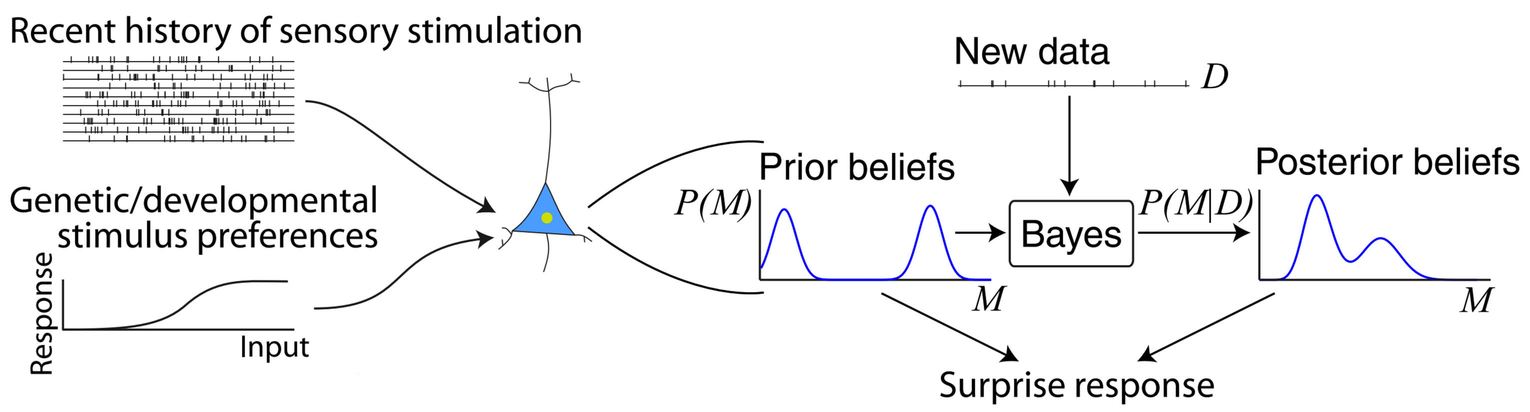
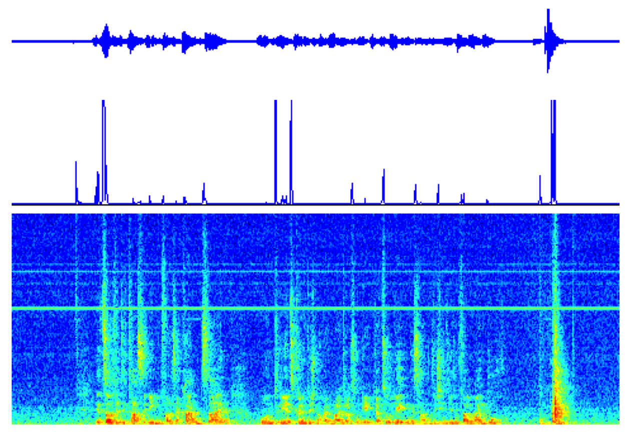
"Wow!" Bayesian Surprise for Salient Acoustic Event Detection

Boris Schauerte & Rainer Stiefelhagen
Karlsruhe Institute of Technology
{boris.schauerte, rainer.stiefelhagen}@kit.edu

Abstract

We propose the use of Bayesian surprise to detect arbitrary, salient acoustic events. We use Gaussian or Gamma distributions to model the spectrogram distribution and 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 efficiently detect arbitrary surprising/salient acoustic events.





