PARAMETRIC HEAR THROUGH EQUALIZATION FOR AUGMENTED REALITY AUDIO

Rishabh Gupta[1], Rishabh Ranjan[1], Jianjun He[2] and Woon Seng Gan[1]

[1]Digital Signal Processing Laboratory, School of Electrical and Electronic Engineering,
Nanyang Technological University, SINGAPORE
[2]Maxim Integrated, 160 Rio Robles, San Jose, CA, US, 95134

ABSTRACT

Augmented Reality (AR) audio applications require headphones to be acoustically transparent so that real sounds can pass through unaltered for natural fusion with virtual sounds. In this paper, we consider a multiple source scenario for hear through (HT) equalization (EQ) using closed-back circumaural headsets. AR headset prototype (described in our previous study) is used to capture real sounds from external microphones and compute the directional HT filters using adaptive filtering. This method is best suited for single source scenarios as one best filter corresponding to the estimated source direction is optimally used for HT filtering. In this paper, we propose parametric HT EQ for multiple-source scenarios in time-frequency domain by estimating a sub-band Direction of Arrival (DOA) using neural networks (NN) and selecting the corresponding HT filters from a pre-computed database. Objective analysis using spectral difference (SD) is used to evaluate the performance of different HT EQ filters with open ear scenario used as a reference. Using dummy head measurements with bandlimited pink noise and real source signals, it was found that the proposed integrated system significantly improves the performance over the conventional HT system in multiple source scenarios.

Index Terms— *Spatial Audio, Augmented Reality, Neural Networks, Hear Through (HT), Directional of Arrival, Time-Frequency Parametric Processing*

1. INTRODUCTION

Augmented reality (AR) audio aims to create an immersive user experience by overlaying spatial virtual content in real acoustic environment and seamlessly fusing the two together [1]. Today, a large number of commercial wearable AR devices such as Microsoft HoloLens, Magic Leap One, Meta 2 and Sennheiser Ambeo headset are being widely used in various applications such as in gaming, education, entertainment etc. [2-6]. In order to create a successful AR audio system, the first important step is to allow the real sound to reach the ears unaltered, which must be followed by spatial rendering of virtual sounds in sync with the dynamic real acoustic environment [7]. An ideal solution would be a set of completely transparent headphones with no headphone isolation. Since headphone isolation is finite and design specific for all headphones, additional processing called active hear through (HT) must be used to compensate for this effect. Ranjan and Gan [8] used adaptive headphone equalization (EQ) to fuse virtual sounds and real sounds using a prototype based on open back headphones. However, headphone isolation was found to be greater than 10 dB for high frequencies (above 5 kHz), which requires additional processing for HT EQ. Majority of previous studies have used in-ear headphones for AR [13]. Harma and Tikander [9-10] have described one such AR headset and ARA mixer, which allows manual tuning by user to compute a perceptual hear-through filter. Rämö et. al. [11] used an all pass filter to avoid the comb filtering effect, whereas in [12] adaptive HT EQ has been used to estimate headphone isolation curves for different fittings. Although in-ear headphones capture pinnae cues, they change the natural ear canal resonance and suffer from occlusion effects. In our previous study, a prototype based on closed back circumaural headphones was used, which does not suffer from issues listed above [13]. However, all the above studies for AR audio reproduction dealt with single source scenario. Moreover, in our previous study, the system assumed a priori knowledge of the sound source direction for applying directional HT EQ.

This study aims to address the above issues using parametric or time frequency analysis for active HT system. Parametric representation is extremely useful in multiple source scenario, since the spectral coefficients obtained as a function of time can be processed independently from each other. Moreover, when microphone signals are transformed in time-frequency domain, the differences in phase and magnitude for different sound positions are quite similar to the variation of spectral cues for humans [14]. This approach has been widely used in past for hearing aids [15], directional audio coding [16], speech enhancement etc. [17]. In this paper, a time-frequency analysis is integrated with direction of arrival (DOA) estimation using neural networks (NN). Recently, NN has been shown to give superior localization accuracy over other techniques [18-19]. Section 2 describes the proposed integrated HT system.

