Australian Journal of Basic and Applied Sciences, 9(16) Special 2015, Pages: 97-103

ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

A Saliency Based Effective Browsing of Visual and Acoustics

[1]R. Venkatesan, [2]J. Reeni and [3]A. Balaji Ganesh

[1]Research Assistant, Department of EIE, Electronic Design System Laboratory,Velammal Engineering College, Surapet, Chennai-600066, India
[2]M.E – Embedded System Technologies, TIFAC-CORE in Pervasive Computing Technologies, Velammal Engineering College, Surapet, Chennai-600066, India.
[3]Professor,Department of EIE,Electronic Design System Laboratory,Velammal Engineering College, Surapet, Chennai-600066, India.

ARTICLE INFO

Article history:
Received 20 January 2015
Accepted 02 April 2015
Published 20 May 2015

Keywords:
Audio-Visual attention STFT algorithm GBVS saliency map DTCWT Artificial Neural Network SOM

ABSTRACT

Browsing of large audio archives in a short duration is a challenging task as our daily life is flooded with auditory information. A complete system is developed to make the task of browsing auditory information simple and easy for access. Developed algorithm is contrasted against previously published auditory saliency maps which treat the two-dimensional auditory time-frequency spectrogram as an image that can be analyzed using visual-saliency models. Rather than proceeding with all the information present in the spectrogram, maximizing the information makes the browsing process further simpler. This focuses on particular event which grabs the attention. Saliency of an event is measured by how much the event differs from the surrounding that precedes it in time. Modeling the auditory attention system will help to predict the salient event and can be applied on event detection. Performance is maximized and shows better results when working on saliency maximized rather than conventional spectrograms. This paper deals with a development of multimodal saliency representation where audio and visual cues compete for the better tracing of information the either way. Extracting sound features like intensity, temporal and spectral contrast helps in triggering the bottom-up attention and creates a time dependent feature vector. Specific unsupervised learning algorithm is used to learn the occurrence of the sound specified.

To Cite This Article: R. Venkatesan, J. Reeni and A. Balaji Ganesh., A Saliency Based Effective Browsing of Visual and Acoustics. ***Aust. J. Basic & Appl. Sci.,*** *9(16): 97-103, 2015*

INTRODUCTION

Automatic detection of acoustic events outperforms human audition but fails in uninterrupted, continuous fatigue long recordings. Many new surveys are devoted to the analysis and synthesis of cognitive perception from the point of view of attention and machine learning. Extraction of attentive events is cognitive mechanisms employed for organizing perceptual method which is functionally correlated to modulations of neuronal activity. It is the process of representing information by reduction and selection mechanisms. Attention is the process of focusing resources on prevailing properties or individual streams. Bottom-up approach of saliency is based on the sensory cues of a stimulus captured by its signal level properties like spatial, temporal and spectral contrast. The model simulates how the concept of auditory saliency is analogous to that of visual saliency. Developing such a model for the auditory domain is relatively novel and recent idea. A model of auditory saliency is interesting since it helps in understanding of auditory attention and perception.

The computational auditory saliency model is analogous to visual saliency models that are used to identify what information grabs the attention of the observers. Researches in visual domain have been done and there are many existing visual saliency models that are used for various applications, such as parameter learning, object recognition and target detection. Intensity, color and orientation, are the main features typically used in the feature set of all visual saliency models. The model given by Itti extracts these three features, along various scales (Itti *et al*.,1998). The different scales are then compared using a center-surround mechanism to obtain the feature maps, where the finer scale is the “center” and the courser scale is the “surround" The feature maps are then combined into a two-dimensional “saliency map" corresponding to the points within an image that are salient. The model of general auditory saliency matches well with results from human subjects on selecting salient sounds from a scene. An

Corresponding Author: R. Venkatesan, Research Assistant, Department of EIE, Electronic Design System Laboratory,Velammal Engineering College, Surapet, Chennai-600066, India
E-mail: ei.venkatesan@velammal.edu.in

advantage of the model is that it can be extended for a variety of applications(Kalinli, 2012).

In the usefulness of the auditory saliency map is brought out by extracting a set of auditory features in parallel from the auditory spectrum of the sound, and fed into a master saliency map in a bottom-up manner. This provides fast selection of conspicuous events in an acoustical scene. This attention model is applied to detect the prominent syllable of musical sound(Kalinli, 2012) in an unsupervised manner.

