we can actively control the overt attention (i.e., the sensor orientation) to optimize the scene perception. For both purposes we profit from the low run-time requirements (depending of course on the configuration, calculating audio surprise requires 1.5 seconds for one minute of audio recordings, in Matlab), which, first, provide a net benefit of required computational resources for audio precessing and, second, allow for the robot's rapid reaction time on salient events. However, previously we did not perform a quantitative evaluation on real-world data and instead focused our evaluations on the behavior of the overall system. In this paper, first, we propose the use of the Gamma distribution with a forgetting factor in place of the Gaussian distribution for surprise calculation. Second, we perform a quantitative evaluation of surprise on acoustic event detection data, which quantitatively demonstrates the practical applicability of our approach.

The remainder of this paper is organized as follows: In the following section 2, we provide a brief overview of related work. Subsequently, in section 3, we describe how we calculate the surprise of an audio signal. Then, in section 4, we present our evaluation results. Finally, we conclude with a brief summary in section 5.

2. RELATED WORK

In recent years computational models of attention have attracted an increasing interest in the field of robotics (see, e.g., [1, 2]) and various other application areas (see, e.g., [3,

The work presented in this paper was supported by the German Research Foundation (DFG) within the Collaborative Research Program SFB 588 "Humanoide Roboter" and the Quaero Programme, funded by OSEO, French State agency for innovation.

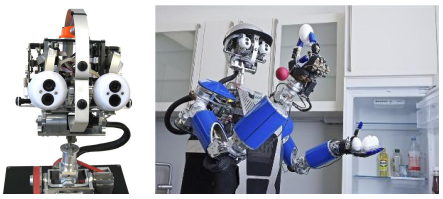

Fig. 1. Our robotic target platform: The Karlsruhe Humanoid Head and the ARMAR humanoid robot (see [18]), which served as platform for our qualitative system evaluations [1].

6–9]). Although in principle all attention models serve the same purpose, i.e. to highlight potentially relevant and thus interesting – that is to say "salient" – data, the vagueness and task-dependence of this problem description leads to a variety of models that may differ substantially in which parts of the signal they mark as being of interest. Unfortunately, in contrast to the fast growing amount of proposed visual saliency models (see [10,11]), only few practically applicable models for acoustic attention exist (e.g., [1,6,7]). Most closely related to our work is the model described by Kayser et al. [6] which is based on the well-known visual saliency model of Itti et al. [12] and, most notably, has been successfully applied to speech processing by Kalinli et al. [7] and, in principle, by Lin et al. [8] to allow for faster human acoustic event detection through audio visualization. However, it has several drawbacks: First, it is computationally expensive, because it requires the calculation and combination of a considerable amount of feature maps. Second, it requires that the spectrogram has elements of the future to detect salient events in the present (due to the inherent down-scaling and filtering in Itti et al.'s model, salient stimuli at the borders are always problematic), which prohibits online detection. Finally, although Itti et al.'s saliency model represents an outstanding historical accomplishment, it can hardly be said to be state-of-the-art (see, e.g., [13–15]). To account for these drawbacks, we introduced acoustic Bayesian surprise in 2011 [1]. It relies on a probabilistic model of the signals' frequency distribution to calculate the "surprise", which in principle measures how unexpected an observed signal is given the preceding observations [16,17]. In this paper, we extend our previous work with respect to two main aspects: First, we propose the use of the Gamma distribution instead of the previously applied Gaussian distribution. Second, we provide a quantitative evaluation, which nicely complements and substantially adds to our previous, mostly qualitative system evaluations (see [1,4,5]).

3. ACOUSTIC SALIENCY MODEL

In the following, we present our definition of acoustic surprise that we use to detect acoustically salient events online and in real-time.

3.1. Time-Frequency Analysis and Bayesian Framework

First, we use the short-time Fourier transform (STFT), short-time cosine transform (STCT) or the modified discrete cosine transform (MDCT) to calculate the spectrogram $G(t,\omega) = |F(t,\omega)|^2$ of the windowed audio signal $a(t)$, where t and ω denote the discrete time and frequency, respectively.

In the Bayesian probability framework, probabilities correspond to subjective degrees of beliefs (see, e.g., [19]) in models which are updated according to Bayes rule as new data is observed. At each time step t, the new data $G(t,\omega)$ is used to update the prior probability distribution $P^{\omega}_{\text{prior}} = P(\cdot|G(t-1,\omega),\ldots,G(t-N,\omega))$ of each frequency and obtain the posterior distribution $P^{\omega}_{\text{post}} = P(\cdot|G(t,\omega),G(t-1,\omega),\ldots,G(t-N,\omega))$, where $N \in \{1,\ldots,\infty\}$ allows additional control of the time behavior by limiting the history to $N \neq \infty$ elements. The history allows to limit the influence of samples over time and consequently "forget" data, which is essential for the time behavior of the Gaussian surprise model.

