DETECTION OF SALIENT EVENTS IN AN ACOUSTICAL SCENE

Kristian Kroschel
Karlsruhe Institute of Technology (KIT)
kristian.kroschel@kit.edu

Abstract: An event in an acoustical scene is salient if it is different from other events in the scene. The event in this paper is the utterance *gut* in German and the sound of keys falling on a hard surface. The other events are white and coloured noise and a constant sinusoidal tone. Humans are able to detect salient acoustic events in a scene with background signals carrying no information due to their stationarity like noise or a constant sinusoid. A lot of research has been done to detect salient events in image processing which is based on the Kullback-Leibler divergence. The basis of this approach is the prior and posterior information carried by the corresponding probability densities which is known from video analysis and transferred to an audio scene. It is shown that the relevant events utterance and the sound of the falling keys can be marked reliably even if they are severely corrupted by irrelevant events in the background like noise or a constant sine wave.

1 Introduction

Assume that your eyes are closed and you listen to the acoustic events in your environment. If there is a vacuum cleaner running you might realize this event when it is switched on or off, but if it is running continuously you will not pay attention to it. This is different if a human starts to speak: you will listen to the utterance to find out whether it carries information adressed to you. If not, you might not be interested to listen and the utterance is not salient any more.

The question is how can we extract saliency from the sensed data? In the case of images, saliency might be based on a change of colour or texture - a red rose in a green environment of leaves - or in video streams the speed of the movement of objects - a police car is driven at high speed - might be salient. A similar event in an acoustic scene is the sound of the siren of a police car with the background of traffic noise.

But how to detect salient events and how to measure saliency? No doubt: the change of a feature in a given environment is the key to detect salient events. But which are the appropriate features? On the field of visual saliency many features have been proposed [1][2][3].

An acoustic signal can be described in the time and frequency domain, respectively. If a signal carries information it will be the sample function $s(t)$ of a random process [4] which is only short-time stationary. Thus we need a description of the process along the time and the frequency axis. An appropriate tool for this description is the spectrogram $S(t, f)$. Using this tool we are able to discriminate between salient events and those without saliency. Imagine you listen to the sound of a sine wave with no changes in amplitude, frequency and phase: it will carry no information. In the spectrogram such a sinusoid corresponds with a line parallel to the time axis with no change in frequency and intensity. Similar to that is a white noise process with a homogenious - ideally a constant - pattern all over the spectrogram. Both signals are not salient and do not evoke attention for a listening human.

No question: without intensity there is no salient event [5]. But intensity alone and even changes in intensity are not sufficient to detect salient events. Instead, statistical properties of the sensed data are exploited and a measure based on Bayesian Surprise [2] is used.

This measure is the basis for the Kullback-Leibler divergence [6] which compares the probability density of the prior state with the probability density of the posterior state when the sensed data are exploited.

2 A Measure of Saliency

Assume that prior information about an event is given by the probability density function $f_{S_{pri}}(s_{pri})$ which is calculated from old data. Receiving new data, the posterior probability density $f_{S_{pri}|S_{pos}}(s_{pri}|s_{pos})$ can be calculated. Applying the Bayes rule [4], the prior information given the posterior data yields

$$f_{S_{pri}|S_{pos}}(s_{pri}|s_{pos}) = \frac{f_{S_{pos}|S_{pri}}(s_{pos}|s_{pri})}{f_{S_{pos}}(s_{pos})} \cdot f_{S_{pri}}(s_{pri}). \tag{1}$$

If the new data determining the conditional density $f_{S_{pri}|S_{pos}}(s_{pri}|s_{pos})$ do not carry saliency, the prior density $f_{S_{pri}}(s_{pri})$ is not affected. Only if the conditional density differs significantly from the prior density the data carry saliency. For the comparison of both densities the Kullback-Leibler divergence is used

$$D_{KL}(f_{S_{pri}|S_{pos}}(s_{pri}|s_{pos}), f_{S_{pri}}(s_{pri})) = \int_s f_{S_{pri}|S_{pos}}(s_{pri}|s_{pos}) \cdot \log_2 \left(\frac{f_{S_{pri}|S_{pos}}(s_{pri}|s_{pos})}{f_{S_{pri}}(s_{pri})} \right) ds_{pri}. \tag{2}$$

The problem is to allocate the probability densities from the available data: what are the prior data, what the posterior data? To calulate both entities, the data stream of the sound is cut into overlapping blocks. The prior data are given by all available data up to the actual data block and the posterior data include in addition the actual data block.

