Liu *et al. Hum. Cent. Comput. Inf. Sci.* (2019) 9:7
https://doi.org/10.1186/s13673-019-0166-9

Human-centric Computing and Information Sciences

RESEARCH Open Access

An audio attention computational model based on information entropy of two channels and exponential moving average

Yu Liu[1], Cong Zhang[1*], Bo Hang[2], Song Wang[1] and Han-Chieh Chao[1]

*Correspondence:
hb_wh_zc@163.com
[1] School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan, China
Full list of author information is available at the end of the article

Abstract

The main down-top attention model usually uses some characteristics of audio signal to extract the auditory saliency map at present. But existing audio attention computational model based image saliency mostly doesn't consider the continuity and attenuation mechanism of the human brain on paying attention to some occurred events in our real environment. To address these issues, we propose a model based on information entropy of two channels and exponential moving average (EMA) to simulate the processing of sound signals from the basilar membrane to the cochlear nucleus, and get the local information entropy of image and audio channel, finally we use the relevant EMA methods to calculate the auditory attention map. Some experimental results on artificial audio signal and real-world audio signal belonging to a public corpus show that the proposed model not only can detect the attention event, but also reflect the mechanism of human brain, and the experimental results of talk show in real environment are better.

Keywords: Audio, Attention computational model, Information entropy, Exponential moving average

Introduction

Audio signal as the most important medium of human communication contains lots of information, so how to automatically detect some important area is the subject of attention by researchers from different fields. Different audio signal heard by human ear may cause our different level attention, and the processing of sounds heard by human ear is a series of complicating actions which are carried out to filter some information and the important content will be noticed by us in our daily environment. Both involuntary perception and voluntary perception are two modes of hearing the sound, the voluntary perception as a purposeful physiological behaviour based on prior knowledge and previous experience can help us clearly hear the sound in a very noisy environment, and the involuntary perception relies on the external environment and directly caused by a difference sound from the surrounding sound will give rise to pay more attention to the different sound in our environment. If we walk in a quiet park, a loud voice will immediately attract our attention. With no doubts, audio attention computational model used to

extract the concerned areas of the signal will reduce the workload of processing signals, and has practical value for the artificial intelligence audio field.

The interest area of audio signal can be determined by the degree of the audio attention, and there are two types of audio attention models commonly used: top-down audio attention model and bottom-up audio attention model [1]. Top-down attention audio is a goal-driven task-based process which prior knowledge and past experience will learn to focus attention on the scene of the goal [2]. Bottom-up model built on character of audio signal is a fast frequently-used detection model, which main method is to calculate the degree of the audio attention by analysing some audio characteristics, and the bottom-up model is more suitable for engineering applications [3]. The study of audio attention was inspired by a summarization of video. In the early 1998, the ideas of extracting the primary visual features were introduced to help to produce video summarization for the first time [4]. And then a hearing saliency map based Itti's method was proposed by Kayser to extract contrast characters of intensity, duration and frequency from the Fourier spectrum of audio signal in multi-scale, and construct a bottom-up audio saliency map [5]. After this, more and more researchers have begun research on the field of audio attention.

Rencently, audio attention model based on image saliency algorithm has become mainstream and usually extracts time domain features and frequency domain parameters to get an image attention map [6], but did not consider the mechanism of the human brain on paying attention to some occurred events in our real environment. With time going by, a high degree of our attention to one occurred event will gradually decrease, and when this event finished we will not immediately lose this attention. Due to our continuity and attenuation characteristics of attention, the short interval between two events can easily absorb our attention in the time dimension and our attention is also constantly changing. For example, we not only pay attention to a single word of a sentence, but also pay attention to the interval between each word. Thus, in our real environment, the current audio attention model can't accurately represent the attention mechanism (continuous characteristics and attenuation characteristics).

The purpose of this paper is to explore a better computing model for auditory system based on fewer characters.The calculation of attention is a very complicated process, which needs to consider the complexity of the algorithm and the auditory characteristics associated with the human auditory system. In order to get a more accurate and faster model, we have carried out a lot of related work. The first thing we did in the research was to reduce the complexity of the algorithm, so we used the information entropy correlation processing method to performance the information, and the last thing we did was that we used statistical correlation methods to reflect the relevant characteristics of the audio signal. Firstly, we expect to get satisfactory results only by processing the audio channel, but the accuracy and adaptability of this audio channel need to be improved. We considered that the image processing methods are widely used in audio signal processing field. we finally introduced the image channel and then combined the results from the audio channel and the image channel. A low computational complexity calculation model was proposed based on information entropy of two channels and EMA in this paper. Using different train sample sets, the simulation experiment result demonstrates that this model not only can detect the attention event, but also reflect

the mechanism of human brain, and we also found that the attention value changes smoother than Kayser calculation model in the time dimension.

The remainder of the article is organized as follows: "Related work" section is the literature survey, investigating the current development of this topic. "Mechanism of audio attention" section presents some basic principles of this paper. In "Computational method" section, the proposed methodology is described in detail. In "Experiment and result analysis" section, the experiment will be conducted. Finally, in "Conclusion and future work" section, we summarize our study and conclude the paper and directions for future work.

Related work

Attention is a concept of neurobiology. As early as 1890, psychologist James first proposed the theory of human attention and laid the theoretical foundation of the human attention [7]. Since then, many researchers have conducted relevant research on the attention model from the perspective of visual attention on the basis of psychological theory. In the early 1998, Itti of the University of Southern California made a pioneering study on the selection and transfer mechanism of visual attention and proposed a visual attention model [4]. In this model, a set of underlying features (such as color, brightness and orientation) are extracted to calculate the saliency maps of each feature, and then the combined saliency map can reflect the area with a large degree of attention is used as the area of interest.

