MODELLING OF HUMAN DIRECTIONAL AND SPATIAL HEARING USING NEURAL NETWORKS

Juha Backman and *Matti Karjalainen*

Helsinki University of Technology, Acoustics Laboratory
Otakaari 5A, SF-02150 Espoo, Finland

ABSTRACT

The modelling and simulation of human directional and spatial hearing has been a difficult task due to the complexity of the auditory system. Although based on "built-in" mechanisms, sound localisation by humans is greatly improved by the rich experience of listening to all kinds of sound sources in various acoustical environments. In this paper we introduce a new way to model and simulate human directional hearing by using artificial neural networks in order to train the desired directional behaviour. Experimental results show that a combination of a dummy head, an auditory preprocessor, and a neural network may learn directional discrimination that in simple cases outperforms human listeners, showing also some ability of generalization. As an introduction we formulate an approach to this problem in general, including possible and potential applications.

1. INTRODUCTION

Detailed modelling and simulation of human directional and spatial hearing is a difficult task due to the complexity of the auditory system. The geometry of the head and external ear of an individual listener has considerable effect on the neural processing that is needed to create accurate directional and spatial percepts [1]. This ability, although based on general "built-in" auditory mechanisms, must be acquired from a rich experience of listening to all kinds of sound sources in various acoustical environments. Thus it is attractive to use methods with a learning ability, such as neural networks, when modelling these phenomena by computational means.

There exist relatively few attempts towards modelling of binaural directional and spatial hearing by computational means (see e.g. [1], [2], [3]). The problem is the same as in advanced modelling of human perception in general: it has been extremely hard to find explicit formulas and rules to describe the complex nonlinear behaviour of these systems. In this paper we approach the problem from the perspective of artificial neural networks that reveal new possibilities to develop advanced models of directional and spatial hearing that are able to learn desired forms of behavior. We first discuss the general framework of the topic, possible approaches and model implementations, as well as interesting applications of such models. In the experimental part of the work we report the results achieved so far, primarily related to directional hearing.

2. FORMULATION OF THE PROBLEM DOMAIN

The human binaural hearing system is a two-channel acoustical signal analyzer that in proper conditions has the ability to separate sound sources and to analyze these sounds in detail, including the estimation of the direction and distance of a source and the properties of an acoustical environment. In free-field conditions (e.g. anechoic chamber) and with a single sound source the detection of source direction is based on two main factors: the inter-aural delay (± 0.7 ms) due to a difference in distance to each ear and a level difference (up to about 20 dB) due to the shadowing by the head of the ear opposite to the source [1].

For the purposes of this study we may divide the auditory system into three subsystems (fig. 1):

1) external ear (source -> ear drum). This is the part of the system that is dependent on source location.
2) auditory model (middle and inner ear filter bank), and directional feature analysis
3) direction estimator (neural network in our case)

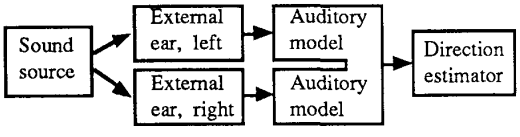

Fig. 1. *Block diagram of an artificial directional listener*

When modelling directional hearing in the form of an artificial directional listener the external ear can be realized as a dummy head. The auditory model may be implemented as a critical-band (Bark-band) filter bank or an equivalent formulation. Directional feature analysis within the auditory model can be of various forms. In our experimental study we used the inter-aural cross-correlation function and a vector of inter-aural level ratios computed from critical bands, as described below.

2.1 General aspects and the scope of this study

Our experiments described here were concentrated on verifying the validity of the model in relatively simple conditions. To probe further into the problem we have to address the following questions:

- A proper time window must be determined where the short-term analysis will be carried out. A strategy of processing must

0-7803-0946-4/93 $3.00

be determined for cases where the source properties, the direction, or some environmental factors are changing in time. Here we will discuss only the steady-state case.

