Monaural room acoustic parameters from music and speech

Paul Kendrick, Trevor J. Cox, and Francis F. Li
Acoustic Research Centre, University of Salford, Salford M5 4WT, United Kingdom

Yonggang Zhang and Jonathon A. Chambers
Advanced Signal Processing Group, Department of Electronic and Electrical Engineering, Loughborough University, Leicestershire LE11 3TU, United Kingdom

(Received 12 October 2007; revised 21 April 2008; accepted 23 April 2008)

This paper compares two methods for extracting room acoustic parameters from reverberated speech and music. An approach which uses statistical machine learning, previously developed for speech, is extended to work with music. For speech, reverberation time estimations are within a perceptual difference limen of the true value. For music, virtually all early decay time estimations are within a difference limen of the true value. The estimation accuracy is not good enough in other cases due to differences between the simulated data set used to develop the empirical model and real rooms. The second method carries out a maximum likelihood estimation on decay phases at the end of notes or speech utterances. This paper extends the method to estimate parameters relating to the balance of early and late energies in the impulse response. For reverberation time and speech, the method provides estimations which are within the perceptual difference limen of the true value. For other parameters such as clarity, the estimations are not sufficiently accurate due to the natural reverberance of the excitation signals. Speech is a better test signal than music because of the greater periods of silence in the signal, although music is needed for low frequency measurement.
 [DOI: 10.1121/1.2931960]

PACS number(s): 43.55.Mc, 43.60.Np, 43.60.Cg [NX] Pages: 278–287

I. INTRODUCTION

Room acoustic parameters, such as reverberation time, are routinely used in the design and evaluation of enclosed and semienclosed spaces such as concert halls, classrooms, and stadia. Normally, these parameters are measured by with artificial test stimuli, such as gun shots, pseudorandom noise, or sine sweeps, because this yields accurate and reliable results. The interest in this paper, however, lies in trying to measure the parameters using the natural sounds occurring in rooms; of particular interest is the use of speech or music as a test signal. Measuring by using natural sounds should make occupied measurement easier because the signals will not disturb the room occupants. Consequently, it should facilitate the monitoring of in-use conditions; this is what motivates the work presented.

Li and Cox[1] developed a machine learning method to determine the speech transmission index from received running speech. This method is quasiblind: source signals do not need to be monitored during measurements but they are required during the training phases of the machine learning algorithm. A key limitation of this method is that it is an empirical approach which requires extensive training before use. Even so, it can be shown that with about a minute of speech, high accuracy can be obtained. With slightly compromised accuracy, the method can be made completely blind because the low frequency statistical properties of speech are not very different from speaker to speaker.[2] This method is termed the "envelope spectrum method" because of the preprocessor used.

The envelope spectrum method was originally developed to be used with narrated speech. For parameters used in the evaluation of concert halls, however, it is natural to examine the use of music as an excitation signal, and this has not been considered with the method before. In comparison to speech, music offers a larger bandwidth of excitation and so enables acoustic parameters to be measured over a wider range of frequencies. In particular, in comparison to speech, music has more sound power in the lower octave bands that are considered in room design. Music, however, is a rather imperfect test signal, as shall be shown later. To work with music, the envelope spectrum method needed to be adapted to deal with the inherent statistical differences between speech and music, and these adaptations are outlined in this paper. Since the development of the envelope spectrum method, the data set of room impulse responses used to train the machine learning algorithm has been improved, and this affects the accuracy of the method. Details of this are also provided in the paper.

An alternative to the envelope spectrum method is one using a maximum likelihood estimation (MLE). This approach was originally developed by Ratman *et al.*[3] The concept is to use decay phases following speech utterances or music notes to estimate the reverberation time. The method is inherently blind because it works off sound decays and uses the shortest decay as the one where the excitation is most impulsive and the decay is least corrupted by noise. The accuracy demonstrated by Ratman *et al.*, however, was insufficient for parameter evaluation. Kendrick *et al.*[4] improved the method by using a more realistic model for the envelope of room impulse responses, one that allowed for nondiffuse spaces, and they demonstrated good accuracy with reverberation time and early decay time (EDT). As de-

tailed below, a further extension is needed to extract other monaural parameters[5] because an accurate estimation of the decay curve energy is needed.

When designing a room for music production, a measure of clarity, such as clarity, center time, or an early to late energy ratio, is usually used alongside reverberation parameters to evaluate the acoustic quality. For speech, the Deutlichkeit serves a similar purpose. This paper examines how the two methods, envelope spectrum and MLE, can be used to estimate parameters which examine the relative balance of the early and late energies. Finally, the paper compares and contrasts the two methods for both speech and music examining, which is the best method and which is the best test signal.

II. METHOD

A. Parameters

This paper considers the following monaural parameters: reverberation time (T_{60}), EDT, clarity (C_{80}), and center time (T_s), which are well established and defined in an international standard.[6] In this paper, the reverberation time is calculated from 25 dB of decay. In typical speech and music, it is difficult to measure the end of reverberant tails because of masking from other utterances and notes, and this often makes the estimation of the more normal T_{30} inaccurate. Unless otherwise stated, the results are shown for the 1 kHz octave band.

