ROOM ACOUSTIC PARAMETER EXTRACTION FROM MUSIC SIGNALS

Paul Kendrick[†], Trevor J. Cox[†] Yonggang Zhang[‡], Jonathon A. Chambers[‡], Francis F. Li,*

[†]Acoustic Research Centre
University of Salford, Salford M5 4WT, UK.
[‡]The Centre of Digital Signal Processing, Cardiff School of Engineering
Cardiff University, Cardiff CF24 0YF, UK.
*Department of Computing and Mathematics
Manchester Metropolitan University, Manchester M1 5GD, UK.

ABSTRACT

A new method, employing machine learning techniques and a modified low frequency envelope spectrum estimator, for estimating important room acoustic parameters including Reverberation Time (RT) and Early Decay Time (EDT) from received music signals has been developed. It overcomes drawbacks found in applying music signals directly to the envelope spectrum detector developed for the estimation of RT from speech signals. The octave band music signal is first separated into sub bands corresponding to notes on the equal temperament scale and the level of each note normalised before applying an envelope spectrum detector. A typical artificial neural network is then trained to map these envelope spectra onto RT or EDT. Significant improvements in estimation accuracy were found and further investigations confirmed that the non-stationary nature of music envelopes is a major technical challenge hindering accurate parameter extraction from music and the proposed method to some extent circumvents the difficulty.

1. INTRODUCTION

The performance of an acoustic space is often quantified using a set of acoustic parameters, which are known to be reasonably correlated to important aspects of human perception. For example, the Reverberation Time (RT) is the time taken for a sound to decay from a steady state condition by 60dB. Early Decay Time (EDT), on the other hand, is six times the time it takes for a sound to decay by 10dB and is often used as it is better correlated with subject reverberance [10]. Standard measurement of these parameters involves excitation of the space with artificial signals; the nature and level of these signals makes occupied measurement problematic and this prompts the use of naturalistic signals such as speech or music to facilitate occupied measurement.

The envelope spectrum of running speech [12] can be used as an indicator of the level of reverberation and background noise in an enclosure. A classic measure of the effect of the room on the envelope is the Modulation Transfer Function (MTF) [13], defined by the level of sinusoidal modulation transferred to a receiver from a single sinusoid modulating a noise source.

Reverberation smoothes the envelope of signals and this smoothing effect is similar to a low-pass filtering operation dependent on the decay time of the room. Machine intelligence function mapping methods can utilise the low frequency speech envelope spectrum to recognises the key features of the MTF and from there estimate an accurate RT value [1]. The bandwidth of speech limits RT estimation to mid-range frequency bands and the measurement is limited to spaces used for speech such as lecture theatres. This has inspired the use of music signals which often have a broader range of excitation frequencies and are used in other spaces such as concert halls. This paper discusses the application of artificial intelligence methods to the extraction room acoustic parameters from music signals.

A music envelope is highly non-stationary due to intensity and tonal fluctuations. The Complex Modulation Transfer Function (CMTF) is defined by the impulse response of the room [3] and can be estimated from music signals using the anechoic and reverberant envelopes with a local stationary hypothesis that removes low frequency fluctuations [2].

Pilot studies with music signals show that sufficient parameter accuracy cannot be achieved using the octave band envelope spectrum. Other pilot studies have shown that utilising the empirical CMTF [2] as pre-processor for music yields only a small improvement in EDT and RT estimation accuracy over the octave band envelope spectrum detector. It is proposed that this low estimation accuracy from music signals is related to the uneven response across the octave. Western music is based on the equal temperament scale. Therefore, the music signal power is focused in discrete narrow frequency bands, each related to a note from the equal temperament scale. The result is a lack of signal excitation between notes and an uneven response caused by bias to particular notes in a piece (major/minor etc). This response causes uneven excitation of the room response. Variation in decay shape for each of the corresponding note bands in the impulse means the effect on envelope doesn't accurately represent the whole octave band MTF which is used in parameter calculation.

