Proceedings of the Institute of Acoustics

EXTRACTION OF ROOM ACOUSTIC PARAMETERS FROM SPEECH USING ARTIFICIAL NEURAL NETWORKS

F. Li University of Salford, School of Acoustics and Electronic Engineering, UK
T. J. Cox University of Salford, School of Acoustics and Electronic Engineering, UK

1. INTRODUCTION

Reverberation Time(RT), Early Decay Time(EDT) and Speech Transmission Index(STI) are important concerns in the design of acoustically critical spaces. Such spaces include concert halls, theatres and cinemas where sound quality plays an important role in the enjoyment of a performance, and lecture rooms, shopping malls and railway stations where speech and public address announcements should be intelligible. Measurement of such objective parameters is essential in room acoustic practices. For one century measurement techniques have evolved, from Sabine's set-up utilising a stop watch, to sophisticated modern measurement systems such as maximum length sequence based digital instruments. There are still a number of unresolved problems in the measurement of room acoustic parameters, including the difficulties in undertaking occupied measurements. It is known that the currently used measurement methods have various limitations in performing occupied measurements. The reason is threefold: high sound pressure level test signals are unacceptable to the audience; the required signal to noise ratios for accurate measurements are difficult to obtain with the presence of audience, and equipment set up and logistical problems tend to be difficult to overcome. If room acoustic parameters could be found from naturalistic sources, such as the speech broadcast over a PA system or music from an orchestra, many of these difficulties would be overcome as measurement could take place under normal in use conditions.

The effects of room conditions are audible down telephone lines and acousticans can make reasonable judgements of RT through listening to naturalistic sounds. These suggest a possibility that room acoustic parameters can be extracted from speech or music utilising artificial intelligence methods - a purpose-trained artificially intelligent ear. Such a method is also of significant academic interest, as it is closely related to separating two convoluted signals and imitating cognitive aspects of human brains.

Motivated by these thoughts, investigations have been carried out. As the first endeavour to extract room acoustic parameters from natural sound source the scope was focused on speech. The artificial intelligence engines chosen were supervised artificial neural networks (ANN). Investigation so far has shown a better than 0.1s accuracy for RT and EDT, and better than 0.01 accuracy for STI using a closed set of speech examples. It has also revealed that the proposed method has potential to be further developed to extract acoustical parameters from arbitrary speech.

The method of using artificial neural network to extract room acoustical parameters is to train the neural network on a large set of reverberated speech examples with prior knowledge of acoustical conditions, i.e. RT, STI and EDT. After training, the artificial neural network is expected to output room acoustical parameters when a case not necessarily seen before is presented to its input. The trained artificial neural network can be regarded as an 'artificially intelligent ear' for a predefined purpose, for example, listening to the speech and extracting the RT. A block diagram of the training process is illustrated in Fig. 1. Reverberated speech utterances with known acoustical parameters are used as training examples. Training examples are pre-processed and conditioned to yield input vectors for the ANN. The output of the neural network and the corresponding acoustical parameter

(teacher) of the input example are compared to obtain the error. The training process is to iteratively update the internal parameters of the neural network (synaptic weights and biases) using certain optimising algorithms so that the means-square-errors between the teachers and the ANN outputs over all the training examples are minimised. In the retrieve phase, speech utterances as received by a microphone in the space to be measured are sent to the trained neural network via the pre-processor. The neural network then gives the required acoustic parameter as shown in Fig. 2.

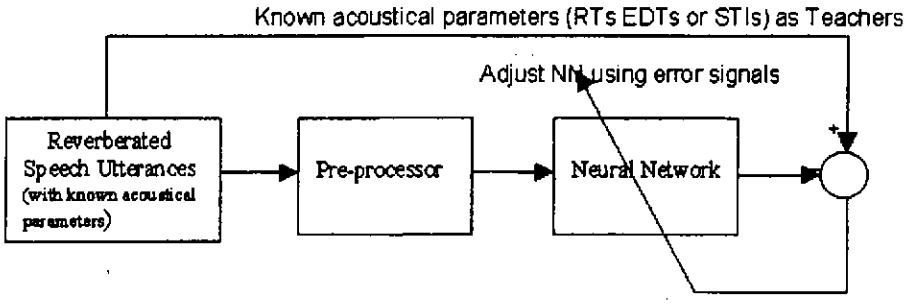

Fig.1 Block diagram of training phase

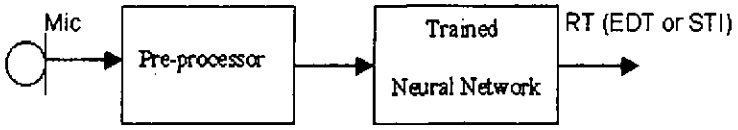

Fig-2 Block diagram of retrieve phase

2. PERCEIVED SPEECH IN ROOMS AND THE RELEVANCE

2.1 Reverberated Speech

The sound transmission characteristics of a room can be described by its impulse response as depicted in Fig. 3. The perceived sound *m(t)* is the convolution of source *s(t)* and the impulse response *h(t)* of the room in a noise free condition.

