TOWARDS THE UNDERSTANDING OF ACOUSTICS OF AUDITORIA IN TRUE IN-USE CONDITIONS

--A review and outlook of techniques for occupied room acoustic measurements

FF Li Acoustics Research Centre, The University of Salford, UK

1 INTRODUCTION

Since Sabine [1] instigated objective measurements of acoustics a century ago, the advancement of room acoustics has accumulated a wealthy body of knowledge about suitable or preferred acoustics for concert halls and auditoria for diverse purposes in terms of objective acoustic parameters [2]. Established measurement techniques, such as Maximum Length Sequence (MLS) and Exponential Sine Sweep (ESS) methods, resort to noisy test signals, and consequently are often restricted to unoccupied spaces - the measurements are invasive and the test stimuli are intolerable to ordinary audiences. Occupancy alters acoustics of a space considerably by affecting the physical propagation of sound, evidenced by a reduction in reverberance. It is known that estimates from unoccupied data are unreliable and problematic, because they do not accurately represent the true in-use conditions [3]. Occupied in-situ measurements would lead to a greater understanding of room acoustics.

The problem of noisy testing signals that hinder the occupied measurements can be mitigated if measurements are made with naturalistic sound sources, e.g. music or speech or more favorably, naturally occurring sources when the spaces are in-use, i.e. live music or speech. Motivated by this thought, naturalist source based measurements have attracted considerable attention not only in the concert hall acoustics research community but also in the hearing aids development, telecommunications and other fields of research [2,3,4,5,6,7,8,9]. Over the past decade a number of new methods were proposed and developed to determine individual acoustic parameters from received speech or music [10,11,12,13,14,15]. These methods potentially alleviate the problems of occupied measurements, but are limited to a few parameters derivable from decay curves. The impulse response of a room is of particular importance, because it gives complete description of the characteristics of point-to-point sound transmission in an enclosure. Virtually all room acoustic parameters can be derived from the impulse responses. Following a review of a series of methods using naturalist or naturally occurring sound sources to enable the measurements of common room acoustic parameters, this paper proposes the use of synthesised music and masked pseudorandom sequence to determine complete impulse responses in-situ.

2 THE MACHINE LEARNING METHODS

2.1 Extracting reverberation parameters from discrete speech utterances

Cremer and Müller suggested a method to measure the reverberation time (RT) of an occupied concert hall from the energy decay curve of a loud stop chord played by an orchestra [2]. This was used by some researchers to show the discrepancies between the estimated and in-situ measured RTs in occupied auditoria [3]. Indeed, decay curves can be estimated from naturalistic sources with abrupt energy drops followed by a period of silence. Speech is a rich source of such instantaneous energy differences and therefore is a useful natural sound source for RT. Utterances of anechoic speech show unequal and non-trivial decay times, which are further elongated by reverberation. Examples are illustrated in Fig 1. If Fig. 1 were plotted in logarithmic scales, one might think a straightforward line fitting or the average of a number of line fittings would be sufficient to determine the RT. Unfortunately, the received signals of speech utterances are much complicated and “noisier” then the idealised scenario: Speech signals are non-stationary stochastic processes, which do not have constant short-term energy like tone bursts. Impulse responses of many rooms are not exactly exponential, two or more decay rates are commonplace in coupled spaces or non-diffused fields. In addition, a reverberation tail and the next utterance can overlap. Estimation of the reverberation time through a straightforward examination of slopes of decay edges is simply too inaccurate to be of practical uses. An Artificial Neural network (ANN) method was therefore proposed and developed to accurately extract RTs and EDTs (Early Decay Time) from received signals of pronounced digits. The work was carried out around the last millennium [10].

