THE USE OF ARTIFICIAL NEURAL NETWORKS IN THE ESTIMATION OF THE PERCEPTION OF SOUND BY THE HUMAN AUDITORY SYSTEM

D. Riordan*, P. Doody and J. Walsh

Intelligent Mechatronic and RFID Technology Gateway, Institute of Technology, Tralee, Co. Kerry, Rep. of Ireland.

*Email: daniel.riordan@staff.ittralee.ie

Abstract- The human auditory system perceives sound in a much different manner than how sound is measured by modern audio sensing systems. The most commonly referenced aspects of auditory perception are loudness and pitch, which are related to the objective measures of audio signal frequency and sound pressure level. Here we describe an efficient and accurate method for the conversion of the sensed factors of frequency and sound pressure level to perceived loudness and pitch. This method is achieved through the modeling of the physical auditory system and the biological neural network of the primary auditory cortex using artificial neural networks. The behavior of artificial neural networks both during and after the training process has also been found to mimic that of biological neural networks and this method will be shown to have certain advantages over previous methods in the modeling of auditory perception. This work will describe the nature of artificial neural networks and investigate their suitability over other modeling methods for the task of perception modeling, taking into account development and implementation complexity. It will be shown that while known points on the perception scales of loudness and pitch can be used to objectively test the suitability of artificial neural networks, it is in the estimation of the perception of sound from the unknown (or unseen) data points that this method excels.

Index terms: auditory system modeling, audio sensors, artificial neural networks, perception of sound, digital signal processing, loudness, pitch.

1806

I. INTRODUCTION

The modeling of the perception of sound by the human auditory system is vital in fields such as speech recognition, speech synthesis and audio quality measurement. The perception of sound is governed by two main perceptual measures; the Perceptual Loudness measure (Phon or Sone Scale) and the Perceived Pitch (Critical-Band Rate or Bark Scale) of an audio signal.

The perceived loudness of an audio signal presented to the ear is influenced by both the signals frequency and sound pressure level (S.P.L). The current method for the conversion from frequency and S.P.L. to perceptual loudness is outlined in ISO 226:2003 – 'Acoustics – Normal equal-loudness-level-contours' [1]. This conversion involves a complex calculation which incorporates three 29 entry look-up tables. The conversion from frequency to pitch was originally presented by Zwicker[2]in table format. Zwicker's table documents the Critical Frequency Bands along with their corresponding center frequency, maximum cut-off frequency and bandwidth. The current most widely used method for this conversion is a function approximation of Zwicker's data created by Traunmuller[3]. These conversions are described in great detail in section 2.

The existing measures for modeling the auditory system's perception of pure-tone audio signal are attempting to model the behavior of the ear (and to a certain extent the filtering effects of the head and torso) in conjunction with the primary auditory cortex. The primary auditory cortex is a biological neural network. Therefore, it may be beneficial to model the conversion from the analytical measures of frequency and S.P.L to loudness and pitch using Artificial Neural Networks (A.N.N.s). A.N.N.s are regarded as a good candidate for the estimation of this perceptual mapping function as their structure is based upon biological Neural Networks. Also, their behavior both during and after the training process, has also been found to mimic that of biological Neural Networks[4].

This chapter describes the development and testing of a system which will use A.N.N. to model both features of sound perception mentioned above. It will also be investigated if a single A.N.N. model can be used to model both of these aspects of sound perception simultaneously, as in Figure 1.
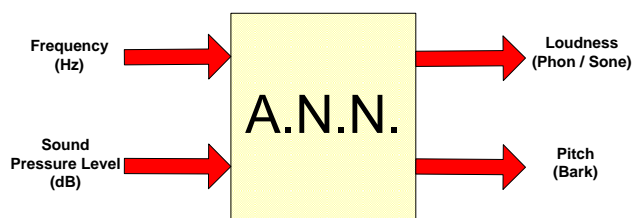
Figure 1: An A.N.N. Model of Pure-Tone Perception

II. THE PERCEPTION OF SOUND BY THE HUMAN AUDITORY SYSTEM

Sound is loosely defined as vibrations which travel through the medium of air (although any medium or combination of media will suffice) as longitudinal waves and are perceived by the human ear. There are two main analytical parameters that define the characteristic of a sound, the Sound Pressure Level (SPL) and the frequency components of the longitudinal waveform.

Similarly, there are two main characteristics which define a sound as perceived by the auditory system, pitch (measured in Bark) and perceived loudness (measured in Phon or Sone). An otologically normal person is a person who has a fully functioning auditory system, free from impairments. For such a person, the magnitude of these vibrations that can be perceived is generally accepted to be those with a SPL of greater than 20μPa or 0 dB. This is known as the Absolute Hearing Threshold (AHT). This value is actually the AHT for a signal of frequency 1 kHz. The AHT is known to vary with the frequency of signal being perceived. [5]

For a similarly otologically normal person, the frequencies of vibrations which can be perceived are those within the range of 20 Hz to 20 kHz and of sufficient SPL. This detectable frequency range generally deteriorates with the age of the listener. This frequency range may also be adversely affected by overexposure to loud sounds causing hearing damage. [5]

a.   Pitch

Critical-Band Rate is a perceptual measure, usually quantified in Bark, of the perceived pitch of an audio signal. This measure is directly related to the frequency of the sound being perceived. The conversion from frequency to perceived pitch is often referred to as 'frequency-warping'. The critical-band rate is a sub-division of the audible frequency range

into 'critical bands'. These 'critical bands' are more closely related to the manner in which the mechanics of the basilar membrane of the human inner ear operate [6].
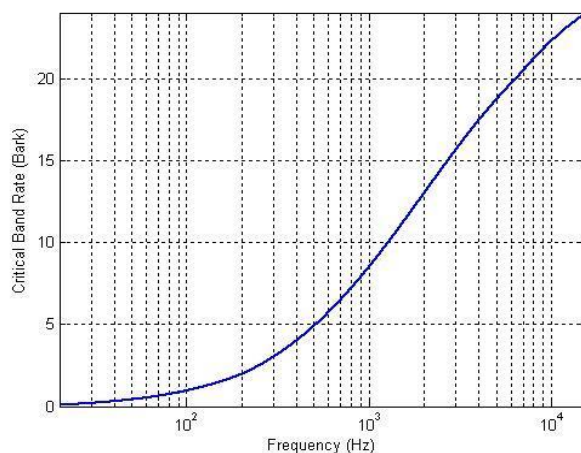

Figure 2: Critical-Band Rate versus Frequency

The conversion from frequency to pitch was originally presented by Zwicker [2] in table format. This table is presented in the Appendix as Table A.1. Zwicker's table documents the Critical-Band number along with their corresponding center frequency, maximum cut-off frequency and bandwidth [5]. A plot of the relationship between Frequency and Critical Band Rate outlined by Zwicker is shown in Figure 2.

Since the first publication of this table in 1961 the conversion from frequency to Critical-Band Rate has been modelled using many function approximations of the data. Resulting equations and algorithms have been proposed by [7], [8] and [9]. The current, most widely used and accepted method for this conversion is outlined by Traunmuller [3].

