Proceedings of the Institute of Acoustics

EXTRACTING STI FROM ARBITRARY RUNNING SPEECH

Trevor. J. Cox School of Acoustics and Electronic Engineering, Salford University, Salford UK.
Francis. F. Li Dept. of Computing and Mathematics, Manchester metropolitan University, UK.

1. INTRODUCTION

Speech intelligibility is an important concern in spaces and transmission channels such as lecture rooms, theatres, auditoria, PA systems and telephone lines. Subjective and objective methods are can be used to quantify speech intelligibility. Subjective methods, based on human perception, use real yet controlled speech sounds, for example phonetically balanced word lists. This enables assessments to be made using quasi-natural speech excitations. Nevertheless, it is known that the involvement of human perception introduces reproducibility and repeatability problems, and makes tests lengthy and expensive to conduct. Objective methods, represented by parameters such as STI, are physical methods using artificial test signals and measurement instruments to obtain more accurate and repeatable results. However, the use of artificial test signals as stimuli hinders occupied measurements due to logistical and technical obstacles. Many dilemmas encountered in occupied measurements could be overcome if naturally occurring sound sources such as music or speech could be used. The endeavour to develop such a non-invasive method is not new, for example Steeneken and Houtgast proposed a method to estimate STIs by comparing the envelope spectra of source and received running speech, but at a cost of compromised accuracy [1]. Although the attractiveness of this approach is well appreciated, it appears to be rarely used in practice.

Inspired by the fact that human hearing can sensitively differentiate reverberation times to a fraction of a second, artificial intelligence methods were applied to room acoustic parameter estimation problems by the authors [2,3,4]. In particular, artificial neural networks (ANNs) with purpose designed pre-processors were developed to extract reverberation parameters and STIs from speech utterances. The work began by looking at a time domain approach with separated utterances, and more recently has moved to running speech as a source excitation. This paper presents the recent refinements of the neural network method to obtain accurate STI values from received running speech especially arbitrary speech. The proposed method exploits the envelope spectrum method developed by Steeneken and Houtgast, refining the method with modern digital signal processing techniques. The significance of the proposed method is that it potentially enables source independent measurement, i.e. STI from received arbitrary speech. Crucially, this extraction is done without monitoring the speech at its source.

2. SPEECH ENVELOPE SPECTRUM METHOD

Consider the method for obtaining the envelope spectra as developed by Steeneken and Houtgast [1]. A short rectangular window with a width of T is moved along a running speech signal *s(t)* of length *L*, the square of the windowed portion divided by the average value of the squared long-time speech gives intensity function *i(t)*.

$$i(t) = \overline{s^2(t)}^{T} \Big/ \overline{s^2(t)}^{L} \qquad (1)$$

The envelope spectrum is defined as the spectrum of this intensity function. The process is illustrated in Figure 1. In the original technique, envelope spectra were estimated by repeatedly

passing squared and accelerated speech through 1/3-octave acoustic filter banks and therefore the resolution was restricted by the hardware filter banks available. 45- 60s speech extracts were found appropriate to secure reasonably stable envelope spectra and fourteen data points at the centre frequencies of 1/3 octave bands from 0.63 Hz to 12.5 Hz were found adequate for speech intelligibility assessments.

