Speech transmission index from running speech: A neural network approach

F. F. Li[a] and T. J. Cox
School of Acoustics and Electronic Engineering, University of Salford, Salford M5 4WT, United Kingdom

(Received 20 August 2001; revised 30 October 2002; accepted 14 January 2003)

Speech transmission index (STI) is an important objective parameter concerning speech intelligibility for sound transmission channels. It is normally measured with specific test signals to ensure high accuracy and good repeatability. Measurement with running speech was previously proposed, but accuracy is compromised and hence applications limited. A new approach that uses artificial neural networks to accurately extract the STI from received running speech is developed in this paper. Neural networks are trained on a large set of transmitted speech examples with prior knowledge of the transmission channels' STIs. The networks perform complicated nonlinear function mappings and spectral feature memorization to enable accurate objective parameter extraction from transmitted speech. Validations via simulations demonstrate the feasibility of this new method on a one-net-one-speech extract basis. In this case, accuracy is comparable with normal measurement methods. This provides an alternative to standard measurement techniques, and it is intended that the neural network method can facilitate occupied room acoustic measurements.
 [DOI: 10.1121/1.1558373]

PACS numbers: 43.55.Mc, 43.58.Gn, 43.60.Lq [MK]

I. INTRODUCTION

Traditionally, two different approaches are used to quantify room acoustics: subjective and objective assessments. Subjective measurements are based on human perception and so usually use music or speech signals. Objective measurements, on the other hand, use artificial test signals such as noise, to ensure reproducibility and repeatability. Much room acoustics research concerns quantifying acoustic quality in terms of objective parameters and this enables designs to be readily made and evaluated. Objective measurements, however, use high sound pressure levels that are usually unacceptable to audiences. This hinders objective measurements under occupied in-use conditions. Occupied measurements are important because it is well established that occupancy affects the acoustic, especially the absorption and background noise levels. It is suggested that many of the problems encountered in occupied objective parameter measurements could be overcome if the naturally occurring signals in a space, such as music or speech, were used as test signals. A technique to achieve this for speech is given in this paper.

Speech Transmission Index (STI) is a common objective parameters used to assess speech intelligibility of spaces and other transmission channels, such as classrooms, theaters, public address and telephony systems.[1–4] STI combines two major phenomena that affect speech intelligibility—reverberation and noise—to extract a single index that gives good correlation with subjective perception.[5] Moreover, STI method utilizes simple artificial test signals and enables portable instrumentation to be implemented.[6] Consequently, STI has been adopted as a mainstream objective parameter for speech intelligibility[7] and is included in standards and performance specifications. Nevertheless, it is known that in certain circumstances STI is not completely successful.[8]

The normal STI measurement method uses artificial test signals and so is not particularly well suited to occupied measurement. The standard method takes about 10–15 minutes to perform,[9] which is a long time to expect occupants to listen to noise and yet continue with their normal activities. For this reason, RaSTI was developed in the mid 1980s, and more recently, another new technique, STI–PA has been proposed.[9] Both RaSTI and STI-PA have reduced measurement time, but the test signal is still noise, and consequently neither are true nonevasive test techniques. Steeneken and Houtgast did, however, propose a method to estimate the modulation transfer function (MTF) and in turn STI from running speech.[10] It was suggested that the MTF is roughly estimated by comparing the envelope spectra of source and received speech signals. This technique works, but at a cost of compromised accuracy. For this reason, practical measurements of STI are rarely made with running speech, but still rely predominantly on artificial test signals.

Inspired by the fact that humans can sensitively differentiate reverberation times, artificial intelligence methods have been developed as a means of extracting objective parameters from speech. Previously, a time domain approach was applied to extract reverberation parameters from separate monosyllable word utterances.[11] The time domain method, however, is not applicable to running speech and has signal to noise ratio problems when estimating octave band parameters. A neural network method to estimate STI from running speech was proposed and a few pilot results published.[12] Since then, many refinements have been made, especially to the preprocessor, to form a more accurate and robust method. This paper will present the details of this new method to accurately estimate STI from running speech excerpts using artificial neural networks (ANNs). ANNs are

[a]Current address: Department of Computing and Mathematics, Manchester Metropolitan University, Manchester M1 5GD, UK.

systems that can perform nonlinear mapping, in this case from running speech to objective parameters. ANNs learn a mapping through experience, in this case by being exposed to many examples of running speech in rooms and the corresponding objective parameter values of the room. The inputs to the neural network are 60 s speech excerpts, preprocessed with an envelope spectrum estimator. This preprocessing exploits and extends the underlying mechanism of the envelope spectrum technique developed for the standard STI method. Refinements are needed to allow better resolutions and maintain a reasonable number of input neurons to facilitate machine learning. As it is the basis for many aspects of the project, the paper starts by discussing the envelope spectrum and STI method.

II. STI METHOD AND ENVELOPE SPECTRA

The STI method is closely linked to the envelope of running speech. Speech needs to retain its original envelope to be intelligible, the more the envelope is modified the poorer the intelligibility. A room acts as a low pass filter of speech envelopes, smoothing the envelope and hence degrading intelligibility. Moreover, ambient noise disturbs speech signals and reduces intelligibility. Both reverberation and noise cause the normalized low frequency envelope spectrum of speech to decrease.[13] Consequently, the MTF was introduced to quantify the combined effect of reverberation and ambient noise by means of modulation index reductions.[2] By properly combining and processing selected frequencies in the MTF, a single index STI is formed.[3] The measurement of the MTF is the core process in determining STI, and three different methods exist. Once MTFs are obtained, the STI can be calculated from 98 MTF data points via a series of weighting, limiting and averaging processes.[3–7] The three methods for obtaining the MTF are as follows.

A. Standard method

The envelope shaping effect of a transmission channel is determined by measuring the MTFs in seven octave bands (125 Hz–8 kHz) using sine-wave modulated noise with its spectrum shaped to be the same as typical long-term speech.[7] The modulation transfer function MTF is determined by the ratio of modulation index of output to input intensity. This process needs to be carried out for each of the 14 modulation frequencies, and hence the process is relatively slow. The advantage of this method is that it works with nonlinear systems, such as many public address systems.