Traunmuller's equation for the conversion from frequency to Critical-Band Rate is

$$z' = \frac{26.81f}{1960 + f} - 0.53$$

If z < 2, $z' = z + 0.15(2 - z)$

If z > 2, $z' = z + 0.22(z - 20.1)$

Else $z = z'$ (1)

Where **z** is the critical-band rate (Bark) and **f** is the frequency (Hz).

b. Loudness

The Perceptual Loudness Measure is a psychoacoustic measure correlating to the physical intensity of an audio signal. Perceived Loudness is usually measured in the units Phon or Sone. As well as being sensitive to the SPL of the signal being observed, the perceived loudness of a signal is also highly dependent on the frequency components of the signal. This has led to the creation of the 'Equal Loudness Curves', shown in Figure 3[5].
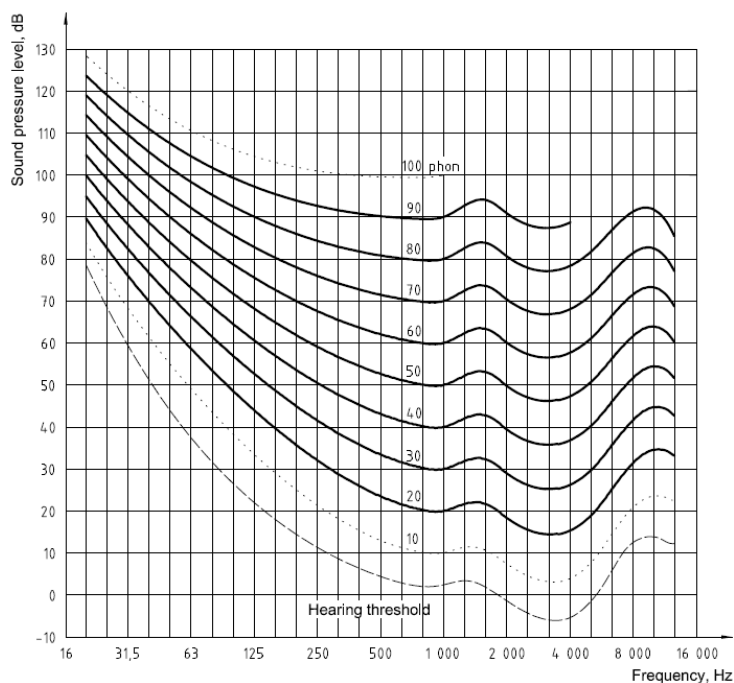

Figure 3: The Equal Loudness Contours (I.S.O. 2003)

The 'Equal-Loudness Contours' depict the sound pressure levels (SPL) which are required to ensure a perceived constant 'loudness' over the audible frequency band. As it can be seen from the contours of Figure 3, for a perceived loudness of 10 Phons at 1000 Hz an SPL of 10dB is required. To maintain a perceived loudness of 10 Phons at 50Hz an SPL of approximately 55dB is required.

The 'Equal-Loudness' contours were initially devised by Fletcher and Munson in 1933. The contours were derived using subjective measures, involving a panel of test subjects. Each listener was presented with a pure tone of 1 kHz of certain intensity and then a second pure tone of a different frequency. The intensity of the second tone was then varied until the listener perceived the 2 tones to be of equal loudness. The results obtained from the various test subjects were then averaged to obtain the final contours [10].

This experiment was repeated in 1956 by Robinson and Dadson, who found their results to differ greatly from those of Fletcher and Munson [11]. Robinson and Dodson's results were accepted as the International Standardisation Organization's (I.S.O.) official standard until replaced by the current standard in 2003. [1].

$$L_N = \left(40 \cdot \lg B_f\right) + 94 \quad (2)$$

where

$$B_f = \left[0.4 \times 10^{\left(\frac{L_p + L_U}{10} - 9\right)}\right]^{\alpha_f} - \left[0.4 \times 10^{\left(\frac{T_f + L_U}{10} - 9\right)}\right]^{\alpha_f} + 0.005135$$

The current I.S.O. standard is documented in I.S.O 226:2003. This document gives information on the conditions under which the subjective testing for the definition of the curves took place. The derived equations which may be used for the conversion of sound intensity data to perceptual loudness data are also included. These consist of equations for the conversion from frequency and SPL to perceptual Loudness (in Phon) and vice versa and are given here as Eq. 2 and Eq. 3 respectively. These equations are accompanied by a 'look-up' table which is required to implement these equations. This 'look-up' table can be found in the Appendix Table A.2 [1].

In Eq. 2, **LN** is the perceived loudness level in Phon, **Tf** is the threshold of hearing, **αf** is the exponent for loudness perception, **Lu** is a magnitude of the linear transfer function normalized at 1000 Hz and **Lp** is SPL. The three factors **Tf**, **αf** and **Lu** each have values determined by the 29 frequencies specified in the lookup table

$$SPL = ((10./af) . \times \log_{10}(Af)) - Lu + 94 \quad (3)$$

where,

$$Af = 4.47 \times 10^{-3} \times (10.^{(0.025 \times Ln)} - 1.15) + (0.4 \times 10.^{(((Tf + Lu)/10) - 9)})^{af}$$

and all symbols represent the same factors as in Eq. 2.

The Sone scale of perceived loudness is very similar to the Phon scale. In fact it is a direct translation of the calculated Phon value. In certain instances, the Sone scale can be a more useful measure than the Phon measure.

The Sone unit of perceived loudness is analogous to the manner in which the human auditory system perceives a change in loudness. In the Phon scale of perceived loudness, a doubling of the perceived loudness is associated with a rise of 10 Phon[6]. Using the Sone scale, the

perceived loudness of two different signals would be in ratio to the resulting positions on the Sone scale. In other words, a perceived doubling of the loudness of a signal would result in a doubling of the units of the perceptual loudness measure on the Sone scale.

$$S = \begin{cases} 2^{(l-40)/10}, if \quad l \geq 40 \\ (l/40)^{2.642}, otherwise \end{cases}$$

(4)

The equation for the conversion from the Phon scale to the Sone scale is shown in Eq. 4, where **S** is the resulting perceived loudness in Sone and **l** is the loudness level in Phon. [12]

III. ARTIFICIAL NEURAL NETWORKS

A biological neural network is an interconnection of processing elements (neurons) responsible for the processing of information in the nervous systems of animals. Each connection between neurons has a certain strength or 'weight', which may be strengthened or weakened, which allows the neural network to 'learn' and thus perform processing operations. It is the use of neural networks that allow animals to perform various tasks with ease which have proved excessively difficult to achieve by computational means. [4]

An Artificial Neural Network (A.N.N.) is a computational method which is modeled on biological neural networks. An A.N.N. consists of an interconnection of processing elements (artificial neurons) which each carry out a simple computational operation. The neurons are interconnected by weighted connections, similar to the connections in biological neural networks. The weights of each connection are updatable during the 'training' process. It is this ability that allows the A.N.N. to learn functions and processes in the same way as biological neural networks. Artificial Neural Networks (A.N.N.s) are a branch of the inductive machine learning subfield of Artificial Intelligence (A.I.) techniques. A.N.N.s are based upon the behavior, structure and architecture of biological neural networks. For this reason A.N.N.s are very suited to the modeling of biological functions which have traditionally been extremely difficult for other computing methods to model.

Their advantages over traditional processing techniques include their ability to "learn" from pre-existing training material. An A.N.N. generally learns in much the same way as biological neural networks learn. When presented with training material the connection strengths within

the A.N.N. are either strengthened or weakened until the desired associations are made. Many A.N.N. architectures and training algorithms have been developed to date, each having specific advantages and disadvantages.

a. Architectures of Artificial Neural Networks

A.N.N. architecture is the arrangement of neurons into layers and the patterns of the interconnection of those layers and the neurons within the layers. The neural nets are often separated into single layer and multi-layer architecture.