3.2. Acoustic Surprise

3.2.1. Gaussian model

Using the Gaussian distributions as model[1], we can calculate the surprise $S_A(t,\omega)$ for each frequency

$$S_{\text{A}}(t,\omega) = D_{\text{KL}}(P^{\omega}_{\text{post}}||P^{\omega}_{\text{prior}}) = \int P^{\omega}_{\text{post}} \log \frac{P^{\omega}_{\text{post}}}{P^{\omega}_{\text{prior}}} dg \quad (1)$$

$$= \frac{1}{2}[\log \frac{|\Sigma^{\omega}_{\text{prior}}|}{|\Sigma^{\omega}_{\text{post}}|} + \text{Tr}\left[\Sigma^{\omega^{-1}}_{\text{prior}}\Sigma^{\omega}_{\text{post}}\right] - I_{\text{D}} + \quad (2)$$

$$(\mu^{\omega}_{\text{post}} - \mu^{\omega}_{\text{prior}})^T \Sigma^{\omega^{-1}}_{\text{prior}}(\mu^{\omega}_{\text{post}} - \mu^{\omega}_{\text{prior}})] \quad ,$$

where μ and Σ is the mean and variance, respectively, of the data in the considered time window, i.e. history. D_{KL} is the Kullback Leibler Divergence and Eqn. 2 results from the closed form of D_{KL} for Gaussian distributions (see [20]). Consequently, an observed spectrogram element $G(t,\omega)$ is surprising if the updated distribution P^{ω}_{post}, which is the result of incorporating $G(t,\omega)$, differs significantly from the prior distribution $P^{\omega}_{\text{prior}}$.

3.2.2. Gamma model

Similar to the approach by Itti and Baldi for detecting surprising events in computer vision [21], we can alternatively use the Gamma distribution

$$P(x) = \gamma(x;\alpha,\beta) = \frac{\beta^{\alpha}x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)} \quad (3)$$

with $x \geq 0$, $\alpha,\beta > 0$, and Gamma function Γ, to calculate the surprise.

[1] If you are interested, you can download and experiment with our public demo implementation, see http://bit.ly/SqDDkn.

Given a new observation $G(t,\omega)$ and prior density $P^{\omega}_{\text{prior}} = \gamma(\cdot;\alpha,\beta)$, we calculate the posterior $P^{\omega}_{\text{post}} = \gamma(\cdot;\alpha',\beta')$ using Bayes' rule

$$\alpha' = \alpha + G(t,\omega) \quad (4)$$
$$\beta' = \beta + 1\,. \quad (5)$$

However, using this update rule would lead to an unbounded growth of the values over time. To avoid this behavior and reduce the relative importance of older observations, we integrate a decay factor $0 < \zeta < 1$

$$\alpha' = \zeta\alpha + G(t,\omega) \quad (6)$$
$$\beta' = \zeta\beta + 1\,. \quad (7)$$

This formulation preserves the prior's mean $\mu = \frac{\alpha}{\beta} = \frac{\zeta\alpha}{\zeta\beta}$ but increases its variance, which however represents a relaxation of belief in the prior's precision after observing $G(t,\omega)$.

Now, we can calculate the surprise as follows

$$S_{\text{A}}(t,\omega) = D_{\text{KL}}(P^{\omega}_{\text{post}}||P^{\omega}_{\text{prior}}) = \int P^{\omega}_{\text{post}} \log \frac{P^{\omega}_{\text{post}}}{P^{\omega}_{\text{prior}}} dg \quad (8)$$
$$= \alpha' \log \frac{\beta}{\beta'} + \log \frac{\Gamma(\alpha')}{\Gamma(\alpha)} \quad (9)$$
$$+\beta' \frac{\alpha}{\beta} + (\alpha - \alpha')\psi(\alpha)\,, \quad (10)$$

where ψ is the Digamma function. Unfortunately, the Gamma and Digamma functions Γ and ψ, respectively, do not have a closed form. But, there exist sufficiently accurate approximations (see, e.g., [22]), which however make the calculation slightly more complex than in the case of the Gaussian model.

3.3. Across Frequency Combination

Finally, we calculate the acoustic saliency $S_{\text{A}}(t)$ as the mean over all frequencies

$$S_{\text{A}}(t) = \frac{1}{|\Omega|} \sum_{\omega\in\Omega} S_{\text{A}}(t,\omega) \quad . \quad (11)$$

We do not use an alternatively possible joint (e.g., Dirichlet) model for the surprise calculation due to its computational complexity. Such a joint model would require the calculation of a general covariance matrix, which given the typically large number of analyzed frequencies makes it impractical for real-time processing.

