NEURAL NETWORK APPROACH TO ANALYZE SPATIAL SOUND

KALLE PALOMÄKI, VILLE PULKKI AND MATTI KARJALAINEN

Helsinki University of Technology, Espoo, Finland
Kalle.Palomaki@hut.fi

Self-organizing maps (SOM) and multilayer perceptron (MLP) neural network approaches are applied to the evaluation of spatial discrimination of real and virtual sound sources. Neural networks are trained with localization cues computed using a binaural model. The ability of the models to simulate human perception of spatial sound is analyzed. Both SOM and MLP showed relatively good ability for generalization when tested with test data from real sources not included in the training data set. Both models were also capable of describing localization blur of virtual sources though still further analysis is needed to find out how well our models correspond to real human auditory localization. The motivation of this study has been to search for methods and techniques to evaluate the quality of spatial sound reproduction.

INTRODUCTION

Objective methods for sound quality measurements simulating human auditory perception have been developed for many years. Such methods for speech and audio quality assessment, originally proposed in [1] and [2], were further developed for example in [3],[4] and [5]. However, those models are monaural and the spatial attributes of perception are neglected. Little has been done to model the perception of spatial sound quality. Modeling of binaural hearing [6, 7, 8, 9, 10] has proven to be difficult since it includes very difficult subproblems such as the precedence effect. Sound localization and spatial discrimination is a complex process, taking place at medium to high neural stages. An exact knowledge of the physiological mechanism behind it does not exist yet.

Artificial neural network approaches, in addition to a binaural model [11], are experimented in this study in order to evaluate spatial discrimination of real and virtual sound sources. Types of networks used are self-organizing maps [12] (SOM, unsupervised learning) and multilayer perceptrons with back-propagation learning algorithm (MLP, supervised learning) [13]. Neural networks are trained with localization cues from the output of a binaural auditory model. The auditory model approximates frequency-dependent interaural time and level difference cues. Different sound source locations are simulated using HRTFs of KEMAR artificial head measured from 710 source directions [14]. Other studies related to neural network modeling of binaural localization are found for example in [15, 16, 17, 18, 19].

1. SOUND LOCALIZATION

The auditory localization of a sound source is influenced by several acoustical cues, such as interaural time differences (ITD) and level differences (ILD) and spectral cues of pinna filtering [20]. In ordinary binaural listening conditions ITD and ILD have usually more importance than spectral cues. When the distance from sound source to ears differs, level and time differences occur between ears. ITD is simply the propagation time difference between sounds arriving to listener’s ears. The shadowing effect of head causes differences of sound pressure levels between ears (ILD) whereby the level is higher at the ear which is on the side closer to the sound source.

ITD and ILD values vary as a function of frequency. For the auditory localization ITD has importance primarily at frequencies below 1.5 kHz and ILD above that. The shadowing effect of the head is strong when the wavelength of sound is close to or shorter than the dimensions of the head. Therefore ILD values are much higher at high frequencies (above 1.5 kHz).

In the median plane where the distance from a sound source to both ears is the same, both ILD and ITD are zero. In this situation spectral cues or head movements have importance when estimating the elevation or front-back discrimination of the sound direction. Spatial discrimination is also difficult in so called cones of confusion where ILD and ITD values vary only slightly, due to the asymmetry of the head. Cones have a symmetry axis passing through the listener’s ears. The region where ILD and ITD values are zero forms a confusion cone equal to
In the simplest form single ILD and ITD values can be calculated for the whole signal bandwidth. However, a more advanced method is to calculate those values for several frequency bands from the outputs of a peripheral auditory model, where the frequency and time resolution of the auditory system is approximated.

2. BINAURAL MODEL

The binaural model of this study has originally been proposed in [11]. It includes a peripheral auditory model of both ears. The model consists of six functional stages corresponding to reasonable functional or physiological parts of the auditory system. The model is shown in Figure 1 and explained below.

