An Audio Attention Computational Model Based on Spatial Cues Gradient

Bo Hang*, Yi Wang, Changqing Kang

School of Mathematics and Computer Science, Hubei University of Arts and Science, Xiangyang, Hubei, 441053, China

Abstract

Present bottom-up audio attention computational model extracts the underlying characteristics of single channel audio such as energy, pitch, zero crossing rate etc., to calculate the audio signal attention level. In audio surveillance, sound source whose directions change rapidly should have higher attention level. But present audio computational models cannot effectively express the audio attention caused by such signals. It may cause some audio events, which should be paid attention to, cannot be detected. To solve this problem, based on the psychological principles that spatial information affects attention, this paper proposes a model that introducing the short-term spatial gradient cues to measure the attention caused by the fast changing of single audio source space direction. This model calculates the mean short-term changes of the spatial cues vector of sub-bands as spatial cues gradient. Compared to the traditional audio attention computational model, the recall of detection of attention audio events increased 4.5 percentage points in experiments.

Key words: Attention Computational Model, Spatial Cues, Gradient, Surveillance.

1. INTRODUCTION

Audio attention level is the degree of attention to audio object triggered by human auditory sense (Chen, Song, Xue, Chen and Wang, 2015). In our everyday environment, the sound heard by human ears is very complex. There are often many sounds simultaneously produced at the same time. For example, when you sit in an office, you can hear the phone ringing, people dialogue, as well as traffic noise outside the office. Because all around sound signal arriving into the cochlea are in a complex, therefore, we need to analyze incoming sound preliminary to separate the auditory input into perceived objects. In order to select the related information from the complex sound streams, the frequency domain information, temporal information and spatial cues are used to conduct a comprehensive analysis.

Audio attention can be divided into two categories, one is a top-down audio attention, and another is the bottom-up audio attention (Alho, Salmi, Koistinen, Salonen and Rinne, 2015). Top-down attention audio is a goal-driven task-based process that prior knowledge and past experience to learn to focus attention on the scene of the goal (Frintrop and Backer, 2005; Liu and Bengson, 2016). The bottom-up audio attention analysis is a fast, saliency driven process (Gao and Vasconcelos, 2007), which detects salient objects by finding the significant differences between neighborhood parts in audio signals. For example, in the traffic noise on street, gunfire suddenly appeared. The gunfire sound is a salient sound which is different from ambient sounds significantly. The comparison between the two types of attention model demonstrates that the bottom-up attention model response more rapidly than the top-down attention model (Connor, Egeth, and Yantis, 2004). So the former is more suitable for real-time scenario such as surveillances.

Kayser proposed a hearing saliency map (Kayser and Petkov, 2005), extracting contrast characters of intensity, duration and frequency from the Froulir spectrum of audio signal in multi-scale, to construct a bottom-up audio saliency map, in order to build a bottom-up attention model based on basic audio features. Based on the model above, Kalinli and Narayan proposed a bottom-up auditory attention model. This model extracted low-level audio features, including audio signal intensity, frequency contrast, temporal contrast, orientation, and pitch, to construct an audio saliency map (Kalinli and Narayan, 2007). Evangelopoulos proposed that audio events are temporal dynamic changing sound objects, including sound change of existed events and sound of new occurring events. Evangelopoulos extracted the maximum average Teager energy, the average instantaneous amplitude and average instantaneous frequency to create a bottom up audio model. This model was used to detect interesting audio fragments in films generate movie abstract or brows video by combining with video attention model (Evangelopoulos, Rapantzikos, Maragos, Avrithis, and Potamianos, 2008; Evangelopoulos, Rapantzikos, Potamianos, Maragos, Zlatintsi, and Avrithis, 2008; Evangelopoulos, Zlatintsi, Skoumas, Rapantzikos, Potamianos, Maragos, and Avrithis, 2009). Zheng extracted audio features including short-term average energy, pitch, the average zero-crossing rate in his audio attention computational model, to represent the strength of sounds, the sharpness of speeches and the degree of the urgency of audio (Zheng and Zhu, 2008).
Bottom-up attention models are based on stimulus driven attention, caused by the dynamic changes of sound objects, including change of existing sounds and appearing of new sound (Kalinli and Narayanan, 2008). The level of attention is determined by the saliency of Sound changes. Sound changes will bring a series of changes of audio features. So we can represent the saliency of audio signal change by the change of audio features. These features are multidimensional features, including time-domain features, frequency domain features and spatial features. In audio attention computational model, the changes of multi-dimensional audio features are calculated to obtain an integrated audio attention level.