3 The Saliency Map Extracted from Audio Data

Audio signals can be described in the time and the frequency domain. A representation in both domains is given by the spectrogram $S[k,n]$ with k denoting the discrete time parameter and n denoting the discrete spectral or frequency parameter

$$S[k,n] = \left| \sum_{i=k-N+1}^{k} s[i] e^{-j2\pi in/N} \right|^2, \qquad 1 \le n \le N \tag{3}$$

with N the length of the observed data $s[k]$ from which the spectrum is calculated.

From the samples $S[k,n]$ of the spectrogram given in Eq.3 the prior density $f_{S_{pri}}(s_{pri})$ and the posterior density $f_{S_{pri}|S_{pos}}(s_{pri}|s_{pos})$ have to be calculated. For this calculation the vectors

$$\mathbf{S}_{pri} = (S[k-1,n], S[k-2,n], \ldots, S[k-K,n])^T$$

and

$$\mathbf{S}_{pos} = (S[k,n], S[k-1,n], \ldots, S[k-K,n])^T$$

are used with K a free parameter which in principle can approach $K \to \infty$. Thus we have

$$f_{\mathbf{S}_{pri}|\mathbf{S}_{pos}}(\mathbf{s}_{pri}|\mathbf{s}_{pos}) = g(\mathbf{s}|\mathbf{S}_{pos}), \qquad f_{\mathbf{S}_{pri}}(\mathbf{S}_{pri}) = g(\mathbf{s}|\mathbf{S}_{pri}) \tag{4}$$

in a simplified representation.

For the saliency depending on the frequency paramater n follows

$$\begin{aligned} S_A[k,n] &= D_{KL}(g(\mathbf{s}|\mathbf{S}_{pos}), g(\mathbf{s}|\mathbf{S}_{pri})) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{s}|\mathbf{S}_{pos}) \cdot \log_2 \left(\frac{g(\mathbf{s}|\mathbf{S}_{pos})}{g(\mathbf{s}|\mathbf{S}_{pri})} \right) d\mathbf{s} \end{aligned} \tag{5}$$

which is known as the saliency map [5].

To simplify the calculation it is assumed that the densities $g(\mathbf{s}|\mathbf{S}_{pri})$ and $g(\mathbf{s}|\mathbf{S}_{pos})$ are Gaussian. In this case only the means μ_{pri}, μ_{pos} and variances σ_{pri}, σ_{pos} of the densities are required. From Eq.5 follows [7]

$$\begin{aligned} S_A[k,n] &= D_{KL}(g(\mathbf{s}|\mathbf{S}_{pos}), g(\mathbf{s}|\mathbf{S}_{pri})) \\ &= \frac{1}{2} \left[\log_2 \frac{\sigma_{pri}^2}{\sigma_{pos}^2} + \frac{\sigma_{pos}^2}{\sigma_{pri}^2} - 1 + \frac{(\mu_{pos} - \mu_{pri})^2}{\sigma_{pri}^2} \right] \end{aligned} \tag{6}$$

with the means and variances as unbiased estimates [4] given by

$$\begin{aligned} \mu_{pri}[k,n] &= \frac{1}{K} \sum_{i=1}^{K} S[k-i,n] \\ \mu_{pos}[k,n] &= \frac{1}{K+1} \sum_{i=0}^{K} S[k-i,n] \\ \sigma_{pri}^2[k,n] &= \frac{1}{K} \sum_{i=1}^{K} (S[k-i,n] - \mu_{pri}[k,n])^2 \\ \sigma_{pos}^2[k,n] &= \frac{1}{K+1} \sum_{i=0}^{K} (S[k-i,n] - \mu_{pos}[k,n])^2. \end{aligned} \tag{7}$$

The measure in Eq.6 depends on time and frequency. Since only the dependency on time is of interest, the dependency on frequency is suppressed by averaging

$$S_A[\ell] = \frac{2}{N} \sum_{n=0}^{N/2-1} S_A[\ell,n] \tag{8}$$

which yields the saliency measure.