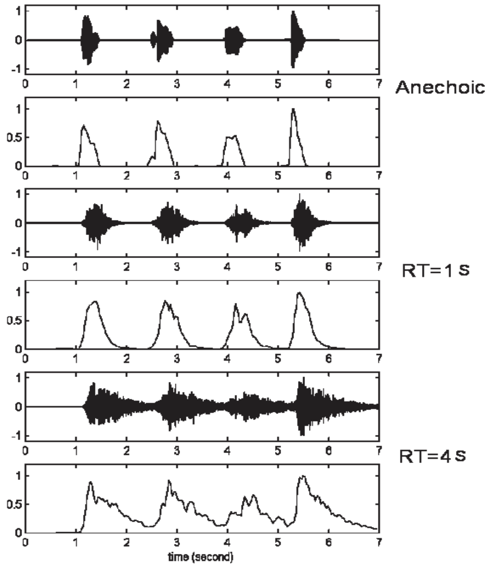

Fig. 1 Signatures of anechoic and reverberated speech utterances and their energy envelopes (normalised amplitude vs time)

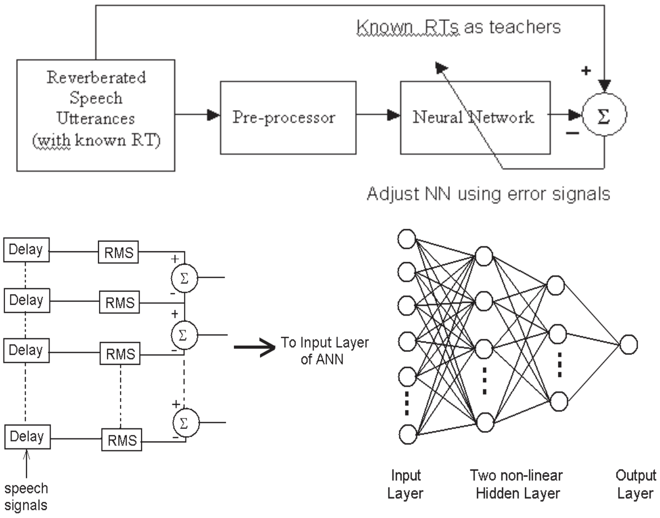

Fig. 2 Block diagram of ANN system in training phase (top), signal pre-processor (bottom left) and ANN architecture (bottom right)

The ANN is a biologically inspired computational paradigm loosely associated with artificial intelligence, but more precisely classified as a soft-computing algorithm. It features massively connected basic processing units called neurons: - simplified models of biological brain cells. The ANN has been a popular machine learning regime over the last decade, capable of statistically learning from a large but limited number of examples (learning phase) and mapping arbitrarily complicated relationships (retrieving phase). Fig. 2 (top) shows a block diagram of the neural network system in its training phase. Reverberated speech utterances with known reverberation times are used as examples for the training. The examples are pre-processed and conditioned (Fig. 2 bottom, left) to yield suitable input vectors for the neural network. The outputs of the ANN and the corresponding true reverberation times (teachers) are compared to obtain the error signal. The training process is to iteratively update the internal synaptic weights of the ANN so that the mean square error between the true and the estimated reverberation times over all the training examples are minimised. In the retrieve phase, teachers and the network adjustment paths as shown in Fig. 3 are removed. Speech utterances as received by a microphone in the space are sent to the trained neural network via the same pre-processor. The ANN then gives accurate estimation of reverberation time. Validation showed the prediction errors are below 0.1 s for RTs and 0.06 for EDTs within the range of RT/EDT=0.5 to 5 s. Since the training phase involves the presentation of the reverberated speech samples to the neural network model, in the application (retrieve) phase, there is not need for the computer to monitor the source. This offers a quasi single-channel method by playback the pre-recorded speech signals as stimuli. Generalisation to arbitrary utterances not being included in the training was attempted but the results were not promising. A neural network of this type simply would handle the complexity of arbitrary speech utterances. Thus the application of this method is limited to the use of pre-recorded anechoic speech and has to use an amplifier and loudspeaker.

2.2 Estimation of speech transmission index from running speech

Speech transmission index (STI) is a standardised method for the assessment of speech intelligibility in spaces [16]. It employs a low frequency (0.63 - 12.5Hz) signal modulated noise to simulate speech and then applies the modulation transfer function (MTF) to quantify the envelope shaping effect on speech due to reverberation. When suitably normalised the MTF also takes into account ambient noise interference. If speech signals could be accurately modeled as an envelope modulated white noise with equal power per frequency bin, the MTF might be obtained by a straightforward subtraction of envelope spectra of received and original speech, Equ 1.