At the beginning, the research on the audio attention model is mainly about the bottom-up calculation model, which calculation usually extract the time domain characteristic parameters of the audio signal and finally merge the saliency maps of each feature. Cai et al. [8] constructed a wonderful audio attention model for TV programs, which calculates the likelihood degree of laughter, applause and cheer in the audio signal, and then measures the product of likelihood and short-term average energy. In order to improve the accuracy of the attention calculation model, some researchers usually extract more types of features from the audio signal. Ma et al. [9] draw attention curves by short-term average energy and short-time energy peaks, and Liu builds an audio attention model by weighting the short-term average energy and the proportion of peak energy [10].

Another attempt to improve the accuracy for detecting the interest area of audio signal is through frequency domain feature. Kayser proposed an audio attention computational model based on extracting contrast characters of intensity, duration and frequency from the Flourier spectrum in multi-scale to construct a bottom-up audio saliency map [5]. 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 [11]. Wang divided the audio signal into sub-frames to calculate the short-term energy, then, basing on the characteristics of auditory stream, energy spectrum of each channel was time-domain filtered in different scales through Gaussian filter groups and the auditory saliency map finally was achieved by linear combination of each frequency channel saliency [12].

There are many models of detecting audio attention referred to the image saliency algorithm [12–14], and some researchers also have explored other audio attention

models based on the top-down models in recent years. In 2016, a significant amount of research based sparse dictionary has been proposed in this top-down model, which highlighted the structure characteristics of noisy acoustic signal, and the attention map could achieve selective attention for certain signal [15]. At the same time, Hang constructed an attention degree calculation model based on spatial clue time domain gradient, which solved the attention detection under the fast changing sound source [6]. And then, Lv proposed a bottom-up and top-down combined auditory attention degree calculation model based on sound source azimuth characteristics and neural network [16].

Mechanism of audio attention

The structure and function of the human air

The sound we heard is processed by multiple complex organs in the human ear. The sound waves collected by the auricle and transmit to the middle ear through the external auditory canal, finally this waves can cause the vibration of the tympanic membrane. These ossicular chains conduct sound from the tympanic membrane to the inner ear, which has been known as the labyrinth. The different positions of the basilar membrane in the cochlea resonate with multi-frequency sound wave. Thousands of hair cells sense the mechanical vibration and convert vibration to electrical signals which are communicated via neurotransmitters to many thousands of nerve cells.

The cochlea is a core transducer component for sound. Anatomical studies have shown that the cochlea has many important tissues, including the scale vestibule, basilar membrane, organ of Corti, and auditory nerve. The different frequencies wave of sound brings out the maxima amplitude on the basilar membrane where the different positions correspond to different optimal frequencies. when the electric potential can saturate when the stimulation intensity reaches a certain level, the inner hair cells show a phenomenon of half-wave rectification to the stimuli. In addition, the cell membrane has a low-pass filtering characteristic which phenomenon is that the electric potential causes the internal potential AC decreases with the frequency increased [17]. The processing of auditory peripheral simulates the characteristics of the basilar membrane and the inner hair cells aims to implement some pre-treatments: audio frequency allotment, half-wave rectification and audio non-linear frequency compression.

The character of bottom-up attention model

The mathematical concept of the audio attention is how much degree of human attention on an area of a sound calculated by some model. Audio attention model can be divided into two principles, one is a top-down model, and another is the bottom-up model. Bottom-up is a saliency-driven approach that relies on a series of characters of sound signal and can make us unconsciously focus on a high-pitched or high-loud part of a sound in a particular scene, and top-down based on prior empirical conditions is a purposeful task-dependent mechanism which can make us clearly hear the other party's talking at a noisy cocktail party.

Bottom-up audio attention directly caused by the environment is the most suitable model for developing real-time process system [3]. There are many environmental factors to do with our ear and many factors might attract us in our actual environment. For example, a loud or fast-paced sound can easily attract us in a quiet environment, and we

can also pay attention to a fast-moving sound in a noisy environment, which cause intensity and time of our binaural sound are difference. At present, the area of audio attention can be got by simulating the characteristic of auditory bottom-up attention, including the short-term average energy, average zero-crossing rate, binaural intensity difference (ILD), and binaural time difference (ITD) of speech [18, 19]. In our real world, some features of a sudden sound can immediately attract our attention for a while, and this process is shown in Fig. 1. The auditory system of us usually has a high degree of attention to the events that occurred a while ago, and the degree of our attention will gradually decrease as we gradually adapt to the environment. At the same time, the short interval between events can also attract us, because there has a certain relationship between two events in the time dimension. So, we find that the audio attention to a certain audio signal usually has two characteristics, one is attenuation and another one is continuity.

Computational method

The framework of calculation method for attention

The proposed computational method based on information entropy of two channels and exponential moving average (EMA) correctly simulate the trigger mechanism of the human auditory system and the persistence of the attention process. Information entropy is an important concept used to represent the average rate at which information are produced in information theory. It has a large value of audio information entropy that an audio signal has the prominent feature: including a high short-term average energy, high average zero-crossing rate, high binaural intensity difference, or short binaural time difference. Compared with other models use many features of audio signal, this attention model based information entropy can avoid the complexity of multi-feature calculation, and has a good detection accuracy through relevant experimental verification in this paper. The block diagram of the calculation model in this paper is shown in Fig. 2. We first use a set of Gammatone filters and Meddis inner hair cell model to perform auditory peripheral process on audio signals. And then we set two channels to get the local entropy, the image channel based image saliency method firstly gets spectrum map which