4. EVALUATION

4.1. Evaluation Measure

In contrast to, e.g., recording eye fixations as a measure of visual saliency (see, e.g., [14]), we can not simply observe and record humans to provide a measure of acoustic saliency. Consequently, we follow a pragmatic, application-oriented evaluation approach that enables us to use existing acoustic event detection and classification datasets. In summary, salient acoustic event detection has to suppress "uninteresting" audio data while highlighting potentially relevant and thus salient acoustic events. However, in contrast to classical acoustic event detection and classification, this leads to a different evaluation methodology in which a high recall is necessary (i.e., we want to detect all prominent events) whereas a high precision is of secondary interest (i.e., we can tolerate false positives as long as we still filter the signal in such a way that we achieve a net run-time benefit when taking into account subsequent processing steps). We can realize this evaluation idea by using the well-established F_β score

$$F_\beta = (1+\beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \quad (12)$$
$$F_\beta = \frac{(1+\beta^2) \cdot \text{true pos.}}{(1+\beta^2) \cdot \text{true pos.} + \beta^2 \cdot \text{false neg.} + \text{false pos.}} \quad (13)$$

as evaluation measure[2], where β "measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision" [23].

4.2. Evaluation Data

We use the CLEAR2007 acoustic event detection dataset for evaluation [24], which was supported by the European Integrated project CHIL and the US National Institute of Standards and Technology (NIST). The dataset contains recordings of meetings in a smart room. For each recording a human user marked and classified (14 classes) acoustic events. Here, it is interesting to note that not all events could be classified by the human user, in which case they were labeled with "unknown".

4.3. Evaluation Parameters

For the time-frequency analysis, we set the window size to contain 1 second of audio data, which has a resolution of 22 kHz, and use 50 % overlap[3]. We also experimented with different window functions (e.g., Blackman, Gauss), but the resulting performance difference between most window functions is relatively small, if the parameters are well defined. We evaluated the performance for the modified discrete cosine transform (MDCT), short-time cosine transform (STCT), and

[2] This procedure is comparable to the use of F_β for salient object detection in image processing applications (see, e.g., [15]).

[3] Please note that the choice of algorithm parameters substantially influences the performance and run-time (e.g., we were able achieve an F_2 score of 0.9414 at the cost of considerably higher run-time requirements). Since ICASSP's page limit does not allow for an evaluation of every parameter's influence, we present results for a reasonable parameter configuration. However, since we make our implementation publicly available, other researchers will be able to experiment with different configurations that might be more suitable for their target application.

Algorithm	F_1	F_2	F_4
STFT + Gamma	0.7668	0.8924	0.9665
STCT + Gamma	0.7658	0.8916	0.9655
MDCT + Gamma	0.7644	0.8894	0.9647
STFT + Gaussian	0.7604	0.8832	0.9531
STCT + Gaussian	0.7612	0.8813	0.9529
MDCT + Gaussian	0.7613	0.8805	0.9538

Table 1. Performance of the evaluated acoustic surprise algorithms on CLEAR 2007 acoustic event detection data. The F_2 and F_4 scores are our main evaluation measure, because for our application a high recall is much more important than a high precision (we provide the F_1 score mainly to serve as a reference). As can be seen, the proposed use of the Gamma distribution improves the performance. Furthermore, this being the first quantitative evaluation, we can see that the proposed use of surprise to detect arbitrary, interesting acoustic events matches our subjective experiences and does indeed perform well in highlighting (salient) acoustic events.

short-time Fourier transform (STFT) to determine whether or not the Gamma distribution is beneficial for every of these transformations. We do this, because one aim is to produce as little run-time overhead as possible, which requires us to ideally rely on the transformation that is used for the subsequent processing steps such as, e.g., sound source localization, event recognition, and/or speech recognition. We optimized the history size and forgetting parameter for the Gaussian and Gamma model, respectively, and report the results for the best choice.

4.4. Results

As can be seen in Tab. 1, quantified using the F_1, F_2, and F_4 score, acoustic surprise is able to efficiently detect arbitrary salient acoustic events. Although in general an F_1 score of roughly 0.77 is far from perfect for precise event detection[3], we can see from the substantially higher F_2 and F_4 scores that we can efficiently detect most (salient) acoustic events, if we tolerate a certain amount of false positives. This nicely fulfills the target requirements for our application domains and comes at a low computational complexity that using, e.g., Gaussian surprise allows us to process one minute of audio data in roughly 1.5 seconds. Furthermore, the detection of salient events can be performed online and we can not just detect salient points in time, but since we calculate the surprise value all frequencies that we subsequently combine, we can also determine which frequencies trigger the detection. We also see that the proposed use of the Gamma distribution provides a better performance compared to the Gauss distribution, see Tab. 1. This is consistent over all considered transformations and with our experience on other parameter configurations[3]. It is also interesting to see that the chosen transformation only has a minor influence on the achievable performance.