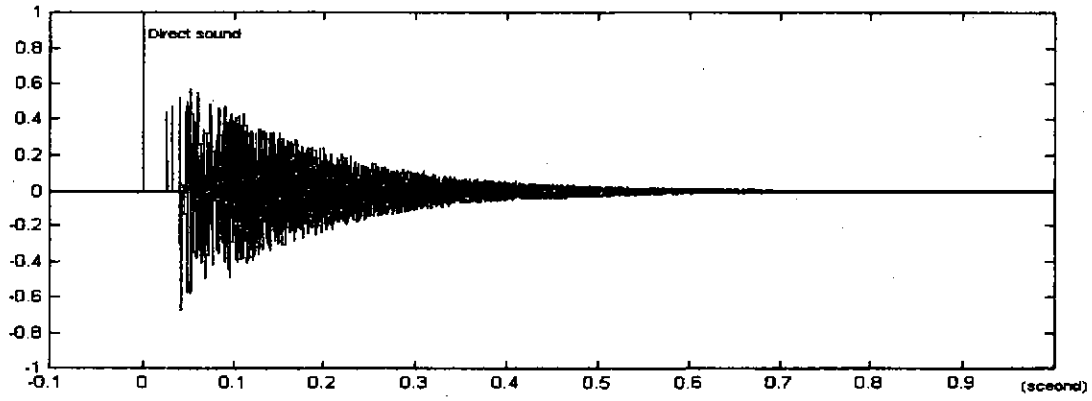

Fig. 3 An example of simulated impulse response of a room (normalised to direct sound)

$$m(t) = s(t) * h(t) \qquad \text{Equ.1}$$

When ambient noise *n(t)* is taken into consideration, Equ.1 becomes

$$m(t) = s(t) * h(t) + n(t) \qquad \text{Equ.2}$$

The convolution indicates that information of the room impulse response is contained in the perceived sound signals. If the impulse response is separated from the convoluted signals, all objective acoustic parameters can be obtained easily. It is therefore natural to consider deconvolution techniques. Classical deconvolution needs entire knowledge of original source signals and is sensitive to the noise in the measured signals. In most cases, it is preferable that room acoustic parameter measurement can be source independent and noise tolerant. Source independence allows extraction of acoustic parameters to be made for any text and any speaker. The so-called blind deconvolution technique is further considered[1], but it is currently underdevelopment and is not accurate enough for room acoustic measurements. Room acoustic parameters are, in fact, simplified features of impulse responses, so complete impulse response extraction is not needed. Extraction of room acoustic parameters can therefore be regarded as finding particular statistical features from reverberated speech. This could be achieved by means of machine leaning, provided that a large enough number of training examples are available and the artificial intelligence algorithms can generalise from these examples.

2.2 Build-up and Decay of RMS Sound Pressure

Speech signals and impulse responses are complicated stochastic processes. In RT and STI extraction, however, only a very low frequency sub-space of the reverberated speech is needed. Indeed others [2], have already considered speech envelopes as an identifier for characteristics of speech transmission channels. But artificial test signals were used instead of natural speech to achieve the required accuracy and repeatability. When excited by noise bursts, the measured short term RMS values of the sound pressure build-up and decay caused by the switch on and off of the test signal are a rough exponential increase and decay [3]. - see Fig. 4.