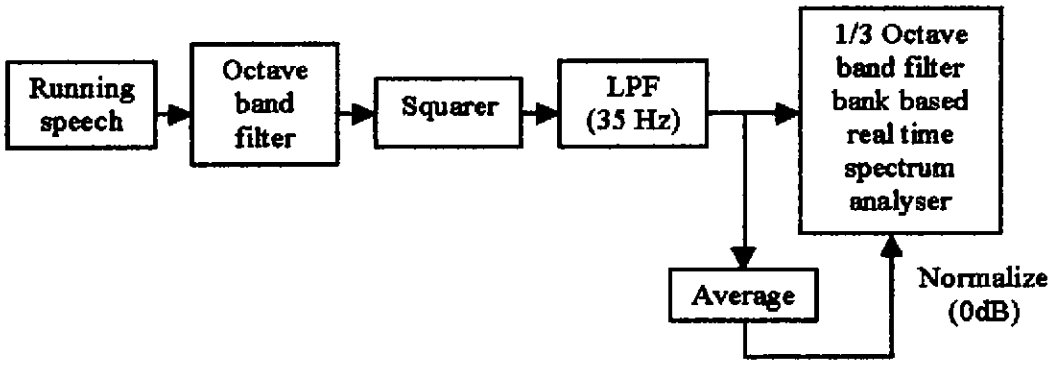


Figure 1. Traditional envelope detector and envelope spectrum analyser

Steeneken and Houtgast suggested that the Modulation Transfer Function (MTF) could be approximately obtained from the difference between the envelope spectra (in dB) of the original and transmitted speech [1]. The validity of this approximation was demonstrated by some measurement results: the STIs obtained from envelope spectra of running speech and those obtained using standard method gave reasonably good agreement (a correlation coefficient of 0.971). However, in both the original paper and the standards [1,5], it is pointed out that extraction MTF from envelope spectra has compromised accuracy. Once the MTF is obtained, the STI can be further calculated by a series of averaging, limiting and weighting [5,6].

The use of envelope subtraction can be viewed as a linear approximation of the true input-output system as shown in Figure 2. In the frequency domain, the transfer function H(jω) of a room describes the input-output relationship of the fine structure of signals in a transmission system, while modulation transfer function MTF(F) approximately relate the (low frequency) envelopes of signals.

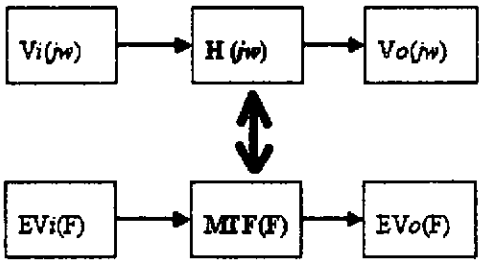


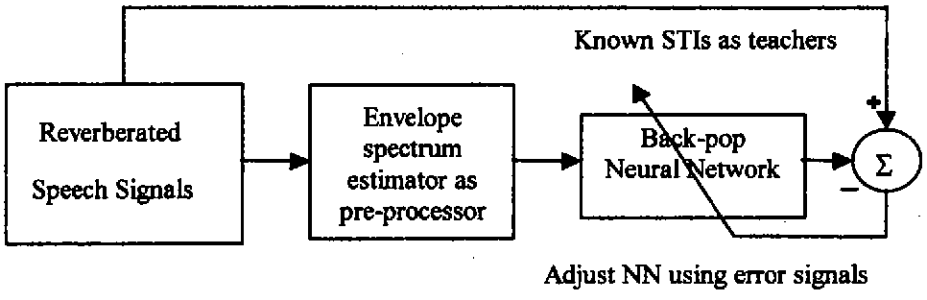
Figure 2. A speech transmission system as an envelope transmission system. V_i and V_o are the input and output spectra, EV_i and EV_o are the input and output envelope spectra..

Modulation transfer functions describe transfer characteristics of low frequency, pure tone modulated, stationary noises. The errors incurred when using speech envelope spectrum subtraction to determine MTFs mainly stem from the deviations of speech from the pure tone