Many of the results in this paper are presented in terms of how many of the estimations fall within one perceptual difference limen of the true value. The difference limen is a measure of the smallest perceivable difference in a stimulus. The difference limen for reverberation time with music signals is around 5% for reverberation times above 0.6 s, and increases to about 12% for shorter reverberation times.[7] There is limited information concerning the EDT difference limen so the criterion for the required accuracy is set at $\pm 5\%$ but with a minimum error of ± 0.1 s, as having accuracy better than this is not usually required.[8] The difference limen for clarity and center time were taken from Cox *et al.*[9]

B. Data sets

To test and develop the methods, a data set of room impulse responses is needed. For the envelope spectrum method, this data set has to include many thousand examples; too many to be obtained from real room measurements. Previously,[1] stochastically generated impulse responses were used. However, in recent years, there have been significant advances in the modeling of rooms by using geometric algorithms. Consequently, a commercial package[10] with a proven track record that utilizes randomized tail-corrected cone tracing was used to generate a training set of examples for teaching the machine learning algorithm used in the envelope spectrum method. The geometric model was also used to generate the first validation set used to examine the performance of both the envelope spectrum and MLE methods.

Rooms with randomized geometries, surface treatments, source, and receiver positions were generated, and the impulse response then calculated by using the geometric room acoustic model. Two room types 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 a variable population density. The receiver grid was spread over the audience area. The algorithm that generated the random rooms was given limits for overall room dimensions, aspect ratios, and material properties to ensure the generated impulse responses were realistic. Room volumes ranged from 75 to 30 000 m^3. Material properties were randomly selected so that the reverberation time of the spaces (calculated by using Sabine's formulation) was set to be less than 4 s; longer reverberation times are seen in the results below because of the nondiffuse nature of many of the spaces. The idea is to generate a wide variety of spaces to allow the machine learning algorithm to learn all possible cases which might occur in reality, and also to rigorously examine the robustness of both methods by testing with a wide range of cases. In addition, a second validation set containing 20 real room impulse responses[11] were used to evaluate performance.

One problem with the use of the geometric room acoustic model was the fact that the calculation was done in octave bands and the results then recombined into an impulse response. This can result in significant discontinuities in parameter values at the edges of octave bands; something which will not arise in real rooms. As shall be shown later, such differences between the simulated and real room responses affect the performance of the methods to extract the parameters.

Music[12] and speech[13] recorded in anechoic conditions were convolved with the room impulse responses to produce the test signals.

C. Envelope spectrum

Figure 1 shows a schematic of the envelope spectrum method used to estimate an acoustic parameter in a single octave band from music. (Li and Cox[1] contains a more detailed description of this approach applied to speech, together with justifications for the technique). An anechoic signal is convolved with a room impulse response to provide an example of a transmitted signal. This is then passed through the low frequency envelope spectrum preprocessor, shown by the dashed box in Fig. 1, and from there to the artificial neural network (ANN). The ANN outputs an estimate of an acoustic parameter. The correct value for the acoustic parameter is calculated from the impulse response by using standard definitions.[6] The difference, or error, between the true and estimated parameters is used to update the internal weights and biases within the ANN. This process is repeated for thousands of different example impulse responses and, gradually, the ANN learns to produce more accurate estimations of the parameter for a wide variety of rooms.

After training, the weights and biases of the ANN are fixed and the performance of the ANN is tested by using impulse responses not previously used in training; these data sets are referred to as validation sets. Approximately 6500 artificially generated impulse responses were used for train-

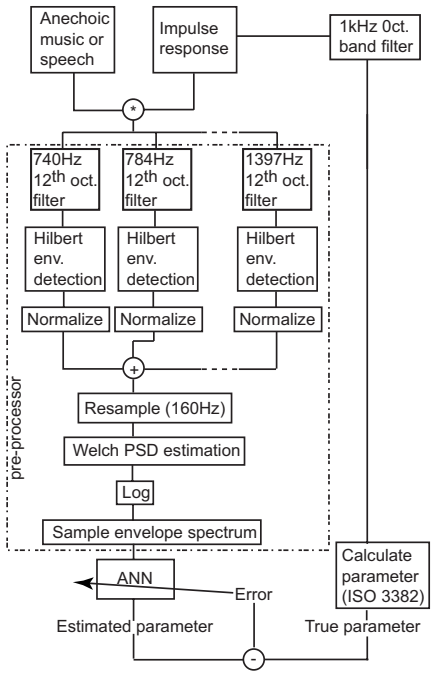

FIG. 1. Schematic of envelope spectrum method for estimating room acoustic parameters from music.