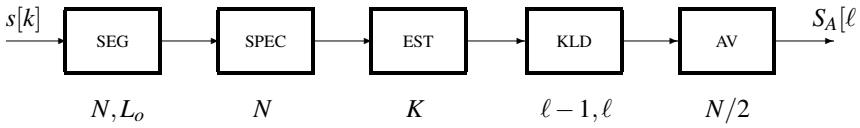

Figure 1 – Signal flow for data processing: SEGmentation, SPECtrum, ESTimation of μ and σ, Kullback-Leibler Divergence, AVeraging.

The calculation of the saliency measure $S_A[\ell]$ is visualized by the block diagram in Fig.1 with the blocks SEG for segmenation, SPEC for calculation of the spectrogam, EST for estimation of the means μ_{pri}, μ_{pos} and variances σ_{pri}^2, σ_{pos}^2, KLD for the calculation of the Kullback-Leibler divergence, and AV for the averaging over the $N/2$ frequency dependent saliency values $S_A[\ell,n]$.

The signal $s[k]$ to be analysed is first cut in block SEG into overlapping data blocks of length N as shown in Fig.2. The parameter N is also the length of the Fast Fourier transform (FFT) which is calculated in block SPEC. The consecutive blocks have an offset of L_o samples

in order to reduce the computational load on one hand and to increase the difference of the densities compared by the Kullback-Leibler divergence on the other hand. The parameter N controls the spectral resolution and the parameter L_o depends on the correlation of the analysed signal. L_o should be large enough so that the data blocks are not correlated any more.

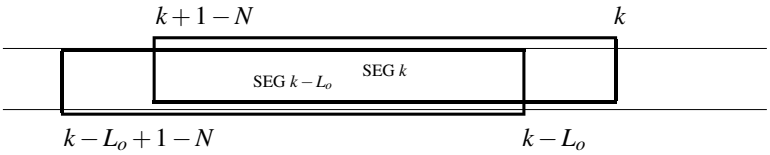

Figure 2 – Segmentation of the analysed signal $s[k]$ into consecutive overlapping data blocks.

By the block SEG the sampling rate is reduced by $1/L_o$ and thus the counter k is replaced by ℓ which is equivalent to down sampling. In block SPEC the block-dependent spectrogram $S[\ell,n]$ according to Eq.3 is calculated by FFT. As shown in Fig.3, the estimates of the means and variances according to Eq.7 are determined in block EST using $0 \leq \ell \leq K$ values for the prior estimates and $1 \leq \ell \leq K+1$ values for the posterior estimates of the spectrogram $S[\ell,n]$, respectively. The parameter K determines the reliability of the estimate, i.e. the larger K is the lower is the variance of the estimate but the larger is the computational load and the time delay. From estimates of the prior and posterior means and variances the values $S_A[\ell,n]$ of the saliency map according to Eq.6 are calculated in block KLD. Finally, these values are averaged over the frequency parameter n using $N/2$ values since $S_A[\ell,n]$ is symmetric with respect to $N/2$ and thus averaging over N does not improve the accuracy of the result.

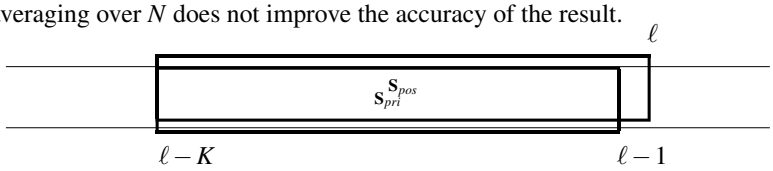

Figure 3 – Data blocks used for estimation of the prior and posterior means and variances.

4 Extracting Salient Events from Noisy Recordings

The signal of interest is the combination of the the German word *gut* by a female speaker and the sound of keys falling on a table. It is assumed that both, the utterance and the sound of the falling keys will be detected as salient events. It is not the task of the saliency detection to identify the salient events as speech and the sound of the falling keys or something similar. The representations of the signal $s[k]$ in the time and frequency domain are given in Fig.4 with the sampling frequency $f_s = 48$ kHz and around 80,000 samples. There is some residual noise visible in the time domain caused by the microphone and the amplifier which is mapped in the spectrogram into a narrow frequency band around $f = 18$ kHz.