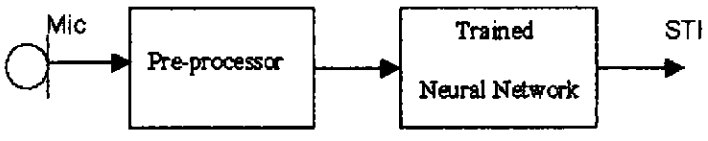
modulated noise. This can be further broken down into two aspects: (i) A speech signal is a complicated non-stationary stochastic process. The spectrum of each individual speech extracts deviates from the noise carrier used for standard MTF measurements. (ii) Envelopes of speech are complicated non-periodic signals, while the standard STI requires periodic modulation. It is known that speech signals are extremely difficult to statistically model. As a result, traditional mathematical tools are not readily available to precisely formulate the relation of MTF and the difference of original and transmitted speech envelopes. Artificial neural networks are therefore considered to statistically learn from examples to accurately map such relationships. Furthermore, explicit knowledge of the input speech signal will not be given to the artificial neural network; this is a more realistic simulation of how humans perceive room acoustics, but makes the STI extraction problem much more difficult.

3. ARTIFICIAL NEURAL NETWORK METHOD

Artificial neural networks can learn to compensate for varying statistical features of speech through examples. In fact, neural networks are particularly suitable for such tasks [6]. The neural network can also incorporate the computation of STI from the MTF using their non-linear function mapping capability. Supervised artificial neural networks are therefore considered to accurately map received speech excerpts onto STIs via an envelope spectrum estimator. The framework of the proposed neural network method is illustrated in Figure 3. In the training phase, the ANN is given a large number of reverberated speech examples and the corresponding STI values in a format of speech example-STI pairs. The neural network learns statistically from examples and performs non-model based regression under a supervised training regime. This is the training phase illustrated in Figure 3(a). The neural network model used here is the well-known multi-layer feed-forward network and is trained by the back propagation algorithm. The training is to iteratively update internal parameters of the artificial neural network, so that the mean square error between the neural network output and the true objective parameter values is minimised over all training examples. Once trained, the ANNs can correctly predict acoustic parameters from received speech samples not necessarily included in the training phase - Figure 3(b).



(a)



(b)

Figure 3. Illustration of generic ANN method for objective parameter extraction

The signal pre-processor plays a very important role in the success of applying neural networks in this case. In this case, a considerable amount of data reduction is required. An improved envelope spectrum estimator derived from classical STI system is adopted.

3.1 The Pre-processor

Envelopes of speech signals are detected as illustrated in Figure 4. Figure 5 is a block diagram of the pre-processor.

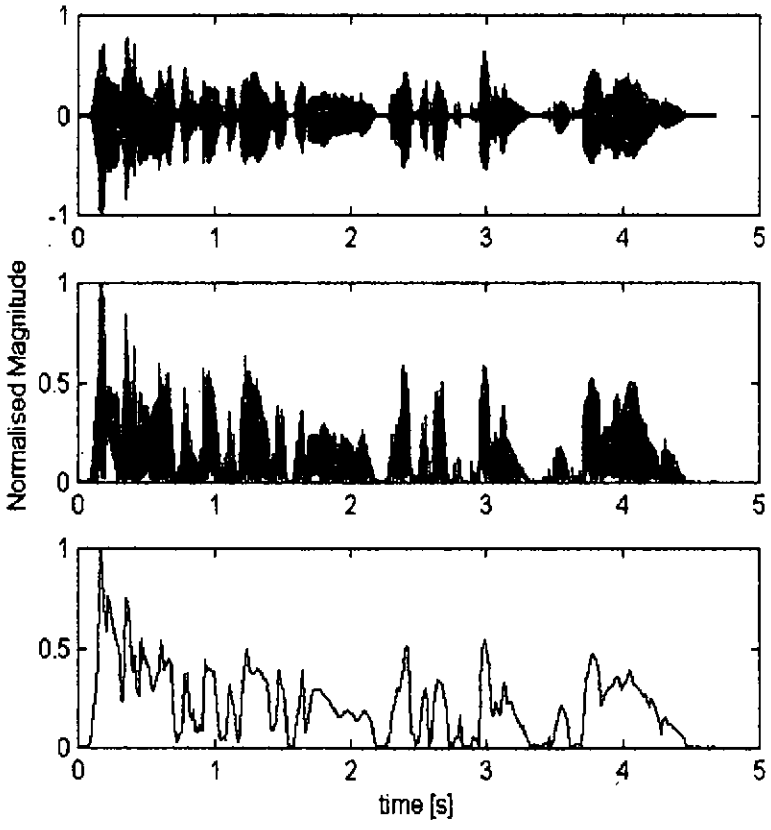


Figure 4. An example of detected envelope;

top: speech extract;

middle: envelope;

bottom: Low pass filtered envelope).