B. Impulse response method

Schroeder[14] systematically reviewed and discussed MTF measurement methods from a signal and system perspective, and reconfirmed a relationship between MTF and room impulse response $h(t)$:

$$\mathrm{MTF}(F)=\frac{\left|\int_0^\infty h^2(t)e^{-2\pi jFt}\,dt\right|}{\int_0^\infty h^2(t)\,dt}\,\frac{1}{(1+10^{(-s/n)/10})}, \qquad (1)$$

where s/n is the signal to noise ratio, and allows the effect of ambient noise interference to be included. This part of Eq. (1) was added by Steeneken and Houtgast.[4] Therefore, an MTF can be obtained by first measuring the impulse response. This is often done using a maximum length sequence signal or a swept sine wave. This approach becomes invalid when a system is nonlinear as the impulse response in Eq. (1) should be linear. This can be a serious limitation. For example, many speech reinforcement and public address systems use compressors to improve intelligibility.

C. Using the speech envelope spectrum

A short rectangular window is moved along a running speech signal, typically 40–60 s long. The square of the windowed portion is divided by the average value of the squared long-time speech signal; this gives the intensity function.[13] An average of the intensity function is taken, and the low frequency envelope spectra of the function found. The part of the envelope spectra important to speech intelligibility lies in frequency band 0–20 Hz. It was suggested that MTFs could be roughly estimated from the envelope spectra of original and transmitted speech.[10] Let $E_X(F)$ be the envelope of original speech and $E_Y(F)$ be the envelope spectrum of received speech, then $\mathrm{MTF}(F)$ is estimated by

$$\mathrm{MTF}(F)\approx E_Y(F)/E_X(F). \qquad (2)$$

This method was validated by empirical results showing that the STI obtained using this approach and measured through standard methods have a reasonably good agreement (a correlation coefficient of 0.971).[10] In both the original paper and the standards,[5,10] however, it is pointed out that MTFs obtained from speech envelopes using Eq. (2) have compromised accuracy.

III. NEURAL NETWORK METHOD

A. Rationale

Room effects are contained in the difference between received, $Y(\omega)$, and transmitted, $X(\omega)$, speech signals. The use of input and output envelope to gain the MTF as suggested above, can be regarded as a linear approximation of a squared linear time invariant filter:

$$\begin{aligned} &Y(\omega)=H(\omega)X(\omega),\\ &E_Y(F)\approx\mathrm{MTF}(F)E_X(F), \end{aligned} \qquad (3)$$

where $H(\omega)$ is the transfer function of the room, which precisely describes the input–output relationship of the signals. The MTF, on the other hand, approximately relates the envelopes of the input and output signals.

Such a relation described by MTFs would be accurate if (a) the envelope of speech were periodic and (b) the spectrum of speech were white with constant power per unit bandwidth.[14] Unfortunately, running speech is a complicated, nonstationary stochastic process, only approximately conforming to these criteria. This makes the mapping relationship between the envelope spectra and MTFs imprecise. Artificial neural networks are therefore considered to perform

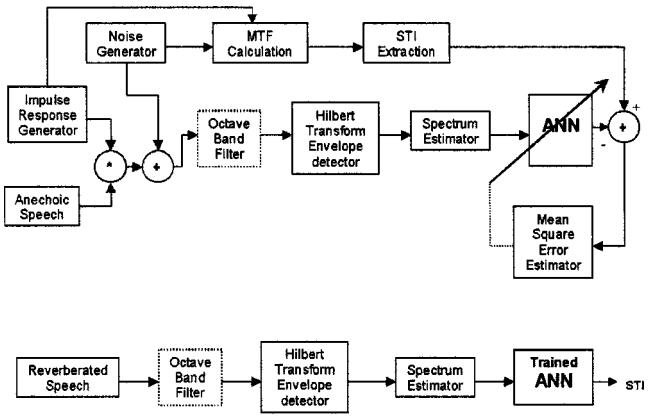

FIG. 1. Block diagram of the training (top) and retrieval (bottom) phases of the system.

this nonlinear mapping. The mapping is likely to be more accurate, because ANNs are inherently powerful nonlinear mapping engines.[15,16]

A speech excerpt can be regarded as a speech envelope modulated noise, once it is chosen, its envelope spectrum and the spectrum of the carrier is fixed and the difference between these and the standard test signal can be obtained. The ANN algorithm is used to perform nonmodel based regression to memorize features of speech samples and compensate the errors found in the estimated MTFs. From MTF to STI is a deterministic nonlinear limiting and linear combination process. ANNs are known to be able to calculate any computable nonlinear and linear functions and therefore should be able to perform these calculations.[17] As one step further, it is possible to combine the accurate MTF estimation and the subsequent calculation of STI value into one neural network system, i.e., use the neural algorithm to map received envelope signals directly onto STI values. This one stage mapping is useful as it simplifies implementation. Such an ANN network method is illustrated as a block diagram in Fig. 1.

Two phases, training and retrieval are involved in the development of the ANNs and applying them to real world problems. In the training phase, the ANNs learn from examples, memorize related information and generalize from a closed set of training data to a class of cases including those not in the training set. This is achieved by presenting a large number of examples to the ANN and utilizing multivariable optimization techniques to minimize the total errors between the actual STI value and the output of the ANN. Examples used to training the ANN are generated using simulation techniques. Convolutions of anechoic speech and simulated impulse responses are used as transmitted speech examples. The expected STI values are obtained from the impulse responses and additional ambient noise. The knowledge of the original envelope spectrum is implicitly built into the ANN as part of training. As a result, there is no need to monitor the original speech in the retrieve phase. This reduces two channel measurement to one channel. However, as shall be discussed later, the drawback is that it is limited to a one-speech-one-net scheme, i.e., a particular neural network

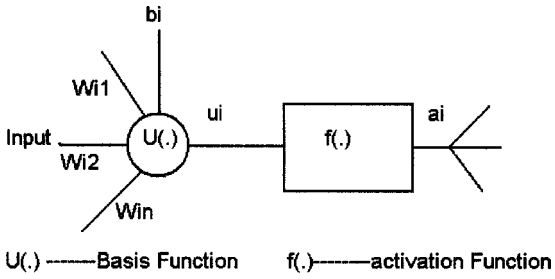

FIG. 2. The neuron model.

learns to memorize features of a particular speech excerpt and works with that excerpt only.