Even in extreme noisy environments, we find ourselves able to easily follow the conversation with others. Delmotte, investigates the role of saliency in the auditory attention process. In order to identify the sounds that grab a listener's attention a model inspired by understanding of the human auditory system and auditory perception was designed (Delmotte, 2012). This algorithm matches well with human performance in detecting salient auditory events in a complex scene. An audio-visual saliency feature extracting method for environment sound signal processing is proposed in (Wang *et al*., 2013). Image saliency maps have been the subject of efforts to model computationally the image regions selected based on the feature properties. Cues for visual saliency are sought in low levels such as intensity, color and orientation. The method they proposed is based on the fusion of image and auditory saliency feature from spectrogram.

The idea of an early filtering process being used by the human auditory system (Mesgarani *et al*., 2012) in order to ensure that only the most important elements of the scene is passed on for higher level processing. The attention process and provides with an auditory saliency map which indicates the areas of a particular auditory scene (Rujiao *et al*., 2013) that stand out perceptually to observers. This information can then be used to improve the performance of current computational models in many different environment such as noise suppression, audio classification, and speech recognition. The auditory saliency is analogous to visual saliency models that are used to identify what areas of an image stand out to observers. More saliency research has been done in the visual domain, and there are many existing applications such as detection or recognition.

The model presented by Kayser is based on the visual saliency models (Kayser *et al*., 2005). Biologically inspiring contrast filters such as those found in center-surround receptive fields were used to calculate contrast. Information is then passed on for higher level processing to filter out less important aspects of the scene and reducing the computational load of the systems. Ittihas emphasis on bottom-up control of attention which is a desirable mechanism not only for biological organisms (Itti, 2005), but also for efficient allocation of resources in artificial systems. The model is applied in attention modeling of automatic target detection to advertising design.

The aural saliency consists of the sound features such as pitch, loudness, spectral and temporal cues. Building on the analogies of early visual and auditory processing, auditory saliency maps of an acoustic scene were developed by the visual paradigm. The auditory spectral representation is processed as an image, by extracting multiple features such as intensity, orientation, frequency and temporal contrast. The model developed by (Bruce, 2005) demonstrates a strategy that predicts human attention deployment on the principle of maximizing information sampled from the scene and facilitates selection based on high level features. Experimental results with applications of simple audio-visual integration models for cognitive scene analysis are provided by (Ramenahalli *et al*., 2013). The auditory salience map consists of the location of the sound source. This model also serves as an auditory salience map because the spatial location of sound stimulus is taken to be the only relevant feature.

The focus of this paper is on the audio-visual saliency representation of data which is the attention product of bottom-up saliency helps in maximizing the information the signal contains for the sensory processing. We formulate the visualization problem as maximizing the salient information between the obtained conventional spectrogram and the visual saliency map. The spectrogram contains a mixture of information which is transformed to a saliency optimized one. Further process of feature extraction and classification through unsupervised training method is modulated. The model presents the formulation of effective browsing of signal which might be an audio or image information.

1. System Architecture Model:

The developed system is a consequence of bottom-up approach with the tuning of the saliency mechanisms that guide the deployment of auditory attention. The features of the audio signal are generated from an auditory spectrogram, a time frequency representation to detect the maximal discrete set of features for the target. The spectrogram is maximized by extracting the optimal information through saliency. This saliency allocates the perceptual resources to the limited region that matters the most thus increasing the robustness and efficiency. When the audio utterance is given as an input to the system the visualization is carried on through commercial spectrogram and then saliency map is computed. Maximum feature extraction is possible with the use of bandpass filters. The filter used in this work is the dual tree complex wavelet transform(DTCWT). The SOM classifies the nearest neighboring neuron which exactly matches to the given input signal and presents the corresponding visual output as a target result. When the visual signal is given as the input to the system it directly goes through the process of the saliency map formation of maximizing the information present in

the image thus enhancing the efficiency of browsing the archived audio file in a large database. The whole setup is presented in figure1 which proposes the visual-auditory saliency model.