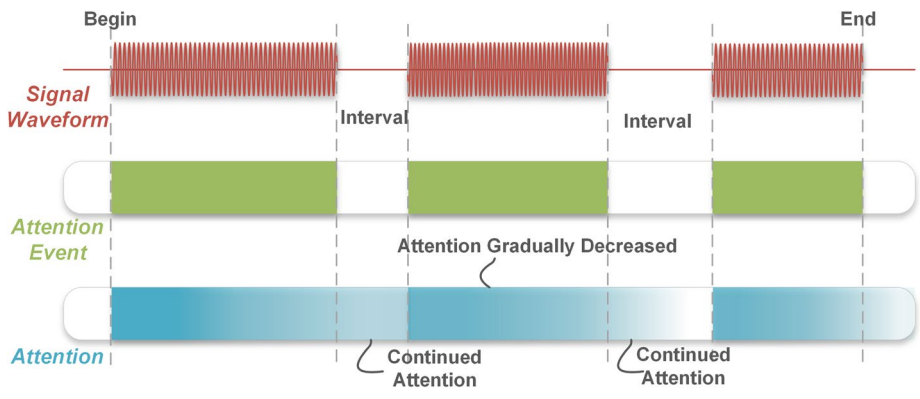

Fig. 1 An illustration of the audio attention processing. The signal waveform marked in red has three-stage signal that represents three attention events triggered by some environment factors, and the begin of three events marked in green can immediately cause high value of attention, but this process will gradually decrease. on the other hand, the intervals between some events can also attract us a period of time, the change in attention values marked in blue

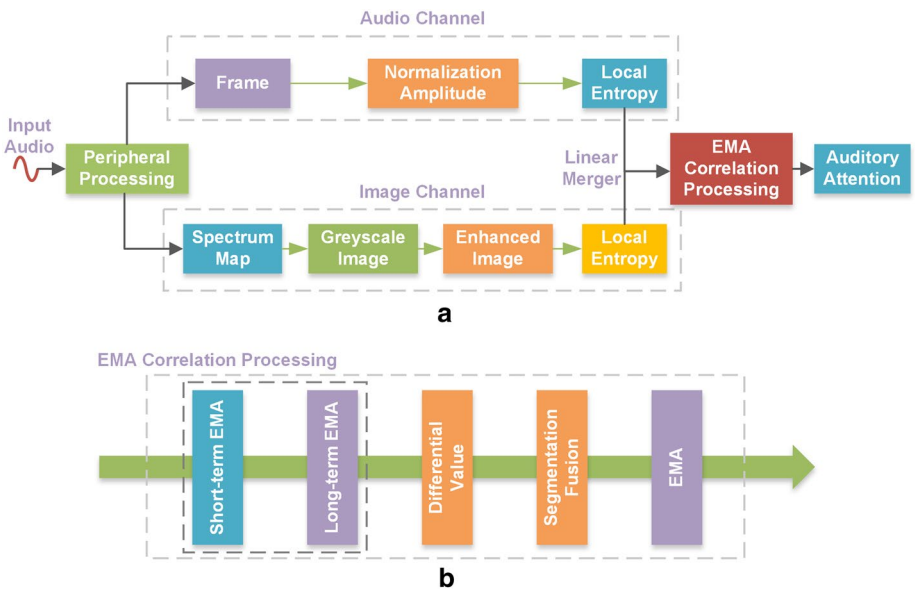

Fig. 2 Schematic illustration of the audio attention computational model. **a** is the framework of this model, and the calculated local information entropy of two channels is linearly merged into one-dimensional vector and then process by EMA correlation processing. The model utilizes two channels to extract more valid information about auditory attention, and the EMA correlation processing represents in **b** can reflect the time dimension characteristics of audio signal

contains the frequency, time and loudness information, owing to some information of the initial spectrum map isn't clear, the image should be converted to grayscale image and enhanced to get a clear greyscale image which can be processed by correlation calculation to obtain local entropy of image channel. In order to get an effective algorithm about audio attention, the other calculation channel is the audio calculation channel, we firstly frame the speech signal and then calculate the local information entropy value of each frame. The local entropy obtained through the audio channel and the image channel should be linearly merged, and the auditory attention can be calculated by an EMA model, which can make full use of the characteristics of the audio signal to reflect the sustainable and attenuation features of the human attention mechanism in time dimension. Finally, we can determine the area-of-interest of audio signal by the degree of attention.

The computational model of auditory periphery

In order to simulate the process of sound signals from the basilar membrane to the cochlear nucleus, the frequency selectivity of the basilar membrane is simulated by a set of bandpass filter banks to extract some sound parameters. At the same time, we simulate the generation and transportation of neurotransmitters in the hair cell-auditory nerve fiber fissures through the inner hair cell and auditory nerve protrusion model. The computational model of auditory periphery realizes the splitting of the audio signal according to different center frequencies, half-wave rectification and non-linear compression. At present, the bandpass filter group for auditory peripheral processing mainly include Mel filter group and Gammatone (GT) filter bank , and the Mel Frequency Cepstral Coefficients (MFCCs) is widely used but MFCC is sensitive to noise and has poor

noise resistance. Some researchers have proved the superior noise immunity of the GT filter banks using the GT filter banks instead of the Mel filter banks, which has a certain inhibitory effect on Gaussian white noise and additive background noise [20, 21]. Finally, the next model of auditory periphery used the Meddis model [22, 23] to describe our inner hair cells and auditory nerve projection.

The Gammatone filter bank to simulate basilar membrane

The cochlear-like map obtained by the GT filter banks is compared with the ordinary spectral map, the low-frequency resolution of cochlear-like map is better than the high-frequency resolution of it [24]. The impulse response of the GT filter can be considered as a Gamma function multiplied by a cosine signal, and the time impulse response formula is shown in formula (1). The value of N is the number of the filter, the parameter t is the time of audio, n denotes the center frequency of the filter, ϕ is the starting phase, α is the order of the filters, A is a constant. The equivalent rectangular bandwidth (ERB) is a psychoacoustic measure of the bandwidth of the auditory filter at each point along the cochlea, in the case of $n = 4$ and $b = 1.1019$ times, the ERB can represent the human auditory filter [25]. For convenience, we set $A = 1$ and $\phi = 0$ [26].

$$\begin{aligned} g(n,t) &= At^{\alpha-1}e^{-2\pi Bt}\cos(2\pi f_n + \phi)u(t), \quad t \geq 0, 1 \leq n \leq N \\ B &= b \times ERB(f_n) \\ ERB(f_n) &= 24.7 \times \left(4.37 \times \frac{f_n}{1000} + 1\right) \end{aligned} \tag{1}$$

in this paper, we set $N = 25$, and the final result of the GT filter banks can be formulated as

$$y(n,t) = s(t) * q(n,t) \tag{2}$$

where * is the convolution, $y(n, t)$ represents the filtered signal, and $s(t)$ is the input audio signal. The frequency responses of the Gammatone filters are represent by $g(n, t)$.