B. Neural network architecture

A nonlinear, multilayer, feed forward network trained by the back-propagation algorithm[17–19] is chosen for this work. Backpropagation is a successful algorithm; the penalty of using this type of neural network can be in excessive training times. The fundamental building blocks of the neural network is the cellules processing neuron unit as depicted in Fig. 2. Typically, it comprises two functions: a linear basis function used to gather input signals, and an activation function $f(\ )$ to nonlinearly process the information. The basis function used is

$$u_i = \sum_{j=1}^{n} w_{ij} x_j + b_i , \tag{4}$$

where w_{ij} is the weight connecting the jth neuron to ith neuron, and x_j is the output of jth neuron. The activation function is

$$a_i = f(u_i) = \frac{1}{1+e^{-u_i}}. \tag{5}$$

The activation function is used for the two hidden layers to provide nonlinear mapping capability. The neural network is constructed in a feed forward fashion by interconnecting a large number of these simple neurons as shown in Fig. 3. The leftmost input layer takes signals from the preprocessor and distributes them to subsequent layers without processing the signals. There are two nonlinear hidden layers. The STI is a normalized index from 0 to 1. As a common ANN design

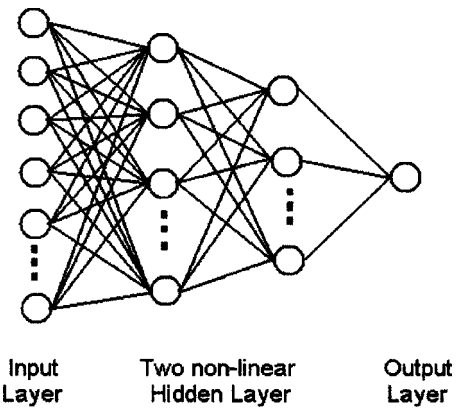

FIG. 3. Multilayer feed forward network architecture.

consideration, a hard limiting nonlinear neuron at the output layer could be to used to clamp the output values so that no output can possibly go beyond the interval of 0 to 1; however, it is found in this application that hard limiting the output reduces the back propagation of errors to the hidden layers in the early stages of training, and hence inhibits convergence. So, a linear summation function without a nonlinear activation function is adopted as the output neuron.

The neural network is trained under a supervised model as an approximator. The training, as shown in Fig. 1, is to iteratively apply the preprocessed speech examples to the input of the neural network and minimize the mean square errors between the teachers (known STIs) and the output of the neural network over all examples in the training set. The optimization is done by iteratively updating the connection weights (w_{ij} and b_i) within the neural network using the well-known back propagation algorithms.[15,17–19] Training is achieved using the delta learning rule:[15,16]

$$w_{ij}^{(m+1)} = w_{ij}^{(m)} + \eta \Delta w_{ij}^{(m)}, \tag{6}$$

where $w_{ij}^{(m+1)}$ is the new connection weights, $w_{ij}^{(m)}$ is the previous weights, $\Delta w_{ij}^{(m)}$ is the estimated maximum gradient according to backpropagation algorithm and η is the learning rate. Learning with too large a learning rate can cause the algorithm to diverge and can also mean important minima are missed, but too small a learning rate results in slow learning and the network is prone to being trapped in local minima. The standard backpropagation algorithm employs a constant learning rate, which is empirically determined. Experiments with the STI problem showed a constant learning rate, when the value is suitable, converges steadily, but is very slow. A modified learning rule, with variable learning rate, is used to speed up the training. The training phase is divided into three periods: output clamping, intermediate training and fine tuning. In the early stage of training, the ANN tends to gradually converge to output values within the [0 1] interval. In the intermediate stage of training, the ANN fits the details from training set. In the final stage, the ANN fine tunes itself to give the best performance for generalization. Different learning rates are used in first and last period. When the output of the ANN is beyond the [0 1] interval, larger steps are used to quickly drive the ANN to produce outputs in the [0,1] region. In the final fine tuning period, smaller steps are used. This modified learning thus can be expressed as

$$\eta = \begin{cases} (1.2-1.3)\,\eta & \text{when output} \notin [0,1], \\ \eta & \text{others}, \\ (0.3-0.5)\,\eta & \text{when error reduction} \rightarrow 0. \end{cases} \tag{7}$$

Such a modification to standard training method is found effective in speeding up the early phase of training and is numerically robust. Using a variable learning rate is common practice in neural network applications[20] and has been shown to speed training for a wide range of applications. Figure 4 shows the typical error reduction found in the early period of training. A rapid drop of ensemble errors can normally be achieved due to the enhanced error back propagation when output is beyond the interval of 0 to 1. Incidentally, If initial convergence is slow, the weights of the neural network are randomized and the training process restarted.

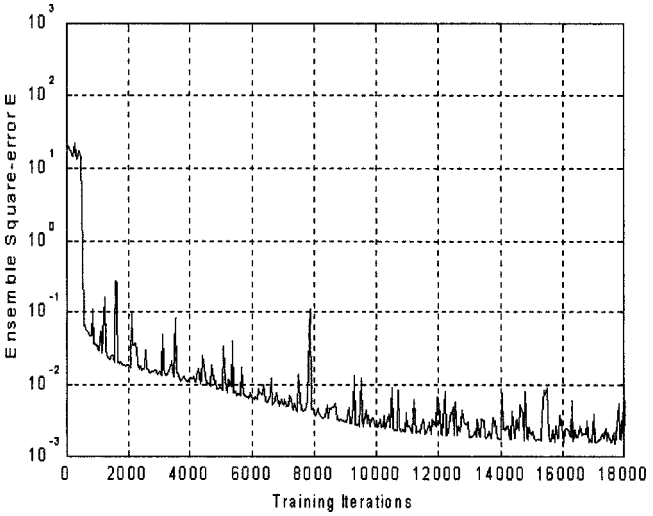

FIG. 4. Typical error reduction found in early phases of training. Each training iteration comprises a block of 50 examples. *y* axis is the sum of squared errors tested by randomly chosen blocks comprising 50 examples.

The termination criteria for neural networks are crucial. The requirement for generalization, i.e., that the network is successful with data not seen in training, indicates that seeking global minimum mean square error for the training set is neither necessary, nor the best solution. Tuning the ANNs to overly fit the training data set does not necessarily guarantee that the best generalization results. Therefore, the training is continuously monitored by the examination of the neural networks' response to a small set of validation data. The training stops when, either the predefined prediction accuracy is achieved or a sign of over-fitting occurs (consistently increasing mean-square-error), whichever the first. It is known that a 0.02 standard deviation is typical when using the STI standard method,[6] and so this is used as a termination criterion.

