Normalized Auditory Attention Levels for Automatic Audio Surveillance

L. Couvreur, F. Bettens, J. Hancq & M. Mancas
Department of Signal Processing,
Faculté Polytechnique de Mons, Belgium

Abstract

In this paper, we define features that can be computed along audio signals in order to assess the level of auditory attention on a normalized scale, *i.e.* between 0 and 1. The proposed features are derived from a time-frequency representation of audio signals and highlight salient regions such as regions with high loudness, temporal and frequency contrasts. Normalized auditory attention levels can be used to detect sudden and unexpected changes of audio textures and to focus the attention of a surveillance operator to sound segments of interest in audio streams that are monitored. The proposed algorithms have been tested on audio material consisting of security-relevant audio events (*e.g.*, gun shot, glass breaking, woman scream, siren sound, *etc*) embedded in sound ambiences in public places (*e.g.*, airport hall, metro station, subway train, sport stadium, *etc*).

Keywords: Public security, audio surveillance, normalized auditory attention levels, audio-based saliency levels, audio-based rarity levels.

1 Introduction

Nowadays, public security represents a major challenge for public authorities and a profitable market for private companies. More and more surveillance equipments are deployed and human resources are enlisted in order to monitor and secure public places (*e.g.*, urban zones, mass transportation hotspots, wide commercial malls, large sport or cultural events, massive community demonstrations, *etc*). Such public security is classically achieved by remotely operating numerous video sensors at key locations in the places to be secured and conveying images via network equipments to screen walls in surveillance rooms. In order to enhance the awareness of the surveillance operators, this security system is more and more often completed with sensors of different natures such as infrared and thermal cameras, millimeter-wave and micro-wave radars, infrared, volumetric, seismic intrusion detectors, badge readers and biometric access controllers, microphones as well as security guards on site equipped with radio communication devices. Unfortunately, this also leads to a dramatic increase of information and often results into a cognitive overload of the surveillance operators.

The research presented in this paper has been performed in the framework of the SERKET (SEcuRity KEeps Threats away) project [1]. It concerns audio signal processing algorithms for detecting abnormal audio events. By abnormal, we mean sudden and unexpected sounds (*e.g.*, gun shot, glass breaking, woman scream, siren sound, *etc*) that are embedded in normal acoustic ambiences (*e.g.*, airport hall, metro station, subway train, sport stadium, *etc*) and render security-relevant issues in public places. Unlike for video with screen walls, mosaicing of several audio streams is not possible due to the transparent nature of sounds and surveillance operators cannot listen simultaneously to several audio channels. Therefore, mechanisms for attracting the attention of a listener to segments of interest in audio streams are of primary importance. Besides, detected segments of interests can be formatted into low-level events with start and stop times, abnormality level and nature of event when possible, which can be temporally and spatially correlated with other low-level security-related events in order to infer higher-level threat scenarios.

2 Proposed method

Automatic detection of abnormal sound events can be viewed as a specific problem in Computational Auditory Scene Analysis (CASA) [2]. Because of limited cognitive resources, listeners have to apply some attentional mechanisms in order to focus their auditory attention to salient sounds or more exactly streams in the time-frequency plan. Although visual attention mechanisms are largely studied, only a few computational models have been proposed for auditory attention [3,4,5]. In this paper, we present models and algorithms for computing normalized auditory attention levels, *i.e.* scores between 0 and 1 that measure the level of auditory attention along audio signals, and for detecting abnormal segments of audio activity. Many concepts in these developments are based on similar research in visual attention.

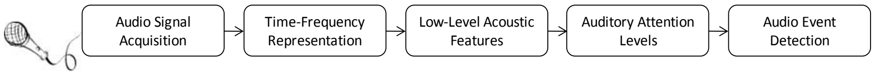
Audio Signal Acquisition
Time-Frequency Representation
Low-Level Acoustic Features
Auditory Attention Levels
Audio Event Detection

Figure 1. Block diagram of algorithm for computing normalized auditory attention levels and detecting salient and rare audio events.