The Meddis inner hair cell model

The inner hair cell is a transducer element of the cochlea which function is responsible for converting the mechanical vibration of the basilar membrane into a potential within the cell membrane. The audio signal filtered by the GT filter banks is processed by the inner hair cell mathematical model, and The functions of this model are half-wave rectification, non-linear compression and adaptive adjustment. The Meddis model is a commonly used in many auditory peripherals composite model. The permeability of membrane changes with the instantaneous sound intensity of sound waves, the Neurotransrmtter penetrates from Pool to Cleft through the cell membrane, the part of Neurotransrmtter in Cleft was collected into the Pool through the reprocessing store, and the other part Neurotransrmtter can be freely lost, and the Factory in the inner hair cells was also constantly making Neurotransrmtter to supplement the depletion.

The Meddis model describes the generating, transmitting and diffusing processing of converting acoustic signals into potential signals, and this mathematical model is simple

and easy to implement on a computer [27]. The differential equations describing the model are shown below:

$$\begin{aligned} k(t) &= \begin{cases} \frac{A+s(t)}{A+B+s(t)}g, & A+s(t) \geq 0 \\ 0, & A+s(t) < 0 \end{cases} \\ \frac{d_{q(t)}}{d_t} &= y(1-q(t)) - k(t)q(t) + xw(t) \\ \frac{d_{c(t)}}{d_t} &= k(t)q(t) - lc(t-rc(t) \\ \frac{d_{w(t)}}{d_t} &= rc(t) - xw(t) \end{aligned} \tag{3}$$

where the osmotic pressure of the cell membrane is $k(t)$. The $s(t)$ is the output of the basement membrane, and A, B, g, y, x, l, r are constants. The $c(t)$ determines the probability of nerve fiber activity, which is represented by h. The result of Meddis model is computed as Eq. (4).

$$V(t) = h \cdot c(t) \tag{4}$$

The two channels of getting local information entropy

The audio attention model generally calculates a significant attention area by linearly combining multi-dimensional features of the signal. In this paper, the signal is processed by the image channel and the audio channel to get the local information entropy, which can make up for the incomplete defects of the single channel calculation model, and the related experiments we did also prove that the model of combining two channels can improve the accuracy of our attention extraction.

The audio channel processing

The audio processing channel is mainly to frame the audio signal, normalize the amplitude and finally calculates the local information entropy. the information entropy is the average rate at which information are produced by a stochastic source of data. If a random signal has a high short-term frequency and uneven energy, the value of information entropy will be larger. On the contrary, the information entropy value of a uniform signal is lower. The signal perceived by the human ear are rarely stable in our actual environment where the characteristics of frequency or loudness have a certain range of changes, and have a higher value of the information entropy. Due to the value of the audio signal amplitude is normalized within -1 to 1, we divide the interval from -1 to 1 into n consecutive cells, such that a vector $Y = \{y_1, y_2, \ldots, y_n\}$ of this interval. The information entropy of each frame can be given by

$$\begin{aligned} H(k) &= -\sum_{i=1}^{n} p_i \log p_i \\ p_i &= \frac{count(y_n)}{\sum_{i=0}^{n} count(y_n)}, \quad y_n = [-1, 1] \\ H_{aud}(t) &= \sum_{k=0}^{t} H(k) \end{aligned} \tag{5}$$

where the $H(k)$ is the value of this information entropy. The p_i is the probability of the amplitude in the y_n interval. The $count(y_n)$ is the number of discrete points in the y_n interval and we set $n = 20$. Finally, $H_{aud}(t)$ is the result of the audio channel processing about getting local information entropy and t is time.

The image channel processing

The main purpose of image channel processing is to obtain a high attention area on the spectrogram where we can get some information different from the audio channel. This channel uses the image saliency correlation algorithm to calculate the local information entropy of the spectrogram, and the first thing we need to do is to get the spectrum of the one-dimensional audio signal with a bank of band-pass filters. A spectrogram is a visual representation of the sound signal as signal vary with time, and a common format of audio signal is a graph with two geometric dimensions: one axis represents time, the other axis is frequency, a third dimension indicating the amplitude of a particular frequency at a particular time is represented by the intensity or color of each point in the image. Here the spectrogram is defined as $Spec(t)$ where t is the time of the audio signal and $Spec(t)$ is in RGB domain, the next calculation of the image channel is described by:

$$Spec(t)_{\text{Gray}} = \frac{(299 \times Spec(t)_R + 587 \times Spec(t)_G + 114 \times Spec(t)_B)}{1000} \tag{6}$$

where $Spec(t)_{Gray}$ is the grayscale image associated with the spectrogram in RGB domain ($Spec(t)_R$ stands for red, $Spec(t)_G$ for green and $Spec(t)_B$ for blue). In order to deal with the grayscale image and get more useful and clear information from this image, we enhance the visual representation of grayscale image, we formulate the process as the following equation:

$$Eh(t) = mean(Spec(t)_{\text{Gray}}) + \frac{(Spec(t)_{Gray} - mean(Spec(t)_{Gray}))}{\frac{contrast}{100}} \tag{7}$$