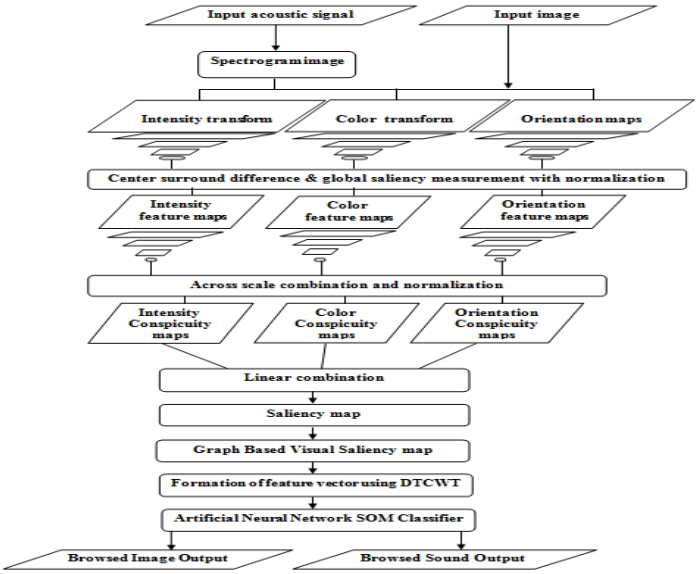


Fig. 1: Architecture for proposed visual-auditory saliency model.

2. Design Methodologies:

Extracting several sound features from the spectrogram to create a saliency map predicts the most attractive area in the scene. The redundancy of two trees produces the real and imaginary coefficients that provide extra information for analysis but at the expense of extra computational power. Any kind of bandpass filters helps in extracting the maximum set of features from the signal. Dual tree complex wavelet transform have the excellent property of directionality selection and shift invariance which is used for feature extraction of the spectrogram in thiswork. The rate of occurrence of the sound specified is learnt through unsupervised learning algorithm called the Self Organized Map (SOM).

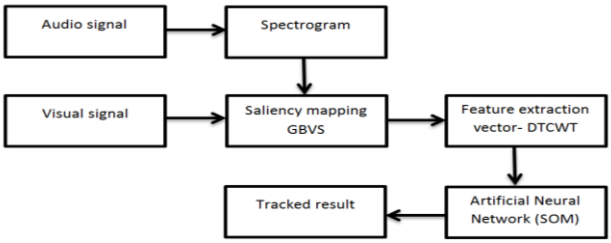


Fig. 2: Block diagram for the proposed system.

A. Visualization of Auditory Signal:

The short time fourier transform (STFT)algorithm does the visualization of sound signal which is one of the most fundamental and powerful tool in audio signal processing with a unified time frequency resolution(Tordini,2013). The computation is fast and simple and unveils the time-frequency structure of the signal with overlapping that maximizes the benefits of STFT analysis (Xinglei *et al*.,2007). Segmenting the signal into narrow time intervals is to be considered as a stationary signal. Original sound signal to be analyzed is given by x(t). Window function is h(t-τ). STFT of x(t) is given as

$$S(\tau,\omega) = \int_{-\infty}^{\infty} x(t)h(t-\tau)e^{-i\omega t}\,dt \qquad (1)$$

This will transform signal x(t) into time-frequency domain (τ, ω). STFT is performed by using a Hamming Window with an overlap of 50%.

B. Saliency Cues:

Bottom-up approaches use features extracted from the signal such as orientation, spectral contrast and temporal contrast. Once the features are extracted it looks for rare, novel, less compressible, maximum areas of information (Jingyu *et al*., 2013). The standard approaches are based on biologically motivated feature selection with the center-surround operations that highlights on the local gradients. A complete bottom-up saliency model is based on graph computation. Graph Based Visual Saliency

(GBVS) consists of a framework of activation and normalization/combination (Harel,2006).Nowadays there are hundreds of different models with various implementations and technical approaches even if initially they all derive from the same idea. It is thus very hard to find a perfect taxonomy which classifies all the methods (Itti *et al*.,2013). This work implies that only some methods can handle top-down information while all bottom-up methods could use top-down approach.

Intensity conspicuity map is obtained by applying to the intensity, given by $F_1=(r+g+b)/3$, where r, g, b are the color components, (Evangelopoulos *et al*., 2013) a local contrast operator that marks a voxel as more conspicuous when its value differs from the average value in the surrounding region:

$$C_I(q)=|_{F_1}(q)-\frac{1}{|Nq|}\sum_{u\in Nq}F_1(u)| \quad (3)$$

Where $q \in Q$ and Nq in eq.4 is the set of the 26-neighbors of q. The 26-neighborhood is the direct extension in 3-D of the 8-neighborhood in the 2-D image space.