There are three other important aspects to specify in the ANNs: size, structures, and preprocessors. In theory, the more neurons and hidden layers a neural network posses the better its function mapping capability will be. Nevertheless, excessive number of neurons and too many hidden layers cause problems in practice,[15] as back propagation tends to be slow and learning becomes inefficient and extremely time consuming. As a practical rule of thumb, small sized ANNs are preferable if they suffice. The suitable size for the network is determined through trial and error. Training tends to be more efficient when input information is coded in a suitable format for the ANN.[15,17] As a common practice, a preprocessor, which functions to perform data reduction and signal conditioning is used to bridge the real world signals and the input layer of the ANN. In this application, an envelope spectrum estimator is the core of the preprocessor.

C. Preprocessor

There are three key issues to be considered in designing the preprocessor for STI extraction. First, useful information should be retained while redundant information should be reduced—only about 0.01% of the original data in the speech

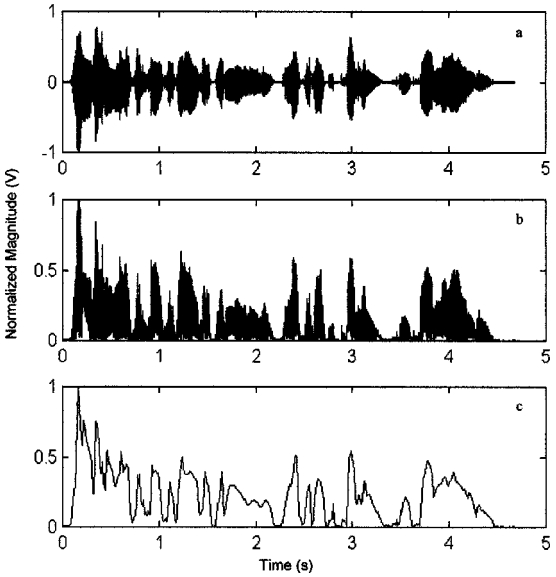

FIG. 5. Envelopes detected by Hilbert transform. Top is speech signal, middle is envelope, and bottom is 80 Hz low pass filtered envelope. Graphs normalized to maximum magnitude.

can be sent to the neural network otherwise the number of input neurons becomes too large. Second, existing knowledge should be used, otherwise the ANN has to model functions that could have been more efficiently processed using traditional means—this minimizes the load on the ANN. Third, the input vectors to the ANN should be normalized to between -1 and $+1$ as this speeds the training. The envelope spectrum estimator is naturally considered as the preprocessor. It can significantly reduce the amount of data in long speech excerpts, but maintains decisive information for STI values.

The best estimate of the long term envelope spectrum needs to be obtained. Traditional methods applied repetitions of the speech to spectral analyzers,[9,13] but nowadays more sophisticated algorithms are available. A Hilbert transform is used as it gives a better estimation of the time signal envelopes.[21] Accordingly, the envelope $ev(t)$ is

$$ev(t)=\sqrt{s^2(t)+s_h^2(t)}, \tag{8}$$

where $s_h(t)$ is the Hilbert transform of speech signal $s(t)$ defined by

$$s_h(t)=H[s(t)]\equiv\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{s(t-t')}{t'}dt'. \tag{9}$$

Figure 5 demonstrates a detected envelope using this method. Such a detector gives an unambiguous definition of signal envelopes; it is superior to techniques where window width and overlap have to be chosen.[21] Another attractive feature of the Hilbert transform based detector is that it can be performed via a fast Fourier transform (FFT) enabling a quicker implementation.

Figure 6 is a block diagram of the proposed preprocessor. One minute of speech is used. Octave bandpass filters are inserted when needed. Only low frequency contents found in the envelope spectra are of interest. Consequently, envelope signals are low pass filtered by a fourth order Butterworth filter with a -3 dB point of 75 Hz and the signal is then resampled at 160 Hz. The decimated envelope signals are then passed onto the power spectrum estimator to obtain the envelope spectrum. This is done using an overlapped Hanning window technique, and the windows are typically 2.5 s long. Envelope spectra are normalized to the average energy of speech signal excerpts (when a sine wave having a RMS value equal to the mean intensity of the speech signal passes through, 0 dB is obtained). The normalization has important practical and physical meaning.

FIG. 6. Block diagram of envelope spectrum preprocessor.

(1) Ensuring envelope spectra are not input signal level dependent.
(2) Expressing the frequency components of speech envelope with respect to the total energy.
(3) Including both speech envelope fluctuation and interference noise levels in the envelope spectra.[13]

Envelope spectra are frequency domain sampled and fed into the input layer of the neural networks. Not surprisingly, it is found that the window width and FFT length of the spectrum estimator has a significant impact on obtaining accurate results. 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. To achieve these sample frequencies requires some zero padding of the time windows. This frequency domain sampling is found adequate in training ANN on single speech excerpt in octave bands. Using a digital implementation of the envelope spectrum estimator enables higher resolution frequency sampling to be achieved, giving a more detailed representation of the envelope spectrum. This is found particularly useful in training ANNs on broadband unfiltered speech and multiple speech examples as discussed in Secs. IV D and IV E.

STI values are determined from MTFs in octave bands of speech interest. However, as speech signals have limited bandwidth, very little energy is found in frequency band above 6.3 kHz. Figure 7 shows the spectrum of a typical anechoic speech excerpt. Because of this signal to noise ration problem, the 6.3 kHz band is used instead of the problematic 8 kHz band.