where $Eh(t)$ is the enhanced image, and contrast is the contrast degree of improving the image quality, $mean(Spec(t)_{Gray})$ denotes the average value of $Spect(t)_{Gray}$ which calculated by Eq. (6). The internal computing process of this channel is formulated by:

$$\begin{aligned} P_i &= \frac{f(i)}{N^2} \\ H(t) &= -\sum_{i=0}^{255} p_i \log p_i \end{aligned} \tag{8}$$

where $H(t)$ is array, where each output pixel contains the entropy value of the 9-by-9 neighborhood around the corresponding pixel in the input image. $f(i)$ is the number of times that a pixel with a gray value of $i (i \in [0, 255])$ appears in a 9-by-9 neighborhood. P_i is the probability value, and we set $N = 9$. Finally, the local entropy value of the image is calculated to one dimensional entropy, and the process can be defined as:

$$H_{img}(t) = mean(H(t)) \tag{9}$$

where $H_{img}(t)$ is the value of the local information entropy that is one-dimensional vector calculated by averaging the columns.

The exponential moving average (EMA) correlation process

In statistics, a moving average(MA) is a calculation to analyze data points by creating series of averages of different subsets of the full data set, it commonly used with time series data to smooth out short-term fluctuations [28]. The EMA is a type of MA that places a greater weight and significance on the most recent data points, and its principles are the same as the human auditory system which give more attention to what happened in the near future.

The EMA for a series audio signal can be formulated as

$$\begin{aligned} EMA(k,n) &= h(k)\cdot a + EMA(k-1,n)\cdot(1-a) \\ a &= \frac{2}{n+1} \end{aligned} \tag{10}$$

where the $h(k)$ is the value of the information entropy. $EMA(k, n)$ is the value of the EMA at the nth frame, and the different values of n will represent the different scales of calculation, which smaller n-value reflects short-term trends in information entropy. The coefficient a is related to n.

In this paper, the short-term exponential moving average is expressed by $EMA(k, s_n)$, the long-term exponential moving average is expressed by $EMA(k, l_n)$, the value of the coefficient s_n and l_n is rated with the duration of the audio. In the same coordinate system, when the short-term exponential moving average line upward through the long-term exponential moving average line, it represents the information entropy value increased in the short-term audio signal, which can be considered as the beginning of an event with high degree of attention. And when the short-term exponential moving average line down through the long-term exponential moving average line, it represents the information entropy value reduced in the short-term audio signal, which is the sign of the human auditory system lose attention to this event. The calculation about differential EMA can be defined as

$$dif(k) = EMA(k,s_n) - EMA(k,l_n), \quad l_n > s_n \tag{11}$$

where *dif* is the difference between short-term EMA and long-term EMA and the higher value of *dif* indicates a higher attention to this time. The values of *dif* greater than 0 mean that we can confirm this frame with high attention, on the other hand the values of *dif* less than 0 mean that this frame can't attract our attention. We also can illustrate *dif* with the column chart that displays data as vertical bars, and there are some segments on the column chart which attention is different from around (the sign of the *dif* value is different from the front and back segments) and relative length of time is very short, so we must deal with these segments to keep the sign of the *dif* same with the surrounding segments. Namely, we have the following fact

$$dif(k)_{\text{deal}} = \text{seg\_}ful(dif(k)) \tag{12}$$

where $dif(k)_{deal}$ is the result of process that some segments have fused. The calculation about the degree of attention is shown in formula (13) where $attention(k)$ represents the degree of attention of the k_{th} frame.

$$attention(k) = EMA(dif(k)_{deal}, 0.1(l_n - s_n)) \tag{13}$$

Experiment and result analysis

To evaluate the performance of our audio attention computational model, we conducted three experiments including different audio signals, and the number of GT filter channels is 25 in three experiments. The first two experiments are mainly used to verify whether the model can well reflect the continuous characteristics and attenuation characteristics and the last experiment is designed to verify the extraction accuracy of this model. We compared our model with the Kayser model in the first two experiments and the region of interest obtained by this model was compared with the actual area of interest obtained by subjective test in the third experiment. The implementations are all from the publicly available sources. The experiment is developed under the PC condition 2.50 GHz of Intel Core i5-3210M CPU and 8G RAM.

Artificial sinusoidal audio signal

We constructed a 2 s audio signal which the first segment is muted in 0 to 0.5 s, then the second segment is a mixed sinusoidal signal with 100 Hz in 0.5 to 1.5 s, and the last segment is also silent. This signal is called signal one, and the difference between our attention calculation model and the Kayser calculation model for detecting signal one is shown in Fig. 3.

The sinusoidal signal portion in Fig. 3a is the interest area that this model needs to detect. The attention graph calculated by Kayser model has some obvious bouncing points at the beginning and end of the sinusoidal signal in Fig. 3b and the value of attention remains unchanged at the middle portion, which phenomenon is similar to