1. The filtering effect of the external ear for different sound source locations is simulated with head-related transfer functions (HRTF) from KEMAR artificial head measured from 710 different sound source directions [14]. The spherical surface around KEMAR is discretized to 14 elevations from $-40^\circ$ to $90^\circ$ with equal size $10^\circ$ increment. At each elevation a full $360^\circ$ was sampled in equal sized azimuth increments which are $5^\circ$ near the lateral plane ($-20^\circ$ to $20^\circ$ elevations) and more than $5^\circ$ elsewhere. The HRTF data is freely available at MIT FTP-site [21].

2. The cochlear frequency resolution and selectivity is modeled with a gammatone filter-bank of 32 band-pass equivalent rectangular bandwidth (ERB) [22] channels.

3. The inner hair-cell (IHC) and auditory nerve firing rate is simulated with half-wave rectification and 1 kHz low-pass filtering in each gammatone band-pass channel.

4. ITD values for each channel are estimated using
interaural cross-correlation (IACC) function [6]. Typically the position of maximum in a single IACC curve indicates the ITD at the frequency corresponding to the ERB channel pair.

5. ILD values are calculated as a difference of loudness levels (in phons) between each IHC-channel.

6. At final stage ILD and ITD data from several directions are analyzed with a neural network.

No precedence effect has been included. The model is described in more detail in [11].

3. FORMING TRAINING DATA SETS FOR SOM AND MLP MODELS

Since the localization data from the binaural model is analyzed with neural networks, it is important to find a reasonable representation for training and testing the networks. This section explains the principles how the training data sets of both SOM and MLP approaches are created from the output of the binaural model.

The binaural model outputs 32 ITD and ILD values which are each calculated from corresponding ERB band-pass channels pairs for an audio signal entering the listener's ears. For network training ILD and ITD data is combined to training vectors, which correspond to a specified sound source directions.

Since ITD has most of its importance in the localization at low frequencies, only 16 ITD values from low frequency channels (below 2.5 kHz) are used. ITD part $d_{az,ele}$ of a training vector is formed as

$$d_{az,ele} = [ ITD_{1}^{az,ele} ITD_{2}^{az,ele} ... ITD_{16}^{az,ele} ]$$  

where $ITD_{ch}^{az,ele}$ is a single ITD value from an output channel of the binaural model. Indices $az$, $ele$ and $ch$ correspond to azimuth, elevation, and output channel of the model.

Respectively, ILD values are used only for high frequency channels from 14 to 32 (1.9 kHz to 20 kHz). Values of consecutive even and odd ITD channels are summed in order to decrease the amount of training data. The ILD part $i_{az,ele}$ of a training vector is formed as

$$i_{az,ele} = [ ILD_{14+15}^{az,ele} ILD_{16+17}^{az,ele} ... ILD_{31+32}^{az,ele} ]$$  

The scalings of ILD and ITD differ substantially. Therefore it is important to normalize the scaling between them to guarantee that they both have balanced weight to the learning process. In practice the normalization guarantees that both ILD and ITD have zero mean and mean of their absolute value is the same with both. ITD and ILD vectors with normalized scaling are denoted as $i_{n}^{az,ele}$ and $d_{n}^{az,ele}$. The composite training vector $x_{az,ele}$ is a concatenation of those,

$$x_{az,ele} = [ d_{n}^{az,ele} i_{n}^{az,ele} ]$$  

This kind of learning data was used with MLP experiments.

However, experiments with SOM showed that the ILD and ITD data from single events with one head position did not carry enough information on elevation and for front-back discrimination. Therefore the effect of "head rotation" was added to the training data. It is actually modeled so that the listener receives localization cues with two head directions combined (see Figure 2) and there is no real rotation. Each training vector contained two vectors of definition (3) from differing azimuth between them. The training vector with rotation effect $x_{rot}$ is given by

$$x_{rot}^{az,az+rot;ele} = [ x_{az,ele} x_{az+rot,ele} ]$$  

where $rot$ denotes the rotation angle between two vectors to clockwise direction.