Not surprisingly, It is found that the window width and FFT length of the spectrum estimator has a significant impact on the accuracy of the process. According to the standard STI method, 14 data points at central frequencies of 1/3-octave bands from 0.63 to 12.5 Hz are used. Such a frequency domain information extraction is found adequate in training ANN on single speech extract in octave bands.

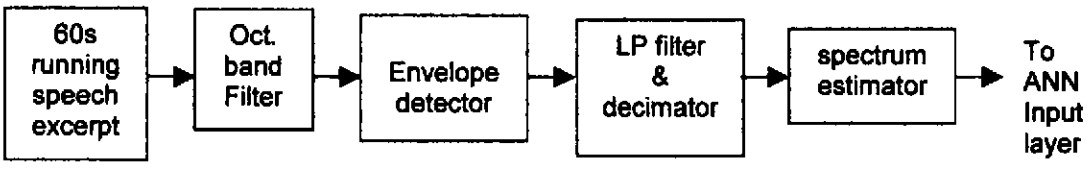


Figure 5. Block diagram of speech signal pre-processor

3.2 Training and validating set

In order to secure good generalization, a large data set is preferable. Simulated impulse responses are used to generate examples via convolution with anechoic speech. White noise interference is used. A stochastic model that hypothetically covers a large number of possible room conditions (a superset of realistic impulse responses) is used. Following the standard approach in ANN research, the data set is split into two halves, one for training and the other for validation. The rigorous tests are warranted by entirely separating training and validating sets.

3.3 Neural networks

Two non-linear hidden layers are used in the neural networks. The optimised number of neurons are determined empirically. The neural network is trained under a supervised regime as an approximator. This is achieved by randomly applying the pre-processed speech examples to the input of the neural network and minimising the mean square errors between the teachers (known STIs) and the output of the neural network over all the examples in the training set. The optimisation is done by updating the connection weights within the neural network using the 'delta learning rule' [7,8]. In our case, a variable training rate is used to speed training [9].

4. TRAINING AND VALIDATION RESULTS

Two sets of neural networks are trained and validated. The first set is intended to train the neural networks to work with one particular speech extract so that the feature of that particular speech stimulus is built into neural network; this is a one-net-one-extract approach. In the retrieve phase, the trained ANNs give very high accuracy in STI estimations under various acoustic conditions. The second set of ANNs are trained on a number of different anechoic speech, intending to generalize to arbitrary speech. This is referred to as a one-net-multiple-extract approach.

4.1 One-net-one-extract

The neural networks used here are expected to learn from examples to generalize all possible impulse responses and noise interferences. It must also learn the statistical feature of that particular speech. The maximum prediction errors found over all the tests are 0.02 STI as shown in Figure 7.