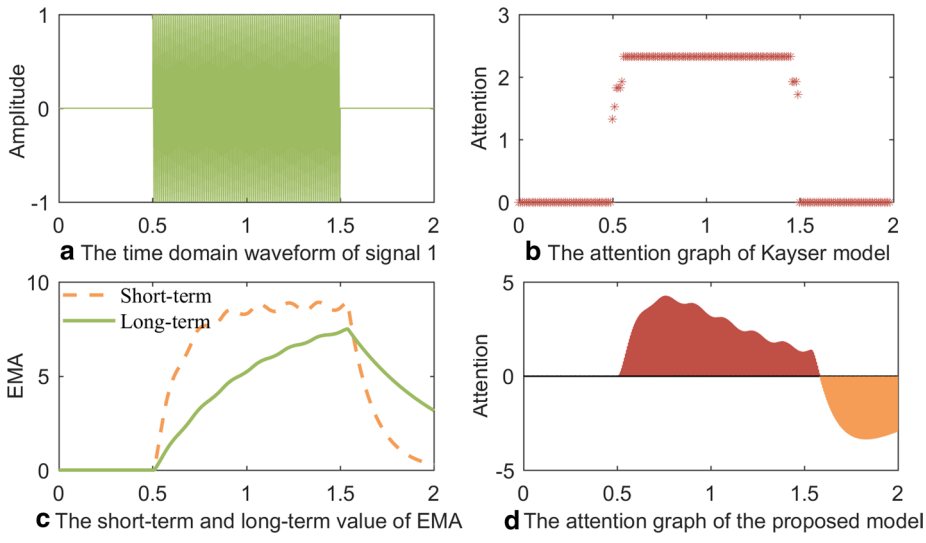

Fig. 3 An illustration of waveform and attention graph of signal 1. The region marked in red is the area of interest and marked in orange is the non-interest area in **d**. In contrast to Kayser model which extracts contrast characters from the spectrum of audio signal in multi-scale, the proposed attention model based on information entropy and exponential moving average smooth reflects the value change of audio attention in the time domain

the results of the Kayser model in other researcher's experiments [14]. As one of the classic models, the Kayser model can detect the start and end points of sinusoidal signals well, but the attention value obtained by the model remains almost unchanged, which does not consider the trend of attention degree with time.

There are two trend lines of short-term and long-term EMA in Fig. 3c, the long-term marked in green is more stability than short-term marked in orange, and the difference of these two lines can reflect the attention of signal 1. We can be seen from Fig. 3d that the value of the proposed model is smoother than Kayser model, and presents a trend of firstly increasing and then decreasing in the time dimension, which is similar to the continuous characteristics and attenuation characteristics of the human auditory system. Both the Kayser model and the model of this paper can accurately detect the sinusoidal signal in signal one, and can quickly detect the starting point of this signal, but our calculation model compared with Kayser has a slightly delay at the end of this sinusoidal signal.

Audio signal of THCHS-30 corpus

In order to verify the effectiveness of the model for the detection of interest region in a real voice scene, an audio signal of THCHS-30 corpus experiments is also carried out. The THCHS-30 corpus [29] of a free Chinese speech database is that an open Chinese speech database published by Center for Speech and Language Technology (CSLT) at Tsinghua University. As shown in Fig. 4, this model has a better effect on the actual environment.

We can see the irregular waveform in the Fig. 4a where there are three main audio parts, and other parts of the waveform are flat. The attention graph of Kayser model can roughly reflect the waveform of audio D32_892, but the shape of attention waveform also has some obvious bouncing points. The signal 2 is an audio signal in a real

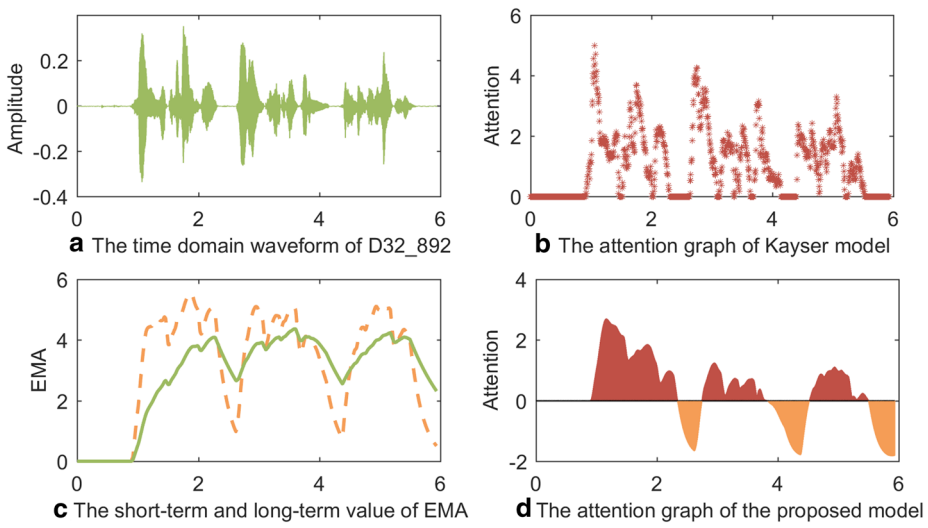

Fig. 4 An illustration of waveform and attention graph of THCHS-30 corpus. **a** Is a waveform of a piece audio numbered D32_892 and the line marked in orange is the short-term EMA and another line marked in yellow is the long-term EMA in **c**, and the region marked in red of **d** is the interest of region

environment including some noise, which can affect the accuracy of our experimental results. We can see from the attention map of Kayser model that the anti-noise performance is not very good, and the model of Kayser can't fully find out the area with high attention.

There are two lines of EMA to show real-time changes in attention, and the line of long-term EMA is relatively smooth to another line marked in orange. The final attention graph of this model is showed in Fig. 4d where there are two categories color which one marked in red is a region of concerns, and the introduction of the exponential moving average correlation algorithm also makes the obtained attention value smoother. Compared with Kayser model, the results generated by the proposed method provide much better continuous characteristics and attenuation characteristics, so it can improve the performance of audio attention computational. This model makes full use of the feature information between image and audio, and it can also smoother display the value change of the auditory attention map in an actual environment.

The accuracy evaluation of testing result

In order to verify the accuracy of the attention model in the actual environment, we select the 20171207 issue of the talk show "Tonight 80's Talk Show" and the 20170321 issue of the talk show "The Ellen DeGeneres Show", which has less background music and manual editing than other video sounds. We intercepted the audio from 15′40″ to 19′00″ in the first talk show as a talk show 1 and another talk show which intercepted from 35′15″ to 38′10″ as a talk show 2, the content mainly includes some of the calm or passionate speech by the host, and some cheering applause from the audience. The calculation result of detecting these talk shows is shown in Fig. 5.