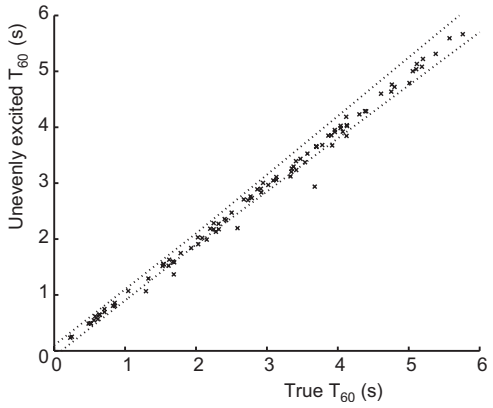

FIG. 2. Unevenly excited reverberation time compared to the true, evenly excited value. The unevenly excited reverberation time has the same octave band frequency response as a typical piece of music. The dashed lines indicate limits derived from the perceptual difference limen.

ing and about 700 examples for validation. These were generated by using the geometric room acoustic model described previously. A second validation set was based on the measured impulse responses.

The ANN is being used to carry out a mapping from the sampled envelope spectrum outputted from the preprocessor, to the desired acoustic parameter; it is essentially learning an unknown nonlinear relationship. An ANN is well suited to this task because of its inherent ability to deal with nonlinear mappings in an efficient manner. A multilayer perceptron with 40 input neurons, 30 and 10 neurons on the first and second hidden layers, respectively, and 1 output neuron was used. The network size was determined by trial and error. A bipolar sigmoid activation function was used. The Levenberg–Marquardt method was used, which offers an order of magnitude decrease in learning time over the back propagation rule.[14] To prevent overfitting, training was stopped when the validation error increased for a number of iterations.

The key stage in the development of this machine learning regime is the development of a suitable preprocessor. The primary role of the preprocessor is to greatly reduce the amount of data so that it can be effectively processed by the ANN. A few minutes worth of audio has to be reduced to a few tens of pieces of data. Finding a preprocessor that can do this while retaining meaningful information regarding the room acoustic parameters is a significant challenge. Reverberation is known to act as a low frequency filter on the envelope of signals. Houtgast and Steeneken[15] showed that the low frequency envelope spectrum can be used as an indicator of the level of reverberation added to speech. It is suggested that the low frequency envelope spectrum also contains information about other decay curve characteristics such as clarity—this hypothesis will be explored below.

As shown in Fig. 1, the first stage of the preprocessor is to break the signal down into one-twelfth octaves across the octave band being considered, where each one-twelfth octave corresponds to a note within an even-tempered chromatic scale. Previous work[1] with speech did not have to do this because it has a flatter frequency response than music across each octave band. Without individually processing each one-twelfth octave, the estimations using music were found to be quite poor.

Excitation within an octave band is uneven with respect to frequency for music signals. This unevenness in the excitation signal must be compensated for by the parameter extraction system. Consider comparing an acoustic parameter calculated from a standard impulse response with even excitation, and the same parameter calculated by using an impulse response filtered to have a frequency response corresponding to the average power spectrum of a piece of music, i.e., an uneven frequency response.

To get the latter impulse response, first, the frequency content of the music extract was estimated by using the Welch power spectrum method using 0.5 s windows and 50% overlap Hanning windows. This frequency response is used to design a linear phase finite impulse response (FIR) filter with the same frequency response as the average spectrum of the music signal. Linear phase is used so that the time response of the impulse response is least distorted, only delayed. This delay is removed after the filtering process (equivalent to zero phase filtering). A short tap length of 301 is used for the FIR filter. The "reverberation time" of the weighing filter was checked to make sure that it did not ring for too long and generally, it was 0.07 ± 0.01 s. The impulse response is passed through this filter and the delay compensated for.

Figure 2 shows the reverberation time for an uneven excitation (excitation with similar spectrum to a music signal) plotted against the reverberation time with even excitation (broadband excitation) for the impulse responses simulated by using the geometric room acoustic model. The dotted lines indicate the required accuracy for parameter estimation and is based on the perceptual difference limen. The result shown is for the 1 kHz octave band. The error intro-

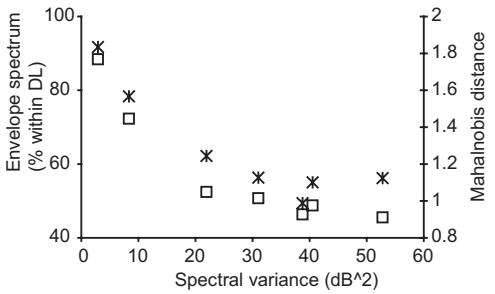

FIG. 3. Variation of the □ Mahalnobis distance and * envelope spectrum method accuracy with the variance of the music spectrum. Six pieces of music and one piece of speech.

duced by the uneven excitation is the same order of magnitude as the difference limen. For the validation set using the 20 real room measurements, the error is somewhat less, being approximately halved. This might be due to the restricted range of reverberation times in the real room dataset, or that the real room impulse responses have less variation in the acoustic parameters across the octave band. For all the other acoustic parameters, EDT, clarity, and center time, the error is larger, at worse being about twice the difference limen.