Single-layer nets usually comprise of an input layer, a single layers of connections and an output layer. The input layer of a neural net does not perform any computation and, therefore, is rarely counted when determining the number of layers in a net. Single-layer nets are often used for pattern classification problems when the output of each output neuron represents a specific class of input pattern. Minsky and Papert proved that single layer neural networks can only be used effectively in problems that are linearly separable. They showed that for more complex problems more complex multi-layer nets need to be used. [13]

Multi-layer nets contain an input layer, any number of hidden layers and an output layer. In multi-layer networks it is common to have a layer of connections between each successive layer; however connection of any individual neuron or layer of neurons to any other is possible.

A layered network architecture allows neurons only to be connected to neurons of the same or subsequence layers. No intra-layer connection is allowed within the input layer. This architecture insures that no closed-loop feedback occurs in the network. Acyclic networks are a form of layered network in which no connection between neurons of the same layer is permitted. Only connections from a neuron to neurons of a subsequent layer are permitted. A special case of the Acyclic Network architecture is the Feed-Forward network. [14]

Feed-Forward A.N.N.s are the most popular form of A.N.N., with the term 'A.N.N.' often being used to describe only Feed-Forward type networks [14]. In this architecture the flow of the signal is always forward through the network towards the output neurons. Connections leading from a neuron to neurons in the same or previous network layers are prohibited in this architecture.
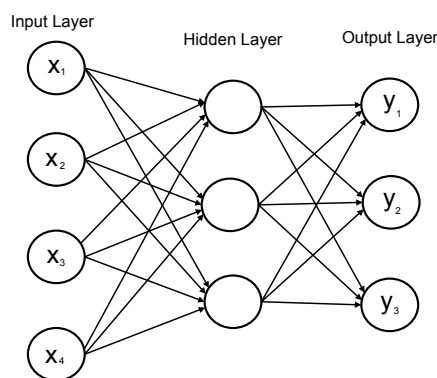
Figure 4: A Feed-Forward A.N.N.

Modular Networks are formed of an interconnection of separately developed A.N.N.s. This allows a large problem to be separated into smaller problems with the developer using an A.N.N. to solve each smaller problem. These smaller A.N.N.s are then combined in a Modular A.N.N. to solve the larger problem. [14]

b. Training Algorithms and Supervised Learning

The purpose of a training algorithm is to optimise a neural network so as the network will perform in the manner desired by the user. Upon creation of an A.N.N., the weight values of the network are often assigned at random. The A.N.N. must then be optimised using a training algorithm to perform a useful computational function. This is achieved altering the weights according to a predefined set of rules [14]. Learning algorithms can be divided into two wide-ranging types; Supervised Learning algorithms and Unsupervised Learning algorithms.

Supervised Learning algorithms are very similar to function approximation algorithms. The A.N.N is provided with a training set from which to learn. Each training set consists of an input vector with a corresponding target output vector. The inputs are presented at the inputs nodes of the network and the resulting output is logged. The difference between this A.N.N.s outputs and the target output vector contained in the training vector is said to be the error vector. The supervised learning algorithm then performs some form of optimisation algorithm in order to minimise this error. Depending upon the type of training algorithm being used and for what purpose the A.N.N. will be used, either the M.S.E. or the number of misclassifications is minimised. This involves a measured alteration of the weights of the connections within the network. Most training algorithms are repeated for a number of iterations until some termination criterion is met. This criterion is often a predefined number of iterations, a goal M.S.E. or number of misclassifications or a minimum reduction of error per iteration.

c. Over-Fitting & Generalization

Generalization is the ability of an A.N.N. to perform well when presented with unseen data based upon what it has learnt during the training process. One of the major problems which occurs during the training of A.N.N.s is the memorization of training material. When Memorization occurs, the A.N.N. has over-fitted to the requirements of the training data. It has exactly learned the input-output values in the training set, but performs poorly when presented with unseen data. Over-fitting of the training data may occur when the network has been excessively trained. If a suitable large A.N.N. is trained repeatedly until its M.S.E. is a minimum, it may have memorized the input output relationship and perform poorly on unseen data. [14]

Over-fitting can be avoided by limiting the amount of training iterations of the training algorithm. By dividing up the training set into a training set and a validation set, the generalization of the A.N.N. can be monitored. Once trained, the A.N.N. is presented with the unseen validation set. The M.S.E. of the resulting output is monitored. This operation is known as Cross-Validation. Often many A.N.N.s of varying architectures are trained to solve a single problem. By implementing the Cross-Validation training technique with each A.N.N., the A.N.N. with the best generalisation can be identified. This is often not the A.N.N. with the best performance on the training data. [15]

Another method to ensure generalization is to limit the degrees of freedom present in the A.N.N.. By limiting the number of neurons in the A.N.N. the net will be unable to memorize the data due to its lack of flexibility. In this way the A.N.N. is forced to generalize the relationship between the input and target values of the training set. [16]

IV. A.N.N. IMPLEMENTATION OF THE PERCEPTUAL LOUDNESS MEASURE (PHON)

a. Motivation for A.N.N. Implementation

ISO 226:2003 "specifies combinations of sound pressure levels and frequencies of pure continuous tones which are perceived as equally loud by human listeners" [1]. The algorithm to calculate the loudness level, LN, given the frequency, f, and the S.P.L., Lp, of an audio

signal is shown in Eq. 2 earlier in this chapter. This equation uses three 29 entry, look-up tables, to perform the calculation outlined in Eq. 2. See Appendix Table A.2[1].

The algorithm outlined in ISO 226:2003 for the calculation of the perceived loudness can only be implemented accurately for 29 discrete values on the frequency scale. Therefore the frequency components of all audio signals need to be approximated by one of the specified 29 frequencies outlined by ISO 266:2003. This can result in a 'digitization' of such features as uniform tones rising steadily in the frequency domain.

The algorithm outlined in ISO 226:2003 is attempting to model the behavior of the biological function. As mentioned in Section 3, A.N.N.s are modeled upon biological neural networks. Also, the behavior of A.N.N.s both during and after the training process has been found to mimic the behavior of biological neural networks. Therefore, it is logical to say that it may be beneficial to model this perception based function using A.N.N. techniques.

b. Development of A.N.N. Architecture

For simplicity, a two layer feed-forward architecture was used for the A.N.N.. Two nodes are required in the input layer to take the values of frequency and S.P.L. of the audio signal. A single node is used in the output layer to accommodate the output of the Loudness Level in Phon. The number of nodes to implement in the hidden layer was decided during the training process based the performance of various networks during the training/testing process. A tan–sigmoid activation function is used in the nodes of the hidden layer. This allows for the use of efficient backpropagation based training algorithms. A linear output activation function in the output layer node. The linear output function is required to allow the output of the A.N.N. may take on any value. This is required as the desired output of the network will be in the range 0 to 90 of the Phon Scale.

c. Training / Testing

The data used in the training of the A.N.N.s to mimic the manner in which an audio signals loudness is perceived, was generated from the Eq 2 and 3. These equations were implemented for all 29 specified frequencies at each SBL level from 1dB to 90dB (those specified to be accurately catered for by the equations), which resulted in 2581 training vectors. Each training vector contained a frequency value (Hz) and an S.P.L. level (dB) as the input values. A corresponding perceptual loudness level (Phon), calculated by the Eq2 and Eq. 3,was included in the training vectors as a target value. With a large quantity of both input values and corresponding target values a supervised training algorithm may be used to train the A.N.N.s.

The A.N.N. was designed, trained and tested using the Matlab® Neural Network Toolbox. The Levenberg-Marquardt backpropagation algorithm was used to train the A.N.N..