It is proposed that to compensate for the unevenness of the octave spectrum, assuming the level of excitation in the impulse is constant across the band being analysed, the following method is employed: The signal is first separated into 12 narrow frequency bands spaced according to the equal temperament scale. Each note envelope is extracted and normalised to the average intensity of that note. By adding together each note envelope before detecting the spectrum, the effect of the MTF is now independent of individual note level.

A large database of realistic room responses is required to teach the system. Previously stochastically generated impulse

responses have been used but in recent decades, there have been significant advances in the modeling of rooms using geometric modeling techniques. A commercial package with a proven track record that utilises Randomized Tail-Corrected Cone-Tracing (RTCC) is used to generate the required number of training and validation examples [11]. This paper starts by introducing the pre-processor and training regime in section 2. Section 3 details experimental results, further investigation and analysis. Section 4 summarises the findings and further directions.

2. THE PROPOSED METHOD

The function mapping algorithm, in this case an artificial neural network (ANN - section 2.3), is trained to recognise acoustic parameters from reverberated signals from various simulated environments in Fig 1.

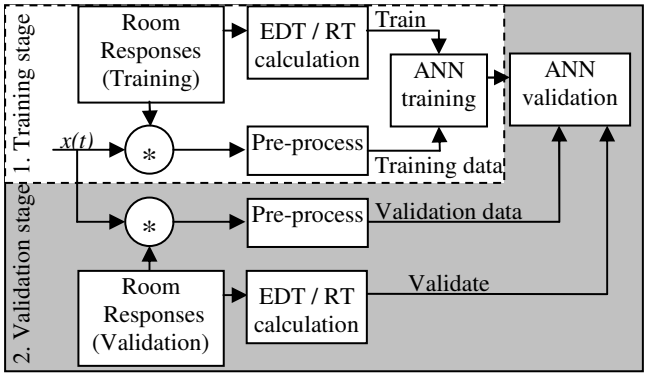

Fig. 1: System Overview

The anechoic signal was convolved with each impulse and then passed through the envelope spectrum pre-processor (section 2.2). The EDT and RT parameters were calculated from the impulses (section 2.1) and collected together with the associated envelope spectra. The data was split into two groups, a training set and a validation set. The training data was used to teach the ANN to map acoustic parameter to associated envelope spectrum. The trained ANN was validated with envelope spectra from the validation set to determine its accuracy.

Two geometrically random room models were used; a box shaped room and a fan shaped design. Each model had a variable source position on stage and an audience area with variable population density. The receiver grid was spread over the audience area. The model generation routine was given limits for room dimensions, aspect ratios and material properties; these have been used to generate 3,208 impulse responses so far, 50 % from each room model.

2.1. EDT and RT

This paper considers the 1 kHz octave band EDT and RT. This is calculated by first filtering the impulse and then applying backwards integration [5]. RT is calculated by fitting a straight line with a least mean square approximation to the decay from -5dB to -35dB and extrapolating the decay time to -60 dB. EDT is calculated in the same manner but only the first 10dB of decay is used. Fig 2 shows the difference caused by non-diffuse early reflections in the EDT and RT measures; notice the difference in the decay shape for the two limits.

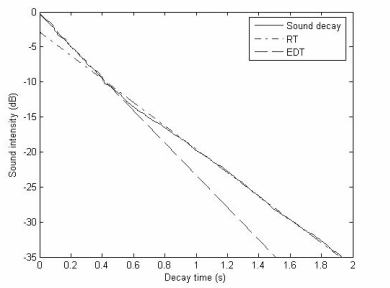

Fig. 2: EDT and RT calculation; the new room model

2.2. Modified Pre-processor

The pre-processor is a modified version of the speech envelope spectrum detector [1] [12]. The reverberant signal is passed though a 12 band BP filter bank where the filter centre frequencies are determined by the equal temperament scale, starting at f#5 (≈740 Hz) in the 1 kHz octave band.