4. ANALYZING LOCALIZATION CUES WITH SOM NETWORKS

The self-organization process of simulated ILD and ITD data is addressed in this section. The theory of self-organization is reviewed shortly in section 4.1 and the experiments are addressed in section 4.2.

4.1. Self-organizing map

The SOM is an efficient neural network tool for visualization and abstraction of high-dimensional data [12]. It converts complex nonlinear statistical relationships between data items into simple geometric relationships on a low-dimensional representation or display. Due to the self-organization process, cells of the network become
236

The self-organization process tunes model vectors to match well to the input patterns [23]. Model vectors are organized so that similar ones are closer to each other than more dissimilar ones.

The learning process does not affect cells independently of each other but as topologically ordered sets. Same kind of correction is imposed in similar kind of cells which are topologically ordered close to each other.

To determine the interaction between neighboring cells, neighborhood set \( N_c \) is defined. At each learning step all cells within a neighborhood set are updated where as cells outside are left intact. The neighborhood is centered around the best matching unit \( m_c \), which is the cell of best match with \( x \). It can be formulated as:

\[
||x - m_c|| = \min_i ||x - m_i||. \tag{5}
\]

The width of neighborhood can be defined with a time varying neighborhood function. Experimentally it has turned out to be preferable that neighborhood \( N_c \) is wide at the beginning of training and shrinks monotonically with time [23].

The updating process of model vectors \( m_i \) corresponding to training vectors \( x_t \) is given by,

\[
m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \tag{6}
\]

where \( t \) is a sample interval of the regression process and \( h_{ci}(t) \) is a neighborhood function which usually decreases with respect of time. It also decreases with increasing distance from the best matching cell unit. A commonly used neighborhood function is Gaussian,

\[
h_{ci} = \alpha(t) \exp(-||r_i - r_c||^2/\sigma^2) \tag{7}
\]

where \( 0 < \alpha(t) < 1 \) is a learning rate factor and \( \sigma(t) \) corresponds to the width of neighborhood. Both are decreasing functions of time.

### 4.2. Self-organization of the localization cues with head rotation

In this section experiments with self-organization of the localization cues is addressed. An objective of the experiment was to investigate the ability of SOM to organize localization cues so that the map shows illustratively the relationship between map position and sound source direction.

In our approach a SOM was trained with localization cues estimated for pink noise real source from 710 directions around KEMAR artificial head. Those directions cover azimuth direction from 0° to 360° for elevations from -40° to 90° (see section 2). Learning vectors were created as explained in section 3 and definition (4). Each learning vector includes cues from a single real source with two head directions so that a listener receives cues combined from two directions (see Figure 2). These two head directions are meant to model the contribution of "head rotation" to localization. Without head rotation SOM organized elevations very poorly. It is important to notice that without head rotation or more generally without head movements a human listener cannot avoid confusions in elevation. Comparing the model with "head rotation" to a listening situation where a listener is restricted not to move the head the model predicts clearly too fine elevation discrimination.

Several head rotation angles were experimented. Head rotation of 30° provided enough information for avoiding front-back confusion inside confusion cones. Smaller angles of rotation didn't gave as perfect results. Wider angles, from 30° to 90° provided good front back discrimination, but since such a large head rotation is somewhat unnatural, 30° was chosen for this study.

Figure 4 shows a map labeled with azimuth angles. The lattice of the map is hexagonal and the grid of cells is a 6x23 rectangle, indicating the relative distances of the model vectors. An average value of the two directions in an individual learning vector is used as label. Positive azimuth angles organize on the right side of the map and negative on the left side. Azimuths from -90° to 90° are on the top half of the map and azimuth from -180° to -90° and 90° to 180° are on the bottom half.