Current audio attention model extracted time domain features parameters (Cai, Lu and Zhung, 2003; Ma, Hua, 2005; Liu and Li, 2007; Zheng and Zhu, 2008; Ma, Hua, 2005; Liu and Li, 2007; Zheng and Zhu, 2008;) and frequency domain parameters (Kalinli and Narayanan, 2008; Evangelopoulos and Zlatintsi, 2009; Kalinli and Narayanan, 2009; Kalinli, 2011), but did not consider spatial information. Thus the current audio attention computational model can not represent the attention mechanism of the human brain on spatial change of the audio objects.

This paper discusses audio features when a sound source position change rapidly, and how to extract the relevant parameters of audio features to measure the quickly change of a sound source position, and then calculate the attention level.

2. AUDIO ATTENTION COMPUTATIONAL MODEL BASED ON SPATIAL CUES VARIATION GRADIENT

2.1. Principle analysis

In order to calculate and analyze audio attention level more accurately, spatial cues are introduced to resolve the problem that the current audio attention computational models do not consider the attention caused by spatial direction of the audio signal changes, create a spatial cues based audio attention computational model.

Location variations of audio image will reflect the changes of spatial cues (Beack and Seo, 2007). And the change is usually a short-term continuous smooth process, shown in Figure 1. Then spatial cues changes in a short time can be extracted to measure the rate of change of the sound source direction. Then audio attention level caused by rapid movement of the sound image can be calculated.

![Figure 1. Continuous trend of different sub-band spatial cues in the time domain](image)

Taking the spatial cues ILD (Inter-aural Level Difference) for instance, by using ILD variation, the audio attention level is calculated when an audio object rotates around the listener in the horizontal plane.

ILD is difference in sound pressure level reaching the two ears (Song and Oh, 2011). In the case of single sound source, if the horizontal orientation angle changes rapidly, the difference of the two ILDs of two time points, of which the time gap is short, will be large. This will result in a high level of attention. As shown in figure 2, when a sound source is moving rapidly from left to right, there are a set of sub-bands ILD values of a previous time point, and a set of sub-bands ILD values of the following time point, and the mean of the differences between these two sets of ILD value.

In the case of the rapid changes of the orientation angle of a sound source, the spatial cues gradients are used to indicate rapid changes in the orientation angle of the sound source. In the real-valued function of a single variable, the gradient is the derivative. For a linear function, the gradient is the slope of the line. As shown in Figure 3, the spatial cue values of the same sub-band in different times constitute a curve. The derivative of the curve at the current time is the slope G (k) of the sub-band spatial cues of the current frame at the current time.
To simplify the slope calculation, we give an approximated solution. Firstly, we calculate the difference of short term sub-band spatial cues. Then the ratio $D(k)$ between the difference and the time interval is calculated as the approximated value of the slope. In this case, the slope value depends on the spatial cue difference and the time interval. So, given a constant time interval, we can use the spatial cue difference to present the slope directly. Since a frame signal is divided into several sub-bands, we can use the mean of all sub-bands spatial cues slope values of a frame to measure current frame temporal gradient, which is the level of short term spatial cue variation. In summary, the mean of short term sub-bands spatial cues variation is used to represent the level of sound source direction change saliency.

$$ILD_{Frame(t)} = G(k)$$

$$D(k) = \frac{ILD_{k} - ILD_{k-1}}{t}$$

$$s_i = 10\log \frac{I_{il}}{I_{ir}} \quad i \in [1,N] \quad (1)$$

Figure 3. A sub-band temporal variation curve and its slope

2.2 Computational method

According to the attention computational model based on spatial audio cues temporal variation curve gradient described above, we compute the mean of the spatial audio cues change to get the attention level caused by a sound source direction change. The detail steps are shown as followed.

Assuming that the signal length of each frame in audio stream is $t_f$, the audio signal of current frame is acquired firstly and transformed from time domain to frequency domain secondly. After transformation, the frequency domain signal is divided into $N$ sub bands. The sub band division is according to the critical frequency band division method (Zwicker, 1961; Moore and Glasberg, 1983). And the frequency domain signal is divided into 21 sub-bands, as shown in table 1.