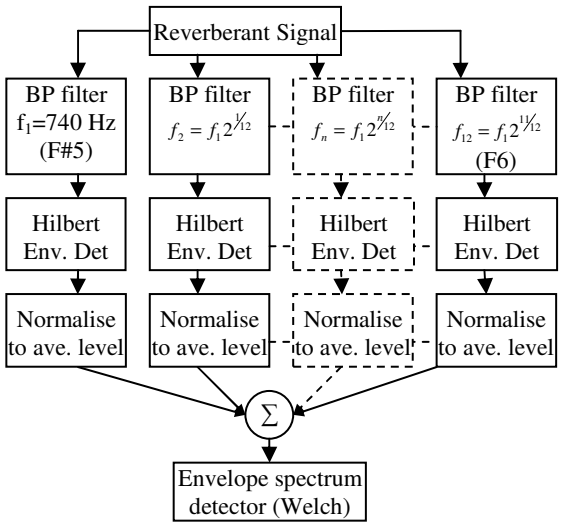

Fig. 3: Modified Pre-processor

All of the note filters are $^{1}/_{12}{}^{th}$ octave in bandwidth except for the first and last which are limited so that they are within the 1 kHz octave band thus preventing leakage from adjacent bands. A Hilbert envelope detector is used with a LP filter at 80 Hz and the envelope resampled at 160 Hz. Welch's power spectral density estimation with 50% overlap and 10s Hanning data windows were used to extract the envelope spectra. The music signals used were all over 2 minutes long. Each note envelope is normalised to its average level so that the envelope spectrum gives a value of 0 dB when a sine wave with amplitude equal to the average level of one note envelope is applied to the spectrum detector. Normalisation has three important consequences:

1. Envelope spectrum is independent of signal level
2. Effects of noise and reverberation are now separate.
3. Effect of MTF is independent of note levels

2.3. The ANN

Although any function mapping routine could be used, the ANN has a proven track record with regard to this work. Therefore, an ANN with a 40-30-10-1 structure was used using a bipolar sigmoid activation function of the form.

$$y = \tanh(xw + b) \quad (1)$$

w - neuron weight, b - bias value, y - output and x - input

The scaled conjugate gradient learning method was used which offers an order of magnitude increase in learning time over back propagation [6]. Student's t-test is performed on with training and validation error sets to detect signs of over-fitting. 3208 rooms were generated, ¾ of these were randomly selected as the training set and the rest as the validation set. The network size was determined in an ad hoc manner.

The reverberant parameter will ultimately be used to gauge predicted subjective performance, therefore the required accuracy of the system is defined by the smallest perceivable change. The difference limen (DL) is a term used to describe the smallest perceivable difference for something. The DL for RT on music signals is around 5% above 0.6s but increases below 0.6s to about 12% [6][7]. There is limited information on the EDT DL so the validation criteria for the required accuracy is set at ±5% but with a minimum error of ±0.1s as having accuracy better than this is not required. The accuracy in this paper is reported as being the percentage that lies within these limits.

2.4. Data separability and the Mahalanobis distance

A repeatable measure of data separability is defined based on the Mahalanobis distance [9] between two groups of data. D_μ is defined as the average Mahalanobis distance between a series of adjacent (in decay time) groups of envelope spectra. A group size of 0.5s is used so that the group size is large enough for a reliable covariance matrix estimate.

$$D_\mu = \frac{1}{N}\sum_{n=1}^{N-1}\sqrt{(\vec{\mu}_n - \vec{\mu}_{n+1})' C_{n,n+1}^{-1}(\vec{\mu}_{n1} - \vec{\mu}_{n+1})} \quad (2)$$

Where μ_1 and μ_2 are the mean vectors of the 2 groups and $C_{1,2}^{-1}$ is the inverse covariance matrix; it is assumed that as group spacing is decreased, D_μ will decrease in a uniform manner. D_μ shows excellent correlation with the system validity (% valid) and is a non-computational intensive repeatable measure of overall data separability used in this paper to quantify data separability without relying on inconsistent ANN results.

3. RESULTS TESTING AND ANALYSIS

3.1. Speech and the updated room model

One minute thirty seconds of running anechoic speech was applied to the system described in section 1, except, the full octave band envelope pre-processor was used [1] instead of the modified pre-processor so that the effect of the more complex room model could be documented. After training the function mapping system on the envelope spectra, the accuracy of extraction for both the RT and EDT is compared. RTCC represents the new room model, model 1 describes exponentially damped Gaussian noise impulse model and model 2 is a stochastic impulse model with time dependent reflection density. The % valid in the validation set after training for each case is shown in Fig 4.