- The fusion or separation of sources happens when several simultaneous sound sources coexist. An important special case is a single source disturbed by (often diffuse) background noise. If separation happens, it is based on a complex analysis of numerous sound features.
- In practical acoustical environments delayed versions of the source signal are received from various directions through reflections and reverberation. The precedence effect (or Haas effect) emphasizes the importance of the direct sound in source localization. Although we did not include this kind of temporal effect to our auditory model some of our experiments demonstrate the ability of the simple model to work in reverberant rooms.
- The distance of the source cannot be estimated reliably in either free-field or reverberant conditions even by humans [1].
- The human auditory system may learn to evaluate complex spatial features and attributes of the source and the acoustical environment, such as the quality attributes of a loudspeaker, the size and form of a room, parameters describing the quality of a concert hall, etc. In computational modelling of such auditory processes the use of learning systems like neural networks is advantageous because any psychoacoustical dimension that reveals consistent behaviour may be used as an output parameter.

2.2 Potential applications of binaural models

In addition to the understanding of basic principles the computational models of binaural hearing reveal several potential applications. Here we mention a few of them:

- direction-selective microphones and speech recognizers,
- environmental orientation aids for the hearing disabled,
- robots and alarm systems reacting to sound environment,
- evaluation of quality parameters of audio systems, especially loudspeakers and effects of listening rooms and the stereo image [2],
- evaluation and measurement of room and concert hall acoustics, and
- binaural estimation of speech intelligibility.

3. THE BINAURAL AUDITORY MODEL

Advanced computational models of the human auditory system have been used e.g. in several monaural speech processing tasks. Similar principles are used also in modelling directional hearing [2], [3]. We have included only the most fundamental aspects in a form that is understandable from the signal processing point of view.

3.1 Inter-aural phase and delay

The horizontal direction of the sound source will affect the time difference of the signals received by the left and the right ear. Due to the geometry of the external ear and pinna this delay is slightly frequency dependent. We considered three candidates to represent this information in our model:

- the direct inter-aural cross-correlation,
- the set of inter-aural correlations from a critical-band filter bank model, and
- the set of interaural delay differences from a filter-bank.

For an orthogonal filterbank it is easy to show that the cross-correlation of the left and right ear signals x(t) and y(t) is equal to the sum of sub-band cross-correlations:

Assume that $x(t) = \sum_i x_i(t)$ and $y(t) = \sum_i y_i(t)$, where i is the filterbank channel index. Then

$$\mathbf{corr}\{x(t),y(t)\} = \mathbf{F}^{-1}\{\mathbf{F}\{\mathbf{corr}\{x(t),y(t)\}\}\}$$
$$= \mathbf{F}^{-1}\{X(j\omega)\cdot Y(j\omega)\} = \mathbf{F}^{-1}\{\sum_n X_n(j\omega)\cdot \sum_m Y_m(j\omega)\}$$

where operators are: F{} is the Fourier transform, $F^{-1}\{\}$ the inverse Fourier transform, and corr{} the cross-correlation. X() and Y() are Fourier transforms of x(t) and y(t).

Due to orthogonality $X_n(j\omega)\cdot Y_m(j\omega) \equiv 0$ for $n \neq m$. Then

$$\mathbf{corr}\{x(t),y(t)\} = \mathbf{F}^{-1}\{\sum_i X_i(j\omega)\cdot Y_i(j\omega)\}$$
$$= \sum_i \mathbf{F}^{-1}\{X_i(j\omega)\cdot Y_i(j\omega)\} = \sum_i \mathbf{corr}\{x_i(t),y_i(t)\}$$

Thus the cross-correlation function of the left and right channel signals represents in a more compact form the same information as the set of band-pass correlations. It is also evident that the analysis of critical-band delay differences (24 Bark channels), e.g. if based on the time position of maximum point in the cross-correlation, may lead in complex cases to more unreliable information than the entire cross-correlation function. Due to this, we decided to use the direct cross-correlation between the left and right ear signals, limited to within a time span of -1.0 ms to +1.0 ms, as a feature vector for phase and delay difference information. The sampling frequency used in our experiments was 22050 Hz that allows for signal bandwidth of 10 kHz and leads to a 45 element cross-correletion vector, denoted here C[].