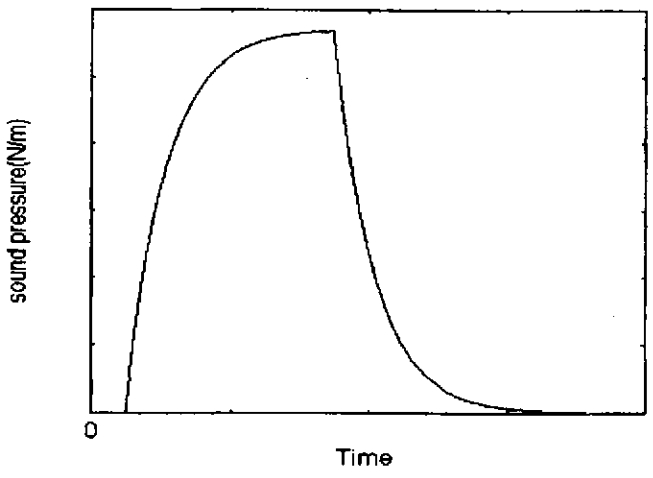

Fig. 4 Build-up and decay of sound pressure

The exponential rise and fall edges comprise information related to acoustical parameters. One can simply uses a ruler to measure the slope from Fig.4 and estimate the reverberation time and EDT through a brief calculation. However, this is the idealised situation. In reality the short term RMS sound pressure or energy as received by a measurement microphone in a room is much more complicated with speech. Therefore, a more sophisticated algorithm is needed to monitor the slopes of the speech envelope and extract the acoustic parameters.

3. ARTIFICIAL NEURAL NETWORKS AND PRE-PROCESSOR

3.1 Artificial Neural Networks

Artificial Neural Network Networks (ANNs) are inspired by human cognitive processes. ANNs possess memory, association, classification and approximation functionality and have the capability

of being trained to store and retrieve information and generalise problems from examples. Given adequate computing power, neural networks can model arbitrarily complicated non-linear processes[4]. In the past two decades, ANNs have been extensively studied and widely used to solve classification, approximation, feature extraction, generalisation and signal processing problems[5]. ANNs are particularly useful for complicated input/output mapping where analytical formulation is difficult but examples are available. When short-term memory mechanisms e.g. tapped delay lines, are used, they are also powerful in tackling temporal signals such as speech[6].

ANNs are massive connectionist networks of a large number of primitive neurons. These neurons are simplified models of neural cells in human brains. A single neuron model used in ANNs is depicted in Fig. 3. This model comprises a simple summation point U(.) collecting and adding up information from the environment and a non-linear processing unit f(.) providing basic processing power. It is trainable because the weights Wij are adjustable subject to certain learn rules. It memorises information by means of the values of weights and supports decision making and non-linear mapping utilising the non-linear activation function f(.). Even a single neuron model as shown in Fig. 3 possesses certain, though limited, capability to learn from examples and to make decisions. ANNs gain their intelligence through massive connections of a large number of such simple neurons. In theory, the more neurons and the more extensive the connections, the more powerful the ANN will be. However, training excessively large networks is difficult [6].

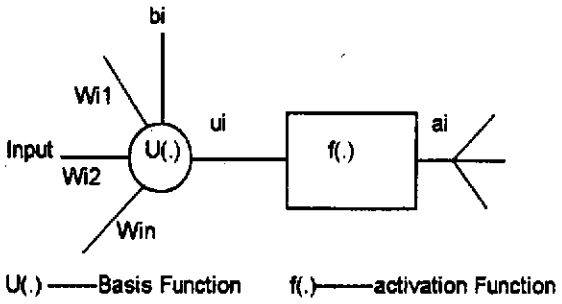

Fig. 3 Model for a neuron

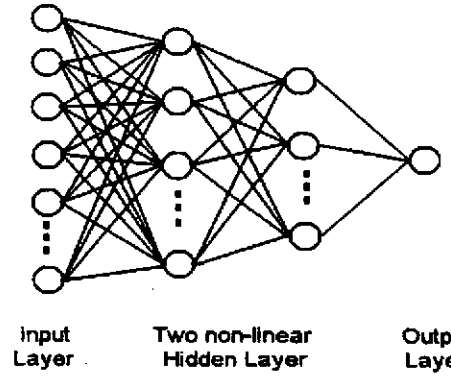

Fig. 4 Architecture of a multi-layer neural network

For this project, the ANN has a two non-linear hidden layered feed forward structure. The input layer is simply an interface between the external data and the first non-linear layer. All the neurons in the non-linear hidden layers have the structure shown in *Fig-3*. The non-linear activation function f(.) is sigmoidal

$$a_i = \frac{1}{1+e^{-u_i}} \qquad \text{Equ. 3}$$

and the basis function used is a linear combination of all the inputs and the bias of the neuron *bi*.

$$u_i(\mathrm{w},\mathrm{x}) = \sum_{j=1}^{n} w_{ij}x_j + b_i \qquad \text{Equ. 4}$$