The Levenberg-Marquardt algorithm is a least-squares error minimization technique used in the supervised training of A.N.N.s. It is noted as having an appropriate trade-off between efficiency and accuracy which would be suitable for function approximation problems with randomized initial weights [15].

The number of nodes in the hidden layer was decided based on the performance received during successive training/testing iteration as follows. The number of hidden nodes was varied from 10 to 60 with each configuration being tested for ten training sessions. Each session consisted of 1000 epochs, each beginning with the weighs of the A.N.N. being randomized. The resulting M.S.E. at the end of each training session is noted and the associated network weights logged.

For each configuration, the A.N.N. with least M.S.E after the ten training sessions is taken as the best initial approximation of the function. These 'best' A.N.N.s are then trained to the maximum amount of epochs as defined by the stop conditions of the Levenberg-Marquardt backpropagation algorithm.

Testing is also carried out to determine the level of generalisation achieved by each A.N.N.. This is done by observing the performance of each A.N.N. configuration when presented with unseen data. For this instance, unseen data will consist for a frequency values other than those included in the look-up table associated with Eq. 2 and Eq. 3.

Table 1: Results from Training of A.N.N. Model of the Perceptual Loudness Conversion (Phon)

No. Nodes in Hidden layer	25	28	30	40	50
Min M.S.E. of 10 1000 epoch Training Sessions	0.01887	0.002901	0.003737	0.00639	0.001862
Max error of Net with min MSE	0.351	0.4469	0.608	0.6866	0.3915
Best MSE result received	0.01418	0.00257	0.00256	0.0009581	0.00113
Standard Deviation	0.1191	0.0507	0.0506	0.031	0.0336
Max error	0.9793	0.4328	0.4576	0.2674	0.3825

d. Results

Table 1 showsa selection of the results achieved during the investigation of the performance of A.N.N.s with varying numbers of neurons in the hidden layer. It can be seen that a number of A.N.N.s suitable for the estimation of the Perceptual Loudness measure in the Phon scale were created. A network comprising of 40 nodes in the hidden layer was developed and trained to provide a M.S.E. of 0.00095812, with a standard deviation of 0.031 and maximum individual error of 0.2883.

It is of benefit with regard the size, execution time and generalisation of the A.N.N. that the number of neurons is kept to a minimum. Therefore, a certain tradeoff between the number of neurons and the accuracy of the results must be made. With this in mind and based on the results shown in Table 5.1, an A.N.N. comprising of 28 neurons in the hidden layer, with a M.S.E. of 0.00257 and standard deviation 0.0507 would be suitable for use in the estimation of the perceptual loudness measure. Of course, where greater accuracy is needed an increase in the number of neurons may be made. If the situation requires a smaller A.N.N. with a shorter execution time, an A.N.N. with fewer neurons may be used at the expense of accuracy.
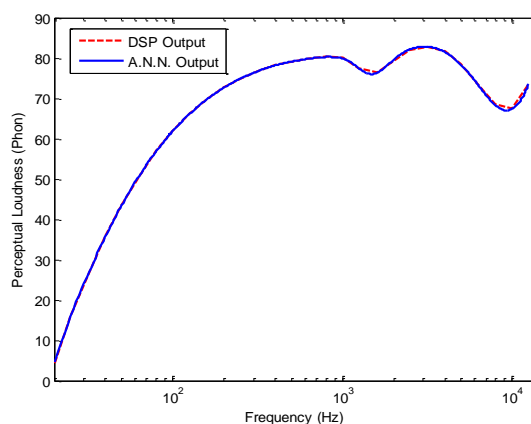

Figure 5: Results from Testing of Perceptual Loudness (Phon) Mapping A.N.N.

Figure 5 depicted a comparison of the performance of the A.N.N. method developed here and the method outlined by Eq. 2 and Eq. 3. Both methods were implemented with a constant S.P.L. of 80dB and a frequency value varying from 20Hz to 12500Hz. The equation based method was presented with those frequency values associated with the look-up table. The A.N.N. was presented with frequency values rising from 20Hz to 12500Hz in increments of 1 Hz. The resulting estimation of the perceived Loudness from both methods is plotted in the

figure. From this figure it can be seen that the A.N.N. shows a very high level of correlation with the values generated by the method outlined in the ISO standard.

It was also found that the A.N.N. method produced highly continuous curve when presented with a constant SPL and varying frequency. This is in contrasted to the discontinuous nature of the results generated by the method outlines by the I.S.O. standard. Figure 6 highlights the digitization effects introduced by the implementation outlined by ISO 226:2003 (labeled 'DSP output'). These effects have been overcome by the A.N.N. method of perceptual loudness evaluation (implemented with 28 neurons in the hidden layer). Figure 5.3 shows the resulting curves when both methods were presented with a constant SPL of 80dB and the frequency was varied from 20Hz to 12500Hz. A good level of generalisation is shown by this A.N.N. configuration as evident by the smooth continuous curve shown in Figures 5 and 6. [17].
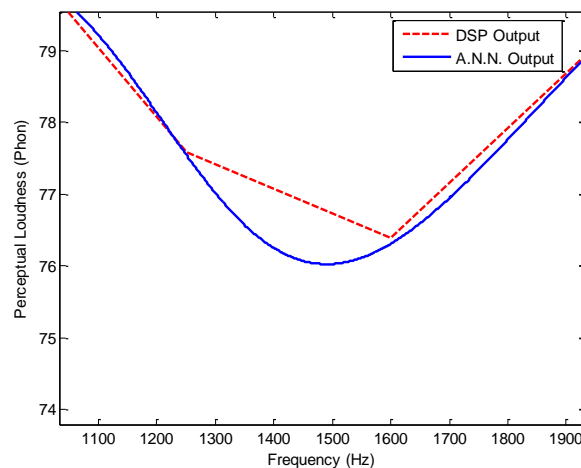

Figure 6: Close-Up of Figure 5.2

V. AN A.N.N. IMPLEMENTATION OF THE PERCEPTUAL LOUDNESS MEASURE (SONE)

a. Motivation for A.N.N. Implementation

The measure of Perceived Loudness may also be measured on Sone Scale. This Sone measure is generally calculated directly from a previously determined Phon measure. The algorithm for this conversion is given in Eq. 4 earlier in this chapter[12].The Sone scale of perceived loudness is often thought to be a more accurate representation of the manner in which

loudness is perceived by the human auditory system. For this reason it may be more desirable to have a direct conversion from Frequency and S.P.L to the Sone scale rather than the Phon scale.

The previous section of this chapter shows that the conversion from frequency and S.P.L. to the loudness measure in Phon can be implemented accurately with an A.N.N. This section will show that an A.N.N. can be used to implement the conversion from frequency and S.P.L. to the loudness measure on the Sone scale.

b. Development of A.N.N. Architecture

" …[Artificial Neural] Networks with just two layers of weights are capable of approximating any continuous functional mapping" [16].The continuity of a function has many different levels. C0 continuity denotes that the function is continuous and it does not generate any discrete behavior. C1 continuity deals with the first derivative of the function and denotes that this derivative is also continuous.

$$@ L = 40 \qquad 2^{(L-40)/10} = 2^0 = 1 \quad (5)$$

$$\lim L \rightarrow 40 \qquad \left(\frac{L}{40}\right)^{2.642} = 1^{2.642} = 1 \qquad (6)$$

The conversion from the Phon measure to the Sone measure presented in Eq. 3 was found to be a C0 continuous function as shown in Eq. 5 and Eq. 6 where both methods give a result of 1 when L = 40 and the limit as L goes to 40 respectively.