D. Training and validation data sets

Artificial neural networks using supervised training need to learn from a large number of example-teacher pairs. A simple stochastic model for impulse response synthesis—multiplying white noise by an exponential decay function—was previous used by Schroeder to investigate MTF measurement methods[14] and was said to be realistic in the late part of reverberation. However, such simple stochastic model does not give frequency dependency and is inaccurate in its description of early reflections. STI by definition is a frequency dependent parameter. An improved stochastic model

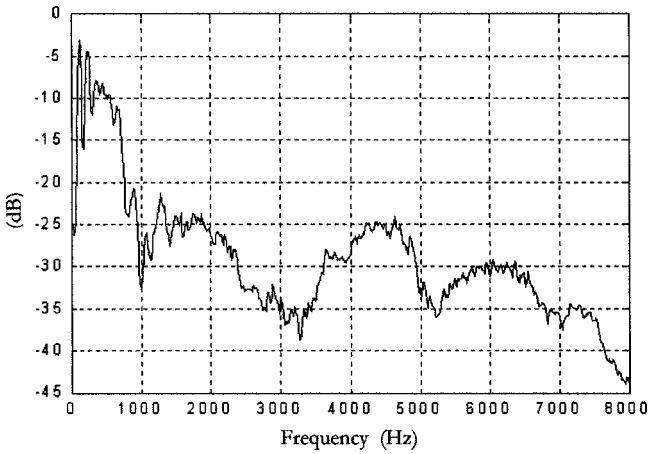

FIG. 7. An example of a speech spectrum.

was developed for this study which incorporates various possible early reflection patterns, frequency dependence and realistic reflection density. Figure 8 shows an example of an impulse response synthesized using the new stochastic model.[11]

The stochastic impulse response synthesiser is coded such that it randomly generates all impulse response patterns while statistically it is controlled by governing diffuse field physical laws. Although governed by the diffuse field laws, the stochastic nature of the generator meant that distinctly nonlinear decays could be generated. When it is run for sufficiently large number of times, it hypothetically generates a superset of impulse responses found in reality. A small number of real impulse responses were also used, not enough to properly validate the systems ability for actual measurements, but some reassurance that the simulations are realistic.

The success of ANN methods is evaluated using validation tests. The validation tests use data not seen in training to test for generalization as is standard practice in ANN research. The data sets are split into two halves, the first half is used to train the ANNs, while the second half is used to validate the trained ANNs. Rigorous validation is achieved by ensuring that cases in validation sets have never been used in training. In this study, six untrained narrators were used to read excerpts from three different text materials. In each different text material two different excerpts were taken. The text extracts were contrasting samples, ranging

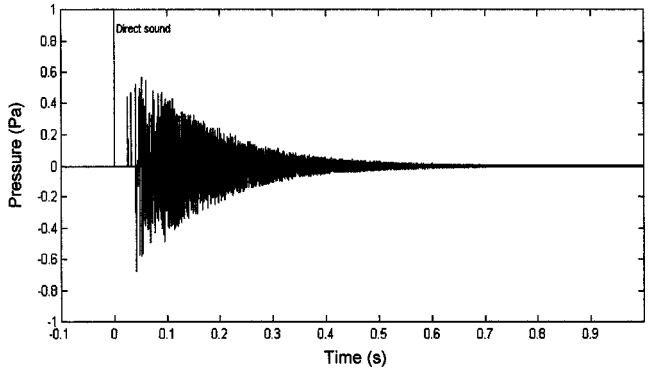

FIG. 8. An example of simulated impulse response.

from a descriptive passage from a classic 19th century novel to a conversation passage from a late 20th century popular novel. The teachers for training the ANNs are the true STI values of these speech examples. STI was calculated from impulse responses as follows. First, the MTFs are calculated using Eq. (1), Then the standard procedure for obtaining STI from the MTF is followed.[3,6,7]

(1) Calculation of mean apparent S/N ratio $\overline{(S/N)}_{\mathrm{app},F}$ in each octave band

$$\overline{(S/N)}_{\mathrm{app},F} = \frac{1}{14}\sum_{F=0.63}^{12.5} \max\left[\min\left[10\log\left(\frac{\mathrm{MTF}(F)}{1-\mathrm{MTF}(F)}\right),15\right], -15\right]. \tag{10}$$

(2) Calculation of overall mean apparent S/N by weighting the $\overline{(S/N)}_{\mathrm{app},F}$ of seven octave bands, and converting to an index ranging from 0 to 1:

$$\overline{(S/N)}_{\mathrm{app}} = \frac{15+\Sigma w k \overline{(S/N)}_{\mathrm{app},F}}{30}. \tag{11}$$

The values of w_k are given in Refs. 3 and 6. In this paper, examples cover reverberation times from 0 to 5 s, signal to noise ratios from 0 dB to noise free. Added noise is white.

IV. APPLICATION OF ANNS TO STI EXTRACTION

A. Training on impulse responses

The capability of ANNs to extract STI values from impulse responses is first explored. The experiment intends to identify (i) if the ANN can generalize impulse responses and (ii) if the ANN can perform the nonlinear calculation needed to obtain STI values. The 14 000 simulated impulse responses are used. These are octave band filtered and used as input signals. The MTF values are obtained via Eq. (1). The 14 MTF values in the 1/3 octave band used in the standard STI method are fed into an artificial neural network. This network has the following neurons in the input, hidden and output layers: 14–10–8–1. Typically, 50 000 iterations (each iteration presents a block of 280 examples) are required for satisfactory results, but this may vary with different initial weights and learning rates. When training is completed, the network is validated. To demonstrate how the network generalizes to different impulse responses, Fig. 9 shows the standard deviation found over all validation tests. It is found that relatively large standard deviations are associated with higher level of noises because the noise interference never repeats. Even so, very low errors can be obtained in this case. This is not surprising, since neural networks can map almost any complicated function. This also shows that the network can generalize from impulse responses seen in the training phase, to ones not seen before in the validation phase. Since the precise relation between impulse response and MTF is known, the use of ANN here does not surpass traditional calculation in terms of accuracy. On the other hand, the ANN

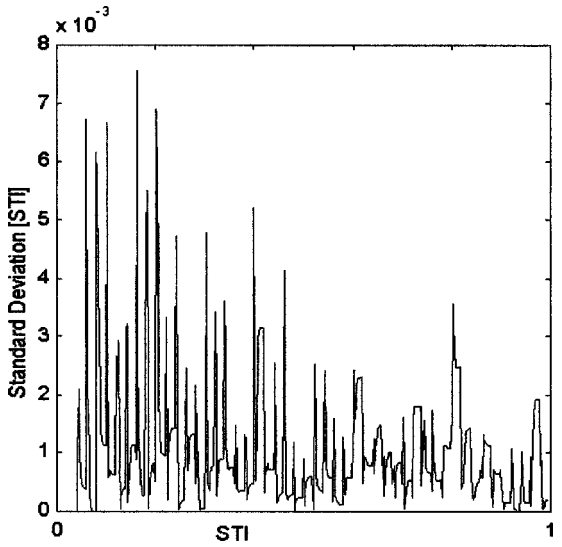

FIG. 9. Standard deviation found over validation tests.

method to map impulse responses to STIs would provide a useful alternative neural computing approach for implementation on very simple hardware.