The one-twelfth octave filtering used in the preprocessor helps to compensate for the uneven excitation. After the one-twelfth octave filters, the envelope of each of the twelve signals is detected by using the Hilbert transform[16] and a normalization to the root mean square in each one-twelfth octave carried out. This normalization reduces the effect of uneven excitation with respect to frequency. The objective is to get a result closer to the one which would be obtained from an artificial test signal with even excitation with respect to frequency. Even with this normalization, however, the unevenness of the power spectrum of the music across the octave band has a significant effect on the accuracy of the parameter estimations. Figure 3 illustrates this for reverberation time. Seven source signals are used comprising six pieces of music and one piece of speech. For each source signal, the ANN within the envelope spectrum method is trained. The validation set, simulated by using the geometric room acoustic model, is used, with the percentage of estimations within one difference limen of the true result being used as a measure of success. This success rate is plotted against the spectral variance calculated across the octave band. The results show that as the excitation becomes more even across the octave band, as the spectral variance decreases, the envelope spectrum method becomes more successful at estimating reverberation time.

For each ANN, the same piece of music or speech is used in both training and validation phases. Li *et al.*[1] presented a method that enabled almost any speech source to be used once the ANN is trained. This facilitated "source independent" estimation of the parameters. Source independence is successful with speech, as the low frequency envelope spectra between talkers are similar. However, due to the highly diverse range of music signals, the low frequency envelope spectrum for a piece of music is quite unique and source independence in the presented framework is not feasible.

Producing each of the envelope spectrum results in Fig. 3 requires training an ANN. Because there is always uncertainty about certain aspects such as ANN size and update rate, a measure to assess the quality of the input data was sought, which is more mathematically rigorous. The Mahalnobis distance[17] is a measure of data separability applied to the input data of the ANN. The measure gives a statistically weighted Euclidian distance between two sets of data. If the distance between the data are large, the data are well separated, and this makes it much more likely that the ANN will be able to carry out the mapping to the acoustic parameters. If the distance is small, however, this indicates ambiguity in the input data set which will make it difficult for the ANN to be successfully trained.

To calculate the Mahalnobis distance, the data are split into groups whose sizes are determined by the difference limen. For example, the difference limen for clarity is known to be about 1 dB, therefore, the range of possible clarity values is quantized into 1 dB steps, and all input data relating to a clarity value within ± 0.5 dB of the center value are assigned to that group. The Mahalnobis distance is calculated between each adjacent group which gives a useful indicator of how separability varies with the parameter. The overall separability of the data set, as shown in Fig. 3, is calculated by averaging the Mahalnobis distances for all clarity groups. (Incidentally, the spectral variance used is not simply a variance of the power spectral density values across the octave band of the signal. To be consistent with the ANN input, it is necessary to first split the data into one-twelfth octave bands and to normalize these as was done in the preprocessor, before recombining the results and calculating the variance.)

As Fig. 3 confirms, pieces of music with more even excitation produce more accurate results with the envelope spectrum method. An alternative way of improving accuracy is to use multiple pieces of music and average the results for the acoustic parameters.

Returning to the preprocessor, the one-twelfth octave normalized envelopes are then recombined, down sampled, and the power spectral density estimated. This is done by using a Welsh algorithm using 50% overlap and 3.5 s Hanning windows (the best window size was determined empirically). After taking a log of the envelope spectrum, the spectrum is converted into a constant percentage bandwidth spectrum. To get 40 bandwidths and hence, 40 spectral samples from 0.2 to 25 Hz, required a bandwidth of $\approx 9/50^{\text{ths}}$ of an octave to be used. These data then form the input data for the ANN.

D. Maximum likelihood estimation

This method has recently been detailed in another paper[4] so only an outline is provided here with the adaptations necessary to obtain other monaural acoustic parameters. The signal is filtered into the octave band being considered. The envelope of the received reverberated music or speech is detected by using the Hilbert transform.[16] A signal segmentation and selection process is undertaken to find parts of the signal that contain free decay—the reverberant tail at the ends of words or notes (decay phases). Decay phases with

sufficient dynamic range for parameter estimation are sought (>25 dB). A MLE is undertaken on these decay segments providing a robust estimation of the envelope of the room impulse response. The estimated impulse response h_e is modeled as a noise signal modulated by an envelope with a double exponential decay

$$h_e(n) = [\alpha a_1^n + (1-\alpha)a_2^n]s(n), \tag{1}$$

where n is the time index, s is the Gaussian noise signal, a_1 and a_2 determine the decay rate of the two exponential functions, and α gives the relative importance of the two exponential decays. A double decay is used to better model the nondiffuse spaces. The fine details (individual reflections) of the impulse response are represented by the zero mean Gaussian noise signal. The MLE is essentially an efficient and robust method of fitting this estimated impulse response to the measured free decay at the end of words or notes. a_1, a_2, α, and the variance of the Gaussian noise are determined via numerical optimization; this is done by forming a likelihood function[18] and maximizing it.

Any piece of music or speech will yield a number of free decay segments from which a number of impulse response envelopes can be found. Previously,[3] the envelope equating to the shortest reverberation time was chosen as the best estimated because this has least contamination from noise. This selection method, however, is not appropriate when trying to estimate other acoustic parameters. For this, a new approach was developed which builds up a best estimate of the decay curve in sections, by considering the energy along the decay phase.