Acoustic Parameter	Percentage valid		
	RTCC	***model 1***	***model 2***
EDT	94 %	99 %	95 %
RT	53 %	95 %	99 %

Fig. 4: Performance of speech

Satisfactory results (>95%) were achieved for models 1 and 2 for both EDT and RT. However, using the RTCC room model, it can be seen that EDT extraction is significantly more accurate. This is due to non-diffuse early reflections; in the same room differing EDT and RT times are common which causes increased problem complexity. Additionally, the envelope responds to the MTF unevenly; modulation frequencies where late reflections have more influence than early ones (RT is a low modulation frequency effect) may not be present or be particularly low in magnitude. This is analogous in the time domain to the masking of late reflections by utterances with early reflections.

3.2. Spectral unevenness of music – effect on the octave band envelope method and results from the modified pre-processor

A number of music signals were generated, each with the same envelope but with differing frequency responses. To do this, music was generated from the same rhythm sequence where all 12 notes in an octave were sounded but had a different combination of note levels for each test signal. The normalised variance of the power spectrum across the octave band was compared with the data D_μ separability of the music envelope spectrum when convolved with the room response database.

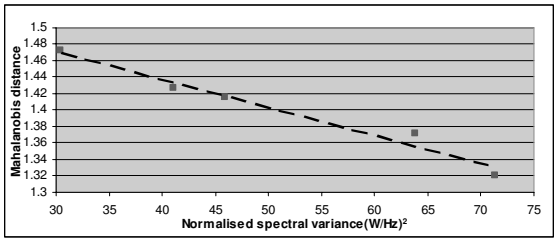

Fig. 5: Increasing spectral variance and data separability

Fig. 5 shows that an increase in the spectral unevenness causes a decrease in the separability of the envelope spectra with respect to decay time and hence a decrease in accuracy of parameter extraction. Music exhibits discrete narrow-band excitation, therefore, the impulse response is only excited in these bands. Across an octave band the room response can vary, this produces an envelope spectrum estimate that is not representative of the broadband MTF. Due to this variation of room response across the octave band, similar envelope shapes exist for different EDT values. The uneven frequency response in music is due to the lack of excitation between notes and the uneven response of the notes particular to the piece of music. The modified pre-processor

is applied to music and compared with the octave-band pre-processor. EDT estimation accuracy is plotted in Fig 6.

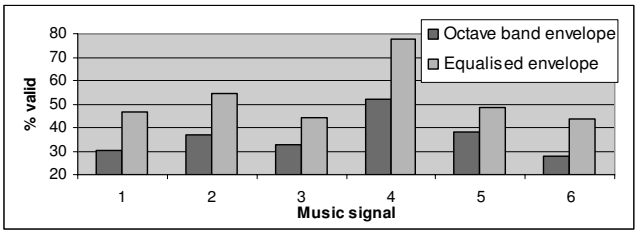

Fig. 6: Performance of music

A big increase in the accuracy of the system is seen when applying the modified pre-processor. This shows that by equalising the response of each note envelope in the overall envelope, the effect of the full octave band MTF is more accurately depicted in the envelope spectra. This method assumes that the impulse response magnitude is uniform across the octave band. Further improvement may be achieved by using the anechoic signal to normalise the envelopes or by improving the equalisation method.

3.3. Envelope Structure of music

Pilot studies show the CMTF pre-processor, which takes into account the non-stationary nature of the music envelope, yields a small increase in EDT estimation accuracy compared with the octave band envelope spectrum method. This improvement may indicate methods for improving the modified pre-processor. Therefore, the effect of the music envelope's non-stationary nature on the parameter estimation accuracy using the octave band envelope spectrum method is investigated.

An octave band music envelope is used to modulate white noise. The same envelope is used to design a filter [4] to apply to a random Gaussian process to produce a random signal with the same power spectrum as the music envelope. This signal is then also used to modulate white noise. Using the octave band envelope detector, both produce identical envelope spectra. Eight tracks of anechoic music and one of speech were used to generate the test signals. For each signal an ANN was trained using the octave band envelope spectra for different rooms. Training was carried out multiple times to get a stable validation error.