Figure 1 shows the general architecture of the method that is proposed in this research for detecting salient and infrequent audio events. We first compute various acoustic features from a time-frequency representation of the audio signals that highlight segments of interest. These features are then normalized according to several mechanisms such that they can be expressed on a common feature-independent scale, namely between 0 and 1, and consistently combined to detect audio events. The proposed method relies primarily on biologically-motivated models for computing the acoustic features and deriving normalized auditory attention levels. Note also that the method is purely bottom-up and do not require any *a priori* acoustic models of the sounds to be detected.

2.1 Time-frequency representation

The first stage of the time-frequency processing consists in applying a fixed array of bandpass linear filters to the audio signal. This filterbank aims at modelling the frequency analysis of the cochlea in the inner ear. It is here implemented as a bank of Gammatone filters with constant unitary bandwidth on an ERB frequency scale (Equivalent Rectangular Bandwith) [6]. In our case, the audio signals are sampled at 8kHz, which leads to 29 bandpass filters. Next, the outputs of the filters are half-wave rectified and power-law compressed by a factor 0.4 in order to model the behaviour of inner ear hair cells and the nonlinear perception of loudness. Finally, the transduced outputs of the cochlear filterbank are considered into overlapping frames of 100ms length and 50ms shift, and the mean value of every frame and every channel is computed. The resulting coefficients form together the so-called *auditory spectrum* of the signal frame.

2.2 Low-level acoustic features

This stage consists in deriving a set of acoustic features for every frame that may help in attracting the listener attention. The auditory spectrum coefficients are naturally good features. In this research, we study also the usefulness of other features, mainly as defined in the MPEG-7 framework for sound description [22] and in the CUIDADO system [23]. Let denote f_b as the central frequency of the b-th Gammatone filter, $1 < b < B$ ($B = 29$ in our case), and $e_{k,b}$ as its output for the k-th time frame. We define additional acoustic features as follows.

The spectral centroid μ_k, spread σ_k, skewness $\gamma_{1,k}$ and kurtosis $\gamma_{2,k}$ are computed as the sample mean, spread, 3rd order and 4th normalized moments of the auditory spectrum considered as a distribution whose values are the central frequencies and probabilities are the normalized spectrum coefficients, that is,

$$\mu_k = \sum\nolimits_b f_b \frac{e_{k,b}}{e_k} \qquad \text{with} \;\; e_k = \sum\nolimits_b e_{k,b} \tag{1}$$

$$\sigma_k = \sqrt{\sum\nolimits_b (f_b - \mu_k)^2 \frac{e_{k,b}}{e_k}} \tag{2}$$

$$\gamma_{1,k} = \frac{1}{\sigma_k^3} \sum\nolimits_b (f_b - \mu_k)^3 \frac{e_{k,b}}{e_k} \tag{3}$$

$$\gamma_{2,k} = \frac{1}{\sigma_k^4} \sum\nolimits_b (f_b - \mu_k)^4 \frac{e_{k,b}}{e_k}. \tag{4}$$

The spectral slope s_k is obtained by linear regression as the rate of (de)-increase of the auditory spectral envelope, normalized by the total amplitude:

$$s_k = \frac{1}{e_k}\left(B \sum\nolimits_b f_b e_{k,b} - \sum\nolimits_b f_b \sum\nolimits_b e_{k,b}\right) \Big/ \left(B \sum\nolimits_b f_b^2 - \left(\sum\nolimits_b f_b\right)^2\right). \tag{5}$$

The spectral decrease d_k represents the amount of decreasing of the auditory spectral envelop and is computed as follows:

$$d_k = \sum_{b=2}^{B} \frac{e_{k,b} - e_{k,1}}{b-1} \Big/ \sum_{b=2}^{B} e_{k,b} \, . \tag{6}$$