Color conspicuity map is based on the color opponent theory that suggests the control of color perception by two opposing systems: a blue–yellow and a red–green mechanism. Such spatial and chromatic opponency
exists for the green-red, red-green, yellow-blue, and blue-yellow color pairs in human primary visual cortex (Evangelopoulos *et al*., 2013):

$C_2(q)=(RG=BY)(q)$

$RG=|R-G|$

$BY = |B-Y|$

Where $R = r-(g+b)/2$, $G = g-(r+b)/2$, $B = b-(r+g)/2$ and $Y = (g+r)/2-|r-g|/2-b$ (4)

Orientation conspicuity map is computed using spatiotemporal steerable filters tuned to respond to moving stimuli. The responses are obtained by convolving the intensity volume with the second derivatives of a 3-D Gaussian filter and their Hilbert transforms and the quadrature response is taken to eliminate phase variation. Energies are computed at orientations θ defined by the angles related to the three different spatiotemporal axes. In order to get a exact measure, the response of each filter is normalized by the sum of the consort and orientation conspicuity is computed by

$$C_3(q) = \frac{\sum_\theta E_\theta(q)}{\sum_u \sum_\theta E_\theta(u)} \quad (5)$$

The orientation is given by $\theta \in \{0°, 45°, 90°, 135°\}$

Different wavelet techniques are applied for signal and image processing methods. The applications of dual tree complex wavelet transform (DTCWT) which has significant advantages over real wavelet transform for processing (Kingsbury,2005)(Kingsbury *et al*.,2011)generates the complex coefficients by using dual tree of wavelet filters to obtain their real and imaginary parts. Complex wavelet useful for de-noising purpose provides high degree of shift-invariance and better directionality compared with the Real DWT (Kingsbury *et al*.,2011) .

The dual-tree complex wavelet transform (CWT) is a relatively recent enhancement to the Discrete Wavelet. The shift invariance and directionality of the DTCWT is applied for feature extraction inimage and signal processing. DTCWT features are computed by representing the frames as a Gaussian pyramid and by convolving each layers of the pyramid with 8x8 DCT bases.

C. Self-Organizing Map:

The self-organizing map (SOM) is an artificial neural network algorithm uses an unsupervised learning to create topographically ordered spatial representations of an input data. In this work SOM is employed for classification of the samples data which employs both training and mapping of target data. Neurons have access to a global time stamp, which allows the gap between the firing time of the best matching unit. The firing time of the current neuron is calculated as given in (Wang *et al*.,2007).

$$f(\Delta t)=\begin{cases} e^{-i\omega t}A^+(1-\frac{1}{\tau^+})^{\Delta t}, & \text{if}\Delta t>0 \\ -\omega ijA^-(1-\frac{1}{\tau^-})^{\Delta t}, & \text{if } \Delta t<0 \end{cases} \quad (6)$$

Difference between the current and the best matching neuron is given by $\Delta t.\ \omega ij$->ωij(t) + f(Δt). A^+ and A^- are bth positive which determines the maximum amount of synaptic strengthening and weakening that can occur. $\tau^+ and\ \tau^-$are time constrains.

Table I shows the corresponding conventional spectrogram image of the audio signal and the saliency map of the obtained conventional spectrogram for the first 4 sample dataset. Table II shows the time of execution for the samples A to B for the computation of the spectrogram, spectrogram GBVS map, the image GBVS map and DTCWT.

2. Results:

This model is tested with the VidTIMIT Audio-Video dataset which consists of recordings of human voice with the faces. The VidTIMIT dataset is comprised of video and corresponding audio recordings of 43 volunteers (19 female and 24 male), reciting short sentences. The recording was done in an office environment using a broadcast quality digital video camera. The audio is stored as a mono, 16 bit, 32 kHz WAV file. A total of 43 dataset was used for the training of target detection with 9 different types of sound uttered by the same individual. Only 4 samples of the results are shown in this paper.

Table I: Sample audio with its corresponding spectrogram and saliency map.

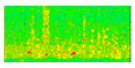
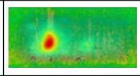
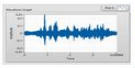
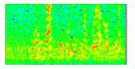
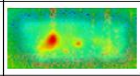
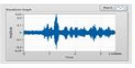
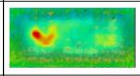
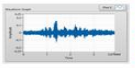
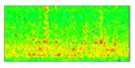
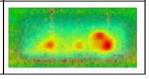
SAMPLES	AUDIO	SPECTROGRAM	SALIENCY MAP
A			
B			
C			
D			

Table II: The time of execution for 4 samples.