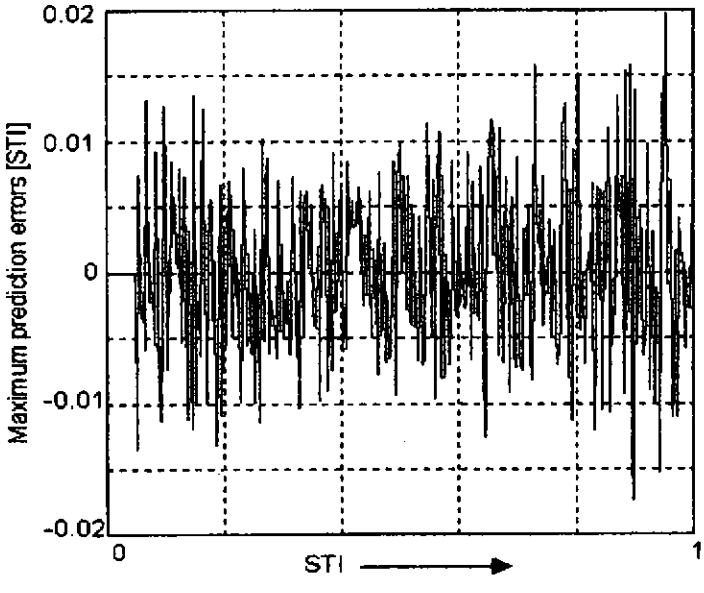


Figure 7. Maximum prediction errors found in validation tests using a one-net-one-extract basis.

This indicates the proposed method can achieve an accuracy comparable to those obtained using the standard STI method, which typically also has a standard deviation of 0.02. The correlation coefficient between the ANN estimated and true STI is 0.9999. This is much higher than the 0.971 obtained when applying real speech to the traditional STI method and having input and output information. Consequently, this method enables STI measurement by broadcasting a standard anechoic speech extract. This is a useful additional measurement capability on top of the standard STI method. It enables a test signal to be broadcast which being naturalistic which will disturb occupants less, and so can form a non-invasive, in-use test method. It would be even more useful, however, if the technique could work with arbitrary speech.

4.2 One-net-multiple-extract

Training one neural network on several speech extracts is explored. This is intended to generalise the network to arbitrary speech so that one network can estimate STI from received speech not seen in training. Furthermore, it is hoped to do this without resorting to monitoring the source. Initial investigation shows that ANN system has some potential to achieve this source independence.

It is necessary to update the traditional envelope spectrum estimator to improve accuracy. Since the development of the original STI method, digital signal processing techniques have advanced greatly and these can be exploited to give better quality information to the neural network. 18 different anechoic speech examples. The examples consist of 3 contrasting texts read by 6 untrained native English speakers. Figure 8 shows the maximum errors found in the validation tests.

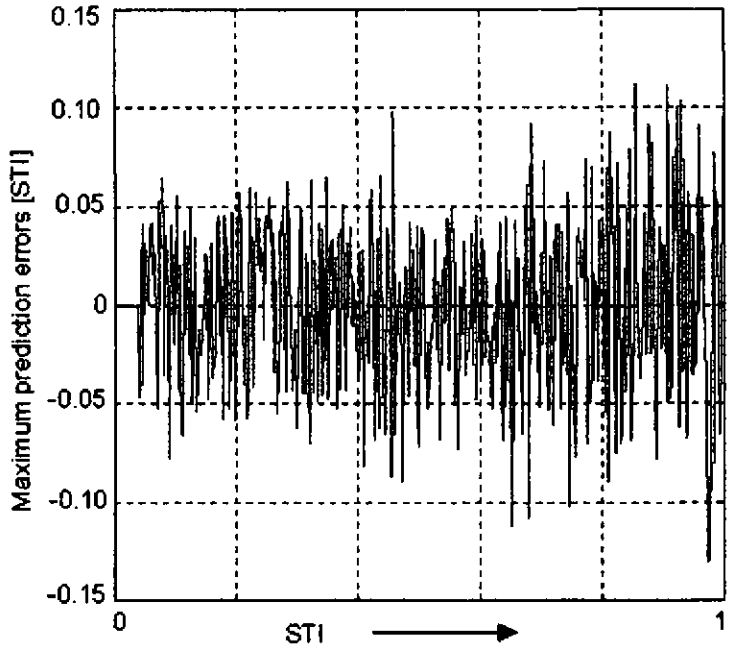


Figure 8. Maximum prediction errors found in arbitrary speech cases.