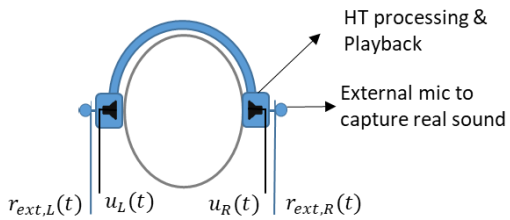

Fig. 1. AR headset structure equipped with two external microphones for HT EQ

2. PROPOSED PARAMETRIC HT SYSTEM

The AR headset prototype used for this study has been described in [13] and shown in Fig. 1 for reference. The external microphone signal are denoted by $r_{ext,L/R}(t)$ and processed ear signals by $u_{L/R}(t)$. In our previous study [13], we derived the HT filter for each source direction and constructed a HT EQ database in the form of FIR filters with 350 taps. In this paper, we aim to derive the HT filter for multiple source scenarios using a time-frequency or parametric processing. Figure 2 shows the block diagram of the proposed integrated AR headset system, where the two main parts are DOA estimation for appropriate HT filter selection and directional HT filtering in time-frequency domain.

2.1 DOA estimation using neural network

Neural Network have been exploited recently to localize the sound in real environment using the binaural recordings as data model [18, 19, 22]. For our proposed AR headset, audio signals captured at the two external microphones are used to extract the binaural features in sub-band to account for frequency dependency of DOA in multiple source scenarios. It should be noted that unlike conventional binaural signals recorded at the ears, the external microphone signals do not contain pinnae cues [13]. This leads us to develop the DOA estimation model only for the frontal horizontal source directions with azimuths from –90° to 90° to avoid front-back confusions. In previous studies, head rotation has been used in conjunction with NN models to minimize front-back confusions [19].

In this work, the aim of the DOA estimator is to obtain the direction of acoustic source(s) for each sub-band at any given time frame from the audio signals recorded at the two external microphones. Estimated directions will be subsequently used to select optimal HT EQ filters. First, a filter bank is applied to $r_{ext,L/R}(t)$ to obtain the sub-band signal, $r_{ext,L/R,s}(t)$ as shown in Fig. 2. For each sub-band, two binaural features, i.e., Interaural Cross Correlation (IACC) and Interaural Level Difference (ILD) are used for DOA estimation, as also being found in previous studies [18, 19, 21]. The IACC at time lag Δt (for each sub-band in a time frame) is computed as the cross-correlation of left right channels normalized by auto-correlation functions [20]:

$$IACC_s(\Delta t) = \frac{\sum r_{ext,L,s}(t) r_{ext,R,s}(t-\Delta t)}{\sqrt{\sum r^2_{ext,L,s}(t) \sum r^2_{ext,R,s}(t-\Delta t)}} \quad (1)$$

For binaural DOA estimation, IACC is usually evaluated for time lags between ±1.1 milliseconds, which results in 101

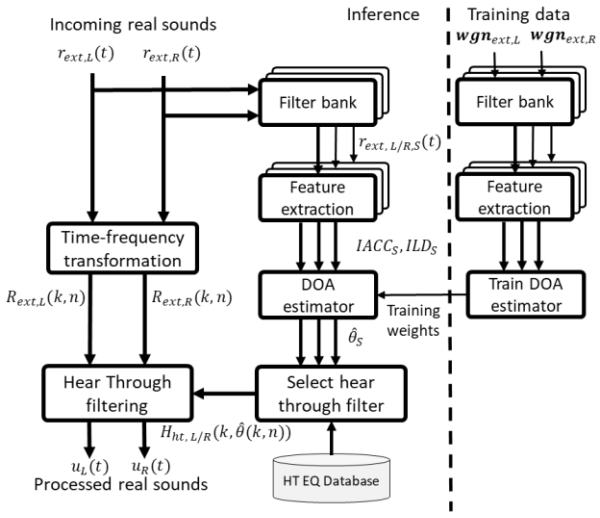

Fig. 2. Block diagram of the proposed parametric HT processing

feature samples at 44.1 kHz. Take note that Interaural Time Difference (ITD) information is embedded in IACC, since ITD is usually computed as the time lag corresponding to maximum IACC value. ILD is computed based on energy ratio of left and right channels:

$$ILD_s = 10log_{10}\left(\frac{\sum r^2_{ext,L,s}(t)}{\sum r^2_{ext,R,s}(t)}\right) \quad (2)$$

With a single ILD value for each sub-band, we have a total of 102-dimension feature for each training example. Therefore, the problem to solve is to find the regression function *f*:

$$\boldsymbol{\theta} = f(\boldsymbol{IACC}, \boldsymbol{ILD}) \quad (3)$$

which obtains the direction given the two binaural features. In this paper, *f* is represented by a neural network with fully connected layers, which learns the mapping between binaural features vector, $(\boldsymbol{IACC}, \boldsymbol{ILD})$ and the corresponding ground truth direction vector $\boldsymbol{\theta}$ for all training examples.