We can see from the Fig. 5a1, b1 that the waveform of the two audio signals are so disordered that we can't directly know which region is the interest area. The long-term exponential average line has a small fluctuation, and the short-term exponential moving average shows more obvious down-traversing and up-traversing phenomena with the attention event occurring in Fig. 5a2 , b2. In order to verify the accuracy of the results of this model, we did some subjective test and manually recorded the highlights of the talk show with the Audition software. To compared the area detected by this model with the manually recorded area, we use the full-rate indicator to quantitatively judge the accuracy of this calculation.

$$\text{Accuracy} = \frac{\text{Lenth of Automatic Time}}{\text{Lenth of Manual Time}} \times 100 \tag{14}$$

It can be seen from Table 1 that talk show 1 has eight segments which contain laughter or hand-clapping and the talk show 2 has six segments of interest. Owing to the less noise of talk show 1, there have a relatively good accuracy for this proposed model, and in the talk show 2, the time block of automatic detection has a slightly delay, so we could add additional time region which locates before the detected time block in production environment to improve the accuracy. The reason why the event of category 1 failed to be detected is that the attention event is short and the playback shows that the amount of laughter of this clip is small. There are 43.1 s about the interested time and 28.4 s about the total extraction time detected by this model in the talk show 1, so the accuracy

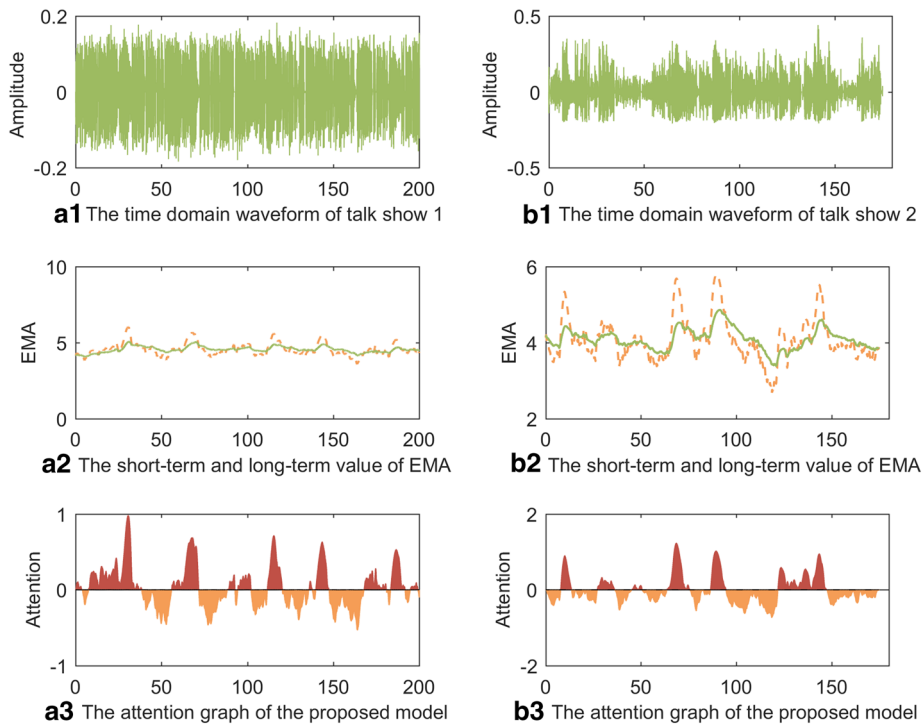

Fig. 5 An illustration of waveform and attention graph of talk show. No. **a** is a talk show 1, and No. **b** is a talk show 2. The short-term exponential moving average line marked in orange represent how much is the value of the information entropy in a short time, and line of the long-term EMA marked in green. **a3** and **b3** are the attention figure of proposed model, which the red histogram is the area of interest

Table 1 Experimental results

Talk show	Category	Manual/s	Automatic/s	Accuracy/%
1	1	12.8–14.2	7.7–38.5	*100*
1	2	25.1–32.7	7.7–38.5	*100*
1	1	49.2–50.8	–	–
1	2	61.8–71.7	56.0–71.7	*100*
1	2	111.3–117.1	111.7–120.3	*69*
1	1	129.5–131.0	–	–
1	2	138.8–144.7	139.9–147.5	*100*
1	1	149.8–150.9	–	–
2	2	7.4–17.7	8.1–13.6	*63.1*
2	2	26.4–30.6	27.3–36.3	*78.5*
2	2	63.4–72.2	63.7–75.6	*96.5*
2	2	85.9–93.2	122.0–147.5	*89.3*
2	2	120.8–132.1	122.0–147.5	*94.3*
2	1	163.2–165.1	–	–

The Category 1 indicates laughter and Category 2 indicates laughter and a burst of hand-clapping. The time block of manual testing with Audition software is in third columns. The time block of this model test is in the fourth column, and the fonts of accuracy appear in italics

about the talk show 1 is 84.9%, and we got 83.4% of talk show 2. At the same time, we find that the time regions extracted by this model are all sequential segments, and there aren't any single jumping points, which result also satisfies the characteristics of the attention event in time dimension.

There are also two problems, which will be our future work. The one is that owing to the different parameters of proposed model may result in differences results of automatic detection in some environments, and another one is that there is some accuracy value of automatic detection is relatively lower, we believe that this is due to the complicated and non-stationary background sounds through analyzing the detection signals.

Conclusion and future work

In this paper, we propose an effective audio attention computational model based on peripheral process, information entropy of two channels and exponential moving average to calculate the audio attention, and we simulate the process of human hear system that can keep attention on the interested event in time dimension. In order to validate this model, three empirical studies are conducted. Compared with Kayser calculation model, this model can't only extract the starting and ending points of the signal of high attention event in the experimental simulation, but also display the change of attention is more in line with human auditory attention mechanism(continuous characteristics and attenuation characteristics). In our actual talk show experiment, this model can detect 84.9% and 83.4% of the clips containing laughter and applause on the talk show. Among the audio attention computational models, this model fully consider the auditory mechanism of human ear. At the same time, compared with many multi-feature models of vision saliency image, this model based on image channel and audio channel has lower computational complexity in terms of computational complexity, and is more suitable for real-time systems. In addition, due to the lack of data sets of audio attention, there are some defects in the accuracy of the evaluation model. In further research, the accuracy of high-interest areas in the actual environment can be further improved by using other audio features or artificial intelligence methods. At the same time, we can try to introduce audio attention computational model into video summary model.