Eq. 7 shows the at L = 40, the first derivative of Eq. 5dS/dL, is equal to 0.069315. Eq. 8 shows that at the limit as L goes to 40 of the first derivative of Eq. 6, dS/dL, is equal to 0.06605. These values are not equal and therefore the function is not C1 continuous.

$$\frac{dS}{dL} = 2^{-4}\left(2^{0.1L}\right)\left(\ln 2\right)\left(0.1\right) = 0.069315\left(2^{(0.1L-4)}\right)$$

$$@ L = 40 \qquad 0.069315\left(2^{(0.1(40)-4)}\right) = 0.069315 \times 1 = 0.069315 \quad (7)$$

$$\frac{dS}{dL} = \frac{2.642(L^{1.642})}{40^{2.642}}$$

$$\lim L \rightarrow 40 \qquad \frac{dS}{dL} = \frac{2.642(40^{1.642})}{40^{2.642}} = 0.06605 \qquad (8)$$

To approximate a non-continuous function efficiently, an A.N.N. with at least 2 hidden layers is required. For this reason a 3-layer Feed-Forward A.N.N. architecture was implemented. "…a three-layer network with threshold activation functions could represent an arbitrary decision boundary to arbitrary accuracy" [16]. Again, a tan–sigmoid function was used in the nodes of the hidden layers and a linear function in the output layer node. Two nodes were required in the input layer to take the values of frequency and S.P.L. of the audio signal. A single node was used in the output layer to accommodate the output of the Loudness Level in Sone. The number of nodes in each of the hidden layers was decided during the training process based upon the performance of various network implementations during the training/testing process.

c. Training/Testing

The data used here in the training of the A.N.N.s, was generated by calculating the Phon values from the equations provided in I.S.O. 226:2003 and then converting these to Sone with Eq. 4. This resulted in 2581 training vectors, each containing a frequency value (Hz), an S.P.L. level (dB) as inputs and a corresponding perceptual loudness level (Sone) as the target value. Again, this training set facilitates supervised training methods.

The A.N.N. was designed, trained and tested using the Matlab® Neural Network Toolbox. The Levenberg-Marquardt backpropagation algorithm was used to train the A.N.N..

The number of nodes in the hidden layer was decided based on the performance received during successive training/testing iterations as follows. The number of neurons in the first hidden layer was varied from 5 to 40 and the number of neurons in the second hidden layer was varied from 1 to 5. Each possible configuration of these variations was tested for ten training sessions, each of 1000 epochs. Each session begins with randomization of the A.N.N.s weights.

For each configuration, the A.N.N. with the least Mean Square Error (M.S.E.) resulting from the ten training sessions is stored as the best initial approximation of the function. These A.N.N.s are then trained to the maximum amount of epochs as defined by the stop conditions of the Levenberg-Marquardt backpropagation algorithm. The results are then logged and analysis to decide upon a suitable A.N.N. configuration for this function approximation.

Each A.N.N. configuration is also tested for the level of generalisation achieved. As before, each network is provided with unseen input data and the resulting outputs are analysed for instances of Over-fitting.

d. Results

The table of results populated during the investigation of the performance of the various A.N.N. configurations can be found in the Appendix, Table A.3. From this table it can be seen that a number of A.N.N.s suitable for the estimation of the perceptual loudness measure in the Sone scale were created. A network comprising of 30 nodes in the first hidden layer and 5 in the second was designed which was trained to provide a M.S.E. of 0.0000678, with a standard deviation of 0.0082 and maximum individual error of 0.0826.
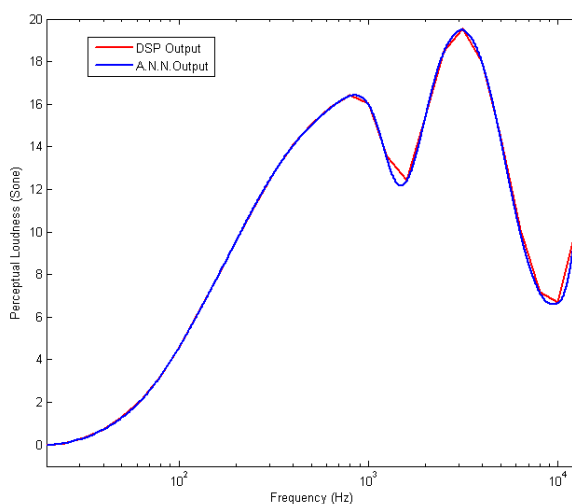

Figure 7: Results from Testing Perceptual Loudness (Sone) Mapping A.N.N.

Figure 7 depicts the performance of the A.N.N. comprised of 30 neurons in the first hidden layer and 5 in the second. This A.N.N. was presented with a constant S.P.L. of 80dB and a frequency value varying from 20Hz to 12500Hz in increments of 1 Hz. For reference the results generated by presenting Eq. 2, 3 and 4 with the same input S.P.L value and those frequencies present in the associated look-up table are also shown in the figure. A high degree of correlation between the A.N.N. based method and the equation based method is again shown. The digitisation effect of the equation based method is still present while the A.N.N. produces a highly continuous curve.
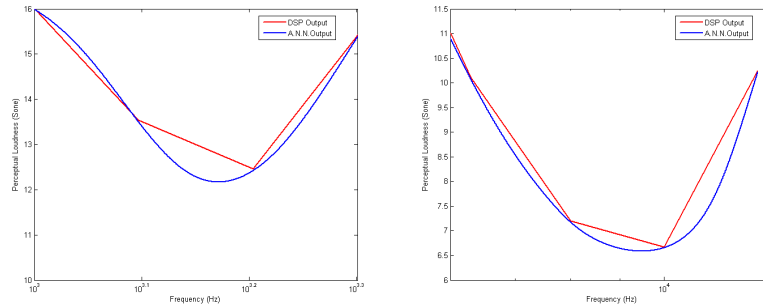
Figure 8: Close-Up of Figure 7 (1)   Figure 9: Close-Up of Figure 7 (2)

Figure 8 and Figure 9 are magnified versions of Figure 5.4 and shown in greater detail the digitization effect which has been overcome by the use of A.N.N.s. The smooth continuous curves shown in Figures 7, 8 and 9 depict the output of a network with a high level of generalisation.

A suitable trade-off between network size and performance may be the network containing 20 neurons in the first hidden layer and 1 in the second. This network yields a M.S.E. of approximately 0.023 and a Standard Deviation of 0.048. Upon investigation it was also found to produce a smooth continuous curve when tested as above. This demonstrates that the A.N.N. possesses a high degree of generalisation. The actual choice of A.N.N. from those presented will be application specific and will be dependent on such features as accuracy required and system requirements.

VI AN A.N.N. FOR THE FREQUENCY TO CRITICAL BAND RATE CONVERSION

a. Motivation for A.N.N. Implementation

The conversion from frequency to pitch was originally presented by Zwicker in table format [2]. Zwicker's table documents the Critical-Band number along with the corresponding center frequency, maximum cut-off frequency and bandwidth.

Since the first publication of this table in 1961, many function approximations of the data, with varying degrees of accuracy, have been presented [7], [18]and [9]. The current most widely used and accepted method for this conversion is outlined by Traunmuller in his paper 'Analytical expressions for the Tonotopic Sensory Scale' [3].

From Zwicker's table outlining the limits of the Critical-Bands, only the Bark value at the specific frequencies listed can be discerned accurately. The Bark values of all other

1823

frequencies values are no-more than educated estimations. While this is generally acceptable in the field of speech processing, there is room for improvement. These improvements may be of use when accurate representations of the perceived pitch are required such as models of the cognitive aspects of sound perception.