The second aspect of directional information, the level difference between left and right channels, should be based on the amplitude ratios of critical-band filterbank channels. We used a previously developed auditory model [4] that is computationally efficient and suitable for the purpose. The algorithm for a single ear channel is

- Fourier transform of a Hamming windowed signal
- Power spectrum warping to Bark scale
- Convolution of the Bark-scaled power spectrum by a spreading function (masking pattern) [5]
- Equal loudness sensitivity correction (this does not have any effect here because only amplitude ratios are used)
- Vector L[] of 24 Bark-channel loudness values

Finally the loudness ratios for each Bark channel pair are computed by the function:

$$R_i(L_{left},L_{right}) = (L_{i,left} - L_{i,right}) / (L_{i,left} + L_{i,right}),$$

that maps loudness ratios to the range [-1.0 1.0]. It is used to return a 24 channel loudness ratio vector R[].

4. NEURAL NETS FOR DIRECTION ESTIMATION

Multilevel feedforward nets [6] with the backpropagation training algorithm are good candidates for direction estimation in our model. They are suitable for function approximation when the output nodes may take on continuous values.

To guarantee the best continuity of mappings it is desirable to use sines and cosines of the sound source direction angles (instead of the angles themselves) to represent the directional information. This requires that the output layer of networks must be

modified from the standard form so that the value range [-1.0 1.0] is supported and convergence of the training algorithm is possible. We replaced the sigmoidal nonlinearity by hyperbolic tangent for the computation of output activations and replaced the normal error expression $a(t\text{-}a)(1\text{-}a)$ by a linear term $(t\text{-}a)$, where a is the activation of an output node and t is the corresponding target value. Network dimensionalities were:

- Input: 69 total, 45 correlation values, and 24 loudness ratios
- Output: 4, sines and cosines of the horizontal and vertical angles of the source direction (2 if only vertical or horizontal)
- One hidden layer: 2, 4, or 8 nodes

5. EXPERIMENTAL SETUP

The data used in the experiment was recorded using a Neumann KU 80 i dummy head. A loudspeaker (coaxial two-way, specially designed for this experiment, on-axis response within ±1.5 dB from 100 Hz to 11 kHz) was mounted on a support so that the distance between the loudspeaker and the center of the line going through the ears of the dummy head was a constant 2 m, and both the vertical and horizontal angles between the dummy head and the loudspeaker could be varied easily and with an accuracy of better than ±0.5°. For practical reasons, the vertical angles were limited to between -20° and +90° from the horizontal plane.

The test signals were approximately one-second samples of white noise, pink noise, 100 µs pulses repeated at 50 ms intervals, two music samples (acoustic guitar, drums), and a one-word speech sample. The signals were reproduced and recorded directly from and to a computer hard disk using 16-bit quantization and a 22.05 kHz sampling frequency.

The recordings were made in an anechoic chamber, and in a hallway with a moderate reverberation time with two different dummy head positions. In these recordings, the vertical angle was first kept at 0°, and the horizontal angle was varied from 0° to 355° by 5° intervals; for the second series of recordings, the horizontal angle was 0°, and the vertical angle was varied from –20° to 90° at 10° intervals. Finally, both the vertical and horizontal angles were varied by 30° steps, with some additional samples at intermediate angles.

6. EXPERIMENTS

The first experiment examined the results of training the network with various combinations of signals recorded in the horizontal plane. We examined the average and the maximum value of the absolute error as a function of the number of nodes and the number of training iterations. The data used both in the training and in the evaluation of the network were identical. The results can be summarized as follows:

1. Increasing the number of the nodes in the hidden layer increases the performance of the system under all conditions. Apparently even higher number of nodes than the maximum of our experiments, 8, would have been beneficial with larger amounts of training data (fig. 2).