Figure 5 shows the same map labeled with elevations. The most negative elevations organize to near the edges of the map and the most positive elevations near the center.

Figure 6 shows an oval-shaped 2-D projection of the map. The graph is oriented so that the diameter from up to down corresponds to the median plane, i.e., azimuth directions 0° and 180°. A longer diameter from left to
Figure 4: Organization of a SOM having 6x23 rectangular shape. Learning data includes the effect of 30° simulated head rotation. Cells of the SOM are labeled with azimuth directions of the best matching sound sources.

right corresponds to the plane where azimuth is close to 90° or -90°.

As seen in the projection the distances between SOM model vectors are much shorter in the median plane (vertical diameter) than from left to right (horizontal diameter). This is a result from the fact that ILD and ITD variations are small close to the median plane.

4.3. Tests with different types of stimuli

In the previous experiments the SOM was trained with cues evaluated from pink noise sound sources as stimuli. The generalization ability of the SOM model can be studied with test cue vectors which are analyzed from varying types of real and virtual sources. The performance of a SOM model responding to a test vector can be evaluated using a distance measure between SOM model vectors and a test vector. The distance should be smallest to the map position corresponding the direction of the test source.

A distance measure vector $d$ between SOM model vectors $m_n$ and the test vector $x_{test}$ can be formulated as,

$$d = \|m_0 - x_{test}\| \|m_1 - x_{test}\| \ldots \|m_n - x_{test}\|$$

where Euclidean distance measure is used. The distance in each case is approximated between original SOM model vectors and a single varying test vector. For presentation, distance vector $d$ can be plotted using the shape of the original SOM. In these experiments contour graph presentation is used.

4.3.1. Test with head rotation

Figure 7a shows the distance contour estimated between SOM models and a single test vector from the original...
training data set (i.e., pink noise input source). A test vector includes cues from a single real source for two directions: $-30^\circ$ az $0^\circ$ ele and $30^\circ$ az $0^\circ$ ele. Contours are labeled with the distance value of the corresponding map position. Smaller numbers mean closer match so that a distance value of 0 indicates perfect match.

For a comparison between different input stimuli a test vector is estimated using a vowel stimulus [a:] from the same direction as the previous one. The distance contour is shown in Figure 7b. The contours are very similar compared to Figure 7a. This indicates ability to generalize to signals with different spectra.

In Figure 7c the test vector is approximated from a virtual pink noise source which is produced with amplitude panning of two loudspeakers (in this case loudspeakers are modeled with two point sources). Loudspeakers are placed so that the angle between them is $60^\circ$ and so that the center point between the speakers is in the same direction as the real source of Figure 7a. They are both fed with exactly the same input signal. The real source has closer match to a single spot on the map than the virtual source. This means that the perceived direction of the virtual source is blurred compared to the real one.

Figure 7d shows a graph for a source with higher elevation angle. In this case pink noise source is in direction $30^\circ$ az $60^\circ$ ele.

4.3.2. Test without head rotation

The same SOM model which has been trained with head rotation can be tested also without head rotation. Therefore the distance measure of equation (8) has to be modified.

Actually each model vector of the SOM with head rotation consists of two halves where the vector dimensions of the halves are 25. The first half is tuned always according to head direction $-15^\circ$ and the second half according
Figure 6: Projection graph of the SOM of Figures 4 and 5.

Figure 7: Distance contour estimated between single test vector and all SOM model vectors when the test vector is analyzed from a) pink noise real source 0° az 0° ele (included in training data), b) a vowel real source 0° az 0° ele (not included in training data), c) pink noise virtual source 0° az 0° ele (not included in training data), and d) pink noise real source 30° az 60° ele (included in training data).

A neural network approach to analyze spatial sound.

