An Auditory System Modeling in Sound Source Localization

Yul Young Park
The University of Texas at Austin
EE381K Multidimensional Signal Processing
May 18, 2005

Abstract

Sound localization of the auditory system is useful in the industrial and military applications such as game, sonar, sound quality measurement. Two main key features that the auditory system utilizes for sound source localization are interaural intensity differences (IID's) and interaural time differences (ITD's). Both become the cues for the estimation of the elevation and azimuth of the sound source. In this study, head related transfer function (HRTF) was used for the outer ear model and gammatone filter bank model for the cochlear. IID and ITD were extracted by subtracting and cross correlating the outputs from the both side cochlea. The elevation and azimuth were then estimated by a neural network using IID's and ITD's. The neural network complemented by evolutionary computation was proposed, and still under testing and revision.

Introduction

Identifying the location of an object generating an acoustic signal of the auditory system has several significant applications in the information processing systems. The object detection and its localization in the sonar application is crucial in the military applications, 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 equip 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 provide insight into 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 such as speech and music perception.

Background

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],[7].

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], [5], [6], [10].

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. An exemplary neural network model is a three-layer feedforward neural network with error backpropagation for the decision making block [3]. On the other hand, the probabilistic estimator models have either maximum likelihood estimator or nearest neighbor estimator as a corresponding block [2],[13],[14]. Except decision making block, both classes share many common sub-systems by and large although there are few variations in detail. The pathway from a sound source to pinna is modeled by head-related transfer function (HRTF). HRTF can be considered as a linear time invariant system that filters source signals and output the signals reaching ear drum. HRTF can vary with the frequency of a source signal,ω, and the source location containing azimuth θ, elevation ϕ, and range γ. For the convenience, range variable was ignored in this study and HRTF is represented by $H(\varpi, \theta, \phi)$ in frequency domain or $h(t, \theta, \phi)$ in time domain [2].

The cochlear function is usually modeled by filter bank which is made of a set of constant-Q band pass filter, half wave rectifiers, and post filtering parts. The function of

the basilar membrane, the inner hair cell transduction, and neural adaptation are modeled by those components, respectively [4]. The outputs of the cochlear model are neural signals on the auditory nerve which contain IID and ITD information.

Experimental Setup

i) HRTF's

The HRTF's by a KEMAR dummy head microphone measurement which is freely available on the internet was used [16]. The impulse response of the system was generated by using maximum length (ML) pseudo-random binary sequences at sampling frequency 44.1 kHz. It contains total 710 points source locations which range over elevations from -40° to +90° with 10° sampling of elevation and have 56, 60, 72, 72, 72, 72, 72, 60, 56, 45, 36, 24, 12, and 1 azimuth sampling points on each elevation. The range of source location is set 1.4m, and the length of HRTF is 512 sample points which correspond to 11.61msec time interval. When we assume normal head radius 9cm and 340m/sec sound speed, the maximum ITD is about 690usec, and 11.61msec period is enough long to accommodate ITD's. To reduce computational complexity, the 9 out of 710 locations were first tested to check the functionality of the proposed system.

ii) The Cochlear Model

The ear model package by Laboratory of Acoustics at Helsinki University of Technology was downloaded and modified for this model [15]. The directional filtered signal by pinna, HRTF, is scaled to 60db SPL and then transmitted through the ear canal to the middle ear. The transmission is modeled by filtering the scaled HRTF using the maximum audible field (MAF) threshold. Next, the cochlear function was modeled by the

consecutive action of gammatone filter bank, half wave rectifier and post processing [4], [15]. Overall block diagram is Fig.1.

Fig.1. Ear model Block diagram

The outcome of the model is frequency-time representation of the auditory nerve signal. 64 channels of filter bank and 512 points sampling data give a 64 by 512 frequency-time pattern data for each location which is used as an input to localization system.

iii) Localization System

Localization system first extracts IID and ITD from the frequency-time pattern data of previous stage (Fig.2). A simple subtraction of the left ear frequency-time pattern from the right frequency-time pattern gives IID, and it contains spectral features on each elevation and partial ITD information in its pattern. ITD information was extract from the cross-correlation of left and right HRTF data directly instead of using frequency-time pattern for better estimate of azimuth.

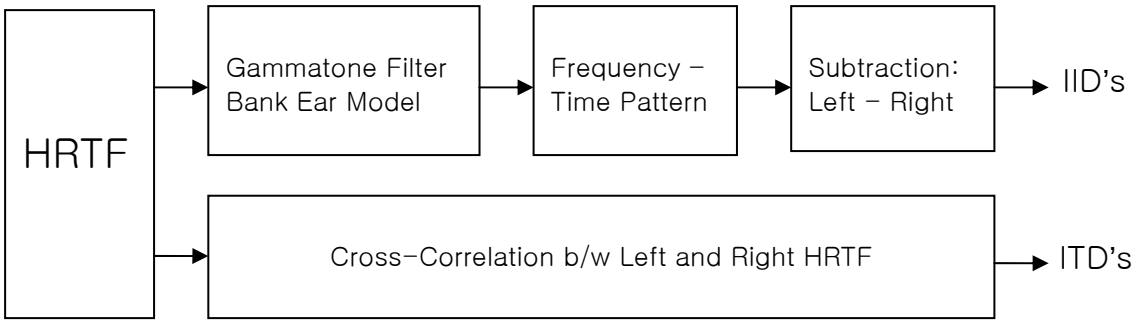

Fig.2. IID's and ITD Extraction

Decision making block is a neural network complemented by evolutionary computation. A three-layer feedforward neural network with backpropagation was implemented and used. To input 64 channel data to the network, 64 input units were used with 64 hidden units and 4 output units as first trial for 9 location data. Among 64x512 data in a frequency-time pattern, 64x32 data was selected to reduce the input data size to save computational complexity. This corresponds to 1/32 down sampling in time domain. Then, 9 location data were concatenated into 64x288 size data, and it is presented to the neural network input. With given input, the backpropagation network was test at several learning rates, momentum, and maximum square error values. A neuroevolution package by Neural Network Research Group as the University of Texas at Austin was tried to be adapted to make the neural network evolved [17].