The spectral roll-off r_k is classically defined as the minimum central frequency so that at least 95% of the auditory spectrum is contained below this frequency. It is derived such that:

$$\sum_{b=1}^{r_k} e_{k,b} \geq 0.95 \sum_{b=1}^{B} e_{k,b} \, . \tag{7}$$

The spectral variation v_k, or spectral flux, represents the amount of variation of the auditory spectrum along time and is obtained from the normalized cross-correlation between two successive spectral vectors:

$$v_k = 1 - \sum\nolimits_b e_{k-1,b} e_{k,b} \Big/ \sqrt{\sum\nolimits_b e_{k-1,b}^2} \sqrt{\sum\nolimits_b e_{k,b}^2} \, . \tag{8}$$

The spectral tristimuli $t_{k,1}$, $t_{k,2}$ and $t_{k,3}$ classically measure the ratio of harmonics in sounds. We adapt here this definition as follows:

$$t_{k,1} = \frac{e_{k,1}}{e_k}, \quad t_{k,2} = \frac{\sum_{b=2}^{4} e_{k,b}}{e_k}, \quad t_{k,3} = \frac{\sum_{b=5}^{B} e_{k,b}}{e_k} \, . \tag{9}$$

The spectral flatness l_k and the spectral crest c_k are measures of the noisiness/sinusoidality of the auditory spectrum. They are computed as the ratio between the geometric mean, or the maximum value, and the arithmetic mean of the auditory spectrum, respectively,

$$l_k = \frac{\sqrt[B]{\prod_b e_{k,b}}}{\frac{1}{B}\sum_B e_{k,b}}, \quad c_k = \frac{\max e_{k,b}}{\frac{1}{B}\sum_B e_{k,b}}. \tag{10}$$

Other instantaneous temporal features are also considered, namely, the first 12 normalized auto-correlation coefficients and the zero-crossing rate. Let denote x_n, $n_{0,k} < n < n_{1,k}$, as the audio signal samples for the *k*-th time frame. The *m*-th order auto-correlation coefficient $a_{k,m}$ for the current frame is defined as

$$a_{k,m} = \sum_{n=n_{0,k}}^{n_{k,1-m}} x_n x_{n+m} \,. \tag{11}$$

These coefficients can be efficiently computed with Fast Fourier Transform. Beside, the zero crossing rate z_k for the *k*-th frame is given as

$$z_k = \frac{1}{n_{1,k} - n_{0,k} + 1} \sum_{n=n_{0,k}}^{n_{k,1-1}} I(x_n x_{n+1} < 0)\,, \tag{12}$$

where $I(.)$ stands for the indicator function and is equal to 1 if its argument is true and 0 otherwise.

All these frame-based feature sequences are added with estimated of their 1st and 2nd order time derivatives that are computed by applying classical finite-difference equations, the so-called delta and delta-delta features. For instance, we obtain these estimates for the low level feature y_k as follows:

$$\begin{aligned} \Delta y_k &= -2y_{k-2} - y_{k-1} + y_{k+1} + 2y_{k+2}\,, \\ \Delta\Delta y_k &= 2y_{k-3} + y_{k-2} - 2y_{k-1} - 2y_k - 2y_{k+1} + y_{k+2} + 2y_{k+3}\,. \end{aligned} \tag{13}$$

2.3 Auditory attention levels

The low-level analysis yields a sequence of vectors of 55×3 acoustic features. In the next stage, we would like primarily to express these features in a common range in order to compare and combine them consistently and secondly to enhance salient and rare values. To do so, we compare normalization mechanisms that rely on a basic assumption: listeners assign attention potential to segments with feature values that significantly differ from the audio ambience within some temporal context. For a given frame, every feature value is normalized with respect to values in neighbouring frames. From a physical and biological point of view, only backward context frames should be considered. However, we observed that forward context frames definitively help in normalization. In our work, backward and forward context lengths are set to 10s ($K_B = 200$ frames) and 3s ($K_F = 60$ frames), respectively.