The output layer neuron used a linear activation function. The dynamic equation of the ANN is:

$$u_i = \sum_{j=1}^{N_{l-1}} w_{ij}(l)a(l-1) + \theta(l) \qquad \text{Equ. 5}$$

$$a_i(l) = f(u_i(l)) \quad 1 \le i \le N_i; \quad 1 \le l \le L \qquad \text{Equ. 6}$$

Where l represents the layer number. (the input layer is layer 0.) The training of the network used the back-propagation method[7,8] : the connection weights were updated according to a gradient-

type learning formula Equ. 7, with the mth training sample $a^{(m)}(0)$ and corresponding teacher $t^{(m)}$ pairs, so as to minimise the energy function E defined by Equ. 8.

$$w_{ij}^{(m+1)}(l) = w_{ij}^{(m)}(l) + \Delta w_{ij}^{(m)}(l) \qquad \text{Equ. 7}$$

$$\text{minimise } E = \frac{1}{2}\sum_{1}^{m}\left[t^{(m)} - o^{(m)}\right]^2 \qquad \text{Equ. 8}$$

The back-propagation training process follows a chain-rule

$$\Delta w_{ij}^{(m)}(l) = -\eta \frac{\partial E}{\partial w_{ij}^{(m)}(l)} = \eta \delta_i^{(m)}(l) f'\left(u_i^{(m)}(l)\right) a_j^{(m)}(l-1) \qquad \text{Equ. 9}$$

where the commonly used term error signal $\delta_i^{(m)}(l)$ is defined as

$$\delta_i^{(m)}(l) = \frac{\partial E}{\partial a_i^{(m)}(l)} \qquad \text{Equ. 10}$$

For simplicity, the bias values *bi* are included in weight vectors by using an extra weights wo.

3.2 Pre-processor

The efficiency of a neural network depends largely on how the input data is presented to the network. A common practice is to insert a pre-processor between the real world signal and the input layer of the neural network. The pre-processor conditions the signals and converts them to a suitable format for the neural network. In many cases, pre-processors can also reduce the number of input data to the neural network. In this project two different types of pre-processors were developed, namely time domain short-term RMS detector and frequency domain pre-processor. The frequency domain pre-processor can reduce input data from 960,000 points to 25 points in dealing with long-term speech signals, providing a data compression rate of 38400:1.

Time domain short-term RMS detector as depicted in Fig.5 is to perform the following functions: (I) Implementation of short-term memory mechanism; (ii) Detection of short-term average RMS value or envelope of speech signals; (iii) Conversion to a information representation format suitable for ANN, and (iv) Normalisation of input signal.

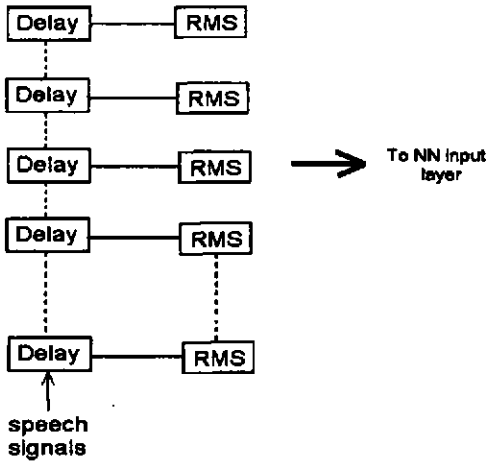

Fig.5 Time domain short-term RMS detector

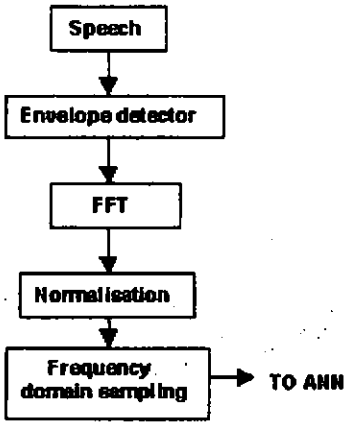

Fig.6 Frequency domain pre-processing

Time domain pre-processing enables presentation of exponential rise and fall edges caused by utterances to the input layer of the ANN. This approach is straightforward and is found suitable for short term speech samples(typically several seconds). To generalise the problem, it is desirable to use longer speech samples to allow characteristics of individual utterances to be averaged out. In such a case, a large number of input neurons will be required and training will become less efficient. To enable better performance when training on long term speech, a frequency domain pre-processor as depicted in Fig. 6 was developed. The averaged rise and fall rates of speech envelope over long time are mapped onto the very low frequency band in the frequency domain. The very low frequency band of room acoustical parameter interest ranges from 0.63Hz to approximately 20 Hz [2]. By frequency domain sampling at some characterised frequency points, relevant information is retained and room acoustical parameters can be extracted by a subsequent neural network.