Consider a single measured decay phase x; this can be split into a number of components,

$$x(n) = h(n) \otimes [\delta(n) + d(n)] + r(n), \tag{2}$$

where $\delta+d$ represent the direct sound, h is the room impulse response, and r are the reflections excited by signals occurring before the start of the selected decay phase or other unwanted noise. The direct sound is split into two components, δ representing ideal impulsive excitement at the start of the frame and a competing noise term d representing the subsequent decay of the musical note or speech utterance.

Calculating the signal energy based on Eq. (2) yields a number of squared and cross terms. However, provided that the energy is estimated over a sufficiently long time window, it may be assumed that the cross terms reduce to zero as variables within these cross terms are uncorrelated and have a zero mean value. Therefore, the signal energy is approximately given by

$$\sum x^2 \approx \sum h^2 + \sum (d \otimes h)^2 + \sum r^2. \tag{3}$$

The first term on the right hand side of Eq. (3) is constant so the energy in a decay phase only changes with d and r. Therefore, by finding the decay phase with the smallest energy for a given time period, the cleanest region of free decay is being found—freest from additional notes masking the decay phase, reverberance of the notes or speech utterances, and the region most like the room impulse response. This preference toward the minimum energy favors the most impulsive musical notes or speech utterances, which is assumed to also tend to reduce the unwanted effects of uneven excitation.

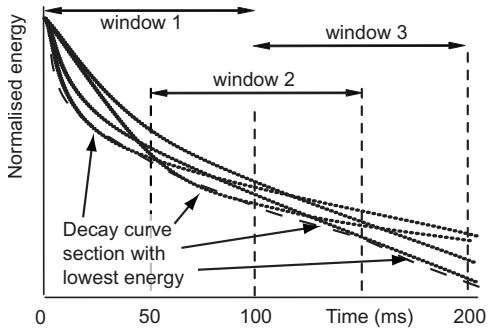

FIG. 4. Illustration of method for reconstructing (– – – – –) best decay curve estimate from (————) four decay curve estimates. In each of the three windows, four decay curves with least energy are chosen. Triangular windowing is used to smooth the response between windows.

Experience has shown that there may be no single decay phase that is "cleanest" for the full length of the impulse response and, consequently, the following process is used, as illustrated in Fig. 4. The decay phase estimates are windowed by using a 0.1 s rectangular window using 50% overlap and the energy is calculated for each window. Figure 4 shows four such estimates. The decay curve with the least energy in each 0.1 s window is selected as the cleanest portion for that range. The final decay curve estimate is constructed from these cleanest portions, using triangular windows to smooth discontinuities in the decay curve. This is then the best estimate of the impulse response envelope which is shown as the dashed line.

Rather than taking a single best estimate of the impulse response envelope from a long piece of speech or music, an average across several best estimates is taken to improve accuracy. With speech, an 8 min recording is split into eight blocks each 1 min long. Within each of these 1 min blocks, the above procedure is undertaken to get a best estimate of the impulse response envelope, yielding eight estimates in total. These estimates are then "averaged" to get the final form of the impulse response envelope. The presence of background noise causes outlying estimates due to overestimation of the rate of decay. Outliers of this nature are common and, therefore, the median is used rather than taking an average using the mean. This operation is illustrated in Fig. 5, illustrating, among others, an outlying decay curve and showing that by using the median, the final estimation has not been overly biased by the outlier. This averaging across eight-estimates helps overcome stochastic variability in the frequency content of single decay phases.

Incidentally, the impulse response should be time invariant, but in many rooms, this is not true. However, as the estimation is over the smoothed envelope of the impulse response, the effect of time variance is probably not going to be a great problem; although this has yet to be formally tested. Once the optimal envelope of the impulse response has been estimated, standard definitions are used to obtain the room acoustic parameters.

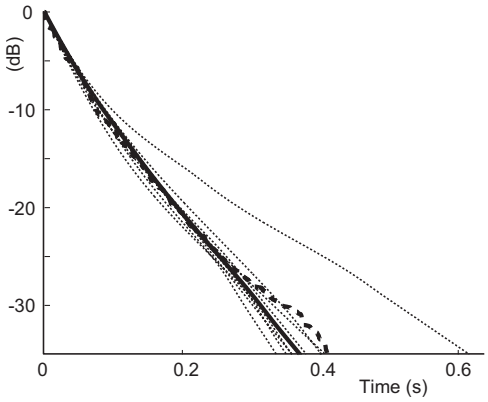

FIG. 5. Impulse response envelopes (……) individual estimates including one outlier shown to right, (————) best estimate formed by taking median of individual estimates, and (– – – – –) true envelope.

III. RESULTS

A. Speech

In this section, performance of the envelope spectrum and MLE methods are compared by using speech excitation. About 9 s of anechoic, male, narrated running speech[13] is used in both cases. By narrated speech, it is meant that the speaker is given a passage of text to be read aloud. This technique tends to slow down the rate at which the narrator speaks and so gives more gaps between utterances where decays can be seen.

Two data sets are used to examine the success of the methods, one based on speech convolved with simulated impulse responses generated by the geometric room acoustic model and the second, a set based on speech convolved with real measured room impulse responses. The performance of the parameter estimation methods vary between these two data sets and comparing the two sheds light on the robustness of the methodologies.