Our network topology consists of an input layer with 128 nodes and input dimension of 102 followed by a single hidden layer with 128 nodes, and finally an output layer, which consists of 13 nodes corresponding to 13 azimuths in the frontal horizontal plane (at 15° resolution). The activation functions of all the layers are "ReLU", except for the output layer for which "SoftMax" function is used. The chosen number of single hidden layer was decided heuristically as increasing hidden layer does not affect the overall accuracy. To avoid overfitting, dropout is used after hidden layer with rate of 0.25. During the training, optimizer was set to "adam" with the learning rate of 0.001 and maximum of 100 epochs were used for training subjected to early stopping in case accuracy did not improve after 10 epochs. Mini batch size of 25 samples were used in training. The above settings were also decided heuristically after trying other settings with no performance improvements. With the above topology, the network is also very fast to train as well as running inference in small CPUs.

The network is trained using simulated data and training samples $\boldsymbol{wgn}_{ext,L/R,S}$ were generated using a 10-second-long white gaussian noise signal filtered with HRTF-like filters $h_{ext,L/R}(n)$, measured at external microphones for left and right

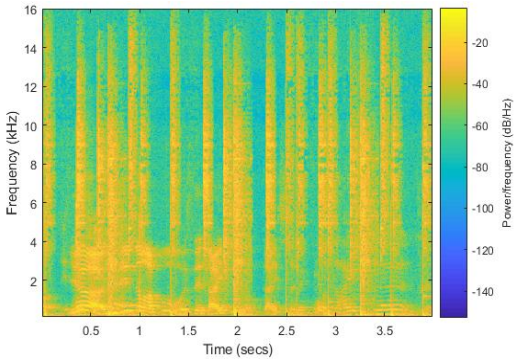

Fig. 3. Spectrogram of two real sources at [0°, 45°] at external mic for left ear

ears. The filtered white noise signal is decomposed into three non-overlapping and non-uniform frequency bands, namely, low: 0.1-1 kHz, mid: 1-5 kHz, and high: 5-16 kHz) with a time window of 200 milliseconds. Therefore, for each source direction, we have total 150 training examples (50 time-frames x 3 frequency bands).

Once the training of the model completes, the learned parameters are passed to inference for DOA estimation as shown in Fig. 2. DOA estimator takes the two binaural features as input and predict the estimated direction $\hat{\theta}_s$ for each sub-band. The estimated sub-band direction is used to select the HT EQ filter, which is then fed to the parametric HT processing block, as presented in the next subsection.

2.2 Parametric hear through processing

According to our previous study under single source scenarios [13], the ideal HT EQ requires a directional dependent filtering. For HT filtering, individual EQ filters are pre-stored for every 15° resolution covering entire 360°. The method using this directional HT filter is termed as idealHT. A zone-based HT filter can also be designed by clustering the directions into three different zones, namely, frontal (-60° to 60°), lateral (60°-120°, -60° to -120°), and rear (120° to 180°, -120° to -180°) [13], and then compute a representative filter for each zone, termed as groupedHT. A single averaged EQ filter can also be computed by taking average of all the directional EQ filters, called avgHT. AvgHT EQ filters were chosen for analysis since it was shown in [13] that this EQ filter has similar performance to the HT mode in commercial headphones such as Sony 1000 XM2.