2.3.1 Saliency-based normalization

The first two normalization algorithms that we consider were initially proposed by Itti *et al.* as part of models of computational visual attention. The first algorithm (based on [9]) consists of the following steps for computing the normalized value $N_1(y_k)$ of low level feature y_k at *k*-th frame within the context window $[k - K_B, k + K_F]$:

1. Scale feature values y_l, $k - K_B < l < k + K_F$, in the range $[0,1]$:

$$y_l' = \frac{y_l - \min_{l\in[k-K_B,k+K_F]}(y_l)}{\max_{l\in[k-K_B,k+K_F]}(y_l) - \min_{l\in[k-K_B,k+K_F]}(y_l)} \tag{14}$$

2. Find all local maxima y_{l^*}', $1 < l^* < L_k$, within the context window such that:

$$y_{l^*}' \geq 0.1 \;\cap\; y_{l^*}' \geq y_{l^*-1}' \;\cap\; y_{l^*}' \geq y_{l^*+1}' \tag{15}$$

3. Normalize feature value y_k:

$$N_1(y_k) = y_{K_B+1}'\left(1 - \frac{1}{L_k}\sum_{l^*=1}^{L_k} y_{l^*}'\right)^2 . \tag{16}$$

The second algorithm (based on [10]) consists of the following steps for computing the normalized value $N_2(y_k)$:

1. Scale feature values y_l, $k - K_B < l < k + K_F$, to $y_l^{(0)}$ using Equation (14) and set $i = 0$.
2. Convolve linearly $y_l^{(i)}$ with a Difference-of-Gaussians (DoG) filter with strong local excitation and broad neighbouring inhibition:

$$y_l' = y_l^{(i)} * \left(\frac{\delta_{ex}^2}{2\pi\sigma_{ex}^2}\exp\left(-\frac{l}{2\sigma_{ex}^2}\right) - \frac{\delta_{in}^2}{2\pi\sigma_{in}^2}\exp\left(-\frac{l}{2\sigma_{in}^2}\right)\right) \tag{17}$$

   where $\sigma_{ex} = 2\%$ and $\sigma_{in} = 25\%$ of the context window width and $\delta_{ex} = 0.5$ and $\delta_{in} = 1.5$. Note that the filter is truncated to the context window width for practical implementation.
3. Add the filter output to the scaled feature $y_l^{(i)}$ and perform half-wave rectification:

$$y_l^{(i+1)} = y_l^{(i)} + y_l' - 0.02\max_{l\in[k-K_B,k+K_F]}\left(y_l^{(i)}\right) \tag{18}$$

$$y_l^{(i+1)} = y_l^{(i+1)} I\big(y_l^{(i+1)} > 0\big) \qquad (19)$$

4. Set $i = i + 1$ and got to 2 unless maximum iterations is reached.
5. Get the normalized feature $N_2(y_k) = y_{K_B+1}^{(i)}$.

2.3.2 Rarity-based normalization

Alternatively, we can apply a normalization mechanism that relies on the information theory concept of self-information and enhances feature values in minority within their temporal context. The application of such algorithm to auditory data was originally proposed in [11] and consists of the following steps:

1. Scale features y_l, $k - K_B < l < k + K_F$, to $y_l^{'}$ using Equation (14).
2. Quantify scaled feature values over N_Q levels uniformly defined between extreme values over the context window (here, $N_Q = 8$).
3. Compute probability of occurrence

$$P(y_k^{'}) = \frac{1}{K_B + K_F + 1} \sum_{l=k-K_B}^{k+K_F} I(y_l^{'} = y_k^{'}) \qquad (20)$$

4. Compute contrast function score

$$C(y_k^{'}) = 1 - \left(\frac{\sum_{l=k-K_B}^{k+K_F} |y_l^{'} - y_k^{'}|}{(K_B + K_F + 1) \max_{l \in [k-K_B, k+K_F]} |y_l^{'} - y_k^{'}|} \right) \qquad (21)$$