$$MTF(F)(dB) \approx Ey(F)(dB) - Ex(F)(dB) \qquad (1)$$

where E_X *(F)* and E_Y*(F)* are the envelope spectra of input and output long-time speech signals of a channel in decibels. Unfortunately, for a real speech signal, this relation becomes a very coarse approximation [17, 18, 19], due to the non-stationarity of speech signals. The envelope spectra are obtained from squared and low-pass filtered speech signals normalised to total signal energy of the speech excerpt. The envelope spectra are typically obtained over 40-60 second speech excerpts to allow statistically meaningful results. Energy of envelope signals lies in a very low frequency band from immediately above DC to about 15 Hz. These frequencies are related to fluctuations of various aspects of acoustic phonetic elements in running speech as illustrated in Fig. 3 (left). Typical envelope spectra of running speech are show in Fig 3(right).

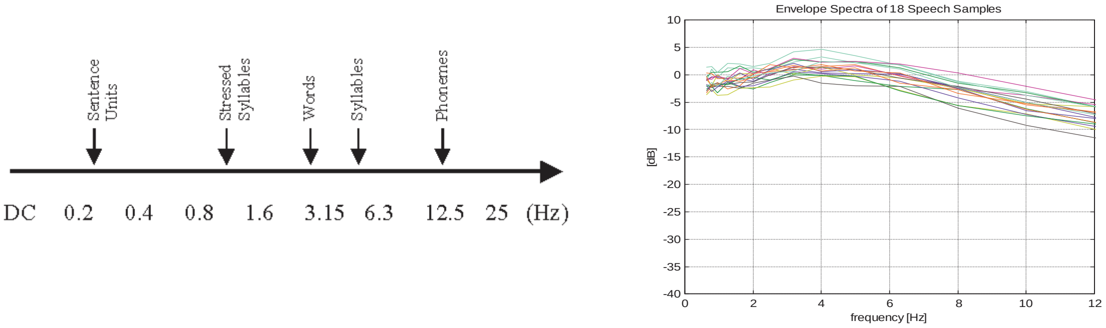

Fig. 3 Contributions to speech envelope spectra and representative envelope spectra of running anechoic speech

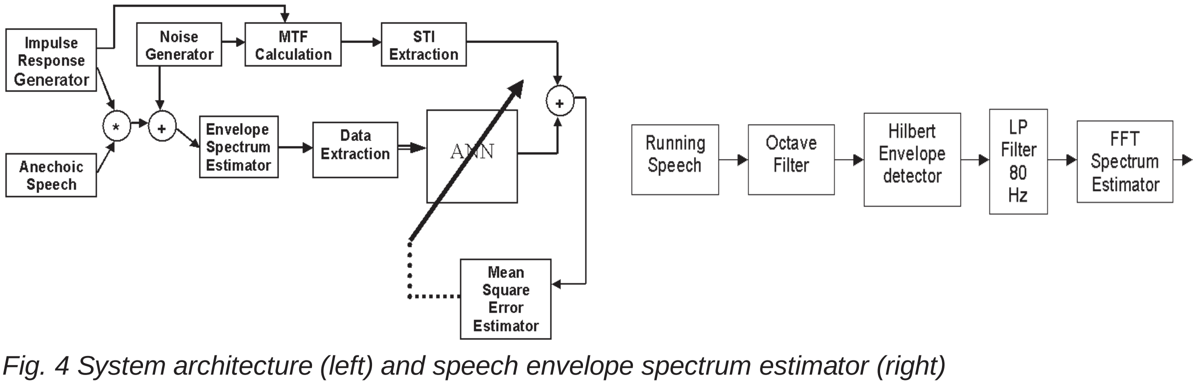

Fig. 4 System architecture (left) and speech envelope spectrum estimator (right)

The discrepancy between the envelope spectra of received and original speech signals contains information about the RT and STI, and therefore machine learning can be applied to extract these parameters from a large number of examples using envelope spectra as a feature space. Artificial neural networks were considered to model such relations to accurately extract STIs from speech signals [11]. The neural network model is shown in Fig. 4. It deploys the idea of envelope spectrum subtraction but uses a multi-layer feed forward ANN trained by the standard back propagation algorithm to memorise features of speech stimuli and generalise from a large set of impulse responses of different acoustic transmission channels. Validation test show that typically the estimation errors are below 0.02 STI.