Traunmuller's equation for the conversion from the frequency scale to the Bark Scale is a function approximation of the information presented in Zwicker's critical-band rate table. This function approximation equation is shown in Eq. 1 earlier in this chapter. "The values calculated in this way agree with the table for f > 100Hz to within ± 0.05 Bark" [3]. This error measurement can only be taken from the frequency values present on the table. Errors associated with frequencies not listed on the table are unknown. Thus the values generated by this equation which are between Zwickers values are, again, an educated guess.

Both Traunmuller and Zwicker, along with many others, are attempting to model the behavior of a fundamentally biological function. Therefore it is logical to say that it may be beneficial to model this conversion using A.I. techniques. The structure of A.N.N.s is based upon biological neural networks and their behavior, both during and after the training process, has been found to mimic that of biological neural networks [15].

b. Development of A.N.N. Architecture

As with the A.N.N. for estimation of perceived loudness, a two layer feed-forward architecture was used for this A.N.N.. A single node was required in the input layer to take the frequency value of the audio signal. A single node was used in the output layer to accommodate the output of the perceived pitch in Bark. A tan–sigmoid function was used in the nodes of the hidden layer and a linear function in the output layer node. The number of nodes in the hidden layer was decided during the training process based the performance of various networks during the training/testing process.

c. Training / Testing

The data used here in the training of the A.N.N.s was taken directly from Zwicker's table of Critical-Band limits. This supplied an input of 25 input frequency values for the network with 25 corresponding output values. This allows for a supervised training algorithm to be used in training of the network.

The A.N.N. was designed, trained and tested using the Matlab® Neural Network Toolbox. The Levenberg-Marquardt backpropagation algorithm was used to train the A.N.N..

The number of nodes in the hidden layer was decided based on the performance received during successive training/testing iterations as described in section 4.3 of this chapter. In this instance the number of hidden nodes was varied from 1 to 20. The results were logged and examined to determine which A.N.N. is the most suitable for the implementation of this function approximation problem.

Each trained A.N.N. configuration is also tested for instances of over-fitting and it's performance on unseen data. This is done here by presenting each A.N.N. with frequency values not present in the data set and ensuring the result is consistent with known values.

Table 2: Results from Training of A.N.N. Model of the Perceived Pitch Conversion

Neurons in Hidden Layer	4	5	10	15	20
Min M.S.E. of 10 1000 epoch Training Sessions	0.000942	0.000725	0.000445	0.000315	0.000156
Max error of Net with min MSE	0.0593	0.0617	0.0407	0.046	0.0372
Best MSE result received	0.000904	0.000416	0.000445	0.000146	4.12E-05
Standard Deviation	0.0308	0.0208	0.0215	0.0123	0.0066
Max error	0.057	0.0554	0.0407	0.0363	0.0194

d. Results

Table 2 shows a sample of the results achieved during the investigation of the performance of various A.N.N.s configurations. It can be seen that a number A.N.N.s suitable for the warping of the frequency scale to Critical-Bate Rate (or the Bark Scale) were created. Based on these results (and other factors to be dealt with later), a network comprising of 10 nodes in the hidden layer would be a suitable candidate for use in the field of auditory system modeling. This A.N.N. was trained to provide a M.S.E. of 0.000445, with a standard deviation of 0.0215 and maximum individual error of magnitude 0.0407. A plot of the output of this A.N.N. when presented with an input of frequency values ranging from 20 – 15500 Hz in increments of

1Hz, is shown in Figure 10. It can be seen to compare very well with a plot of the data listed in Zwickers table shown in the same figure. The curve generated by this A.N.N. can be seen to be an extremely smooth continuous curve ensuring that this A.N.N. has a high level of generalization.
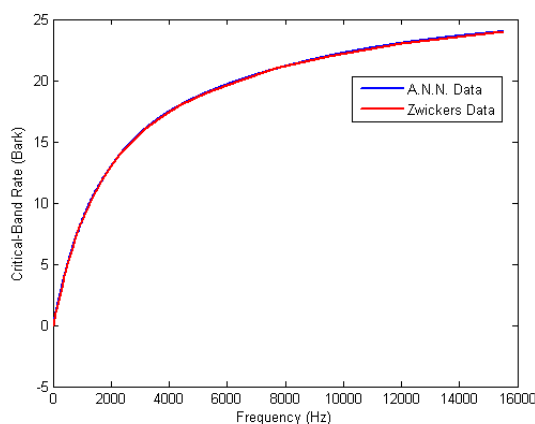

Figure 10: Results of Testing Perceived Pitch (Bark) Mapping A.N.N.

While networks with higher numbers of hidden nodes provided a better M.S.E. on the training data provided, an effect known as over-fitting was witnessed during testing. In an attempt to match the training data more closely, the values generated at instances between the training points became more non-uniform, as demonstrated in Figures 11 and 12. This results in a poor performance of the A.N.N. on unseen input data. The plots in Figures 11 and 12 show the continuous output from the 15 neuron network, when presented with a continuous input varying from 20 – 15500 Hz. Large variations can be seen in the region 100 – 1000 Hz, even though all of the expected outputs of the 25 entry training set have been met to within ±0.0363.
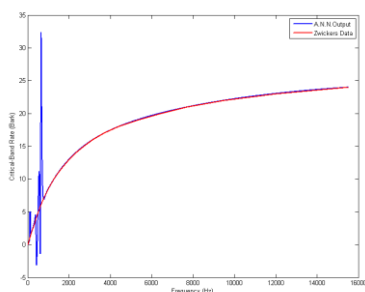

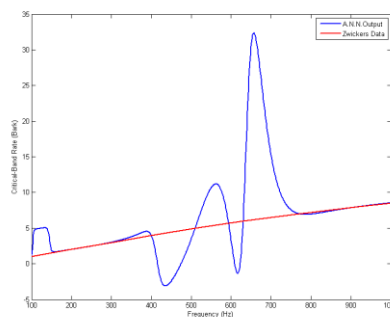Figure 11: Over-Fitting   Figure 12: Close-Up of Fig 11

For a neural Network to perform uniformly for unknown inputs, the A.N.N. will need have a high degree of generalization. A degrading of the generalization will cause over-fitting. Over-

fitting of a neural network occurs when the flexibility, or degree of freedom, of the network is too great. The degrees of freedom of an A.N.N. can be controlled by a process called 'structural stabilization'. This involves limiting the amount of changeable factors (neurons or weights) in a network. The fewer the changeable factors in the network the less likely it is that over-fitting will occur. Therefore the fewer neurons contained in a network, the less over-fitting is likely to occur [16]. This leads to good generalisation within the network. [19]

VII AN ALL-IN-ONE A.N.N. PURE-TONE PERCEPTION MODEL

a. Motivation for A.N.N.

The three previous sections of this chapter have shown that A.N.N. can be used to model the individual features of the human auditory system. This section will present the development of a single A.N.N. with the ability to generate both the Perceived Loudness and Pitch of a audio signal simultaneously.

b. Development of A.N.N. Architecture

As this A.N.N. is being designed to generate both the Perceived Pitch and Loudness, a minimum of three layers will be required in the network. This is due to the non-linear characteristic of the conversion from frequency and S.P.L. to the Sone Scale of Perceived Loudness. For simplicity, a network with two hidden layers was implemented. Two neurons were required in the input layer to take the frequency and S.P.L. values of the audio signal. Two nodes were also required in the output layer to accommodate the output of the perceived pitch in Bark and Perceived Loudness in Sone. A tan–sigmoid function was used in the nodes of the hidden layers and a linear function in the output layer node. The number of nodes in the hidden layers was decided during the training process based the performance of various networks during the training/testing process.

c. Training / Testing

A training set of 2581 vectors was compiled from the data used to train the A.N.N.s described in the previous chapters. Each training vector contains two input values, frequency and S.P.L., and two corresponding target values, the pitch in Bark and the loudness in Sone.