4. TRAINING AND SOME RESULTS

4.1 Sample Generation

To apply supervised neural networks to a practical problem, it is essential that the example data sets for training and validation test be available. It is impossible to obtain the required amount of data by real room measurements. Simulated room impulse responses were therefore used as an alternative. Validation sets were strictly formed from examples which have not been used to train the ANNs for a more rigorous test.

4.2 Training the ANNs to Extract RT and EDT

Anechoic speech utterances ONE, TWO and THREE read by male and female narrators were used. The convolution of the anechoic speech and 10,000 different room impulse responses with RTs and EDTs from 0.1 to 5 seconds were performed to generate a large data set. Time domain short-term RMS type pre-processor was adopted. Validation result shows that the ANN can accurately extract RT and EDT from a closed set of reverberated speech. Fig. 7 shows a gradual reduction of the prediction error over 10 randomly chosen validation samples as training proceeds. After 1 million iterations, for any test datum input, the RT and EDT error is less than 3.3 ms, providing a maximal percentage error of 6.6% for RT and 5.1% for EDT. Fig.8 shows a worst case plot

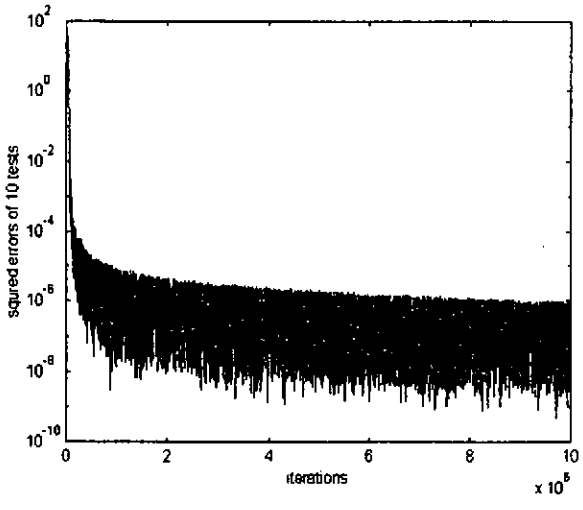

Fig. 7 Ensemble validation test vs. training

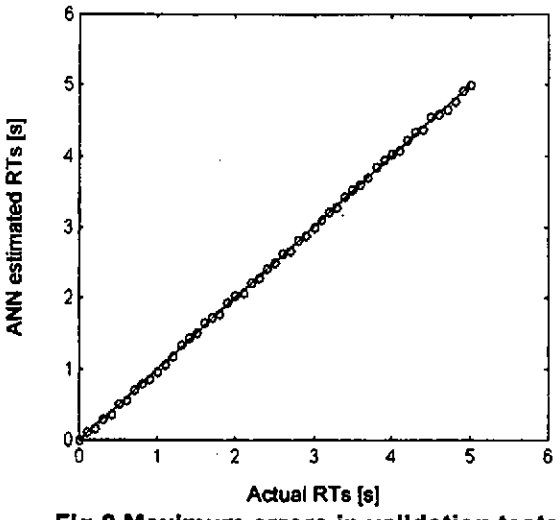

Fig.8 Maximum errors in validation tests

4.3 Training the ANNs to Extract STI

Speech Transmission Index can be extracted from speech utterances in a similar fashion[9]. Frequency domain pre-processor was used for STI extraction. The speech signal is two repetitions of utterances 'ONE-TWO-THREE-FOUR'. Training/validation samples cover RTs from 0 to 7 seconds and S/Ns from 10 to 60 dB. A total number of 158,400 examples are used for training and validation. Fig. 9 plots the maximal percentage errors found in the validation test.

