An Auditory System Modeling in Sound Source Localization

Yul Young Park
The University of Texas at Austin
EE381K Multidimensional Signal Processing
Mar. 25, 2005

Abstract

The auditory system can be considered as an audio signal processing system with two input sensors. With this small number of input information, it shows remarkable performance in sound localization capability. Previously, several models were proposed to explain the sound localization capability of the auditory system. Most attempts can be categorized into two classes: neural network model and probabilistic estimator model, and they have their own merit and demerit depending on the algorithm. In this project, a neural network model expanded by evolutionary computation will be investigated to give better performance over the conventional models.

Introduction

Identifying the location of an object generating an acoustic signal of the auditory system has several significant respects in some information processing systems. The object detection and its localization in the sonar application is crucial in the military situation, and the identification of the speaker location can provide a useful cue to improve signal noise ratio (SNR) in hearing aid and microphone array applications. Also, the sound localization capability can gear current game industry with more vivid 3-D virtual reality. At the same time, sound localization of the auditory system, as a primary feature detector, can deliver a clue to the temporal and spatial resolution of the auditory system, and can be applied as a basic block to building more complex cognitive function of the brain like speech and music perception.

Sound Localization of the Auditory System

The auditory system extracts several cues from the neural representation of stimuli which are passed along the auditory signal pathway. The head, shoulder, upper body, and pinna give the transformed characteristics of the stimuli, and the middle ear causes filtering and amplification of the input. The signal arrived at the inner ear is decomposed into its frequency components by the hair cells in the cochlear. Now, this frequency information is converted into the neural signal, called action potential, and passed to the primary auditory fiber. Then, the neural signal is transmitted along the auditory nerve, cochlear nucleus, superior olivary complex, inferior colliculus, medial geniculate body, and finally to the auditory cortex [1],[6].

From this signal transmission, sound source to the primary auditory fiber pathway primarily provides an appropriate transformation of the spectral/temporal characteristic of the stimulus, and the cross-connected pathway starting from the cochlear neucleus to auditory cortex processes that signal to extract the source location information, which is called interaural time difference (ITD) and interaural intensity difference (IID). Then, ITD provides the azimuth of the source location and IID gives the elevation information [1], [4], [5], [9].

Previous Methods

The auditory pathway is conventionally thought to be composed of a cascade of sub-systems. Depending on the decision making at the final stage, there are two categories in general: neural network model and probabilistic estimator model.

A three-layer feedforward neural network model of sound localization in cat was proposed [3]. It modeled the pathway from a sound source to pinna by a transfer function which is so called head-related transfer function (HRTF) and the cochlear by simply dividing the input into 128 frequency components. Those 128 frequencies were input to the three-layer feedforward neural network which consisted of 128 input units, 4 to10 hidden units, and 11 by 17 two dimensional output units. The relationship between the layers was described by Eq.(1) and Fig.1. The input data characteristics were represented in the weight coefficients, w_{jk}, by training network with a training dataset.

$$(1) \qquad s_k = \theta_k + \sum_j w_{jk} x_j \qquad \textit{and} \qquad y_k = f(s_k) = \frac{1}{1+\exp(-s_k)}$$

The broadband input noise location in the training data set ranges over azimuths between -30° and +30° and elevations from -30° to +90° with 15° and 7.5° sampling of azimuth and elevation, respectively. The network training was done by adjusting the weight coefficients, w_{jk}, to reduce the error between the desired outcome, $\{d_k(t)\}$ and actual network output, $\{N_k(t)\}$, so called gradient descent method (2).

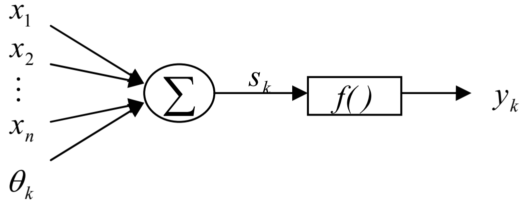


Fig.1. Structure of one model neuron with input $\{x_i\}$, threshold θ_k, activation function *f( )*, and output y_k.