To head direction of 15°. Thus any of the model vectors can be expressed as

\[ m_n = [m_n^{az-15°}, m_n^{az+15°}] \]  

where \( m_n^{az-15°} \) and \( m_n^{az+15°} \) are estimated according to head direction -15° and 15°. The distance measure with only one head direction can be formulated as

\[ d = \begin{bmatrix} ||m_0^{az-15°} - x_{test}'|| \\ ||m_1^{az-15°} - x_{test}'|| \\ ||m_n^{az-15°} - x_{test}'|| \end{bmatrix}^T \]

where \( x_{test}' \) is a test vector.

Figure 8 shows the distance contours without head rotation. Confusion cone regions are oriented along straight lines to 30° from vertical direction. This kind of tilt in the orientation occurs when same SOM model which has been trained with "head rotation" is tested so that "head rotation" effect is ignored. The observed change in tilt is clearly a consequence of ignoring the latter part of SOM model vectors in the distance calculation. Distance contours of constant value are spread up along confusion cone regions. In Figures 8a and b real and virtual sources from direction -5° az 10° ele are compared. The distance graph of the real source (Figure 8a) has two spots which have a distance value less than 1. The match is closest to the largest of those spots and thus it predicts perceived sound source direction. Another spot is a likely occurring confusion. With virtual source (Figure 8b) the largest spot noticed in Figure 8a has vanished. The spot having distance value less than 1 is now in location which corresponds to a larger value of elevation. This predicts that confusion on the elevation is likely to occur. The result was expectable since -5° az is close to the median plane where ILD and ITD vary little and thus they don't provide much information about elevation.

Figures 8c and d show distance graphs for direction...
−30° az 0° ele. Now the closest match region with virtual source is more clearly spotted to one direction and thus discrimination of elevation should be better than it was in the previous case.

5. MLP EXPERIMENTS

In this section an MLP approach of binaural localization cue analysis is addressed. The difference compared to the SOM approach is that the learning process is supervised, as is partially true in real human auditory localization where supervisory information is provided by vision. In the experiments the MLP output builds up an array where each single output cell corresponds to a separate sound source direction, i.e., spherical space around the head is discretized to these outputs.

5.1. Multilayer perceptron

The multilayer perceptron with backpropagation learning algorithm [13] is an efficient neural network solution for function approximation when output values as a function of inputs have a finite number of discontinuities.

Figure 9 shows a schematic diagram of the network structure used in the experiments. It is a three-layer perceptron network containing an input layer, a hidden layer, and an output layer. All of them consist of several neurons having specified properties. Each neuron has a number of inputs $x_1, ..., x_j$, a threshold $\theta_k$, and an output $y_k$. However, in the input layer neurons have only one input. In hidden and output layers each input of a neuron is typically connected to all of the outputs of the previous layer. The connections are associated with weights $w_{ij}$ which determine the strength of each connection. In a neuron the weighted inputs are summed to value $s_k$,

$$s_k = \theta_k + \sum_j w_{jk} x_j \tag{11}$$

which is then passed through a saturating activation function. In the input and hidden layer neurons of our networks it is the hyperbolic tangent sigmoid (see Figure 10a) function

$$y_k = \frac{2}{1 + e^{-2s_k}} - 1 \tag{12}$$

and in the output layer the logistic sigmoid (see Figure 10b) function

$$y_k = \frac{1}{1 + e^{-s_k}} \tag{13}$$

where $y_k$ corresponds to a neuron output. As seen in Figures 10a and b hyperbolic tangent function (in output layer) ranges between -1 and 1 whereas logistic sigmoid (in input and hidden layers) ranges between 0 and 1.

The connectivity of neurons is shown schematically in Figure 9. There are connection only in forward direction. Each input of the neuron is connected to all outputs of neurons in previous layer and each neuron outputs only a single value. The weights $w_{ik}$ together with thresholds $\theta_k$ (not shown in Figure 9) determine the properties of the network. With the error backpropagation algorithm they can be tuned so that they correspond to the desired input-output relation.