Consider first the results for the data set based on the simulated impulse responses from the geometric room acoustic model. The envelope spectrum method produces as good or better parameter estimation accuracy than the MLE method for these data. In general, the MLE method finds it harder to accurately estimate the early parts of the impulse response because it cannot separate the decay due to the room reflections from the decay of the speech utterances. Because the envelope spectrum method is based on an experiential method, the ANN learns to compensate for the overestimation of the early reverberation that occurs due to the inherent reverberance of the speech utterances. Consequently, the envelope spectrum method is more accurate than the MLE method for parameters such as clarity, center time, and EDT, but for reverberation time, there is little difference in the accuracy of the two methods.

Figure 6(a) shows the results for reverberation time, showing similar accuracy for both methods. To improve the accuracy of the methods further, especially at middling reverberation times, a longer section of speech can be used, or further averaging of the estimated parameters across many lengths of recorded speech can be done.

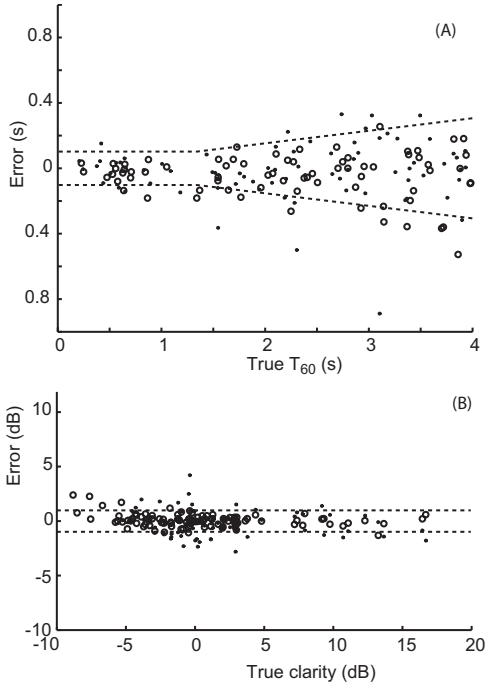

FIG. 6. Error in parameter estimation verses the true value: (A) reverberation time and (B) clarity. The dotted lines indicates the difference limens. (●) MLE and (○) envelope spectrum method. Validation set using simulated impulse responses. Speech excitation.

Figure 6(b) shows the results for clarity, as an example of one of the other parameters where estimation is more difficult. Clarity requires the accurate estimation of the energy arriving in the first 80 ms after the direct sound, and the MLE method finds it difficult to precisely obtain this value because of the natural reverberance of many speech utterances. In contrast, a method based on machine learning, such as the envelope spectrum method, can learn to compensate for errors in the estimation. Hence, for medium and large clarity values, the envelope spectrum method is more accurate than the MLE method. For very low clarity values, there are no results for the MLE method because the segmentation method failed to find any decay phases with sufficient dynamic range; these are very reverberant rooms where the start of the current utterance significantly masks the end of the decay of the previous utterance. To obtain a MLE estimation at such low clarity values requires larger time gaps between the utterances. As the clarity increases, the accuracy of the MLE estimations increases. While the envelope spectrum method does provide an estimation at these low clarity values, the accuracy of the estimation suffers because there is an insufficient information about the late part of the decay which is masked by subsequent utterances.

Tests on real room measurements, however, yield a slightly different story to that found with the simulated impulse responses. This validation set uses impulse responses measured in real rooms convolved with anechoic speech. Both the envelope spectrum and the MLE methods provide roughly the same accuracy for all the parameters. Figure 7 shows the results for reverberation time and clarity for the real room measurements, which can be compared to those shown in Fig. 6 for the validation set using simulated im-

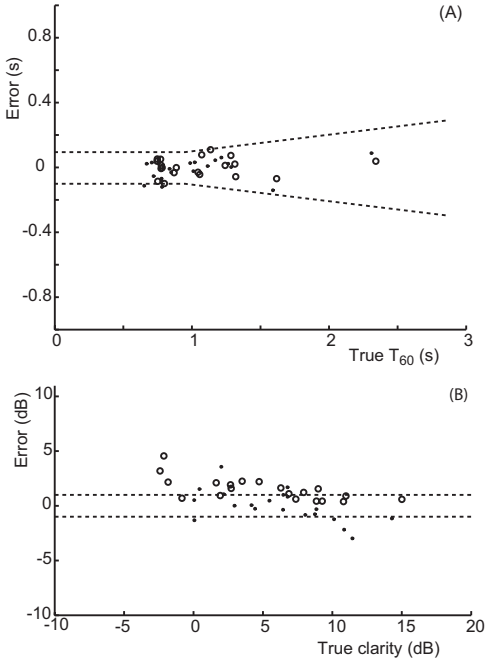

FIG. 7. Error in parameter estimation verses the true value: (A) reverberation time and (B) clarity. (●) MLE and (○) envelope spectrum method. Validation set using real room measurements. Speech excitation.