Figures 9 is another illustration of the accuracy obtained when using arbitrary speech. The maximum prediction error for STI found in all cases is 0.13, and the correlation coefficient between actual and predicted STI is 0.995. Problems arise because the low frequency statistics of the anechoic speech are not sufficiently stable for accurate estimation. These low frequency statistics alter as the text, speaker and utterance are changed. Consequently, better accuracy can be obtained by averaging over several different speech extracts to stabilise the low frequency statistics, especially using both male and female speakers. When averaging over three different

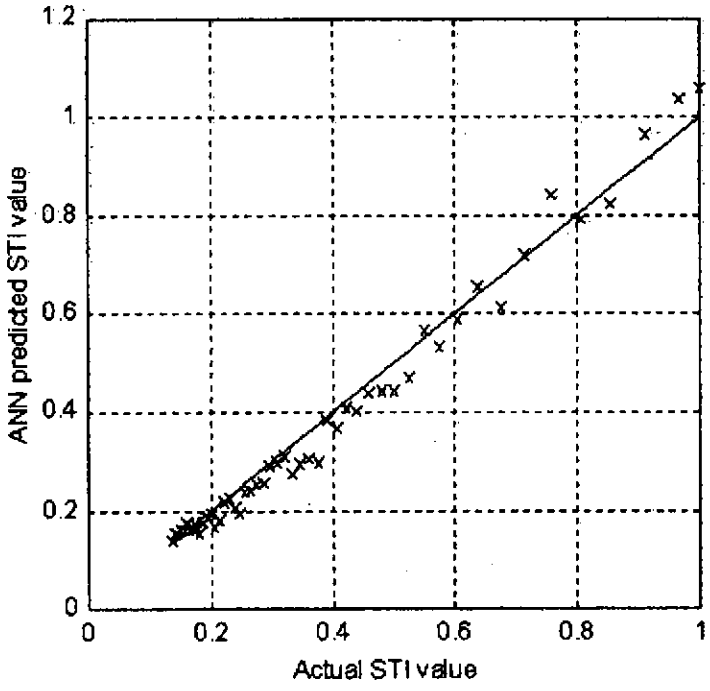


Figure 9. An example of 50 tests with an arbitrary speech extract not included in training.

speech extracts read by different narrators, prediction errors can normally be reduced to less than 0.1. When averaging is carried out over even more extracts, accuracy can be further improved. Figure 10 shows the ANN estimated values obtained by averaging over 6 speech excerpts read by different narrators. In this case, estimation error drops down to 0.05, which although not as good as the standard method, is probably accurate enough for STI evaluation. This method, however, has its limitations. It is unrealistic to change narrators so frequently in many in-use cases. Moreover, averaging may not always produce better accuracy, since one may have several narrators having similar reading styles. A more sophisticated approach is therefore sought.