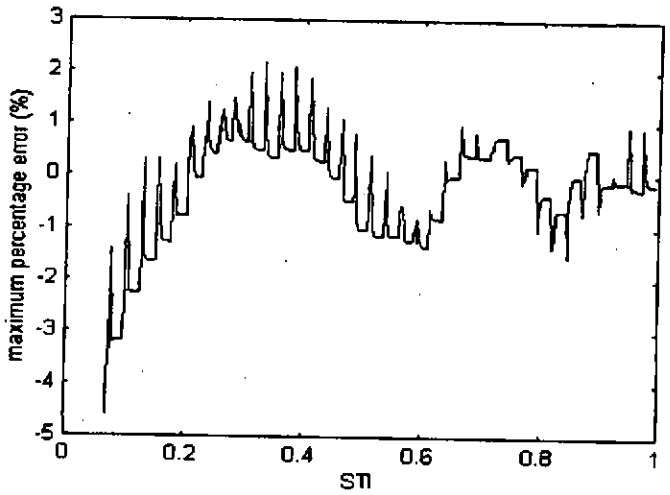

Fig. 10 Maximum prediction error for STI

4.4 Source Independence

New utterances 'FOUR' and 'FIVE' were used to test the ANN trained on utterances 'ONE, TWO, THREE" to identify its independence to the source. The results are presented in Table-1. For each reverberation time 10 different tests were performed and the error listed in the table are the worst cases. It can be seen that the trained neural network does show certain independence to utterances used, though, in the middle of the reverberation time range, relatively large errors occur. To investigate the feasibility of extract room acoustic parameters from arbitrary speech, long term speech (60s) in conjunction with frequency domain pre-processor was used. This approach enabled monitoring and averaging out speech signals over a long period in frequency domain and has shown improved robustness to different speeches and speakers. A better than 0.2s resolution was achieved with an open set of speech signals.

Table-1

	With One NEW Utterance		With TWO NEW Utterances	
Expected value [s]	NN predicted value [s]	Percentage error [%]	NN predicted value [s]	Percentage error [%]
0.50	0.50		0.50	
1.00	1.03	3.0	1.00	—
1.50	1.58	5.3	1.46	2.7
2.00	2.59	30	2.95	48
2.50	3.26	30	4.29	72
3.00	4.07	36	4.64	55
3.50	5.38	53	4.10	17
4.00	3.76	6	5.04	26
4.50	4.24	6	5.52	23
5.00	6.23	25	6.05	21

5. CONCLUSION AND DISCUSSION

This paper has presented a novel method to extract room acoustical parameters from speech utterances using artificial neural networks. The proposed method can correctly identify room reverberation time, early decay time and speech transmission index from a closed set of speech examples. This enables acoustical parameters to be measured using pre-recorded speech signals and so hopefully facilitating occupied measurements. The proposed method can generalise different impulse responses. There is potential for this method to be further developed to realise source independent measurement, extracting acoustical parameters from arbitrary speech signals. It is anticipated that if more words and speech samples are presented in the training phase, the trained networks can be made more source independent. This may require the development of more sophisticated pre-processors and neural network models. Although training the artificial neutral networks is somewhat tedious, this is a one-off process. Once the neural networks are trained and connection weights obtained, implementation is straightforward and extraction of parameters can be made in real time. Moreover, the trained artificial neural networks can be implemented on a hardware platform at a low cost. This may lead to the development of a new type of intelligent instrumentation.

ACKNOWLEDGEMENT

This project is funded by the Engineering and Physical Sciences Research Council, UK (GR/L89280). Insightful suggestions from Dr. P. Darlington are gratefully appreciated.

REFERENCES

[1] H. Attias and C. E. Schreiner, Blind Source separation and Deconvolution: The Dynamic Component Analysis Algorithm, Neural Computation 10, pp 1373-1424, 1998
[2] H. J. M. Steeneken and T. Houtgast, A Physical Method for Measuring Speech Transmission Quality, J. Acoust. Soc. Am.67(1), Jan. 1980
[3] M. Barron, 'Auditorium Acoustics and Architectural Design', E & FN Spon, an imprint of Chapman & Hall, 1998
[4] G. Cybenco, Approximation by superpositions of a sigmoidal function, Mathematics of Control, Signals and Systems, 2:303-314,1989.
[5] S. Y. Kung, Digital Neural Network, Prentice-Hall information and system science series, 1993
[6] S Haykin. Neural Networks: A Comprehensive Foundation, 2nd edition, Prentice Hall, 1999
[7] D. E. Rumelhart, G. Hinton et.al. 'Learning internal representations by error propagations' in 'Parallel distributefd processing', Vol1 MIT Press, 1986
[8] M. Riedmiller,' Advanced supervised learning in multi-layer perceptrons-from back propagation to adaptive Algorithms' Special issue on Neural network(5),1994
[9] F. Li and T. J. Cox, " Predicting Speech Transmission Index from Speech Signals Using Artificial Neural Networks", Proceedings World Muticonference on Systemics, Cybernetics and Informatics, Vol. VI., pp. 43-47, Orlando, Florida, USA, July 2000