In the time domain the utterance in the first half and the sound of the falling keys in the second half of the data record are clearly visible in Fig.4. The samples in the interval 2,640 $\leq k \leq$ 18,960 which is with $t = k/f_s$ equivalent to 0.055 s $\leq t \leq$ 0.395 s belong to the utterance and at $k = 41,571$ and $k = 57,893$ which is equivalent to $t = 0.866$ s and $t = 1.206$ s the peaks of the sound of the falling keys are found. The low level of the residual background noise is almost not audible for the human ear.

The spectrogram $S[k,n]$ is shown in the center of Fig.4 with $t = k/f_s$ along the time axis and $f = n \cdot f_s/N$ along the frequency axis. The utterance covers the spectral components around

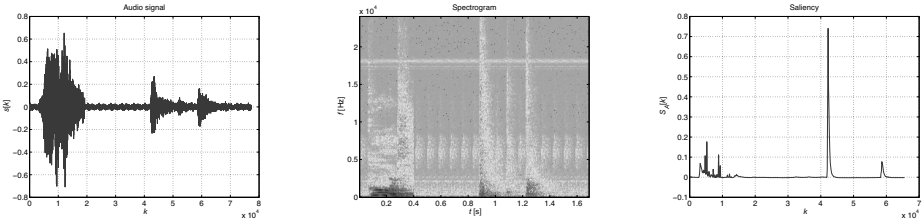

Figure 4 – Acoustic events. Time domain (left), spectrogram (center), saliency measure (right).

$t = 0.225 \pm 0.175$ s whereas the sound of the falling keys are found at $t = 0.866$ s and $t = 1.206$ s which corresponds with the time domain representation on the left. The basis of the spectrogram is the FFT [4] with $N = 512$ samples.

The magnitude of both, the utterance and the sound of the falling keys is maximum at low frequencies and decays with increasing frequency up to half of the sampling frequency at $f = 24$ kHz. At about $f = 18$ kHz is a sound of a constant low-amplitude which corresponds with the residual sound known from the representation in the time domain.

On the right side of Fig.4 the saliency is plotted in the linear scale. For the calculation $N = 512$ was used for the block length and FFT, respectively. $L_o = 128$ was chosen for the offset between data blocks and the calculation of the means and variances is based on $K = 100$ samples for the prior and posterior density functions.

At the position of the utterance and of the sound of the falling keys the saliency $S_A[k]$ is significantly larger than the values at the other positions. $S_A[k]$ shows a set of maxima for the utterance and two distinct maxima for the sound of the falling keys. It is interesting that only the beginning of a salient event is monitored. Therefore the saliency measure decays after the initial peak and drops at around $k = 1000$ samples despite the fact that the utterance is not yet finished. Between the salient events the saliency measure is not zero but keeps a more or less constant low level because of missing novelty. Many reasons can be given for this effect: first, there is some background noise, even if it is low so that it would not be registered by a human listener. Second, the parameters to calculate the saliency measure are based on estimates with a limited accuracy. The residual background noise at $f = 18$ kHz which was visible in the time domain representation and the spectrogram has obviously no influence on the saliency.

In the sequel the influence of corruptions will be investigated. Coloured noise, white noise and a sinusoidal signal will be added to the previously investigated sound signal. The coloured noise is a white noise process shaped by a Butterworth filter [8] of order $n_0 = 6$, the cutoff frequency $f_c = 6$ kHz or $f_c/f_s = 0.125$ and the amplitude $a_n = 0.1$. This results in the signal to noise ratio SNR=3 dB averaged over die two components, utterance and the sound of the falling keys. In Fig.5, again the corrupted signal $s[k]$, the spectrogram $S[k,n]$ and the saliency measure $S_A[k]$ are shown from left to right.