(2) $$Error = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{L}[N_k(t) - d_k(t)]^2$$

where T is the number of patterns in the training set and L is the number of output units.

This model focused on the importance of spectral cues in monaural and binaural input presentation rather than ITD, and finally modeled a auditory space map in the superior colliculus. It showed that the best performance in elevation estimation occurred with 6.30° average error in binaural input and the notches presented in the 5k to 18kHz region of the input were crucial to sound localization. Furthermore, the responses of neurons in the hidden unit and output units were studied in the same way that was for the neurophysiological characterization of auditory neurons, and the response maps of some neurons in the network corresponded to those of the neurons in the dorsal cochlear nucleus.

Related to the previous results, Zakarauskas developed a mathematical model in the monaural localization [10]. He proposed two operators, $D_n'(\theta,\phi)$ and $C_n'(\theta,\phi)$:the first and second partial derivative of the HRTF with respect to the two neighboring frequencies f_n and f_{n+1} respectively, which seemed to be necessary for the neuron playing a role in sound localization. The simulation of the proposed operators was performed on the input data which were sampled 10° both azimuth and elevation. By solving the minimization problem Eq.(3), the elevations for two operators were estimated, and it produced the better outcome for the second derivative case, where the hit rate was 566 out of 614 test data and the average error ranged over elevations from 0.3° to 28° depending on the source spectra.

(3) $$\underset{\min \theta,\phi}{Error(\theta,\phi)} = \sum_n \left| D_n - D'_n\ (\theta,\phi) \right| \quad or \quad \sum_n \left| C_n - C'_n\ (\theta,\phi) \right|$$

Where D_n and C_n are the first and second finite differences between the observed intensity levels at two neighboring frequencies f_n and f_{n+1}.

On the other hand, the probabilistic estimator models share many common sub-systems by and large although there are few variations in detail. An example of such system is show in Fig.2 in block diagram. The pathway from the sound source to the pinna is modeled by HRTF, and the inner ear is described by various cochlear models [2],[12],[13]. The superior olivary complex was approximated as a cross-correlator and subtractor, and the higher level up to the auditory cortex was modeled as probabilistic decision making such as maximum-likelihood estimator (MLE) [2], [13], [7].

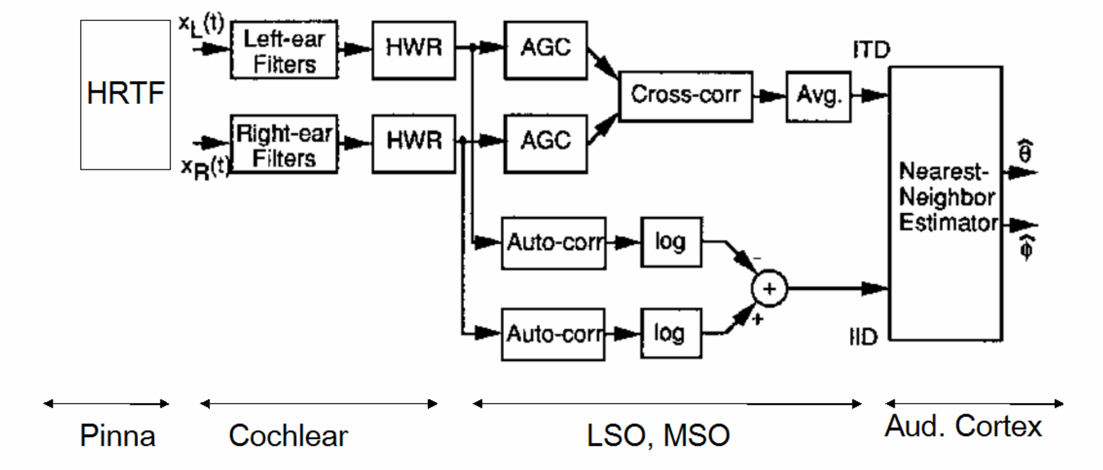


Fig.2. A typical sound localization system, modified from [2]