Abbreviations

MA: moving average; EMA: exponential moving average; AC: alternating current; ILD: binaural intensity dierence; ITD: binaural time dierence; GT: Gammatone.

Authors' contributions

YL described the proposed algorithms and wrote the manuscript. CZ, BH, HC and SW made a theoretical improvement on the algorithm and made a preliminary design of the experiment. All authors read and approved the final manuscript.

Author details

[1] School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan, China. [2] School of Computer Engineering, Hubei University of Arts and Sciences, Xiangyang, China.

Acknowledgements

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Funding

This research was supported by National Natural Science Foundation of China (Grant No. 61272278), Hubei provincial major science and technology projects (Grant No. 2018ABA099), Natural Science Foundation of Hubei Province (Grant No. 2018CFB408) and Hubei provincial education department science and technology research project (Grant No. Q20181807).

Publisher's Note

Received: 5 November 2018 Accepted: 28 January 2019
Published online: 26 February 2019

References

1. Alho K, Salmi J, Koistinen S, Salonen O, Rinne T (2015) Top-down controlled and bottom-up triggered orienting of auditory attention to pitch activate overlapping brain networks. Brain Res 1626:136–145
2. Liu Y, Bengson J, Huang H, Mangun GR, Ding M (2016) Top-down modulation of neural activity in anticipatory visual attention: control mechanisms revealed by simultaneous eeg-fmri. Cereb Cortex 26(2):517–529
3. Connor CE, Egeth HE, Yantis S (2004) Visual attention: bottom-up versus top-down. Curr Biol 14(19):850–852
4. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20(11):1254–1259
5. Kayser C, Petkov CI, Lippert M, Logothetis NK (2005) Mechanisms for allocating auditory attention: an auditory saliency map. Curr Biol 15(21):1943–1947
6. Hang B, Wang Y (2016) Spatial cues gradient in time domain based audio attention computational model. J Syst Simul 28(10):2369–2377
7. James W (1950) The principles of psychology. Am J Psychol 2(4):761
8. Cai R, Lu L, Zhang HJ, Cai LH (2003) Highlight sound effects detection in audio stream. In: International Conference on multimedia & expo
9. Ma YF, Hua XS, Lu L, Zhang HJ (2005) A generic framework of user attention model and its application in video summarization. IEEE Trans Multimed 7(5):907–919
10. Liu A, Li J, Zhang Y, Sheng T, Yang Z (2007) Human attention model for action movie analysis. In: International conference on pervasive computing & applications
11. Zheng Y, Zhu G, Jiang S, Huang Q, Wen G (2008) Visual-aural attention modeling for talk show video highlight detection. In: IEEE international conference on acoustics
12. Wang X, Xia X, Zhang X, He P (2013) Research on a novel saliency map computational model of auditory attention. J Signal Process 9:1142–1147
13. Liu Y, Zhang M, Zheng F (2013) Cognitive neural mechanisms and saliency computational model of auditory selective attention. Comput Sci 40(6):283–287
14. Zhang X, Xia X, Wang X (2014) A saliency map extraciton model for auditory attention. J Sichuan Univ 2:292–298
15. Chen X, Xia X (2016) Auditory saliency calculation based on sparse dictionary. Comput Syst Appl 25(4):201–205
16. Lv F, Xia X (2017) Study on computational model of auditory selective attention with orientation feature. Acta Automatica Sinica 43(4):634–644
17. Russell IJ (1983) Origin of the receptor potential in inner hair cells of the mammalian cochleaevidence for Davis' theory. Nature 301(5898):334–336
18. Kalinli O (2016) Analysis of multi-lingual emotion recognition using auditory attention features, pp 3613–3617
19. Kaya EM, Elhilali M (2012) A temporal saliency map for modeling auditory attention, pp 1–6
20. Long LK, Zhou P, Yang HY (2017) Speech feature extraction algorithm based on Gammatone filter bank and subband power normalized. Meas Control Technol
21. Qi J, Wang D, Jiang Y, Liu R (2013) Auditory features based on Gammatone filters for robust speech recognition, pp 305–308
22. Mu L, Peng Y, Qiu M, Yang X, Hu C, Zhang F (2016) Study on modulation spectrum feature extraction of ship radiated noise based on auditory model, pp 1–5
23. Meddis R, Hewitt MJ, Shackleton TM (1990) Implementation details of a computation model of the inner haircell auditorynerve synapse. J Acoust Soc Am 87(4):1813–1816
24. Shao Y, Wang D (2008) Robust speaker identification using auditory features and computational auditory scene analysis, pp 1589–1592
25. Onizawa N, Koshita S, Sakamoto S, Abe M, Kawamata M, Hanyu T (2016) Gammatone filter based on stochastic computation, pp 1036–1040
26. Ngamkham W, Sawigun C, Hiseni S, Serdijn WA (2010) Analog complex Gammatone filter for cochlear implant channels, pp 969–972
27. Lei W, Yuan P, Lin ZQ, Jiang XH, Lin MU, Zhang FZ (2012) The application of computational auditory peripheral model in underwater target classification. Acta Electronica Sinica 40(1):199–203
28. Moving average. https://en.wikipedia.org/wiki/Moving_average#cite_note-1. Accessed 23 Sept 2018
29. THCHS-30. http://www.openslr.org/18/. Accessed 23 Sept 2018