For stereo signals, the left and right channels are divided into 21 sub-bands respectively.

Calculate spatial audio cues of all sub bands in the current frame. According to the ILD calculation, we can get the ILD value $s_i$ of the sub-band $i$. In the formula (1), $I_{il}$ and $I_{ir}$ are the left and right channels energy values of the sub-band $i$ in the stereo signals:

$$s_i = 10\log \frac{I_{il}}{I_{ir}}$$

The current frame is frame $k$. And there are $N$ sub-bands in a frame. A $N$-dimensional vector $S_k = \{s_1, s_2, \cdots, s_{N-1}, s_N\}$ of the current frame can be obtained.
### Table 1. Sub-band division method.

<table>
<thead>
<tr>
<th>No. of sub-band</th>
<th>Lower freq. limit (Hz)</th>
<th>Upper freq. limit (Hz)</th>
<th>Bandwidth</th>
<th>No. of sub-band</th>
<th>Lower freq. limit (Hz)</th>
<th>Upper freq. limit (Hz)</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>100</td>
<td>50</td>
<td>12</td>
<td>1480</td>
<td>1720</td>
<td>240</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>200</td>
<td>100</td>
<td>13</td>
<td>1720</td>
<td>2000</td>
<td>280</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>300</td>
<td>100</td>
<td>14</td>
<td>2000</td>
<td>2320</td>
<td>320</td>
</tr>
<tr>
<td>4</td>
<td>300</td>
<td>400</td>
<td>100</td>
<td>15</td>
<td>2320</td>
<td>2700</td>
<td>380</td>
</tr>
<tr>
<td>5</td>
<td>400</td>
<td>510</td>
<td>110</td>
<td>16</td>
<td>2700</td>
<td>3150</td>
<td>450</td>
</tr>
<tr>
<td>6</td>
<td>510</td>
<td>630</td>
<td>120</td>
<td>17</td>
<td>3150</td>
<td>3700</td>
<td>550</td>
</tr>
<tr>
<td>7</td>
<td>630</td>
<td>770</td>
<td>140</td>
<td>18</td>
<td>3700</td>
<td>4400</td>
<td>700</td>
</tr>
<tr>
<td>8</td>
<td>770</td>
<td>920</td>
<td>150</td>
<td>19</td>
<td>4400</td>
<td>5300</td>
<td>900</td>
</tr>
<tr>
<td>9</td>
<td>920</td>
<td>1080</td>
<td>160</td>
<td>20</td>
<td>5300</td>
<td>6400</td>
<td>1100</td>
</tr>
<tr>
<td>10</td>
<td>1080</td>
<td>1270</td>
<td>190</td>
<td>21</td>
<td>6400</td>
<td>8000</td>
<td>1600</td>
</tr>
<tr>
<td>11</td>
<td>1270</td>
<td>1480</td>
<td>210</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\Delta T$ is the duration of several frames in a short term, for example 100ms. The current signal spatial audio cues vector difference $D_k = [d_1, d_2, \ldots, d_{N-1}, d_N]$ is computed via (2).

$$D_k = S_k - S_{k-\Delta T}$$  \hspace{1cm} (2)

Then, the mean $\mu_k$ of spatial audio cues vector difference $D_k$ is computed as shown in formula (3):

$$\mu_k = \frac{1}{N} \sum_{i=1}^{N} d_i$$  \hspace{1cm} (3)

$\mu_k$ could be positive or negative, depending on the position or moving direction of the sound object. But the attention level must be positive value. So we can get the audio attention level $a_k$ via the absolute value of spatial audio cues gradient. Besides that, High energy signals should also be pay more attention. And if the signal energy is too low it can not cause attention. Therefore the energy of current frame should be introduced into the attention computational model.

$$a_k = |\mu_k| \times E$$  \hspace{1cm} (4)

#### 2.3 Normalization

Next step is normalization. There are $n$ non-attention frame with attention level values $a_1, a_2, \ldots, a_{n-1}, a_n$, saved in a non-attention queue $Q_b$, which is the background queue, as shown in figure (4):

![Non-attention (background) queue](image)