According to the conventional auditory pathway model, ITD is approximated well as a cross-correlation of cochlear outputs of both ears. IID, which is obtained from the difference between the signal spectra in both ears, is not as easy to model. For example, the IID of the sound source located at the median plane (azimuth =0 case) is zero, and it can give no information about the source elevation [2]. A system combining the monaural and binaural cues was shown to give improved results [13], [11]. The comparison of the models is presented in the Table 1.

Evolutionary Computation

In general, backpropagation neural network has several drawbacks such as the local minimum, generalization, and the fixed network architecture. Neuroevolutionary computation which evolves the artificial neural network by genetic algorithms can be a possible solution to such problems. Among several mothods, NeuroEvolution of Augmenting Topologies (NEAT) by Stanley has some useful features [8]. It starts from the minimal structure, and continue to add its nodes and connections incrementally on searching the optimal solution by examining the fitness. Each genome in NEAT includes a list of connection genes, and each connection gene contains in/out node number, connection weights, status bit to show the expression of gene, and an innovation number which is a tag to identify the particular gene. The evolution through the mutation and crossover of genes is tracked by historical marking which is a sequence composed of the innovation number of each gene in a genome. Historical marking enables the system to divide the population into species according to their topological similarity, and to analyze its topology easier when it crossover. Then, the fitness test is done within a species first, and then survived genomes from each species are tested to find global optimum in the entire population. Thus, NEAT can provide not only a solution for the problems of backpropagation neural network, but also an relatively efficient computation algorithm. Furthermore, it can give more biologically compatible neural network structure through complexification.

Conclusion

The conventional system does not reflect the plasticity of the brain as well as feedback from the upper level such as the feedback from the inferior colliculus to the outer hair cell through the cochlear nucleus. Thus, if those components can be added to the conventional model, improved performance is expected. The decision block in the conventional system will be composed of a three-layer backpropagation neural network and an evolutionary computation to provide plasticity and feedback in the system. The block diagram of the suggested system is in Fig.3.