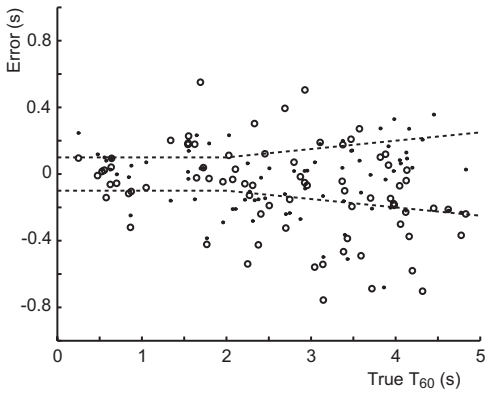

FIG. 8. Error in reverberation time estimation verses the true value. The dotted lines indicates the difference limen. (●) MLE and (○) envelope spectrum method. Validation set using simulated impulse responses. Music excitation.

pulse responses. For reverberation time, the MLE method becomes more accurate for the real room measurements in comparison to the validation set using simulated impulse responses. (Note that although the graphs imply this is also true for the envelope spectrum method, the difference is not statistically significant). For EDT, the envelope spectrum method is slightly more inaccurate and the MLE method is slightly more accurate when using the real measurements in comparison to the simulated ones. For clarity and center time, the MLE method gives similar accuracy for the real and simulated validation datasets, but the envelope spectrum becomes less accurate and furthermore, a bias error is introduced.

The loss in accuracy with the envelope spectrum method when estimating some parameters probably occurs because the simulated room impulse responses used to generate reverberated speech for training the ANN are not completely representative of real room impulse responses. Consequently, the data used for training and validation have some significant statistical differences. The introduction of a bias, as shown with some parameters and illustrated in Fig. 7(b), is good evidence for this. As an ANN works to minimize the mean square error, a well trained ANN should not generate a bias error, unless something is wrong, such as differences between the validation and training sets. It might be anticipated that as the accuracy of the geometric room acoustic models improves, then this problem should disappear because the training set will better match reality.

The MLE performs better with real room measurements when compared to the results for the reverberated speech by using simulated impulse responses. The real impulse responses have a greater reflection density than the simulated ones and, consequently, there is a smoother transition from the early to the reverberant sound field. It is suggested that this might improve the fitting of the simple model of envelope-shaped Gaussian noise [Eq. (1)] to the real room data, and this could therefore explain the improved accuracy in the parameter estimation.

B. Music

1. Envelope spectrum versus MLE

In this section, the performance of the envelope spectrum and MLE methods are compared using music. Six different pieces of music[12] are used, 2–4 min in length. As noted above, the accuracy of the estimation changes from piece to piece, for instance, the spectral variance affects the accuracy of the envelope spectrum method estimation (Fig. 2). Consequently, the results from the parameter estimations are averaged across all six pieces of music to improve accuracy and to reveal underlying trends.

Consider first the validation set generated by using simulated impulse responses from the geometric room acoustic model. For reverberation time, the MLE and envelope spectrum methods produce similar accuracy, as shown in Fig. 8, although the MLE is marginally more accurate. However, for the other parameters—EDT, clarity, and center time—the envelope spectrum method is more accurate than the MLE method. Again, the MLE method finds it difficult to accurately estimate the early parts of the impulse response due to the presence of the reverberance of the musical notes themselves. Figure 9 shows the results for these three parameters. For the envelope spectrum method, most of the parameter estimations are within the difference limen. For the EDT and center time, there is a tendency for the MLE method to overestimate the value by an amount larger than the difference limen. For clarity, there is a tendency for the MLE to overestimate low values of clarity and underestimate large values, in other words, the range of clarity estimated by the MLE is smaller than the true range.

With real room measurements, the results are again somewhat different. Figures 10 and 11 show the error in the four parameters as a function of their true value. In comparison to the use of the simulated validation set, the estimation of reverberation time and EDT has become more accurate with real room measurements for the MLE, while the enve-

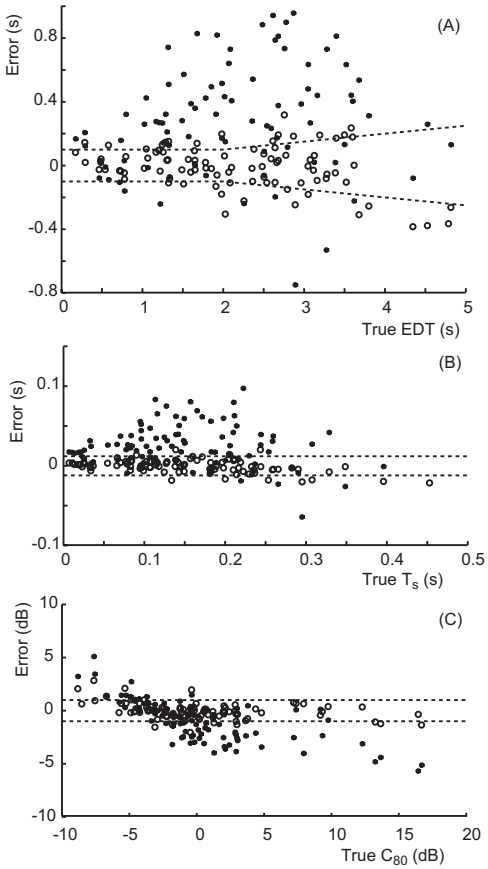

FIG. 9. Error in parameter estimation verses the true value: (A) EDT, (B) center time, and (C) clarity. The dotted lines indicate the difference limens. (●) MLE and (○) envelope spectrum method. Validation set using simulated impulse responses. Music excitation.