2. If the number of the nodes was high enough, very good results were to be obtained; the best average errors were around 0.1-0.2 degrees after 8000 training iterations with 8 nodes, and apparently with longer training the errors could have been reduced still further. A typical example of the estimated vs. actual angle of incidence for different signal types in anechoic conditions is shown in fig. 6.

3. Including more different types of test signals in the training data decreased the performance of the network, and significant differences between the various types of signals were apparent. The best localization results were to be obtained using noise or pulses, and music and speech yielded significantly worse results with the same amount of training (fig. 3). Similarly, anechoic data yielded the best results, although the difference to results obtained with reverberant data only was relatively small, but the learning results obtained using both anechoic and reverberant data were significantly worse (fig. 4).

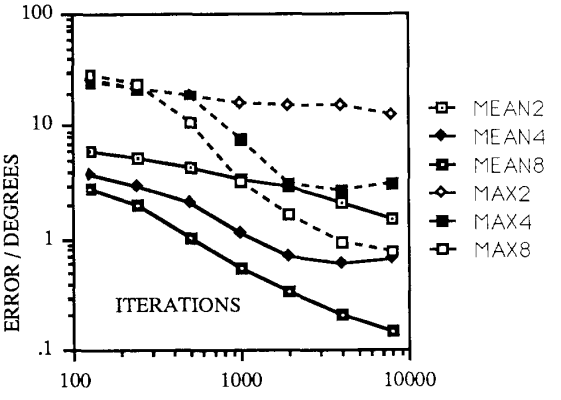

Fig. 2. *Mean and maximum error of learning directional hearing in anechoic environment as a function of number of iterations and number of hidden nodes.*

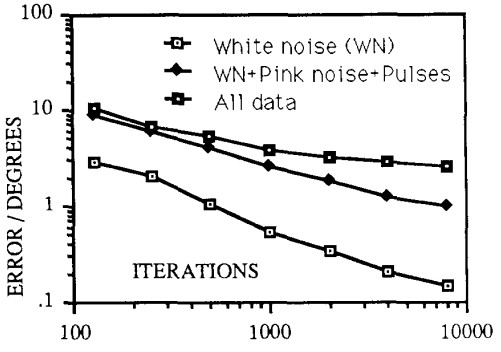

Fig. 3. *Mean error of learning in anechoic environment as a function of number of iterations and number of excitation signal types (8 hidden nodes).*

4. Using both correlation and loudness ratio data improved the results over the situations in which only one of these was used. Interestingly, in anechoic conditions the cross-correlation only yielded better results than using only the loudness ratio, but in reverberant conditions using both correlation and ratio data yielded worse results than using only ratio data (fig. 5).

The second experiment examined the ability of the network

to generalize the localization to data not present in the original training set. This was tested in three ways: leaving some angles out from the training set, and testing the ability of the network to interpolate to these angles, leaving some signal types out from the training data, and performing the evaluation with data recorded in different acoustical conditions than the training data. Increasing the number of the nodes improved the results, but in some experiments increasing the amount of training iterations degraded the performance with a high numbers of the hidden nodes, as the network apparently started to learn the different angles of incidence as distinct special cases.