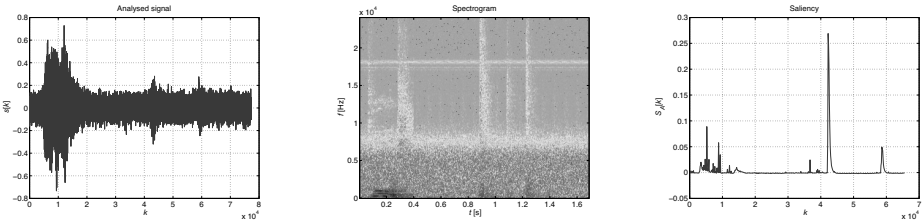

Figure 5 – Acoustic events corrupted by coloured noise. Time domain (left), spectrogram (center), saliency measure (right).

From the recording the noise is clearly audible and also visible in the time and frequency domain, respectively. In the time domain the amplitudes of the utterance rise above those of the coloured noise significantly which is not true for the sound of the falling keys. In the spectrogram in the center of Fig.5 the low frequency components of the utterance and the sound of the falling keys are totally covered by the coloured noise up to the cutoff frequency at $f_c = 6$ kHz. Above this limit the spectrogram is more or less identical with the spectrogram of the uncorrupted signal.

In the saliency measure on the right side of Fig.5 the positions of the utterance and the falling keys are marked at the correct positions. There are no significant differences between the saliency measures of the uncorrupted signal shown in Fig.4 and the one corrupted by coloured noise in Fig.5. In general the magnitudes of the saliency measure for the sounds corrupted by coloured noise are lower than those in case of no corruption. Additionally a very low component is seen at $k = 36,900$ in case of coloured noise whereas this component is not found in the salience measure $S_A[k]$ of the uncorrupted signal. Nevertheless, the Kullback-Leibler divergence shows a high level of robustness with respect to corrupting noise.

The next interferer is white noise with the signal to noise ratio of SNR=-3 dB shown in Fig.6. The sound signal of the falling keys are almost totally covered by the noise. This is also seen in the spectrogram where only the strong low frequency components of the utterance and the falling keys are visible.

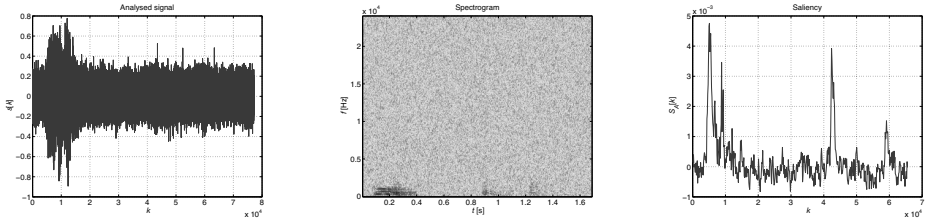

Figure 6 – Acoustic events corrupted by white noise. Time domain (left), spectrogram (center), saliency measure (right).

Despite the strong influence of the white noise the saliency measure on the right of Fig.6 shows clearly the positions of the utterance and the sound of the falling keys along the time axis. In general the amplitudes are much more reduced with respect to those in case of no corruption and coloured nose. The first spike at the beginning of the utterance is very high in comparison with the other spikes and compared to the saliency measures of the uncorrupted sounds and those corrupted by coloured noise. The values of the saliency measure $S_A[k]$ between the two acoustical events are much higher with respect to the spikes indicating salient events in comparison to the cases with no corruption and coloured noise in Fig.4 and Fig.5, respectively. Nevertheless, a threshold could be found to separate the acoustical events from the corrupting white noise.

As the last interferer a sinusoid of frequency $f_0 = 480$ Hz or $f_0/f_s = 0.01$ and amplitude $a_s = 0.75$ is investigated from which a section is shown in Fig.7. This section is positioned between the utterance and the sound of the falling keys as can be seen from Fig.8. The signal to noise ratio is $SNR = -17.5$ dB for the utterance and the falling keys with respect to the corrupting sinusoid. By this the amplitudes of the salient events almost do not rise above the noise amplitudes.

Due to the resolution the frequency of the sinusoid cannot be read from the time domain representation as Fig.8 shows. But in the spectrogram $S[k,n]$ in the center of Fig.8 a sharp spectral line is visible at $f = 480$ Hz which does not change with time. The rest of the spectrogram

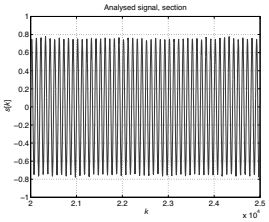

Figure 7 – Section of the corrupting sinusoidal signal.

is identical with the spectrogram of the uncorrupted sound signal in Fig.4.