In multiple source scenarios, a single HT filter can no longer be used. In this paper, we propose to apply HT EQ filtering in the time-frequency domain as it allows us to apply different HT filters at different frequency band and thus, exploiting the varied temporal-spectral characteristics of real sounds. Furthermore, frequency domain filtering is also more efficient in real-time implementation. A commonly used time-frequency transform is the Short Time Fourier Transform (STFT), with time window length of length L samples. The STFT yields an output $R_{ext,L/R}(k,n)$ at nth time frame and kth frequency bin, where $1 \leq n \leq N$ with N being the total number of time frames, and $1 \leq k \leq K$ with K being the total number of frequency bins. As shown in Fig. 2, a HT filter, denoted by $H_{ht,L/R}(k,\hat{\theta}(k,n))$ is chosen for each direction from a database of frequency responses of pre-computed HT filters based on the estimated angle $\hat{\theta}(k,n)$ derived from the DOA estimation. Finally, the HT EQ can be written as:

$$U_{L/R}(k,n) = R_{ext,\ L/R}(k,n)H_{ht,\ L/R}(k,\hat{\theta}(k,n)), \quad (4)$$

where $U_{L/R}(k,n)$ is the processed real signal (in time-frequency domain) that is played back through the headphone after the inverse time-frequency transform.

3. RESULTS AND ANALYSIS

This section provides details of the experiments used to evaluate the performance of different HT filters using bandlimited pink noise and two real signals, namely speech and music.

3.1 Signal synthesis

Two un-correlated pink noise signals of length two seconds each are used to synthesize test signals for the two experiments. These signals are first filtered by three bandpass filters with frequency ranges of 0.1-1 kHz (low), 1-5 kHz (middle), and 5-16 kHz (high). The obtained bandpass signals are filtered with appropriate impulse response $h_{ext,L/R}(n)$ for two direction pairs (0°, 30°) and (-15°, 75°). A total number of 12 test cases are created from the combinations of these bandpass signals, where 6 cases of overlapping frequency bands and 6 cases without overlapping frequency bands. In addition to the above cases, a broadband drum and a narrowband speech signal each of length four seconds was taken. The signals were convolved with $h_{ext,L/R}$ chosen for two directions selected randomly from a set of 13 frontal azimuthal angles (-90° to 90° in steps of 15°). All the possible permutations (156 in total) were taken for creating the test tracks. Fig. 3 shows the spectrogram (left channel) for one of the test signal track created using impulse response at two azimuthal angles 0° and 45°. It can be seen that the signal has content in almost all frequencies below 16 kHz, with slightly more signal power below 1 kHz.

Each of the three synthesized tracks as explained previously is decomposed in time-frequency domain with frame lengths $L = 8820$ and $K = 16384$ for STFT analysis.

3.2 Hear through equalization results

The spectral difference (SD) is used in this study to evaluate the performance of the proposed parametric HT system [13]. The formula used to calculate combined SD for each of the N time frames (frame index is omitted for brevity) is given by:

$$SD_{combined} = \frac{SD_L P_L + SD_M P_M + SD_H P_H}{P_L + P_M + P_H} \quad (5)$$

where the SD and power of the low, middle, and high frequency bands are computed as

$$SD_{L/M/H} = \sqrt{\frac{1}{K_{L/M/H}} \sum_{K_{L/M/H}} \left| 10log \frac{R^2_{ref,L}(k) + R^2_{ref,R}(k)}{\hat{R}^2_{ref,L}(k) + \hat{R}^2_{ref,R}(k)} \right|^2} \quad (6)$$

$$P_{L/M/H} = \sum_{K_{L/M/H}} \left(\left| R_{ext,L}(k) \right|^2 + \left| R_{ext,R}(k) \right|^2 \right), \quad (7)$$

where $R_{ref,L/R}(k)$ is the frequency spectrum of the target

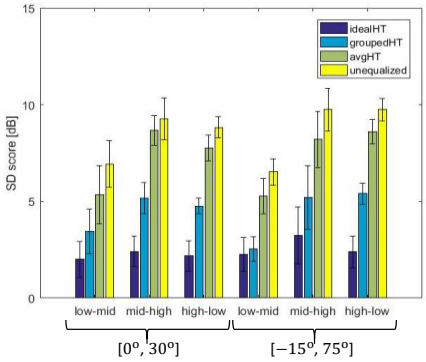

Fig. 4. Mean SD with standard deviation for four HT filters with two bandlimited pink noise sources in non-overlapped frequency regions