SAMPLES	TIME OF EXECTUTION (sec)			
	A	B	C	D
Spectrogram	0.79889	0.81538	0.82454	0.97548
Spectro GBVS Map	0.98462	1.08462	1.04658	1.09477
Image GBVS Map	1.44137	1.55109	1.56987	1.64788
DTCWT	0.26402	0.22439	0.23286	0.40589

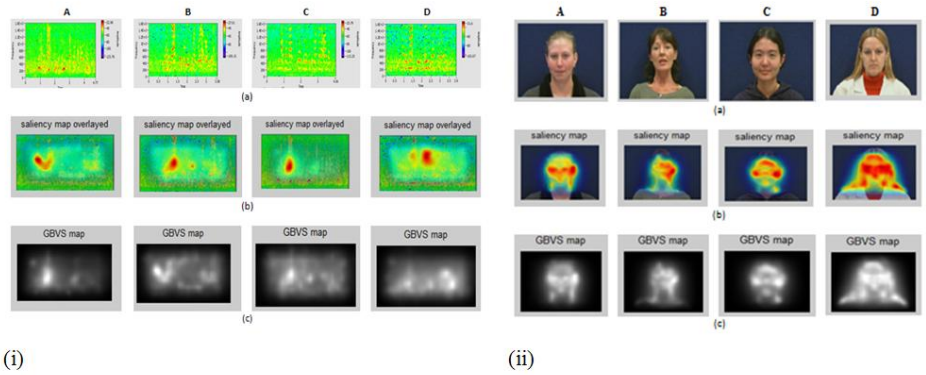


Fig. 3(i): The spectrogram and saliency feature extraction of sound signal for 4 samples A to D. (a) Spectrograms of the sound signals. (b)Saliency overlay map for the spectrogram. (c)Saliency features processed with GBVS algorithm.Fig.3(ii).The saliency feature extraction of visual signal for 4 samples A to D. (a) Original visual signals. (b)Saliency overlay map for the image. (c)Saliency features processed with GBVS algorithm.

Fig. 3, demonstrates the results of the samples A to D. The images of fig.3i(a) shows the processed sound signal which is been converted to a spectrogram and visualized. This used the STFT algorithm for the spectrogram conversion. The images of Fig.3(b) shows the saliency map extracting the saliency features of the sound signal sample. The image of Fig.3(c) shows the saliency maximized GBVS map with the removal of background noise signal.

Confusion matrix is developed with the results obtained from the SOM classifier by using the dataset. It gives the table of actual and the predicted samples that is listed in table III. Using the data found from Table III the sensitivity and the precision values are calculated and listed in Table IV.

The setup is demonstrated by designing an appropriate graphical user interface which gives a list of voice recordings, image data and the name list. The speakers audio information is obtained by selecting the name or the image of the speaker and vice versa. Fig.4 depicts the GUI for the sample A. If there is a new speaker to be added or a recording has to be deleted from the database the GUI gives the option to do it.

Table III: Sample confusion matrix for SOM classifier with 1000 training epochs

Predicted / Actual	A	B	C	D	E	F	G	H	NC
A	7	0	0	1	0	1	0	0	0
B	0	8	0	0	0	0	0	0	1
C	0	0	7	0	1	0	1	0	0
D	1	1	0	7	0	0	0	0	0
E	0	0	0	0	9	0	0	0	0
F	0	0	0	0	0	9	0	0	0
G	0	0	0	1	0	0	8	0	0
H	0	0	0	0	0	0	0	9	0

Table IV: Sample accuracy rate, sensitivity and PSNR values for few set of samples.

Samples	A	B	C	D	E	F	G	H
Accuracy rate (%)	77.8	88.9	77.8	77.8	100	100	88.9	100
Sensitivity (%)	87.5	88.9	100	77.9	90	90	88.9	100
PSNR (dB)	28.8	27.6	27.3	28.7	29.1	28.2	27.2	27.6

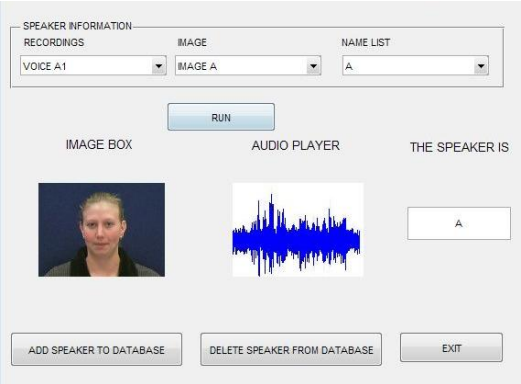


Fig. 4: The GUI setup for the selection of input and for displaying the output at the user end.