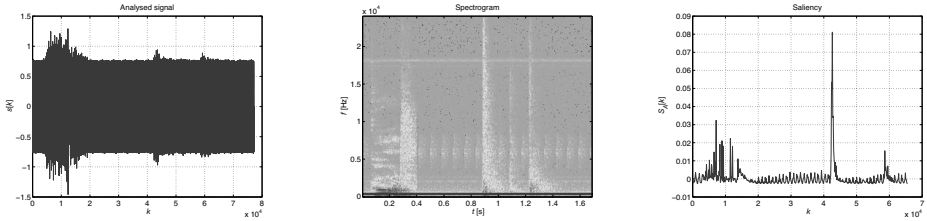

Figure 8 – Acoustic events corrupted by a sinusoid. Time domain (left), spectrogram (center), saliency measure (right).

The amplitudes of the saliency measure on the right side of Fig.8 are lower than those for the uncorrupted signal and the signal corrupted by the coloured noise but larger than those in case of white noise. Furthermore, small amplitudes are visible between the salient events which are caused by the sinusoidal interferer. But still the location of the salient events are clearly visible.

5 Conclusion

In this paper it has been shown that the saliency measure $S_A[k]$ based on the Kullback-Leibler divergence is a useful tool to extract saliency from an acoustic scene. Background noise and sinusoids can be suppressed significantly which is shown by comparing the spectrogram and the saliency diagram $S_A[k]$ which is calculated from the saliency measure $S_A[k,n]$ averaged over the frequency parameter n.

Whereas the spectrogram can be used to identify salient events in the environment of sinusoidal interferers this applies only partially for coloured noise and fails totally with white noise. But in all these cases the saliency measure reliably detects salient events.

It has been shown that besides other parameters the type of corruption influences significantly the salincy measure. For practical applications it would be of interest to reduce this influence. Furthermore the question should be answered whether the magnitude of the saliency measure tells something about the importance of the events generating these magnitudes. That means for the example discussed in this paper whether the utterance is more important than the sound of the falling keys or vice versa.

6 Acknowledgment

The investigation presented in this paper is the result of the cooperation between the Karlsruhe Institute of Technology (KIT) and the Fraunhofer Institute of Optronics, System Technology and Image Exploitation. Whereas at the IOSB the focus was on the exploitation of salient

events by optical means the researchers from the KIT investigated saliency in acoustic scenes.

The author thanks the colleagues from the IOSB for the hospitality, the helpful hints within the progress of the project and the fruitful cooperation on an interesting field of research.

References

[1] ITTI, L, KOCH, C.: *Computational Modelling of Visual Attention*. Nature Reviews, Neuroscience, vol. 2, pp. 194-203, March 2001

[2] ITTI, L., BALDI, P.F.: *Bayesian Surprise Attracts Human Attention*. Advances in Neural Information Processing Systems, 2006

[3] SCHAUERTE, B. et al.: *Multimodal Saliency-based Attention for Object-based Scene Analysis*. 2011 IEEE/RSJ, Intl. Conf. on Intelligent Robots and Systems, pp. 1173-1179, 2011

[4] KROSCHEL, K., RIGOLL, G., SCHULLER, B.: *Statistische Informationstechnik*. Springer, Heidelberg, Dordrecht, London, New York, 2011

[5] KAYSER, C. et al.: *Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map*. Current Biology, vol. 15, pp. 1943-1947, 2005

[6] KULLBACK, S.: *Information Theory and Statistics*. Dover Publications, 1968

[7] HERSHEY, J.R. et.al.: *Approximating the Kullback-Leibler Divergence between Gaussian Mixture Models*. Proc. Interntl. Conf. on Acoustics, Speech and Signal Processing, ICASSP, 2007

[8] KAMMEYER, K.-D., KROSCHEL, K.: *Digitale Signalverarbeitung*. 10th edition, Springer-Vieweg, Wiesbaden, 2022