5. Compute normalized feature

$$N_3(y_k) = \frac{-\log\big(P(y_k^{'}) \cdot C(y_k^{'})\big)}{2 \log(K_B + K_F + 1)}. \qquad (22)$$

2.4 Auditory event detection

In order to derive a detection binary signal (1 for presence of audio event, 0 otherwise) from the normalized attention signals, we adopt a very simple decision scheme that applies Otsu's thresholding [12] for every context window. This method builds a histogram of the values and finds a threshold maximizing the between-class variance. The reference frame for a given context window is assigned a binary score depending whether it is over or below the estimated threshold decision. The detection binary signal is further applied morphological opening and closure operations in order to remove abnormally short detected segments and fill up gaps between potentially correctly related detected segments.

3 Experimental results

The experimental audio material in this research consists of recordings of security-relevant audio events, namely gun shot, glass breaking, woman scream and siren sound that are mixed with recordings of normal acoustic ambiences in public places, namely airport hall, metro station, subway train and sport stadium. In order to assess the robustness of the proposed attention levels, audio events were embedded in audio ambiences at several time locations, actually every 15s along 3 minute recordings, and with various ambience-to-event ratios, namely -10dB, -5dB and 0dB.

As already mentioned, the proposed attention levels were primarily developed for monitoring audio signals. As an example, Figure 2 shows the audio signal of a gunshot event embedded in audio ambience of a metro station together with the cochleogram as obtained by stacking the auditory spectrum vectors and the three normalized auditory attention levels. We clearly observe that the second saliency-based normalization mechanism is potentially more efficient and better isolate the segment of interest. Although the rarity-based mechanism performs well, it suffers from more spurious peaks and noisier activity in absence of events.

Table 1: Detection rates at frame level of auditory detection algorithms for various ambient-to-event ratios (averaged over all ambience and event conditions).

Normalization Method	Detection Rates [%]		
	Precision	Recall	False Alarm Rate
Ambience-to-Event Ratio = -10dB			
N_1	45.5	52.5	10.8
N_2	72.6	70.8	4.6
N_3	71.6	65.0	3.9
Ambience-to-Event Ratio = -5dB			
N_1	43.1	47.5	10.8
N_2	70.2	67.1	4.9
N_3	68.9	58.0	4.5
Ambience-to-Event Ratio = 0dB			
N_1	38.3	44.6	12.4
N_2	63.5	63.1	6.3
N_3	62.3	50.1	5.2

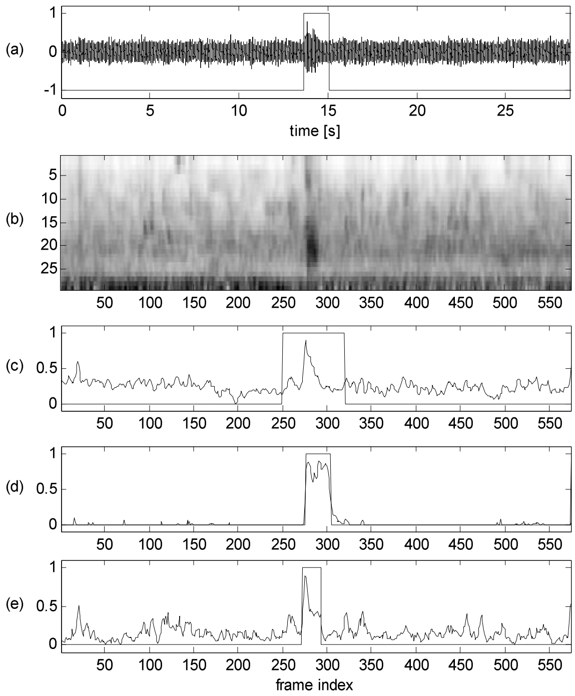

Figure 2. Example of a gunshot event embedded in an audio ambience of a metro station (0dB ambience-to-event ratio): (a) audio signal with reference segmentation, (b) cochleogram, (c)-(d)-(e) normalized attention levels N_1, N_2 and N_3 with detection signals.