5.2. Directional distribution as desired output for MLP

In the MLP experiments the training data doesn’t include the effect of head rotation and a single learning vector includes cues estimated with only one head direction. The

Figure 8: Distance contours are estimated without head rotation for a test vector from a) training data, pink noise real source −5° az 0° ele, b) pink noise virtual source −5° az 0° elev (not included in training data), c) pink noise real source −30° az 0° (included in training data), and d) pink noise virtual source −30° az 0° (not included in training data).
output layer of the network has 145 model neurons which can be reorganized to form a 29x5 array representing a sector (-70° az to 70° and -20° to 20° ele) of the space around KEMAR. The desired output is approximately to represent the directional distribution of the sound sources or source combinations corresponding to the training vector. The form of distribution that we have used is originally presented in [15]. It is a simplified version of the representation of direction in the superior colliculus [24].

When a sound arrives from a particular direction the evoked response in the desired output is set to be a double Gaussian form centered on a point in the neuron array which corresponds to that particular direction [15]. Output activity \( d_{i,j} \) in model neuron \( i, j \) for a sound originating from the direction corresponding to model neuron \( k, l \) is

\[
d_{i,j}(k,l) = \exp \left\{ - \left[ \frac{(i-k)^2}{\sigma_k} + \frac{(j-l)^2}{\sigma_l} \right] \right\} (14)
\]

where \( i, k = 0, ..., 10 \) and \( j, l = 0, ..., 16 \)

of the network output can be expressed in matrix form as

\[
D = \begin{bmatrix}
    d_{1,1} & d_{1,2} & \cdots & d_{1,n} \\
    \vdots & \vdots & \ddots & \vdots \\
    d_{m,1} & d_{m,2} & \cdots & d_{m,n}
\end{bmatrix}
\]

(15)

In our experiments the output grid side is 29x5 which represents azimuth values from -70° to 70° and elevations from -20° to 20°. This means that distance between consecutive output cells is 5° in azimuth and 10° in elevation. Only a sector from front section of the whole space around the head was taken to experiments, because otherwise the computational load of the network would have been too heavy.

In a simple case the training vector \( x_{az,ele} \) (see definition 3) includes cues from a source which is localized precisely in one direction (no blur). Distribution \( D_{az,ele} \) is then calculated to describe only that single direction, which means that the target values of a network (see definition 14) are calculated with fixed values of \( k \) and \( l \) corresponding to that direction. Figure 11 shows an example distribution with \( k = 11, l = 2 \). It corresponds to a source direction of -15° az 0° ele.

Experiments were carried out where the training data included localization cues for real sources of several types, such as noise signals and spoken vowels. They provided good generalization ability for other real sources but much worse for more complex virtual sources. Therefore the next section presents a method for enhancing the generalization for virtual sources.

5.3. Method for generating training data for enhancing generalization ability

A network trained only with real sources did not provide good generalization when tested with virtual sources which were produced with two loudspeakers in similar manner than in section 4.3.2. The main reason for poor generalization is expected to come from the larger variation of ITD and ILD of virtual sources compared to real sources.

Figure 11: An example for target of network outputs which is double Gaussian distribution function. Outputs represent localization space for azimuths form -70° to 70° and elevations from -20° to 20°. The distribution is centered at a position -15° az 0° ele which corresponds to \( i = 11 \) and \( j = 2 \) output cell position.

Figure 9: Multilayer perceptron network.

Figure 10: Transfer functions of a) hyperbolic tangent sigmoid, and b) logistic sigmoid.
Figure 12: Contour graph presentations of the MLP-network outputs for different test vectors and also for example desired output. a) Desired output from direction 30° az 0° ele. A test vector is evaluated for b) pink noise real source source from direction 30° az 0° ele (included in training data), c) Finnish [a:] vowel real source 30° az 0° ele (not included in training data), d) pink noise virtual source 30° az 0° ele (not included in training data), e) pink noise real source 65° az 10° ele (included in training data), and f) pink noise virtual source 65° az 10° ele (not included in training data).

sources (see [11]). This variation occurs mainly above channel 10 which has center frequency of 1.2kHz.