Figure 4. Non-attention (background) queue

Given $A_n = \{a_1, a_2, \ldots, a_{n-1}, a_n\}$, the mean $\mu_{Ak}$ of the set $A_n$ is calculated as followed:

$$\mu_{Ak} = \frac{1}{n} \sum_{i=1}^{n} a_i$$  \hspace{1cm} (5)

$\sigma_{Ak}$ is maximum value among the background queue $Q_b$ and current frame $a_k$.

$$\sigma_{Ak} = \text{Max}(a_k - \mu_{Ak}), a_k \in (A_n \cup \{a_k\})$$  \hspace{1cm} (6)

To eliminate the background sound affect, the normalized attention level $M_k$ of current frame $k$ is computed as equation (7):

$$M_k = \text{Max} \left( \frac{a_k - \mu_{Ak}}{\sigma_{Ak}}, 0 \right)$$  \hspace{1cm} (7)

Given a threshold $M$, if $M_k \geq M$, the frame $k$ is an attention frame; if $M_k < M$, the frame $k$ is a non-attention frame and will be put into the queue $Q_b$ to update the background queue for next frame attention level computation.
3. EXPERIMENT RESULTS

Purpose: The purpose of this experiment is to test the effectiveness and performance of the proposed audio attention computational model based on spatial audio cues changing gradient.

Principle: The experiment principle is that we use the proposed attention computational model and comparing model to compute the attention level curve of audio signal, and label the attention segments and non-attention segments to verify the performance of proposed model in this paper. The synthetic audio sequences are synthesized by foreground and background signal with different ratio.

Experimental materials: there are two categories experimental audio sequences. One is audio signal synthesized by foreground and background audio based on ITU-T P.56 specification. These sequences are used to verify the effectiveness of the spatial cues based attention model for the single sound source rapid oriental variation attention audio signal detection, and select the appropriate threshold. Environmental sounds include three categories: noisy environment, relative noisy environment, quiet environment. Foreground sounds include speech (normal speech and screaming) and non-speech (sound of trains, motorcycles, barking dogs, etc.). Another category is audio segments from classic movies, including all kinds of sounds, with or without rapid direction change.

Test sequences standard: foreground signals level is adjusted to -26dBov according to the RMS method described in the ITU-T P.56. In addition, according to same method the level of background signal is adjusted to 15, 5 and 0 dB based on signal noise ratio (SNR). Then the foreground and background signals were mixed to obtain audio sequences for test, which can be used for evaluating the effectiveness of the audio attention computational model proposed in this paper. As shown in Table 2.

<table>
<thead>
<tr>
<th>Environment background (SNR)</th>
<th>Durance (s)</th>
<th>Events No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy (0dB)</td>
<td>60</td>
<td>10</td>
</tr>
<tr>
<td>Normal (5dB)</td>
<td>60</td>
<td>10</td>
</tr>
<tr>
<td>Quiet (15dB)</td>
<td>60</td>
<td>10</td>
</tr>
</tbody>
</table>

Experiment Methods: Firstly, the attention audio events segments in audio stream are artificially marked. Then the audio attention level values of all frames are calculated to get the audio attention curve. Different attention threshold values are selected for automatic detection of audio events. At last, recall and precision under different threshold are calculated. Recall and precision are defined respectively as following:

Recall rate = (total number of events detected correctly / total number of events) * 100%
Precision = (total number of events detected correctly / total number of events detected) * 100%

In the experiments, the frame length is \( t_f = 20\text{ms} \).

Experiment results: For different threshold values, we get different recall and precision values. The results are shown in Table 3.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Quiet Recall</th>
<th>Precision</th>
<th>Normal Recall</th>
<th>Precision</th>
<th>Noisy Recall</th>
<th>Precision</th>
<th>Average Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>43.48%</td>
<td>100.00%</td>
<td>34.48%</td>
<td>100.00%</td>
<td>59.32%</td>
</tr>
<tr>
<td>0.4</td>
<td>100.00%</td>
<td>83.33%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>90.00%</td>
<td>100.00%</td>
<td>91.41%</td>
</tr>
<tr>
<td>0.5</td>
<td>100.00%</td>
<td>100.00%</td>
<td>90.00%</td>
<td>100.00%</td>
<td>80.00%</td>
<td>100.00%</td>
<td>90.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>0.6</td>
<td>80.00%</td>
<td>100.00%</td>
<td>80.00%</td>
<td>100.00%</td>
<td>30.00%</td>
<td>100.00%</td>
<td>63.33%</td>
<td>100.00%</td>
</tr>
<tr>
<td>0.7</td>
<td>60.00%</td>
<td>100.00%</td>
<td>50.00%</td>
<td>100.00%</td>
<td>30.00%</td>
<td>100.00%</td>
<td>46.67%</td>
<td>100.00%</td>
</tr>
<tr>
<td>0.8</td>
<td>50.00%</td>
<td>100.00%</td>
<td>40.00%</td>
<td>100.00%</td>
<td>20.00%</td>
<td>100.00%</td>
<td>36.67%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