Beside monitoring, the auditory attention levels can be used for spotting segments of interest to security operator and a simple detection algorithm was proposed (see section 2.4). In order to assess the detection performance, we estimate the precision and recall metrics at frame level, which are defined as the ratio between the number of frames correctly detected as part of audio events to the total number of detections (including false alarms) and the total number of frames referenced as part of audio events, respectively. We also estimate the false alarm rate as the ratio between the number of incorrectly detected frames to the total number of frames. First, we observe that the saliency-based normalization mechanism N_1 constantly performs worse than the two other methods with significantly lower precision and recall and higher false alarm rate. Next, we note that the N_2 saliency-based method is slightly better that the N_3 rarity-based method with higher precision and recall but at the cost of a higher false alarm rate.

4 Conclusions and perspectives

In this paper, we presented several acoustic features and normalization mechanisms for estimating auditory attention levels along audio signals. The acoustic features rely mostly on the auditory spectrum as computed by a cochlear model. The normalization methods are inspired by visual attention techniques and use the concept of saliency and rarity. Promising results were obtained in terms of both monitoring of audio signals and detection of security-relevant audio events.

The proposed methods are definitively adapted for transient and short acoustic events (*i.e.*, less than a few seconds) and not suitable for detecting long stationary audio activity (*e.g.*, alarm signal of a few minutes). This is coherent with human auditory reaction as the listeners get used to permanent audio stimulus. However, this may be critical in surveillance applications and this situation should be handled. Beside, further improvement can be expected including *a priori* knowledge about the sounds to be detected. This can be performed by estimating statistical models of acoustic features related to these sounds and making them contribute to the derivation of the normalized levels. The drawback of this approach is the need for collecting numerous examples of these events. Finally, we believe that normalization could be improved if applied directly on multi-dimensional features instead of combining the normalized scores obtained for one-dimensional features separately.

References

[1] SERKET-ITEA, http://www.research.thalesgroup.com/software/cognitive_solutions/Serket.

[2] D. L. Wang & G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, John Wiley & Sons, 2006.

[3] Y. F. Ma, L. Lu, H. J. Zhang & M. Li, A User Attention Model for Video Summarization, *Proc. of ACM International Conference on Multimedia*, Juan-les-Pins, France, Dec. 2002.

[4] S. N. Wrigley & G. J. Brown, A Computational Model of Auditory Selective Attention, *IEEE Trans. on Neural Networks*, vol. 15, no. 5, pp. 1151-1163, Sep. 2004.

[5] C. Kayser, C. Petkov, M. Lippert & N. Logothetis, Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map, *Current Biology*, vol. 15, no. 21, pp. 1943-1947, Aug. 2005.

[6] M. Slaney, An Efficient Implementation of the Patterson-Holdsworth Auditory Filterbank, *Apple Computer Technical Report*, no. 35, 1993.

[7] ISO 15938-4, *Multimedia Content Description Interface, Part 4: Audio*, ISO/IEC Standard, 2001 (revised 2004).

[8] G. Peeters, A Large Set of Audio Features for Sound Description, *IRCAM Technical Report*, Apr. 2004.

[9] L. Itti, C. Koch & E. Niebur, A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[10] L. Itti & C. Koch, A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention, *Vision Research*, vol. 40, no. 10-12, pp. 1489-1506, May 2000.

[11] M. Mancas, L. Couvreur, B. Gosselin & B. Macq, Computational Attention for Event Detection, *Proc. of Workshop on Computational Attention and Applications*, Beilefeld, Germany, Mar. 2007.

[12] N. Otsu, A Threshold Selection Method from Gray-Level Histograms, *IEEE Trans. on Sys., Man., Cyber.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.