To avoid these problems a method for enhancing the generalization behavior was developed. Our goal was to generate learning data including sources which are similarly blurred than virtual sources of our test data and also to find proper directional distribution capable of describing those blurred sources. From the viewpoint of localization cues this means that training data having similar ILD variations than blurred virtual sources has to be used in the learning process. Finding a proper directional distribution describing virtual sources is a difficult task, therefore we could not use them for training. Instead we used a method where blurred sources were produced somewhat artificially combining localization cues of real sources and also combining corresponding directional distributions from several different directions.

Here the method of generating artificial blurred sources and corresponding directional distributions is described formally. First a base training vector $x_{az0,ele0}$ corresponding to any direction $az0$, $ele0$ is selected. The base vector is now of the form

$$x_{az0,ele0} = [x_{az0,ele0}^1 ... x_{az0,ele0}^{25}]$$

where $x_{az0,ele0}^1$ to $x_{az0,ele0}^{25}$ describe the scalar elements of the vector. Then $m$ other vectors are taken from random directions $az1, ele1$ to $azm, elem$ in the following way

$$x_{az1,ele1} = [x_{az1,ele1}^1 ... x_{az1,ele1}^{25}]$$
$$x_{az2,ele2} = [x_{az2,ele2}^1 ... x_{az2,ele2}^{25}]$$
$$...$$
$$x_{azm,elem} = [x_{azm,elem}^1 ... x_{azm,elem}^{25}]$$

(17)

The new "artificial" learning vector is combined from the elements of those vectors so that elements $1,...,n$ are taken from base vector $x_{az0,ele0}$ and elements from $n+1$ to 25 are taken from vectors of formula 17 in the same order as they are expressed. In this way the generated vector can be expressed as

$$x' = \begin{bmatrix} x_{az0,ele0}^1 & x_{az0,ele0}^2 & ... & x_{az0,ele0}^{25} \\ x_{az1,ele1}^1 & x_{az1,ele1}^2 & ... & x_{az1,ele1}^{25} \\ x_{az2,ele2}^1 & x_{az2,ele2}^2 & ... & x_{az2,ele2}^{25} \\ ... \\ x_{azm,elem}^1 & x_{azm,elem}^2 & ... & x_{azm,elem}^{25} \end{bmatrix}^T$$

(18)

where index 1,...,n,...,25 corresponds to frequency channel of localization cue values in a training vector.

Then the directional distribution of desired output is
formed to correspond the learning vector of definition 18. The spatial distribution is superimposed from distributions evaluated for directions az0, ele0 to azm, elem. It can be expressed as a weighted sum of distributions $D_{az0, ele0}$ to $D_{azm, elem}$ in the following way

$$D' = \frac{1}{n} D_{az0, ele0} + \frac{1}{25} D_{az1, ele1} + \ldots + \frac{1}{25} D_{azm, elem}$$

(19)

where the weights of $D_{az0, ele0}$ to $D_{azm, elem}$ are chosen according to the number of vector elements included in $x'$ from the corresponding directions az0, ele0 to azm, elem. The distribution $D_{az0, ele0}$ has the largest weight value 1/n, since n vector elements from base vector $x_{az0, ele0}$ are included in the artificial learning vector 18.

5.4. Results of the MLP experiment

The generalization ability of a MLP model was tested mainly with virtual sources. The aim in these experiments was to train the network so that it gives reasonable approximation for directional distribution of a virtual source. Training data was approximated from four types of real sources which are pink noise, white noise, Finnish [e]-vowel, low-pass noise with 1.5kHz cutoff frequency, and high-pass noise with cutoff at 3kHz. However, none of these real source types were capable to describe high-frequency ITD and ILD variations of virtual sources. Therefore somewhat artificially created learning data was added to the training data set as described in the previous section. A suitable number of hidden layer neurons to provide optimal generalization was found to be 20. The optimal learning length seemed to depend on several things such as the amount of learning data, and the number of hidden neurons. The network chosen for this experiment was learned 2000 epochs. The criteria for the selection was the lowest relative error when tested with real and virtual test sources. Relative error for the training data set was 20.7% and for real test source 24.0%.