Again, the A.N.N. was designed, trained and tested using the Matlab® Neural Network Toolbox. The Levenberg-Marquardt backpropagation algorithm was used to train the A.N.N..

The number of neurons in the hidden layers was decided based on the performance received during successive training/testing iterations as described in section 5.3. In this instance the number of neurons in the first hidden layer was varied from 5 to 40 and the number of neurons in the second hidden layer was varied from 1 to 5.

Testing was also carried out to determine the level of generalisation achieved by each A.N.N. on both the estimations of Perceived Loudness and Pitch as before.

d. Results

The table documenting the performance of the A.N.N.s during training is presented in the Appendix, Table A.4. It can be seen from this table that a number of A.N.N.s have been developed and trained which are suitable for the implementation of both the perceived pitch and loudness measures.
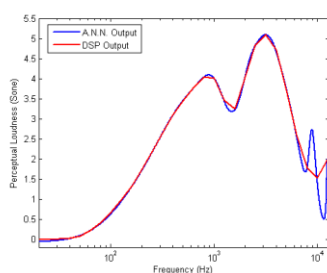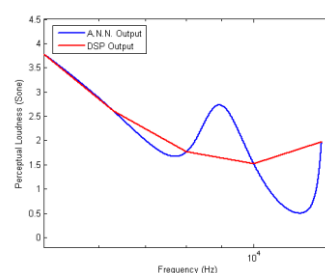

Figure 13: Over-Fitting   Figure 14: Close-Up of Figure 13

When the inevitable trade-off between network size and performance is taken into account, the network containing 30 neurons in the first hidden layer and 2 neurons in the second seemed to be a viable choice. For the estimation of perceived Pitch, this A.N.N produces a M.S.E. of less-than 0.00011 and standard deviation of the 0.0361 with values obtained from Zwicker's table. Similarly good results of 0.000367 and 0.1634 are produced for the estimation of Perceived Loudness in the Sone scale when compared with results outlined in ISO:226:2003. Upon further investigation it seems over-fitting has occurred with this A.N.N.. Figures 13 shows the estimation of Perceived Loudness resulting from inputs of 60 dB S.P.L. and a frequency varying from 20Hz to 12500Hz. Figure 14 is a magnified version of Figure 13 which highlights the irregularities which are not supported by the training data.
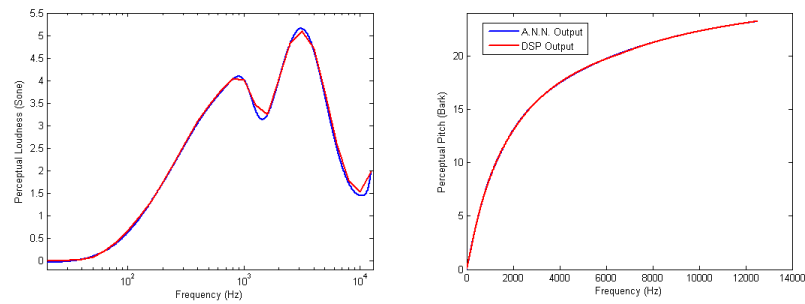
Figure 15 & Figure 16: Generalisation in All-In-One Perceptual Model A.N.N.

Investigations into another suitable A.N.N. with 20 nodes in the first hidden layer and 2 nodes in the second showed that network possessed a high level of generalisation. Figure 15and 16 show the estimations of Perceived Loudness and Pitch, respectively, produced for the same input values mentioned above, for the A.N.N. containing 20 nodes in the hidden layer.

VIII Conclusions

The results which have been presented here, clearly show that the conversion from frequency and S.P.L to Perceived Loudness and Critical-Band Rate (or Bark) can be implemented using an A.N.N.. It has also been shown that the use of A.I. techniques presented here has certain advantages over the existing and accepted methods.

The use of A.N.N.s in the estimation of perceived loudness has been shown to eliminate the need to approximate the frequency value of the signal to one of 29 specified frequencies. The values generated for frequencies between those specified are generated purely by the A.N.N. and cannot be validated without subjective testing. Some validation can be inferred by the fact that A.N.N.s have been noted to possess very similarly characteristics to that of biological neural networks and be adept at modeling biological functions.

Similarly, the implementation of the frequency to Critical-Band Rate conversion through A.N.N.s is shown bridge the gap between the 25 specified critical band values specified by Zwicker. While this has been done in the past by many function approximation attempts, an A.N.N. approach might prove to be a more suitable method. Again due to the nature of the A.N.N. it is well suited to the modeling of biological functions. Therefore the intermediary values generated by the A.N.N. implementation may be more representative of the true operation of the auditory system.

APPENDIX

Table A.1: Subdivision of the Audible Frequency Range into Critical Bands [2]

Band Number	Center Frequency (Hz)	Cut-off Frequency (Hz)	Bandwidth (Hz)
0		20	
1	50	100	80
2	150	200	100
3	250	300	100
4	350	400	100
5	450	510	110
6	570	630	120
7	700	770	140
8	840	920	150
9	1000	1080	160
10	1170	1270	190
11	1370	1480	210
12	1600	1720	240
13	1850	2000	280
14	2150	2320	320
15	2500	2700	380
16	2900	3150	450
17	3400	3700	550
18	4000	4400	700
19	4800	5300	900
20	5800	6400	1100
21	7000	7700	1300
22	8500	9500	1800
23	10500	12000	2500
24	13500	15500	3500

Table A.2: Look-Up Table for the Estimation of the Perceptual Loudness Measure[1]

Hz	$\propto_f$	Lu	Tf
20	0.532	-31.6	78.5
25	0.506	-27.2	68.7
31.5	0.48	-23	59.5
40	0.455	-19.1	51.1
50	0.432	-15.9	44
63	0.409	-13	37.5
80	0.387	-10.3	31.5
100	0.367	-8.1	26.5
125	0.349	-6.2	22.1
160	0.33	-4.5	17.9
200	0.315	-3.1	14.4
250	0.301	-2	11.4
315	0.288	-1.1	8.6
400	0.276	-0.4	6.2
500	0.267	0	4.4
630	0.259	0.3	3
800	0.253	0.5	2.2
1000	0.25	0	2.4
1250	0.246	−2.7	3.5
1600	0.244	−4.1	1.7
2000	0.243	−1.0	−1.3
2500	0.243	1.7	−4.2
3150	0.243	2.5	−6.0
4000	0.242	1.2	−5.4
5000	0.242	−2.1	−1.5
6300	0.245	−7.1	6
80000	0.254	−11.2	12.6
10000	0.271	−10.7	13.9
12500	0.301	−3.1	12.3

Table A.3: Results of Testing of A.N.N. implementation of the Conversion from Frequency and S.P.L. to Perceived Loudness (Sone)

No. Nodes in Hidden layer	15-1st 1-2nd	15-1st 2-2nd	15-1st 5-2nd	20-1st 1-2nd	20-1st 2-2nd	20-1st 5-2nd	30-1st 1-2nd
Min M.S.E. of 10 1000 epoch Training Sessions	0.115115	0.0333068	0.000247075	0.0053773	0.00222906	0.000607566	0.0125564
Max error of Net with min MSE	2.9058	1.1325	0.0919	0.487	0.5437	0.1266	0.8621
Best MSE result received	0.00561793	0.0278857	0.00013442	0.00230188	0.000305534	0.00044021	0.0121992
Standard Deviation	0.075	0.167	0.0116	0.048	0.0175	0.021	0.1105
Max error	0.6493	1.0743	0.1	0.2807	0.1106	0.142	0.8494