B. Training on octave band STI

Octave band STI is defined as $(\overline{(S/N)}_{\mathrm{app},F}+15)/30$, where $\overline{(S/N)}_{\mathrm{app},F}$ is defined in Eq. (10). The training phase and retrieval phases are as illustrated in Fig. 1. Again, envelope spectra values at 14 designated modulation frequency points are extracted and fed into the input layer of the neural networks. It is empirically found that a 14−20−8−1 network performs well for all octave bands. It is found that very high accuracy is possible. No signs of over-training were found before the maximum error in validation tests dropped to below 0.01 STI. (The reason why a more strict termination criterion was used here is because the extracted octave-band STI values will eventually be used to calculate full STI.) This method is found robust to different speakers, text types and mode of reading, as long as they are individually trained on those particular speech excerpts, i.e., each speaker or text requires a different network. This demonstrates that the ANN can be used to memorize speech spectra and compensate the error in measurement when natural running speech is used as excitation.

C. Full STI

As the full band STI is a linear combination of STI values in octave bands according to Eq. (12), this can be implemented with a fixed network structure as shown in Fig. 10. Seven neural networks representing corresponding octave bands are individually training as described above, they then form a bank of trained neural networks. The outputs of these are processed by the additional linear layer. As the weights for octave bands are known, no further training for this weighting layer is required.

The maximum prediction error found is 0.018 STI as shown in Fig. 11. Unfortunately, accurate error values for Steeneken and Houtgast's running speech method[10] are not

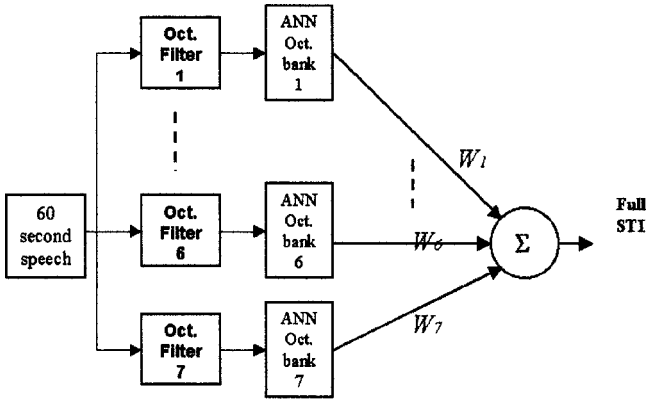

FIG. 10. How the ANNs trained to give octave band STI are combined to give full STI values.

available from their brief conference paper, but it is known that the correlation coefficient between their running speech method and the true STI was 0.971. The correlation coefficient between the ANN estimated and true STI is 0.9999, considerably higher.

D. Training on unfiltered speech

Artificial neural networks together with the preprocessor were trained on unfiltered speech to extract STI values. The motivation here is to form a more compact structure which would then be more efficient to implement in instrumentation. In this case, octave band filters are not inserted. Speech was fed into ANN via the spectra estimator directly. ANNs have been successfully used to enhance accuracy of spectrum analysis[22] and this philosophy is followed. Due to the squaring operation in forming the envelope, high frequency components of the received speech appear also at low frequencies, these additional components being generated by the cross terms in the squaring operation. This gives the ANN access to additional information to base its parameter estimation on.

It is therefore sought to train the ANN to learn from unfiltered speech and automatically associate contributions from different octave bands to give reasonably accurate STI estimations. Envelope spectrum values for the 14 1/3-octave frequency bands used in the standard STI method (the energy being summed over the 1/3-octave band), are used as input vector for the ANN. Full STI values were used as teachers. Gradual convergence was shown in training process, however, ensemble errors were not reduced to a satisfactorily low value even after training for a long time. The test result showed a maximum prediction error of 0.07.

It has been empirically identified that the traditional STI envelope spectrum analysis[1,9,12] is inadequate for extracting STI from unfiltered speech. To decrypt the intermodulated information, a higher resolution envelope spectrum estimator is needed. Consequently, the Hilbert transform detected envelope, low pass filtered at a cutoff frequency of 80 Hz, is used to allow more information to be used. High resolution power spectrum estimator gives envelope spectra at a 0.3125 Hz frequency step giving 40 linearly sampled envelope spectra from 0.3125 Hz to 12.18 Hz. The 40 data point are sub-

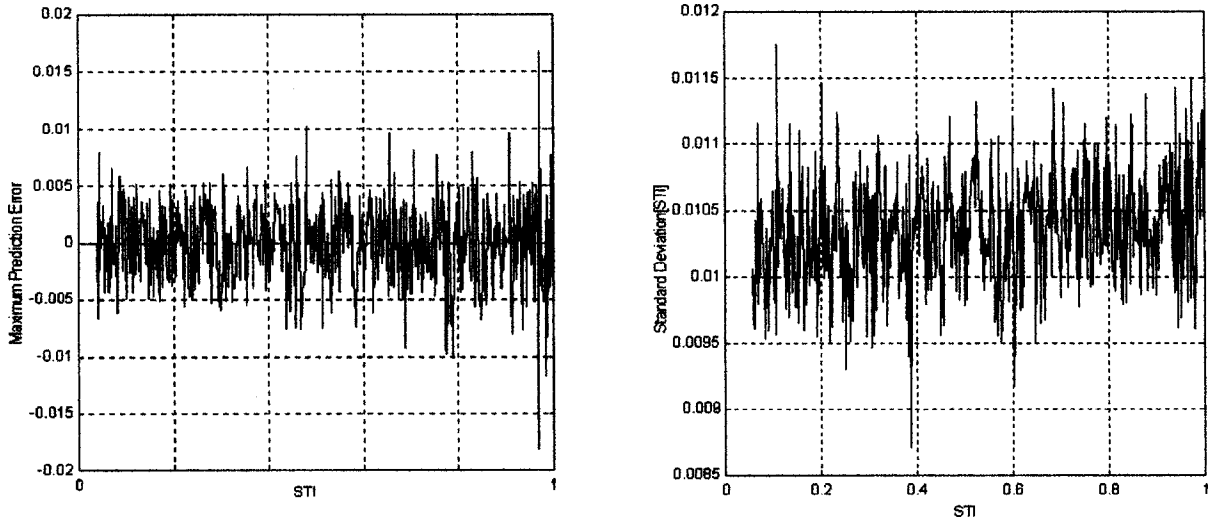

FIG. 11. Maximum prediction errors and standard deviations found in validation tests. STI obtained using the structure shown in Fig. 10.

sequently used to form the input vector for the ANN. A network size of 40–20–10–1 was found suitable. Validation results are shown in Fig. 12. The correlation coefficient between the ANN estimated and true STI is 0.999 97 and the maximum prediction error is 0.0197 STI.