When those values are close to each other the generalization is assumed to be close to optimal. If the network was trained longer the error for train data decreased but for test data it increased which indicates over learning in that case.

Figure 12a shows a target distribution for the direction 30° az 0° ele so that it can be used as a standard for comparison for distributions produced in following test cases.

Figure 12b shows a network output distribution for a test vector evaluated for a pink noise real source from direction 30° az 0° ele. It is taken from the original training data set. The generalization ability for test vectors outside training data set is demonstrated with Figure 12c. This vector is evaluated for [a:]-vowel real source from the same direction than the previous one. Since the distributions produced with these sources seem to be similar, the model predicts the same direction for the both sources.

Figure 12d shows the output distribution for a virtual source where two speakers are positioned so that the angle between them is 60° and also the center point between the speakers is in 30° az 0° ele. The distribution of the virtual source in Figure 12d blurred comparing it to real source of Figure 12b.

Figures 12e and f show the network output for a real and virtual source near the edge of the map. The pink noise real source is in direction 65° az 10° ele and the center point between speakers creating the virtual source is in the same direction. The distribution of virtual source is in 12f is oriented ≈ 5° toward the median plane compared to the real source of Figure 12d. This was expectable since the result follows the theory an experiments presented in [20]. Compared to the real one there is also change of ≈ 10° in elevation of the virtual source.

Figure 13 demonstrates the model’s poor ability to learn directions near the median plane. Figure 13a shows a real source in direction 0° az 0°. Since all ITD and
ILD values from the median plane are zeros the model is not able to learn to discriminate elevations in this case. In Figure 13b the same real source is rotated to direction $-5^\circ$ az $0^\circ$ where the ability to discriminate elevation becomes better.

6. DISCUSSION AND CONCLUSIONS

The SOM and MLP neural network methods were experimented in order to analyze their modeling capability of binaural localization of spatial sound. The SOM method was capable to find directional dependence between localization cues and the map position. However, the aid of "head rotation" was needed for avoiding confusions in elevation. For the analysis of sound localization the SOM approach provided a possibility to calculate distance graphs between SOM model vectors and test vectors. Since the organization of the SOM has a certain direction dependence, the best matching region between test vector and SOM model vectors describes the direction of the sound source. In the case of simple virtual sources the distance graph was also capable to describe the blur in localization. However, an open question remains how well this blur noticed using our model correspond the blur of human auditory perception.

The advantage of the MLP method over SOM was that the directional distribution of sources is controllable by the selection of desired output values. However, good or even reasonable generalization ability demanded a much more complex training data set. The test of true validity of directional distributions produced with our MLP-model would have demanded careful psychoacoustic listening test which was out of the scope of this paper.

 Ability to learn a direction close to the median plane with both of our models was worse compared to real human auditory perception since the cue vector does not include enough information in such case. Therefore our interest to develop the model in future is to add extra information from the sources to learning data in the form of sum cross correlation of ERB channels and composite loudness level spectra CLL [11] calculated for ERB channels.

As a conclusion, there are probably numerous ways of formulating an artificial listener, composed of binaural auditory model and neural networks, that can map spatial sound signals to directional source distributions in a useful manner. Since it is not known in detail what kind of distributions the human listener perceives, the construction of an equivalent computational model is not possible. We believe, however, that it is feasible to develop an artificial listener that can judge the attributes of spatial hearing, at least in carefully controlled experimentations, and yield useful objective measures for spatial sound quality. Extensive further work is needed to reach this goal.

REFERENCES