Experimental Results

Disappointedly, although more simplified model and reduced data set was used, the system couldn't be finished. HRTF data and frequency-time pattern were obtained and shown to give necessary IID and ITD information as in Fig.3.

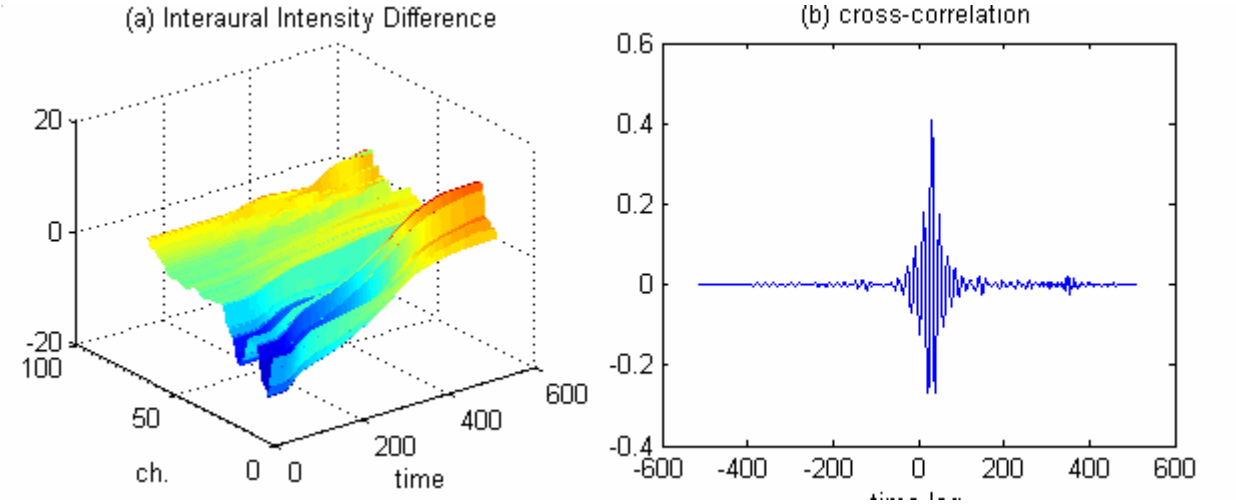

Fig.3. Network input signal of source location at elevation 0° and azimuth 45°. (a) IID by subtraction of left and right side frequency-time pattern (b) ITD by cross-correlation of left and right HRTF's.

Network training results cannot be generated due to failure of training. Before the network is made to evolved, it was trained with fixed condition for test. However, with the given training data, the computation was too huge to produce a solution, even a local minimum. Just, 100 iterations with 0.01 MSE took more than 3 hour simulation and could not converge to a local minimum.

Discussion and Future Work

If this training is deployed on the 150 initial populations, the linear estimation of simulation would be 3x150 hours for nothing, which should be avoided. Thus, the reduction of input data dimension seems to be critical. If the input data dimension is reduced, it causes smaller input unit size and corresponding hidden unit size. Then, the reduced system may relieve the system of huge amount of computation. Probably, a data clustering method to the input data would help to reduce the input data dimension. At the same time, more simple fitness evaluation without backpropagation network may be

tried. Once this stage is cleared satisfactorily, the evolution of the network from initial population will be pursued.

Acknowledgments

First of all, I appreciate many suggestions and helps from Dr B. Evans. Dr. T. Kite gave me an initial direction and Mr. P. Calamia generously provides his HRTF data and source codes of his MS thesis.

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] C. J. Plack and A. J. Oxenham, “Basilar-membrane nonlinearity and the growth of forward masking,” *J. Acoust. Soc. Am*. vol.103, pp.1598-1608, March, 1998.
[5] J. Blauert, *Spatial Hearing*, MIT Press, Cambridge, MA, 1983.
[6] J. C. Middlebrooks and D. M. Green, “Sound Localization by human listeners,” *Annual Review of Psycholgy*, vol. 42, pp. 135-159, 1991.
[7] J. O. Pickles, *An Introduction to the Physiology of Hearing*, Academic Press, London, 1988.
[8] 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
[9] 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.
[10] L. A. Jeffress, “A Place theory of sound localization,” *J. Comp. Physiol. Psychol*., vol. 41, pp. 35-39, September, 1947.
[11] P. Zakarauskas and M. S. Cynader, "A computational theory of spectral cue localization," *J. Acoust. Soc. Am.*, vol. 94, pp. 1323-1331, September, 1993.
[12] 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.
[13] 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.
[14] 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.
[15] http://www.acoustics.hut.fi/software/HUTear, Ear model Package, Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, accessed on May 2005
[16] http://sound.media.mit.edu/KEMAR.html, HRTF data, MIT Media Lab, MIT, accessed on April 2005
[17] http://nn.cs.utexas.edu/, Neuroevolution Package, Neural Network Research Group, The University of Texas at Austin, accessed on April 2005