The ANN model for STI extraction presented in this section is speech dependent: the information about the speech is built in the network during its training phase. Pre-recorded anechoic speech materials are needed to achieve the high accuracy. Envelope spectra over 40-60s are generally stable for speech signals from different speakers and speech materials as illustrated in Fig. 3 (right). As a result, the STI might be estimated from envelope spectra of received speech only. The ANNs have been trained to generalise to not only arbitrary impulse responses but arbitrary speech materials as well. The results are generally promising but at the cost of compromised estimation accuracy with errors up to 0.11 STI.

2.3 Estimating reverberation time from running speech

Running speech is apparently a more attractive source for the measurement. It is therefore useful to extend the above running speech method for STI to RT extraction. The modulation transfer function can be express as a reverberation term and a noise term. If exponential decay is assumed, this can be written as

$$MTF(F) = [1 + (2\pi FRT/13.8)^2]^{-1/2} \cdot \frac{1}{1+10^{(-S/N)/10}} \qquad (2)$$

where F is modulation frequency and RT is reverberation time. It is apparent that in noise free cases, the MTF and RT has a nonlinear one-one mapping relation. Therefore the ANN method for STI can be used to extract RTs or EDTs in at least noise free cases. Simulation and validation proved that this would work. If a 40 dB signal to noise ratio is maintained, RT can be accurately determined (error<0.1 s) from running speech using its envelope spectra and proposed machine-learning method. For noisy cases, the accuracy might be slightly compromised.

2.4 Blind estimation using eigenvalues as a feature space

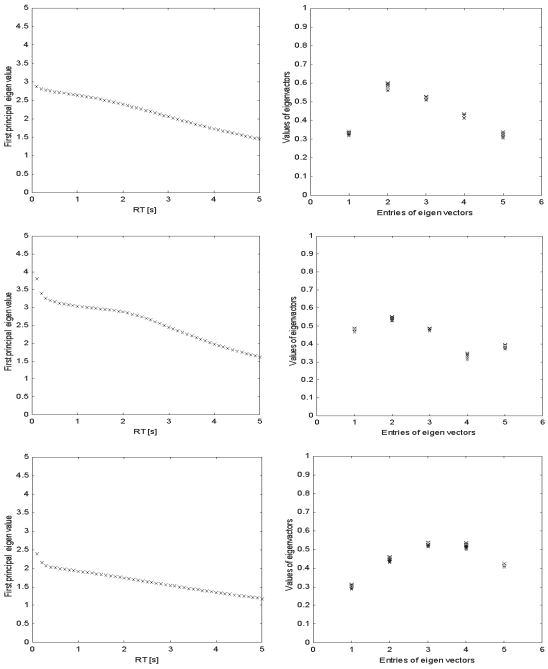

Fig. 5 Three running speech excerpts are convolved with 50 impulse responses with RTs from 0.1 to 5s. Eigenvalues and vectors of the first principal component of envelope signals are observed. The left plots show the eigenvalues against RTs. The right-hand side graphs are over-plots of the eigenvectors for 50 examples with different RTs.

This section looks into the use of an additional feature space for speech to achieve better accuracies in bind estimation, i.e. estimation from arbitrary speech sources not being used in the training phase. It is found that eigenvalues and eigenvectors of the received speech envelopes are correlated with reverberation and the features of anechoic speech respectively [12, 13]. For a given running speech excerpt, the principal component eigenvalue of its envelope monotonically decreases when RT increases, Fig.5 (left). Different speech materials show distinctive eigenvectors, not being affected by RTs, Fig.5 right. This means that the eigenvectors provide a useful feature space for the original speech. Since principle component eigenvector can be calculated using unsupervised machine learning known as Principal Component Analysis (PCA) neural networks, this suggests a feasibility of a hybrid ANN model shown in Fig. 6, in which an unsupervised model used to obtain feature space of original speech is added to the supervised models described in previous sections. Eigenvectors of first and second principal components obtained by the PCA sub-network from envelope signals and the envelope spectra data are both fed into a supervised neural network. The maximum prediction error found in the model was 0.087 for STI extraction with arbitrary speech.