E. Multiple speech excerpts: source independence

So far, a set of ANN models have developed to accurately extract STIs from received speech. The systems work with a specific speech excitation, as the statistical feature of that particular speech is learned and imbedded in the ANN. This means prerecorded speech signals have to be used. As one step further, the feasibility to learn from different speech excitations and generalize to arbitrary speech (source independent) extraction of STI is explored. This would enable the instrument to be much more flexible, with the potential to work with live speech.

The very low frequency envelope spectrum of anechoic running speech is known to be a generally stable.[13] However, speech is very complicated stochastic process and the envelope spectra are not sufficiently stable to be regarded as constant for STI extraction. Figure 13 shows an over-plot of 18 envelope spectra of anechoic speech signals read by six untrained native English narrators. A maximum difference of approximately 7 dB is found.

The problem with arbitrary speech is that the attack and decay of the anechoic speech mixes with the reverberance in the room. To take a simple example, a word pronounced with a long decay (e.g., ''bus'') in a dry room, can have the same envelope as a short decay word (e.g., ''stop'') in a reverberant space. While using a long speech extract can help average out random variations, consistent differences in pronunciation will affect the envelope spectrum. These differences in the envelope spectrum caused by pronunciation can be indistinguishable from the changes due to reverberance in the room. For this reason, contradictions in the data set are seen, and the neural network fails to properly converge because it is asked to map similar envelope spectra to different STI values.

To deal with these contradictions, additional information must be fed to the ANN. One possibility is to feed additional information from frequency ranges not previously used ($>$25 Hz). As the envelope spectra are normalized, a speech having lower level spectrum in certain frequencies must result in

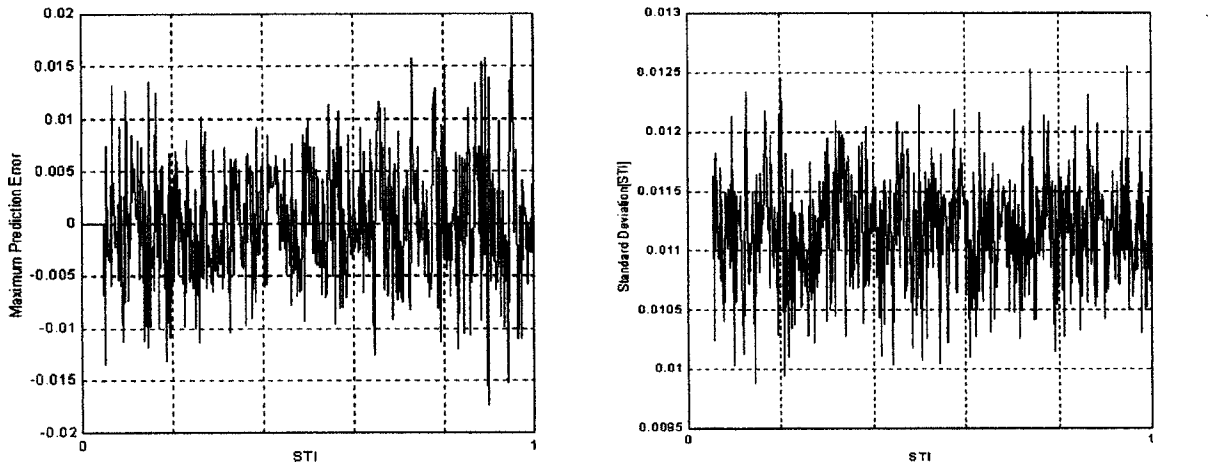

FIG. 12. Maximum prediction errors and standard deviations found in validation tests. STI obtained using one neural network from unfiltered speech.

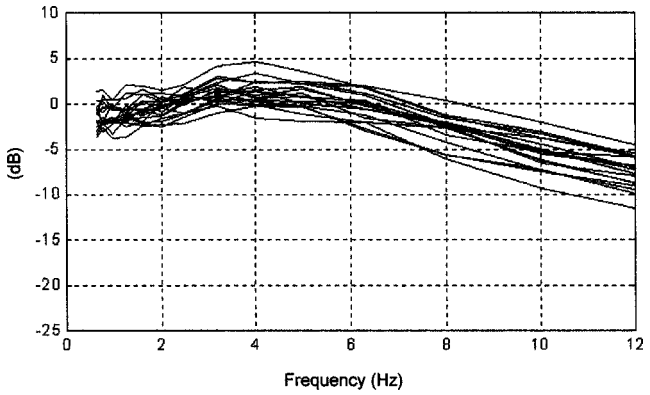

FIG. 13. Over plot of envelope spectra of 18 speech excerpts in 1 kHz band. Envelopes obtained using a digital implementation of a traditional envelope spectrum analyzer.

higher level spectrum at other frequencies as there is no leakage. In addition, it is assumed that in the vast quantity of data being filtered out by the preprocessor, there is information concerning how the speaker pronounces words, and this information is needed for the neural network to resolve the contradictions in the data set.

The envelope spectra are taken up to 80 Hz; these are estimated using Welch's average periodogram method.[23] Frequency contents are sampled at a 0.5 Hz step up to 80 Hz, providing 160 inputs for the ANN. The ANN has a 160–40–20–1 architecture. The 18 different anechoic speech examples and three different texts read by six narrators are used in the training.

Figure 14 illustrates the errors found with the validation tests. The maximum prediction error for STI found in is 0.13, and the correlation coefficient between actual and predicted STI is 0.9948. Better accuracy can be obtained by averaging over several different speech excerpts. When averaging the estimated STI over three different speech excerpts read by different narrators, prediction errors can normally be reduced to less than 0.1. Consequently, accurate enough extraction from arbitrary speech with only output speech knowledge has not been achieved.