3. Conclusion:

An audio-visual saliency feature extraction method for efficient image and audio browsing is proposed in this paper. The saliency maximized implementation shows better performance than with the conventional spectrogram. An unsupervised algorithm is implemented with learning set and its feature vectors for tracking. Thus Dual tree complex wavelet transform is found to a very efficient tool in extracting the features. The module can be applied in variety of applications which also includes speaker identification and person voice authentication.

In the future, the implementation is taken to the next level of real time processing of the audio-visual attention model which grabs the attention of the speakers along with the localization of the sound source.

ACKNOWLEDGMENT

The authors wish to thank Department of Science and Technology for awarding a project under Cognitive Science Initiative Programme (DST File No.: SR/CSI/09/2011) through which the work has been implemented.

REFERENCES

Bruce, Neil, John Tsotsos, 2005. "Saliency based on information maximization." *Advances in neural information processing systems on*, 3. IEEE, 2555-2561

Delmotte, Varinthira Duangudom, 2012. "Computational auditory saliency." *Information Sciences and Systems (CISS), 2013 47th Annual Conference on*, 1(6): 220-212.

Evangelopoulos, G., A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, Y. Avrithis, 2013. "Multimodal Saliency and Fusion for Movie Summarization Based on Aural, Visual, and Textual Attention," *Multimedia, IEEE Transactions on*, 15(7): 1553-1568.

Harel, Jonathan, Christof Koch, Pietro Perona, 2006. "Graph-based visual saliency." *Advances in neural information processing systems*, 22(6): 123-191.

Itti, Laurent, G. Rees, J. Tsotsos, 2005. "Models of bottom-up attention and saliency." (2005), *Neurobiology of attention International Conference on*. 3. IEEE, 255-261.

Jingyu Wang, Ke Zhang, K. Madani, C. Sabourin, 2013. "A visualized acoustic saliency feature extraction method for environment sound signal processing", *TENCON 2013 - 2013 IEEE Region 10 Conference* (31194), 1-4: 22-25 .

Itti, L., C. Koch, E. Niebur, 1998. "A model of saliency based visual attention for rapid scene analysis", *IEEE Transactions on Pattern Analysis and Machine* on, 19-7. Page(s): 215 – 222.

Bruce, N., J. Tsotsos, 2005. "Saliency Based on Information Maximization"(2005), *in* Audio, Speech, and Language Processing, IEEE Transactions on, 15-5 Page(s): 1545-1559.

Ramenahalli, S., D.R. Mendat, S. Dura-Bernal, E. Culurciello, E. Nieburt, A. Andreou, 2013. "Audio-visual saliency map: Overview, basic models and hardware implementation," *Information Sciences and Systems (CISS), 47th Annual Conference on*, 1(6): 20-22.

Rujiao Yan, T., Rodemann, B. Wrede, 2013. "Computational Audiovisual Scene Analysis in Online Adaptation of Audio-Motor Maps," *Autonomous Mental Development, IEEE Transactions on*, 5(4): 273-287.

Selesnick, R., Baraniuk, Kingsbury, 2005. "The dual tree complex wavelet transform", *IEEE Signal Processing Magazine*, 22(6): 123-151.

Tao Hong, N., Kingsbury, M.D. Furman, 2011. "Biologically-inspired object recognition system with features from complex wavelets," Image Processing (ICIP), 18th *IEEE International Conference on*, 261(264): 11-14 .

Tordini, Francesco, 2013. "Toward An Improved Model Of Auditory Saliency." *IEEE International Journal on*, 121-143.

Xinglei Zhu, G.T., Beauregard, L. Wyse, "Real-Time Signal Estimation From Modified Short-Time Fourier Transform Magnitude Spectra"(2007) *in* Audio, Speech, and Language Processing, IEEE Transactions on, (15-5) Page(s): 1645 – 1653.

Yonggang Wang, Deng Li, Xiaoming Lu, Xinyi Cheng, Liwei Wang, 2014. "Self-Organizing Map Neural Network-Based Nearest Neighbor Position Estimation Scheme for Continuous Crystal PET Detectors," *Nuclear Science, IEEE Transactions on*, 61(5): 2446-2455.