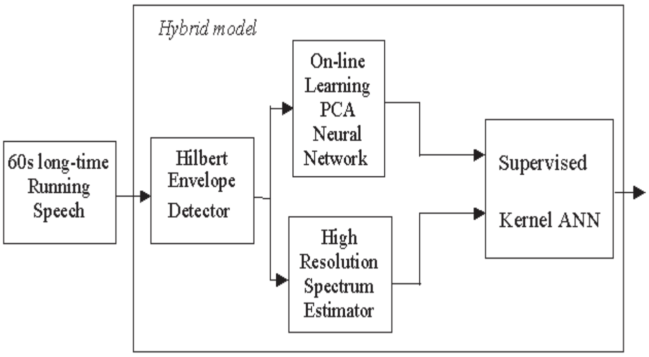
Hybrid model
60s long-time Running Speech
Hilbert Envelope Detector
On-line Learning PCA Neural Network
High Resolution Spectrum Estimator
Supervised Kernel ANN

Fig.6 Hybrid model for blind estimation of STI

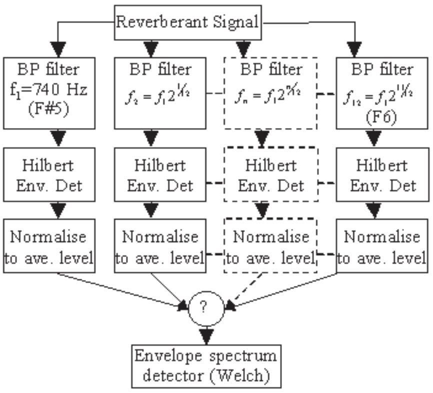

Fig. 7. Envelope spectra for music signals

2.5 Using music as stimuli

Speech stimuli have limited frequency contents. They can only be used to determine acoustic parameters in mid-frequency sub bands from 250Hz to 4 kHz. Music, especially orchestral music is considered as stimuli to obtain room acoustic parameters in all frequency bands of music interest. The follows the envelope spectrum method outlined above, but the accuracy of estimation is not as good as that from speech signals. The spectrum of a speech signal is pretty “full” from 250 to about 4500Hz, with no significant discontinuity. Traditional orchestral music, however, follows equal temperament scales. Signal power is centred around discrete and narrow sub bands, each related to a note from the scale. The result is a lack of excitations between notes and uneven spectra biased to particular notes in a piece (major/minor etc). Therefore, the signal to noise ratios are poor between the notes. A note matching filter bank is developed to address this issue [20]. For each octave band, the signal is further separated into 12 narrow frequency bands spaced according to the equal temperament scale. Envelope signals for each note are calculated and normalised to the average intensity of that note. Fig. 7 gives an example of 12 sub bands within 1kHz octave band. The reverberant signal is passed though the filter bank for 12 notes where the filters’ centre frequencies are determined by the equal temperament scale, starting at f#5 (≈740 Hz) in the 1 kHz octave band. Envelope spectrum of the octave band is estimated from the combination of all envelope signals calculated note by note. Thus machine leaning on envelope spectra in room acoustic parameter estimation can be extended to music stimuli. Following this method, over 95% of the reverberation estimates show errors of less than +/-5%, which is generally below perception limen.

3 MAXIMUM LIKELIHOOD ESTIMATION OF DECAY CURVE

Motivated by the needs of intelligent hearing aids and other speech communications devices, the reverberation time was estimated by performing a maximum likelihood estimation (MLE) on decays following speech utterances with an exponential decay model [6]. However, non-exponential or weakly non-exponential decays are fairly common, the single exponential model limits the accuracy of estimation. A more sophisticated decay model was adopted to improve the accuracy in blind room acoustic parameter extraction [15]. The MLE is a parametric estimation method in statistics. In essence, if there exists a parametric model for a statistical process, in the form of a probability density function *f*, then the probability that a particular set of parameters θ are the parameters that generated a set of observed data $x_1, x_2, \dots x_n$, is known as the likelihood *L*. This is denoted by

$$L(\theta) = f(x_1, x_2 \dots x_n \mid \theta) \tag{3}$$

An analytic model of an underlying process is first determined and a likelihood function formulated. The parameters that result in a maximum in the likelihood function are, by definition, the most likely parameters that generated the observed set of data. In a decay curve estimation context, let a room impulse response *h[n]* be modeled as a random Gaussian sequence *r[n]* modulated by a decaying envelope, *e[n].*

$$h[n] = e[n]r[n] \tag{4}$$

where *n* is the sample number. The envelope is represented by a sum of exponentials:

$$e[n] = \sum_{k=1}^{M} \alpha_k a_k{}^n \tag{5}$$

where a_k represent decay rates, α_k are weighting factors and M is the number of decays. If two decay rates are chosen, it can be weighted by a single factor.