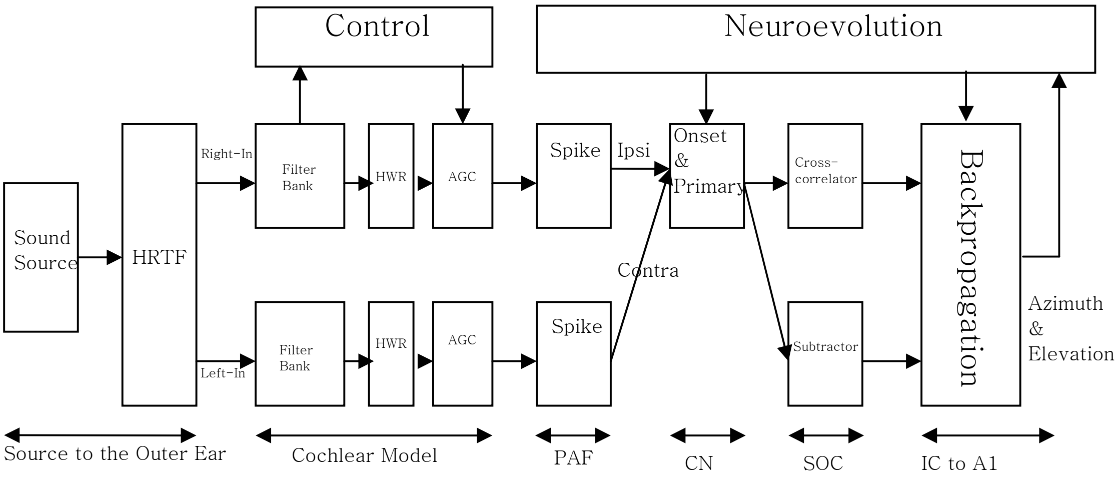
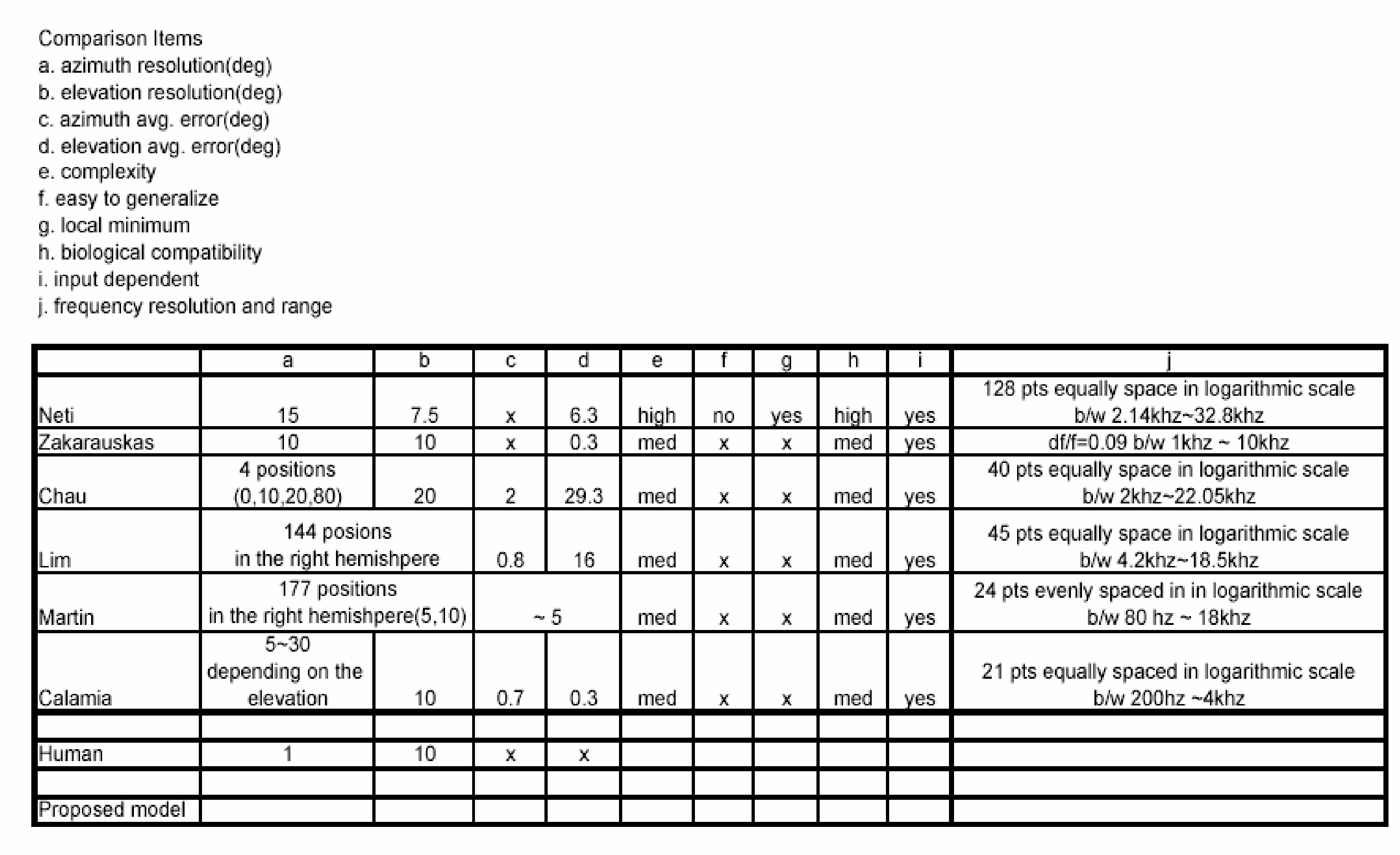