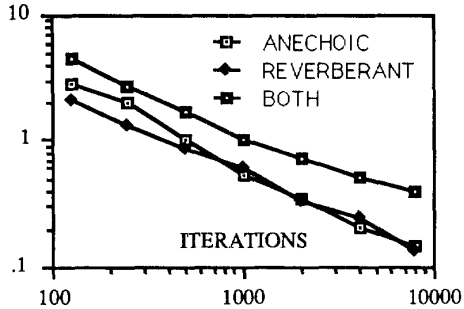

Fig. 4. *Mean learning error of a 8 hidden node network with white noise data from anechoic room, reverberant room, and both rooms.*

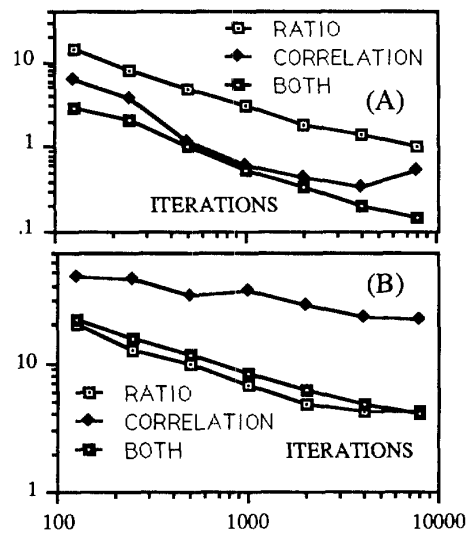

Fig. 5 *Mean learning error with ratio data, correlation data, and both feature vectors as a function of iterations in (A) anechoic room with white noise and (B) in reverberant room with all excitations.*

Our third experiment studied the ability of the network to localize in both vertical and horizontal directions. For 0° horizontal angle and with the vertical angle as the only variable, the accuracy of localization was comparable to the results obtained in the horizontal plane, i.e. average error was around 0.1°. When both the horizontal and vertical angle were variable, the average error increased to approximately 1°, but when considering the complexity of the problem, this result can be regarded to be rather good.

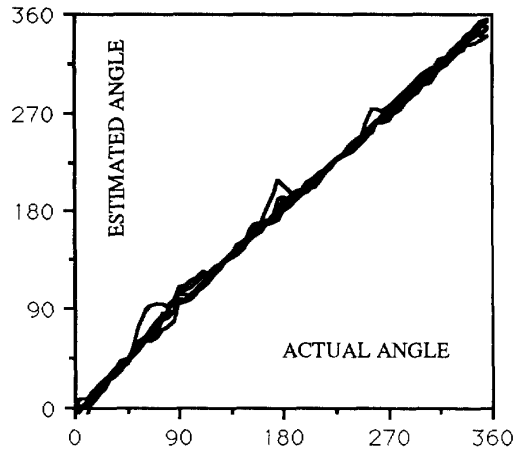

Fig. 6. *An example of estimated angle as a function of actual angle with data from an anechoic room,using all the different signals.*

7. DISCUSSION AND CONCLUSIONS

We have demonstrated that estimating the direction of a sound source from binaurally recorded material using an artificial neural network is possible in different acoustical conditions. At best, the system described in our work has an accuracy exceeding human directional hearing, but although the system has demonstrated some ability to generalize, in that respect the performance is still rather poor. To improve the system, we apparently need more advanced preprocessing methods resembling more closely the human auditory system.

REFERENCES

[1] J. Blauert, Spatial Hearing. MIT Press, 1983.

[2] E.A. Macpherson, "A Computer Model for Binaural Localization for Stereo Imaging Measurement," J. Audio Eng. Soc., Vol. 39, No 9, Sept. 1991.

[3] Richard F. Lyon, "A Computational Model of Binaural Lo - calization and Separation," Proceedings of ICASSP 83, vol. 3, pp. 1148 - 1151.

[4] Matti Karjalainen, "A New Auditory Model for the Evaluation of Sound Quality of Audio Systems," Proceedings of ICASSP 85, vol. 2, pp. 608 - 611

[5] M.R. Schröder, "Objective Measure of Certain Speech Signal Degradations Based on Masking Properties of Human Auditory Perception," in Frontiers of Speech Communication Research (ed. Lindblom&Öhman), Academic Press 1979

[6] Proceedings of the IEEE, Special Issues on Neural Networks, I: theory and modeling, September 1990, II: mantementical andlysis, related topics, implementations, and applications, October 1990.