$$e[n] = \alpha a_1{}^n + (1-\alpha)a_2{}^n \tag{6}$$

where a_1 and a_2 represent the two decay rates and *α* is a weighting factor that changes the level of contribution from each individual decay. This enables the representation of an energy response with a non-uniform decay rate and by changing the value of *α,* the model can adapt to best fit the decay phases. The likelihood of a sequence of independent, identically distributed, Gaussian variables occurring is given by [21]:

$$L(r,\sigma,\mu) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{(r[n]-\mu)^2}{2\sigma^2}\right)} \tag{7}$$

where *µ* is the mean and σ^2 the variance of the Gaussian process. The room impulse response model has no DC component, so *µ=0*. For the decay phases found in reverberated sounds *s,* the envelope is of interest. Thus the probability of the sequence, which has a zero mean and is modulated by an envelope *e,* is given by;

$$L(s;\sigma,e) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi}e[n]\sigma} e^{\left(-\frac{s[n]^2}{2e[n]^2\sigma^2}\right)} \tag{8}$$

This can be rearranged to give:

$$L(s,\sigma,e) = e^{\left(\sum_{n=0}^{N-1}\frac{-s[n]^2}{2e[n]^2\sigma^2}\right)} \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \prod_{n=0}^{N-1}\frac{1}{e[n]} \tag{9}$$

The proposed decay model Equ. 6 is substituted into Equ. 9. It is more convenient to work with a logarithmic likelihood function, since the multiplication becomes summation. The final and logarithmic likelihood function is

$$\begin{aligned}\ln\{L(s,\sigma,a_1,a_2,\alpha)\} = &-\sum_{n=0}^{N-1}\frac{\left[\alpha a_1{}^n + (1-\alpha)a_2{}^n\right]^{-2} s[n]^2}{2\sigma^2} \\ &- N/2 \ln\left(2\pi\sigma^2\right) - \sum_{n=0}^{N-1}\ln\left[\alpha a_1{}^n + (1-\alpha)a_2{}^n\right]\end{aligned} \tag{10}$$

Maximizing the log likelihood function with respect to the decay parameters *α*, a_1 and a_2 yields the most likely values for these parameters.

The MLE method for RTs was tested using a good number of real room cases and simulated impulse responses. For RTs in the range of 1-3 seconds, the estimation errors are typically below 0.1 second for from both music and speech. The method has been used successfully in a number of in-situ measurements recently.

4 SYNTHESISED MUSIC AS PROBE STIMULI

The impulse response of a room is of particular importance, because it completely characterises the sound transmission properties from the source to a receiver. Today, room acoustic measurements typically start from impulse responses and then derive individual parameters. Methods discussed the previous sections estimate individual acoustic parameters from either decay curves or modulation transfer functions, and the accuracy and valid frequency bands are source dependent. Dual channel FFT with music or speech source and the prolonged averaging of low level (in the background) maximum length sequence can theoretically measure impulse responses in-situ. The random nature and uneven energy distribution across the frequency bands of interest means that the dual channel FFT with naturalistic sources needs prolonged averaging. The background MLS method requires prolonged averaging too to gain a sufficient signal to noise ratio, say, 35 dB for RT30 calculations. Rooms are generally linear, but airflow and temperature fluctuations make it time-variant over a long period of time as needed for such prolonged measurements. Averaging over a long period of measurements leads to erroneous results [22]! This section will report on the development of a new method to enable occupied measurements of complete impulse responses with synthesised music: while the audience is enjoying the music in a natural setting, measurement can be taken in a non-invasive fashion.