5. CONCLUSION

We have extended our previous work on acoustic surprise. Most importantly, we introduced the use of the Gamma distribution in combination with a "forgetting" factor. Complementing our previous experiments in which we focused on the behavior of the system (i.e., a humanoid robot) as a whole, we have demonstrated that acoustic surprise performs well in detecting arbitrary acoustically salient events using acoustic event data from the CLEAR 2007 corpus. We would like to note that this performance comes with a low computational complexity, which is a key feature that makes the integration of auditory attention into an actual system beneficial.

6. REFERENCES

[1] B. Schauerte, B. Kühn, K. Kroschel, and R. Stiefelhagen, "Multimodal saliency-based attention for object-based scene analysis," in *Proc. Int. Conf. Intell. Robots Syst.*, 2011. 1, 2

[2] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer, "Multimodal saliency-based bottom-up attention: A framework for the humanoid robot iCub," in *Proc. Int. Conf. Robot. Autom.*, 2008. 1

[3] B. Schauerte, J. Richarz, T. Plötz, C. Thurau, and G. A. Fink, "Multi-modal and multi-camera attention in smart environments," in *Proc. Int. Conf. Multimodal Interfaces*, 2009. 1

[4] B. Kühn, B. Schauerte, R. Stiefelhagen, and K. Kroschel, "A modular audio-visual scene analysis and attention system for humanoid robots," in *Proc. 43rd Int. Symp. Robotics*, 2012. 1, 2

[5] B. Kühn, B. Schauerte, K. Kroschel, and R. Stiefelhagen, "Multimodal saliency-based attention: A lazy robot's approach," in *Proc. Int. Conf. Intell. Robots Syst.*, 2012. 1, 2

[6] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map," *Current Biology*, vol. 15, no. 21, pp. 1943–1947, 2005. 1, 2

[7] O. Kalinli and S. Narayanan, "Prominence detection using auditory attention cues and task-dependent high level information," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 17, no. 5, pp. 1009–1024, 2009. 1, 2

[8] K.-H. Lin, X. Zhuang, C. Goudeseune, S. King, M. Hasegawa-Johnson, and T. S. Huang, "Improving faster-than-real-time human acoustic event detection by saliency-maximized audio visualization," in *Proc. Int.*

Conf. Acoustics, Speech and Signal Processing, 2012. 1, 2

[9] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Auditory attention – focusing the searchlight on sound," *Current Opinion in Neurobiology*, vol. 17, no. 4, pp. 437–455, 2007. 1

[10] J. K. Tsotsos, *A Computational Perspective on Visual Attention*, The MIT Press, 2011. 2

[11] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundation: A survey," *ACM Trans. Applied Perception*, vol. 7, no. 1, pp. 6:1–6:39, 2010. 2

[12] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998. 2

[13] B. Schauerte and R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction," in *Proc. European Conf. Comp. Vis.*, 2012. 2

[14] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. Int. Conf. Comp. Vis.*, 2009. 2, 3

[15] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned Salient Region Detection," in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2009. 2, 3

[16] L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems*, 2006. 2

[17] P. F. Baldi and L. Itti, "Of bits and wows: A bayesian theory of surprise with applications to attention," *Neural Networks*, vol. 23, no. 5, pp. 649–666, 2010. 2

[18] T. Asfour, K. Welke, P. Azad, A. Ude, and R. Dillmann, "The Karlsruhe Humanoid Head," in *Humanoids*, 2008. 2

[19] D. Gillies, "The subjective theory," in *Philosophical Theories of Probability*, chapter 4. Routledge, 2000. 2

[20] J. R. Hershey and P. A. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2007. 2

[21] L. Itti and P. F. Baldi, "A principled approach to detecting surprising events in video," in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2005. 2

[22] J. M. Bernardo, "Algorithm as 103 psi(digamma function) computation," *Applied Statistics*, vol. 25, pp. 315–317, 1976. 3

[23] C. J. van Rijsbergen, *Information Retrieval*, Butterworth, 2nd edition, 1979. 3

[24] CLEAR2007, "Classification of events, activities and relationships evaluation and workshop," http://www.clear-evaluation.org. 3