No. Nodes in Hidden layer	30-1st 2-2nd	30-1st 5-2nd	40-1st 1-2nd	40-1st 2-2nd	40-1st 5-2nd
Min M.S.E. of 10 1000 epoch Training Sessions	0.00246688	0.000198157	0.00443119	0.00285045	0.000331487
Max error of Net with min MSE	0.3518	0.0909	0.5202	0.537	0.1145
Best MSE result received	0.000292076	6.77975E-05	0.00191928	0.000200586	0.000317402
Standard Deviation	0.0171	0.0082	0.0438	0.0142	0.0178
Max error	0.0863	0.0826	0.3794	0.0752	0.1114

Table A.4: Results of Testing of A.N.N. All-in-One Pure-Tone Perception Model (Continued on next page.)

No. Nodes in Hidden layer	15-1st 1-2nd	15-1st 2-2nd	15-1st 5-2nd	20-1st 1-2nd	20-1st 2-2nd	20-1st 5-2nd	30-1st 1-2nd
Total No. of Weights in A.N.N.	33	57	58	97	83	137	108
Min M.S.E. of 10 1000 Epoch Training Sessions	0.524409	0.0369397	0.0819293	0.00123143	0.0174609	0.0005344	0.00596207
Max Error of A.N.N with min MSE	6.0128	2.1516	2.4887	0.3272	1.3866	0.1967	0.9597
Best Overall MSE result received	0.452939	0.0238607	0.0135987	0.00024784	0.0156354	0.00034912	0.00150447
Errors on Pitch results							
M.S.E.	0.0291	0.0015	0.0014	1.53E-04	6.71E-04	1.85E-04	4.45E-04
Max error	0.9297	0.1209	0.0882	0.0244	0.0649	0.0361	0.04505
Standard Deviation	0.1707	0.0388	0.0368	0.0124	0.0211	0.0136	0.0211
Cross Correlation	0.99976	0.99999	0.99999	1	1	0.99875	1
P-Value	0	0	0	0	0	0	0
Errors on Loudness Results							
M.S.E.	0.8768	0.0462	0.0258	3.43E-04	0.0306	5.13E-04	0.0026
Max error	5.7148	1.7156	1.3666	0.1088	0.1805	0.1191	0.3957
Standard Deviation	0.9365	0.215	0.1608	0.0185	0.0273	0.0226	0.0506
Cross Correlation	0.99263	0.99961	0.99978	1	0.99828	1	1
P-Value	0	0	0	0	0	0	0

1833

Table A.4 (Continued): Results of Testing of A.N.N. All-in-One Pure-Tone Perception Model

No. Nodes in Hidden layer	30-1st 2-2nd	30-1st 5-2nd	40-1st 1-2nd	40-1st 2-2nd	40-1st 5-2nd
Total No. of Weights in A.N.N.	177	153	257	208	337
Min M.S.E. of 10 1000 Epoch Training Sessions	0.00014918	0.00417786	4.38E-05	0.00915041	4.41E-05
Max Error of A.N.N with min MSE	0.0845	0.8099	0.0505	1.4267	0.0764
Best Overall MSE result received	6.35E-05	0.00023822	4.10E-05	0.000532855	2.97E-05
Errors on Pitch results					
M.S.E.	3.55E-05	1.09E-04	7.49E-06	1.69E-04	4.72E-06
Max error	0.0224	0.0361	0.0083	0.0881	0.0135
Standard Deviation	0.006	0.0104	0.0027	0.013	0.0022
Cross Correlation	1	1	1	1	1
P-Value	0	0	0	0	0
Errors on Loudness Results					
M.S.E.	9.15E-05	3.67E-04	7.46E-05	8.97E-04	5.46E-05
Max error	0.0669	0.1634	0.0419	0.2077	0.0717
Standard Deviation	0.0096	0.0192	0.0086	0.03	0.0074
Cross Correlation	1	1	1	0.99999	1
P-Value	0	0	0	0	0

REFERENCES

[1]   International Standards Organisation, "ISO 226:2003 : Acoustics - Normal Equal-Loudness-Level Contours," ed, 2003.

[2]   E. Zwicker, "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgrupen)," *The Journal of the Acoustical Society of America* vol. 33, p. 248, 1960.

[3]   H. Traunmuller, "Analytical Expression for the Tonotopic Sensory Scale," *Journal of the Acoustical Society of America,* vol. 88, pp. 97 - 100, 1990.

[4]   J. A. Freeman and D. M. Skapura, *Neural Networks - Algorithms, Applications and Programming Techniques*: Addison-Wesley Publishing Company Inc., Massachusetts, 1991.

[5]   W. M. Hartman, *Signals, Sounds and Sensation*. New York: AIP Press, 1997.

[6]   H. Fast and E. Zwicker, *Psycho-Acoustics: Facts and Models*, 3rd ed. Berlin: Springer, 2007.

[7]   A. J. Fourcin, "Speech Processing by Man and Machine – Group Report," *Life Sciences Research Report 5 of the Dahlem Workshops,* vol. Berlin, pp. 307-351, 1977.

[8]   V. L. Tromov, "A Model to Describe the Results of Psychoaustical Experiments on Steady-State-Stimuli," *Analiz Rechevykh Chelovekom,* pp. 36 - 49, 1971.

[9]   E. Zwicker and E. Terhardt, "Analytical Expressions for the Critical Bandwidth as a function of Frequency," *Journal of the Acoustical Society of America,* vol. 68, pp. 1523 - 1525, 1980.

[10]   H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," *Journal of the Acoustic Society of America,* vol. 5, pp. 82 - 108, 1933.

[11]   D. W. Robinson and R. S. Dadson, "A re-determination of the equal-loudness relations for pure tones," *J. Appl. Phys,* vol. 7, pp. 166 - 181, 1956.

[12]   R. A. W. Bladon and B. Lindblom, "Modelling the Judgement of Vowel Quality Differences," *Journal of the Acoustical Society of America,* vol. 69, pp. 1414 - 1422, 1980.

[13]   M. L. Minsky and S. A. Papert, *Perceptrons: An Introduction to Computational Geometry*. Massachusetts: MIT PRess, 1969.

[14] K. Mehrotra*, et al.*, *Elements of Artificial Neural Networks*. Massachusetts: MIT Press, 1997.

[15] R. D. Reed and R. J. Marks, *Neural Smithing – Supervised Learning in Feed-Forward Artificial Neural Networks*. Massachusetts: MIT Press, 1998.

[16] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York: Oxford University Press, 1995.

[17] D. Riordan and P. Doody, "An Alternative Method for the Estimation of the Perceptual Loudness Measure," *Proceedings of the 2007 International Conference on Artificial Intelligence and Pattern Recognition (AIPR-07),* pp. 222 - 224, 2007.

[18] M. R. Schroeder*, et al.*, "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear," *Journal of the Acoustical Society of America,* vol. 66, pp. 1647 - 1652, 1979.

[19] D. Riordan and P. Doody, "An Artificial Neural Network Implementation of the Frequency to Critical-Band Rate Conversion," *Proceedings of the 2007 International Conference on Artificial Intelligence and Pattern Recognition (AIPR-07),* pp. 225 - 228, 2007.