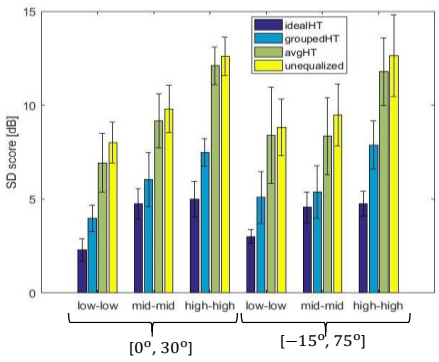

Fig. 5. Mean SD with standard deviation for four HT filters with two bandlimited pink noise sources in overlapped frequency bands

open ear reference, $\hat{R}_{ref,L/R}(k)$ is the spectrum of the sound recorded at the ear from headphone playback of the processed real sound using the derived HT EQ filters, $K_{L/M/H}$ denotes the total number of frequency bins in the low, middle, and high frequency band, respectively. Take note that the power weighted combined SD accounts for the spectrum variation in different frequency bands. Subsequently in this paper, SD refers to $SD_{combined}$. The mean and standard deviation of the SD across the N time frames are computed for each HT filters and discussed below.

Figures 4 and 5 show SD for two band-limited pink noise signals using non-overlapped and overlapped frequency bands, respectively. Overall, SD for all HT EQ filters is higher for signals with overlapped frequency bands (Fig 5) than non-overlapping frequency bands (Fig 4). This is due to closer mapping of estimated direction and directional HT EQ filters for non-overlapping signals as opposed to the overlapping case, where a single direction is estimated for each sub-band. Moreover, all HT filters have lower SD in lower frequencies than mid and high frequency regions as all the equalized responses are very close to the reference HRTF in low frequencies. Among the four HT filters, idealHT filters have the lowest mean SD values for all cases, with less than 3 dB and 5 dB for non-overlapping and overlapping signals, respectively. It is likely that idealHT EQ filters accurately model the directional-specific high frequency reflections of the pinna [13] and hence perform the best. However, storing parametric idealHT EQ filters for each direction requires large storage space, which can be reduced by using groupedHT and AvgHT filters, since they use a single filter for each zone and all directions, respectively. Parametric groupedHT shows better performance than avgHT since it is based on the average of nearest directions in a zone rather than an overall average as in AvgHT. Thus, groupedHT EQ can be used in cases where DOA estimation is not very accurate, although further investigation is required to evaluate the performance in this case. The overall trends for HT EQ performance for both direction pairs (0°, 30°) and (-15°, 75°) are similar. In Fig. 6, we show the SD variation using two real sources for all possible source positions combinations. The performance of HT EQ filters for real signals is similar to the pink noise cases in Figs. 4 and 5, with idealHT EQ performing best with mean SD less than 3 dB.

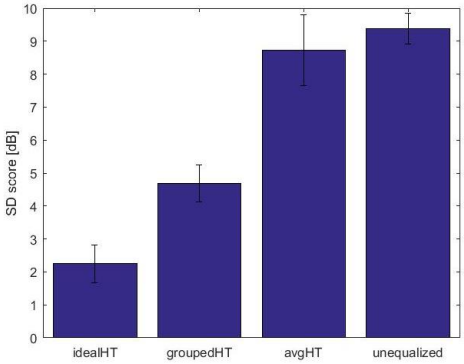

Fig. 6. SD for two real sources

4. CONCLUSION

In this paper, a parametric approach was used to apply hear-through equalization filters for multiple source scenario using an AR headset prototype. Incoming signals were filtered through a filter bank and NN based DOA estimation was used to estimate direction using IACC and ILD parameters as features for training and evaluation. This DOA information was passed to HT filtering block to select the corresponding pre-computed parametric HT filters. SD was computed with signals recorded at open ear as reference. It was found that both proposed parametric HT filters (idealHT and groupedHT) were superior to both AvgHT EQ filters and unequalized response for all test cases including the bandlimited pink noise at all frequency ranges and real signals such as speech and drums. Specifically, the proposed parametric idealHT performed best with mean SD of less than 3-4 dB. With a simple and fast NN topology for DOA estimation and efficient signal processing implementation for HT filtering in time-frequency domain, a real-time system is viable and is currently being explored. The real-time HT system will be evaluated in real-life scenarios. Furthermore, ambient sounds could also be identified using NN and diffuse HT EQ could be used in such cases.