A hybrid method is proposed where in the lower frequency range (up to 4 kHz) synthesised musical notes formed by narrow band chirps (sine sweeps) are used as stimuli. So the test signals are musical and in the mean time the use of sweeps ensures continuity in all frequencies. In the higher frequency region, although it is possible to use the overtones of synthesised notes, harmonics intrinsically contain less energy and can hardly give a satisfactory signal to noise ratio without a large number of averaging. An MLS signal high pass filtered and masked by music is applied to enable fast measurement in the higher frequency bands. Simulation and some initial validation tests suggest the method gives good overall accuracy.

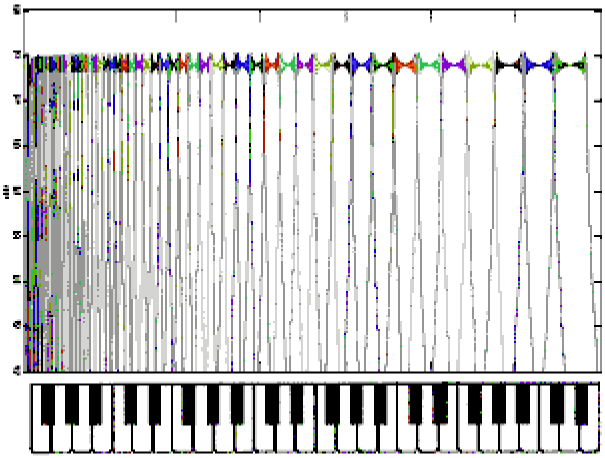
Fig. 8 Illustration of Chirp-Notes and filters

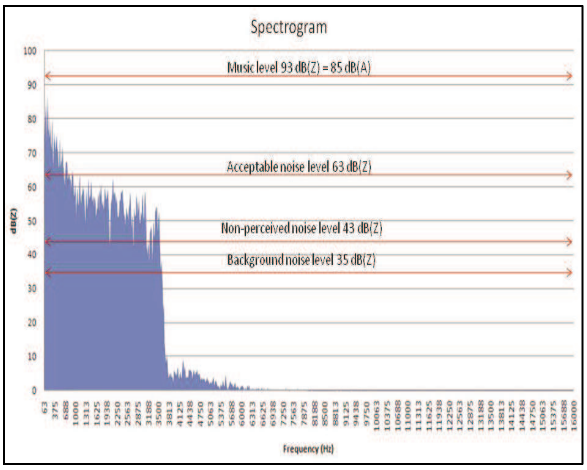

Fig. 9 Masked MLS sequence in operation

Fig. 8 illustrates the use of the “chirp-notes”: each note is represented by a narrow chirp sweeping up or down centering around the pitch of a musical note. Fade in and out filters (windows) are needed to avoid the known “ringing effect”. Due to the large number and regular repeating of such “chirplets”, the lumped effect of the side lobes appears as noise. The design of effective windows to mitigate the side lobe or ringing effect when applying these chirplets is the key technique for this method to succeed.

Fig. 9 shows an example of the MLS signal being masked by orchestral music excerpts. A purposely chosen orchestral music piece (loud passage) is low-pass filtered and compressed to function as the masking signal. A high-pass filtered MLS signal (masked) is applied to take the measurement in the background. Normally, an MLS signal 20-30dB above the background noise in higher frequency region in auditoria can be masked by music signals using this technique. With a few averages, the required signal to noise ratio can easily be obtained. Some listening tests were carried out. Subjects reported that the chirp-notes sounded like woodwind instruments and the masked MLS signals were hardly noticeable. By combining these two methods impulse responses can be obtained with musical stimuli. Simulation and pilot site validation (with random chirp-tones) has shown its potential. More work is needed to commission music pieces with the chip-tones.

5 CONCLUDING REMARKS

The past decade has seen some efforts of developing DSP techniques to enable occupied acoustic parameter measurements. Most recently with synthesised musical notes (in a very near future music pieces will be composed using these notes) and masked MLS, impulse responses can be measured while audience is listening to music. It is hoped that in the near future some more site measurements can be carried out with this technique to enable the collection of occupied and unoccupied data for comparisons. With a large dataset, it is likely to establish a regression model for more precise prediction of occupied acoustic parameters from unoccupied ones.