Fig. 3. A proposed sound localization system block diagram

References

[1] B Grothe, "New roles for synaptic inhibition in sound localization." *Nature Rev Neurosci.* vol. 4, pp. 540-50, July, 2003.
[2] C. Lim and R. O. Duda, "Estimating the Azimuth and Elevation of a Sound Source from the Output of a Cochlear Model," in *Proc. IEEE Asilomar Conference on Signals, Systems, and Computers,* Pacific Grove, CA, 1994.
[3] C. Neti, E. Young, and M. Schneider, " Neural network models of sound localization based on directional filtering by the pinna," *J. Acoust. Soc. Am*. vol.92, pp.3140-3156, August, 1992.
[4] J. Blauert, *Spatial Hearing,* MIT Press, Cambridge, MA, 1983.
[5] J. C. Middlebrooks and D. M. Green, "Sound Localization by human listeners," *Annual Review of Psycholgy*, vol. 42, pp. 135-159, 1991.
[6] J. O. Pickles, *An Introduction to the Physiology of Hearing*, Academic Press, London, 1988.
[7] K. D. Martin, "Estimating azimuth and elevation from interaural differences," Proc. 1995 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, 1995
[8] K. O. Stanley and R. Miikkulainen,. "Efficient Evolution Of Neural Network Topologies," in *Proc.IEEE Congress on Evolutionary Computation* , pp.1757-1762, Piscataway, NJ, May, 2002.
[9] L. A. Jeffress, "A Place theory of sound localization," *J. Comp. Physiol. Psychol.*, vol. 41, pp. 35-39, September, 1947.
[10] P. Zakarauskas and M. S. Cynader, "A computational theory of spectral cue localization," *J. Acoust. Soc. Am.,* vol. 94, pp. 1323-1331, September, 1993.
[11] P. T. Calamia, *"Three-Dimensional Localization of a Close-Range Acoustic Source Using Binaural Cues,"* Master's Thesis, University of Texas at Austin, Austin, Texas, 1998.
[12] R. F. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Proc. IEEE International Conference Acoustics Speech and Signal Processing*, Paris, France,1982.
[13] W. Chau and R. O. Duda, "Combined Monaural and Binaural Localization of Sound Sources," in *Proc. IEEE Asilomar Conference on Signals, Systems, and Computers,* Pacific Grove, CA, November, 1995.



Comparison Items
a. azimuth resolution(deg)
b. elevation resolution(deg)
c. azimuth avg. error(deg)
d. elevation avg. error(deg)
e. complexity
f. easy to generalize
g. local minimum
h. biological compatibility
i. input dependent
j. frequency resolution and range

	a	b	c	d	e	f	g	h	i	j
Neti	15	7.5	x	6.3	high	no	yes	high	yes	128 pts equally space in logarithmic scale b/w 2.14khz~32.8khz
Zakarauskas	10	10	x	0.3	med	x	x	med	yes	df/f=0.09 b/w 1khz ~ 10khz
Chau	4 positions (0,10,20,80)	20	2	29.3	med	x	x	med	yes	40 pts equally space in logarithmic scale b/w 2khz~22.05khz
Lim	144 posions in the right hemishpere		0.8	16	med	x	x	med	yes	45 pts equally space in logarithmic scale b/w 4.2khz~18.5khz
Martin	177 positions in the right hemishpere(5,10)		~ 5		med	x	x	med	yes	24 pts evenly spaced in in logarithmic scale b/w 80 hz ~ 18khz
Calamia	5~30 depending on the elevation	10	0.7	0.3	med	x	x	med	yes	21 pts equally spaced in logarithmic scale b/w 200hz ~4khz
Human	1	10	x	x						
Proposed model										

Table 1. Comparison among various models