We can see the trends of average recall and average precision, as shown in figure 5.

![Figure 5. Trends of average recall and average precision: the blue line with diamond point is the trend of average recall; the pink line with square point is the trend of average precision](image-url)
It can be seen from table 3 that when the threshold is low, under different ambient noise, the precision decreased with ambient noise increasing. At the same time, recall did not change significantly. When the
threshold is high, under different ambient noise, the recall decreased with ambient noise increasing. At the same
time, precision did not change significantly. And as shown in figure 5, the average recall rate decreased with the
threshold value increasing. At the same time, the average precision increased up to 100%. According to the
statistics, when the threshold is 0.4-0.5, both recall and precision values are high. So we select 0.5 as the
threshold of the attention computational model.

Comparison experiments: We subtract some segments from movie, and analysis these segments via the
attention model of this paper and literature (Zheng and Zhu, 2008), as shown in figure 6 and figure 7.

From the experiments above, we can obtain the statistic results as following:

<table>
<thead>
<tr>
<th>Table4. Comparison test results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparing model</td>
</tr>
<tr>
<td>No. of correct detection in sequence1</td>
</tr>
<tr>
<td>No. of correct detection in sequence2</td>
</tr>
<tr>
<td>Total No. of correct detection</td>
</tr>
<tr>
<td>Recall</td>
</tr>
<tr>
<td>No. of wrong detection in sequence1</td>
</tr>
<tr>
<td>No. of wrong detection in sequence2</td>
</tr>
<tr>
<td>Total No. of wrong detection</td>
</tr>
<tr>
<td>Precision</td>
</tr>
</tbody>
</table>

As can be seen from the table 4, the proposed method in this paper can effectively detect attention audio
signals caused by fast-moving of single sound source. The proposed model improved the recall by 4.5
percentage points comparing with the traditional model. There are some problems, which will be our future
work. Firstly, when the noise is relatively large the position of a single sound source rapid change sometimes
missed. In addition, if there are multi-sound-sources simultaneously, such as applause, missing detection could
happen because surface sound source has little effect on the spatial cues changes gradient. Finally, the precision
ratio declined. Through analyzing the wrong detection signals, we believe that this is due to the complicated and
non-stationary background sounds. When there is signal with low energy level but high energy difference in two
channels, the listener and computational model will give different judgment.

4. CONCLUSIONS

An audio attention computational module for single sound source rapid moving is proposed. Spatial cues
short term gradient is introduced to measure the attention caused by rapid single sound source moving. The
experiment results showed that the audio attention model proposed can effectively detect the sound source
changes salient events. Compared with conventional mono audio attention model, the average recall of attention
audio detection is improved. But the precision declines, which needs further study.

Acknowledgements

This work was supported by National Nature Science Foundation of China (61201247) and Nature
Science Foundation of Hubei Province (2011CDB322).

REFERENCES

Alho Kimmo, Juha Salmi, Sonja Koistinen, Oili Salonen, and Teemu Rinne (2015) "Top-down Controlled and
Bottom-up Triggered Orienting of Auditory Attention to Pitch Activate Overlapping Brain
Networks", Brain research, 1626, pp. 136-145.
Beack, Seungkwon, Jeongil Seo, Taejin Lee, and Dae-young Jang. (2007) "Spatial Cue Based Sound Scene
Control for MPEG Surround", Proc.of IEEE International Conference onMultimedia and Expo., pp. 1886-
1889.
Evangelopoulos, Georgios, Athanasia Zlatintsi, Georgios Skoumas, Konstantinos Rapantzikos, Alexandros