6 REFERENCES

1. W.C. Sabine, Collected papers on Acoustics, prepared by T.J. Lyman, Dover, New York, (1964).
2. L. Cremer and H. A. Muller, Principles and Applications of Room acoustics, Applied Science, (1982).
3. T. Hidaka, N. Nishihara, and L. L. Beranek, Relation of acoustical parameters with & without audiences in concert halls & a simple method for simulating the occupied state. J.Acoust.Soc.Am. 109, pp. 1028–1042, (2001).
4. Steeneken HJM & Houtgast T, The temporal envelope spectrum & its significance in room acoustics, Proc. 11th ICA, 7, pp. 85-88, 1983.
5. J. D. Polack, H. Alrutz, M. R. Schroeder, The Modulation Transfer-Function of Music Signals And Its Applications To Reverberation Measurement. Acustica. Vol. 54, pp. 257-265, (1984).
6. R. Ratnam, D. L. Jones, B.C. Wheeler, W.D. O'brien Jr., C.R. Lansing and A.S. Feng, Blind estimation of reverberation time" J. Acosut. Soc. Am. 111(5), pp. 2877-2892, (2003).
7. D. Griesinger, Beyond MLS –occupied Hall measurement with FFT techniques, 101th AES convention, paper No. 4403, (Nov., 1996).
8. R. Ratnam, D.L. Jones and W.D. O'brien Jr, Fast algorithm for blind estimation of reverberation time, IEEE Signal processing Letter, Vol 11 No. 6, pp. 537-541, (June, 2004).
9. Y. Zhang, et.al. Blind estimation of reverberation time in occupied rooms, in Proc. EUSIPCO 2006 (2006).
10. T.J. Cox, F.F. Li, and P. Darlington, "Extraction of room reverberation time from speech using artificial neural networks," *J. AES*, Vol. 49, No. 4, pp. 219-230, (2001).
11. F. F. Li and T. J. Cox, "Speech transmission index from running speech: A neural network approach," J. Acoust. Soc. Am., Vol. 113, Issue 4, pp.1999-2008, (2003).
12. F. F. Li and T. J. Cox, "A hybrid neural network for blind identification of speech transmission channels", in P. Liatsis Edits Recent Trends in Multimedia Information Processing, World Scientific Publishing, ISBN 981-238-242-7, (2002).
13. F. F. Li and T. J. Cox, "A Neural Network Model for Speech Intelligibility Quantification", J. Applied Soft Computing, Vol. 7, Issue 1,pp. 145-155, (January, 2007).
14. P. Kendrick, T. J. Cox, F. F. Li, Y. Zhang, J. A. Chambers, Monaural Room Acoustic Parameters from Music and Speech, J. Acoust Soc. Am. Vol. 124(1). Pp. 278-87 (July, 2008).
15. P. Kendrick, F. F. Li, T. J. Cox, Y. Zhang, J. A. Chambers, Blind Estimation of Reverberation Parameters for Non-Diffuse Rooms, Acta Acustica united with Acustica, Vol. 93, No. 5, pp. 760-770(11), (September/October 2007).
16. IEC 60268-16:1998, Sound system equipment, Part 16: Objective rating of speech intelligibility by speech transmission index, (1998).
17. M. Schroeder, Modulation transfer Functions: Definition and Measurement, Acustica, Vol. 49, pp. 179-182, (1981).
18. H. J. M. Steeneken and T. Houtgast, The temporal envelope spectrum of speech and its significance in room acoustics, 11th ICA conference publication, Paris, (1983).
19. T. Houtgast and H. J. M. Steeneken, "Envelope spectrum and intelligibility of speech in enclosures," IEEE-AFCRL 1972 Speech Conference Proceedings, pp. 392-395, (1972).
20. P. Kendrick, T. J. Cox, Y. Zhang, J. A. Chambers and F. F. Li, Room Acoustic Parameter Extraction from Music Signals Proc. ICASSP 2006 Volume 5, Page(s):V – 5, , (May, 2006)
21. E. Weisstein, Wolfram Mathworld: Eric Weisstein, (2006).
22. M.Serafini, F. F. Li and T. J. Cox, "Occupied Measurement of Room Impulse Response-Feasibility and Limitations" in proc. Inter-noise 2010, (June 2010).