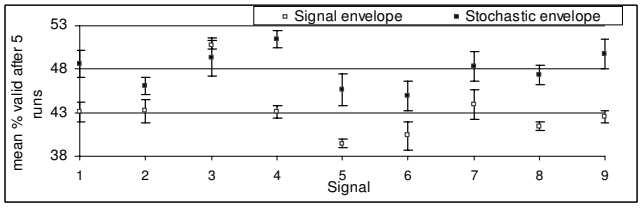

Fig. 7: Difference in accuracy due to non-stationarity

There is a statistically significant difference (5% level) between the validation error of stationary and non-stationary envelopes. Whilst the magnitude of the difference (≈4%) is small, it may be useful in improving the accuracy of the modified pre-processor.

4. CONCLUSIONS

This paper has proposed a method that compensates for the uneven spectra of music stimuli by applying a pre-processor that normalises the level of each note on the equal temperament scale before detecting the envelope spectra. This enables room acoustic parameters including RT and EDT to be reliably extracted from received music signals. Empirical results show large improvement in EDT estimation accuracy. Results also reconfirm that the highly non-stationary nature of the music envelope degrades the performance of the system but methods such as the local stationary hypothesis may be avenues towards further improvement. This paper also shows that the increased complexity of the room model reduces the estimation accuracy. However, the statistical test for over fitting described in this paper indicates that by increasing the number of example impulses, the validation accuracy may also be increased. Future work will focus on improving the spectral equalisation method, methods compensating for envelope non-stationarity and new algorithms combining the note envelope extraction method with the maximum likelihood estimation method for RT [14].

5. ACKNOWLADGEMENTS

The authors would like to acknowledge the support of the Engineering and Physical Sciences Research Council, UK (EPSRC GR/L898280) for funding this project.

6. REFERENCES

[1] T. J. Cox, F. Li and P. Darlington, "Speech transmission index from running speech: A neural network approach", *J. Acoust. Soc. Am.* 113, 1999 (2003)

[2] J.D. Polack, H. Alrutz, M.R. Schroeder,"The Modulation Transfer function of Music Signals and its Application to Reverberation Measurement*,"Acustica*,Vol.54, pp.256-265, 1984.

[3] M.R. Schroeder, "Modulation Transfer Functions: Definition and Measurement," *Acustica*, Vol. 49, pp. 179-182, 1981.

[4] N.D. Venkata and B.L. Evens, "Optimal design of real and complex minimum phase digital FIR filters," *ICASSP*, 1999.

[5] H. Kutroff, *Room acoustics*, Spon Press, London, 2000.

[6] M.F. Moller, "A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning," *Neural Networks*, 6:525--533, 1993.

[7] T.J. Cox, W.J Davies, Y.W Lam, "The Sensitivity of Early Sound Field Changes in Auditoria,"*Acustica*, 79, 27–41. 1993

[8] J. Nannariello and F. Fricke "The Prediction of Reverberation Time Using Optimal Neural Networks" *Building Acoustics* Vol 9, No.1 pp 5–28, 2002

[9] F. Fessant, P. Akin, L. Oukhellouu, S. Midenet "Comparison of supervised selforganizing maps using Euclidian or Mahalanobis distance in classification context"*IWANN2001*, 2001.

[10] V.L. Jorden. "Acoustical criteria for Auditoriums and Their Relation to Model Techniques", J. Acoust. Soc. Amer. Vol 47 pp 408-412

[11] B.-I. Dalenbäck "Verification of Prediction Based on Randomized Tail-Corrected Cone-Tracing and Array Modeling" , *137th ASA/2nd EAA* Berlin March 1999

[12] T. Houtgast, H. J. M. Steeneken, "Envelope spectrum and intelligibility of speech in enclosures", *IEEE-AFCRL Speech conference proceedings,* pp. 392-395

[13] T. Houtgast, H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility", *Acustica*, Vol 28, pp 66 ,1973

[14] Y. Zhang, J. A. Chambers, "Blind Estimation of Reverberation time", *to be submitted ICASSP* 2006