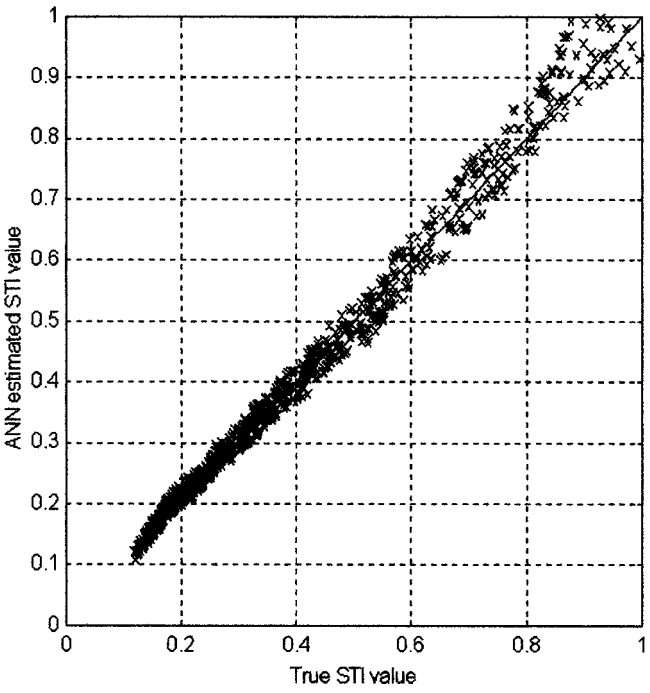

FIG. 14. Comparison of ANN predicted (worst cases) and actual STI over 18 different speech extracts.

V. DISCUSSIONS AND CONCLUSIONS

A neural network method to improve the accuracy and repeatability of STI measurements with running speech is proposed and validated via simulations. This method can significantly improve the accuracy of STI measured with natural running speech, hence facilitating measurement in occupied conditions. The proposed neural network method works with both received broadband and octave band speech signals, providing an accuracy comparable to measurements made using artificial test signals, typically a standard deviation of less than 0.02, when a one-net-one-speech excerpt case is considered.

Source independent extraction of STI from speech, was explored. It seems that the proposed ANN method has a certain capability to learn from examples and adapt to different speakers and texts. The actual STI and ANN estimation show reasonable agreement when testing with speech excerpts not previously seen by the ANN. Further investigations are needed to fully develop such a technique to gain sufficient accuracy for a practical measurement system.

Only a few real impulse responses have been used in training and validations, to fully validate this method and evaluate its use, more on-site validations will be needed. Nevertheless, this method proposed and validated here mainly with simulations, provides a promising avenue towards accurately measuring STI from natural occurring sound sources. The work has only considered the case of natural sound reproduction into rooms, but there is no reason why this cannot work with public address systems. To achieve this, suitable examples of sound reproduced by a public address system would have to be included in the training set.

ACKNOWLEDGMENTS

This paper was funded by the Engineering and Physical Sciences Research Council, UK, under Grant No. GR/L89280. The authors would also like to thank colleagues at Salford University who helped make the anechoic speech recordings.

[1] T. Houtgast and H. J. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," Acustica **25**, 355–367 (1971).
[2] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acustica **28**, 66–73 (1973).
[3] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech transmission quality," J. Acoust. Soc. Am. **67**, 318 (1980).
[4] T. Houtgast and H. Steeneken, "Predicting speech intelligibility in rooms from the modulation transfer function. part I. General room acoustics," Acustica **46**, 60–72 (1980).
[5] ISO TR 4870, "Technical report: Acoustics—The construction and calibration of speech intelligibility tests," 1991.
[6] "Technical Note, Device for measuring the speech transmission index," J. Acoust. Soc. Am. **71**, 1612 (1982).
[7] IEC 60268-16:1998 (also BS EN 60268-16 and BS 60268-16), "Sound

system equipment, Part 16: Objective rating of speech intelligibility by speech transmission index,'' 1998.
[8] P. W. Barnett and P. H. Scarbrough, ''From word scores to RASTI and back—An experimental study of the relationship between RASTI and word scores,'' Proc. Inst. Acoust. UK **22**, 73–82 (2000).
[9] H. J. M. Steeneken, J. Verhave, S. McManus, and K. Jacob, ''Development of an accurate, handheld, simple-to-use meter for the prediction of speech intelligibility,'' Proc. Inst. Acoust., UK **23**, 53–59 (2001).
[10] H. J. M. Steeneken and T. Houtgast, ''The temporal envelope spectrum and its significance in room acoustics,'' Proceedings of the 11th ICA, 7, Paris, 1983, pp. 85–88.
[11] T. J. Cox, F. Li, and P. Darlington, ''Extraction of room reverberation time from speech using artificial neural networks,'' J. Audio Eng. Soc. **49**, 219–230 (2001).
[12] F. F. Li and T. J. Cox, ''Extraction of speech transmission index from speech signals using artificial neural networks,'' Proceedings of the 110th AES convention, Amsterdam, paper 5354, 2001.
[13] T. Houtgast and H. J. M. Steeneken, ''Envelope spectrum and intelligibility of speech in enclosure,'' IEEE–AFCRL Speech Conference, 1972.
[14] M. R. Schroeder, ''Modulation transfer functions: definition and measurement,'' Acustica **49**, 179–182 (1981).
[15] S. Y. Kung, *Digital Neural Network* (Prentice-Hall, New York, 1993).
[16] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. (Prentice-Hall New York, 1999).
[17] G. Cybenco, ''Approximation by superpositions of a sigmoidal function,'' Math. Control, Signals Syst. **2**, 303–314 (1989).
[18] D. E. Rumelhart, G. Hinton, and J. R. Williams, ''Learning internal representations by error propagations,'' *Parallel Distributed Processing* (MIT Press, Cambridge, MA, 1986), Vol. 1, Chap. 8.
[19] M. Riedmiller, ''Advanced supervised learning in multi-layer perceptrons-from back propagation to adaptive algorithms,'' Int. J. Comput. Standards Interfaces **16**, 265–275 (1994), special issue on neural networks.
[20] R. A. Jacobs, ''Increased rate of convergence through learning rate adaptation,'' Neural Networks **1**, 295–307 (1988).
[21] H. Kuttruff, *Room Acoustics*, 4th ed. (Spon Press, 2000), p. 249.
[22] F. Luo and R. Unbehauen, *Applied Neural Networks for Signal Processing* (Cambridge University Press, Cambridge, MA, 1996), pp. 74–120.
[23] M. H. Hayes, *Statistical Digital Signal Processing and Modeling* (Wiley, New York, 1996), pp. 415–420.