ACKNOWLEDGEMENTS

This work is supported by the Singapore Ministry of Education Academic Research Fund Tier-2, under research grant MOE2017-T2-2-060.

5. REFERENCES

[1] R. Ranjan, "3D audio reproduction: natural augmented reality headset and next generation entertainment system using wave field synthesis," *Ph.D. Thesis, Nanyang Technological University, Singapore*, 2016.

[2] https://www.microsoft.com/en-us/hololens Last Accessed: Oct. 27, 2018

[3] https://www.metavision.com/ Last Accessed: Oct. 27, 2018

[4] https://www.magicleap.com/magic-leap-one Last Accessed: Oct. 27, 2018

[5] https://en-sg.sennheiser.com/finalstop Last Accessed: Oct. 27, 2018

[6]https://www.sony.com.sg/electronics/headband-headphones/wh-1000xm3 Last Accessed: Oct. 27, 2018

[7] W. S. Gan, J. He, R. Ranjan, and R. Gupta, "Natural and augmented listening for VR, and AR/MR," *ICASSP 2018 tutorial,* Calgary, Canada, Apr. 2018 [Online]. Available: http://sigport.org/2958. Last Accessed: Oct. 27, 2018

[8] R. Ranjan and W. S. Gan, "Natural Listening over Headphones in Augmented Reality Using Adaptive Filtering Techniques," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 1988-2002, 2015

[9] A. Härmä, et al., "Augmented reality audio for mobile and wearable appliances," *J. Audio Eng. Soc.*, vol. 52, pp. 618–639, 2004

[10] M. Tikander, M. Karjalainen, and V. Riikonen, "An augmented reality audio headset," in *Proc. 11th Int. Conf. Digital Audio Effects (DAFx-08)*, Espoo, Finland, 2008

[11] J. Rämö and V. Välimäki, "Digital augmented reality audio headset," *J. Elect. Comput. Eng.*, vol. 2012, Article ID 457374, 13 pages, 2012. https://doi.org/10.1155/2012/457374.

[12] J. Liski, R. Väänänen, S. Vesa, and V. Välimäki, "Adaptive Equalization of Acoustic Transparency in an Augmented-Reality Headset", in *Proc. AES Int. Conf. on Headphone Technology*, Aalborg, Denmark, Aug. 2016

[13] R. Gupta, R. Ranjan, J. He, and W. S. Gan "On the use of closed back headphones for active hear-through equalization in augmented reality applications" in *Proc. AES AVAR Conference*, Redmond, USA, Aug 2018

[14] V. Pulkki, S. Delikaris-Manias, and A. Politis (Edited), *Parametric time-frequency domain spatial audio*, Wiley, 2018

[15] J. Ahonen, V. Sivonen, and V. Pulkki, "Parametric spatial sound processing applied to bilateral hearing aids". In *Proc. AES Spatial Audio conf.,* Mar. 2012.

[16] J. Ahonen, G. Del Galdo, F. Kuech, and V. Pulkki, "Directional analysis with microphone array mounted on rigid cylinder for directional audio coding," *J. Audio Eng. Soc.*, vol. 60, no. 5, pp.311–324, May 2012

[17] P. Pertilä, and J. Nikunen, "Microphone array post-filtering using supervised machine learning for speech enhancement". *Proc. Interspeech*, 2014

[18] M. Lovedee-Turner, and D. Murphy, "Application of Machine Learning for the Spatial Analysis of Binaural Room Impulse Responses," *Applied Science*, vol. 8, no.1, pp.105-122, 2018.

[19] N. Ma, G. Brown, and T. May, "Exploiting deep neural networks and head movements for binaural localization of multiple speakers in reverberant conditions". In *Proc. Interspeech*. pp. 160-164, Jun. 2015

[20] V. Pulkki, M. Karjalainen, and J. Huopaniemi, "Analyzing Virtual Sound Source Attributes Using Binaural Auditory Model," *J. Audio Eng. Soc.*, vol. 47, no.4, pp. 203–217, Apr. 1999.

[21] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE Trans Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2112–2121, 2014.

[22] E. Koshkina, and J. Bouse, "Localization in Static and Dynamic Hearing Scenarios: Utilization of Machine Learning and Binaural Auditory model". In *Proc. of 21th International Student Conference on Electrical Engineerin*g, 2017.