Motivation

- identify subsets within sensory inputs that are likely to contain important information
- focus complex processing operations on the selected, potentially relevant information
- in general, drastically reduce the computational requirements to process data
- real-time processing and reflex-like reactions despite computational restrictions

Principle

- an observed spectrogram element $G(t,\omega)$ is "surprising" if the updated (using Bayes' rule) distribution P^{ω}_{post} differs significantly from the prior distribution $P^{\omega}_{\text{prior}}$

$$S_A(t,\omega) = D_{\text{KL}}(P^{\omega}_{\text{post}} \| P^{\omega}_{\text{prior}}) \quad (1)$$

$$= \int P^{\omega}_{\text{post}} \log \frac{P^{\omega}_{\text{post}}}{P^{\omega}_{\text{prior}}} dg \quad (2)$$

with Kullback-Leiber divergence D_{KL}

- surprise at time t over all frequencies

$$S_A(t) = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} S_A(t,\omega) \quad (3)$$

- the unit of surprise is called "wow" [4]

Models

Gaussian:

$$S_A(t,\omega) = \frac{1}{2}[\log \frac{|\Sigma^{\omega}_{\text{prior}}|}{|\Sigma^{\omega}_{\text{post}}|} + \text{Tr}\left[\Sigma^{\omega\,-1}_{\text{prior}} \Sigma^{\omega}_{\text{post}}\right] - I_D + \quad (4)$$

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

- mean μ and variance Σ of the data in the considered time window, i.e. history
- advantage: exact closed form solution exists (highly efficient to calculate)

Gamma:

$$S_A(t,\omega) = \alpha' \log \frac{\beta}{\beta'} + \log \frac{\Gamma(\alpha')}{\Gamma(\alpha)} \quad (5)$$

$$+\beta' \frac{\alpha}{\beta} + (\alpha - \alpha')\psi(\alpha) \quad (6)$$

- $\alpha, \beta > 0$, and Gamma function Γ and Digamma function ψ
- advantage: better control over the history using the decay/forgetting factor $0 < \zeta < 1$ and update rule

$$\alpha' = \zeta\alpha + G(t,\omega) \quad (7)$$

$$\beta' = \zeta\beta + 1 \quad (8)$$

Evaluation

Idea:

- we can not simply observe humans to provide a measure of acoustic saliency
- pragmatic, application-oriented approach: use existing acoustic event detection and classification datasets
- salient acoustic event detection has to suppress "uninteresting" audio data while highlighting potentially relevant and thus salient data segments

CLEAR2007 acoustic event detection **dataset:**

- recordings of meetings in a smart room
- a human user marked and classified (14 classes) acoustic events
- not all events could be classified by the human user (i.e., "unknown" class)

F_β score as **evaluation measure:**

$$F_\beta = (1+\beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \quad (9)$$

- we want to detect all prominent events, i.e. a high recall is most important
- we can tolerate false positives as long as we achieve a net run-time benefit when focusing subsequent algorithms, i.e. a high precision is of secondary interest
- "β times as much importance to recall as precision"

Results:

	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

Applications

Robotics [1]:

- the robot can efficiently detect, investigate, and react on arbitrary, unexpected - i.e., interesting - events
- focus and make better use of the robot's limited computational ressources

Intensive Care [2]:

- use Gaussian surprise to detect (sudden) patient agitation based on facial features

Biological Motivation

- spectrogram ~ basilar membrane [3]
- surprise ~ early sensory neurons [4]

Code?

- Gamma and Gaussian surprise implementation public (BSD license) at `http://bit.ly/ZjzXqr`
- comes with a ready to go audio example

References

[1] B. Schauerte, B. Kühn, K. Kroschel, R. Stiefelhagen, Multimodal Saliency-based Attention for Object-based Scene Analysis, in IROS, 2011

[2] M. Martinez, R. Stiefelhagen, Automated Multi-Camera System for Long Term Behavioral Monitoring in Intensive Care Units, in MVA, 2013

[3] Schnupp J., Nelken I., King A, Auditory Neuroscience, MIT Press, Cambridge, MA, 2011.

[4] L. Itti and P. F. Baldi, Bayesian surprise attracts human attention, in NIPS, 2006.

Acknowledgment

This work was partially funded by the French state agency for innovation (OSEO) within the Quaero programme and the German research foundation (DFG) within the collaborative research program "Humanoide Roboter".

ICASSP 2013
Vancouver Convention & Exhibition Centre
May 26 - 31, 2013 • Vancouver, Canada