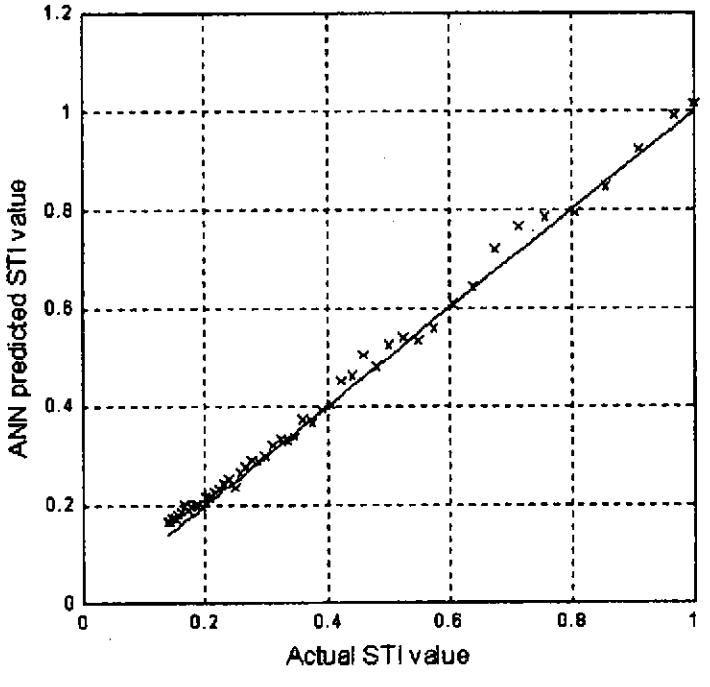


Figure 10. An example of 50 tests averaged over 6 arbitrary speech extracts not included in training.

If more information regarding the occurrences of words, syllables and phonemes could be given to the neural network, the network would have greater ability to adapt to different texts and different reading styles. This can be easily achieved by giving the network input speech knowledge, but it is of fundamental interest to try to get a system to work with output knowledge only. After all, trained humans can estimate speech intelligibility without direct knowledge of particular anechoic speech extracts. There are various ways of achieving this. For example, a simple threshold detector can be used to locate the location of syllables or utterances, and this additional information fed to the neural network. Using these types of techniques, the maximum prediction error reduces to 0.06 - 0.08 STI. If averaging over more than one speaker is used in conjunction with this technique, prediction errors can be further reduced. Problems arise, however, because standard techniques for syllable detection do not work in noise or highly reverberant conditions. Consequently, the improved accuracy only occurs for high values of STI.

5. DISCUSSIONS AND CONCLUSIONS

A neural network method that estimates STI from received running speech signals is developed and validated via simulations. The accuracy is comparable to that achieved by measurements with standard artificial test signals when limited to a one-net-one-extract approach or a closed set of speech examples. Source independent extraction of STI with knowledge only of the received speech is explored. It seems that the proposed ANN method has a certain capability to adapt to different speakers and texts. The actual and ANN estimated STIs show reasonable agreement when testing with speech extracts not seen previously by the ANN. Currently the pre-processor must reject 99.99% of the speech data to make the input vectors to the neural network sufficiently small to allow training. The question that remains unanswered, is whether in this rejected data, there is information to enable the neural network to deal with the arbitrary speech, and whether a method to extract this data can be developed.

6. ACKNOWLEDGEMENT

The work was funded by the EPSRC under grant GR/L89280.

7. REFERENCES

1. H. J. M. Steeneken and T. Houtgast, The temporal envelope spectrum and its significance in room acoustics, Proc. 11th ICA, Vol. 7, Paris 1983 ,P 85-88.
2. T. J. Cox, F. Li and P Darlington., 'Extracting room reverberation time from speech using artificial neural networks', J.Audio.Eng.Soc. 49(4) 219-230. (April 2001)
3. F. Li, T. J. Cox, "Predicting speech transmission index from speech signals using artificial neural networks", Proceedings World Multi-conference on Systemics, Cybernetics and Informatics, SCI 2000' Vol. VI part 2, July, Orlando Florida, USA, July, 2000, pp 43-47,
4. F. Li, T. J. Cox, "Extraction of Room Acoustic Parameters from Speech Using Artificial Neural Networks" Proceedings of IoA Reproduced Sound 16, 2000.
5. IEC 60268-16:1998, (also BS EN 60268-16 and BS 60268-16), Sound system equipment, Pary 16: Objective rating of speech intelligibility by speech transmission index, 1998
6. H. J. M. Steeneken and T. Houtgast, A Physical Method for Measuring Speech Transmission Quality, J. Acoust. Soc. Am.67(1), Jan. 1980
7. S. Haykin. Neural Networks: A Comprehensive Foundation, 2nd edition, Prentice Hall, 1999
8. S. Y. Kung, Digital Neural Network, Prentice-Hall information and system science series, 1993
9. F. F. Li and T. J. Cox. Speech Transmission Index from running speech: A neural network approach. J.Acoust.Soc.Am. Submitted for publication.