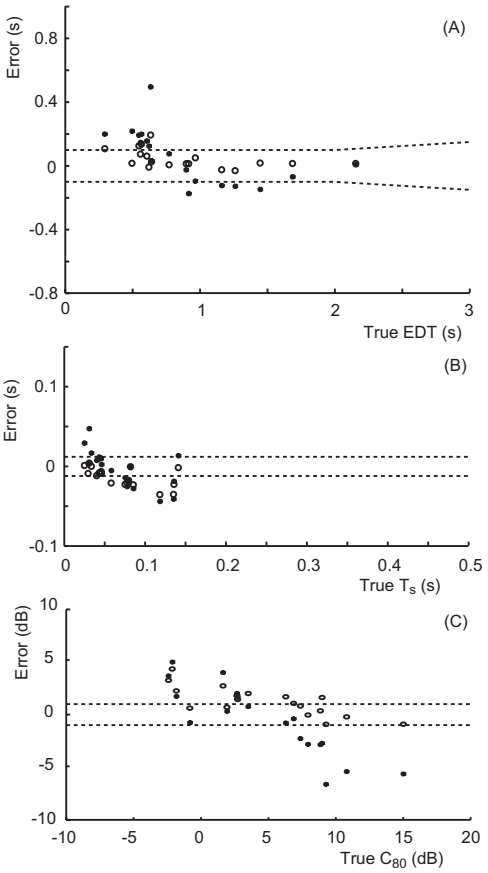

FIG. 11. Error in parameter estimation verses the true value: (A) EDT, (B) center time, and (C) clarity. The dotted lines indicate the difference limens. (●) MLE and (○) envelope spectrum method. Validation set using real room measurements. Music excitation.

lope spectrum method gives similar accuracy. The estimation of clarity is still problematic with the MLE method; there is a tendency for overestimation of low clarity values for the envelope spectrum method. For center time, the MLE method gets somewhat more accurate with real room measurements, but for the envelope spectrum method, a bias error is introduced with the parameter values being underestimated.

The introduction of a bias within the envelope spectrum results is probably again indicative of differences between the training data set which used simulated room impulses responses and the real room measurements (see discussions on speech above). The envelope spectrum method has problems estimating reverberation times. It is suggested that this arises because of the masking of the later parts of the impulse response by subsequent notes. The envelope spectrum method carries out an evaluation on the whole music passage, and in a piece of reverberated music, the early decay of notes is going to be more prominent than later decay portions. Consequently, the envelope spectrum method struggles to accurately estimate reverberation times because the information about late decay are lost in the vast amount of data from the whole music passage. (To explore this further, companding was carried out on the signal before it was fed to the envelope spectrum preprocessor. Companding biases the signal toward late decay and, as expected, this improved the reverberation time estimation, but at the expense of accurate EDT estimation). This was not such a problem with speech because this has more periods of silence. The MLE deliberately pulls out free decay sections and selects those with sufficient decay, therefore, it does not suffer from the same problem. It is suggested that by applying a similar signal segmentation process to the envelope spectrum method, the accuracy of reverberation time estimation might be improved.

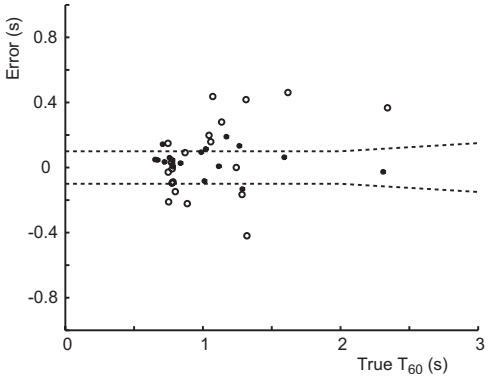

FIG. 10. Error in reverberation time estimation verses the true value. The dotted lines indicates the difference limen. (●) MLE and (○) envelope spectrum method. Validation set using real room measurements. Music excitation.

TABLE I. Standard errors for parameters when estimated using the maximum likelihood and envelope spectrum method. Standard errors calculated using responses where $T_{60} < 2.3$s to allow comparison of real and simulated.

Room impulses	Method	Source signal	Standard error RT (s)	EDT (s)	C80 (dB)	T_s (ms)
Simulated	MLE	Music	0.15	0.20	2.8	13
		Speech	0.14	0.25	1.5	8
	Envelope spectrum	Music	0.19	0.07	0.79	3
		Speech	0.09	0.06	0.56	3
Real	MLE	Music	0.08	0.14	2.5	15
		Speech	0.06	0.12	1.5	10
	Envelope spectrum	Music	0.18	0.10	1.5	12
		Speech	0.06	0.08	1.05	12

For the MLE method, the improvement in estimation of reverberation time parameters when using real measurements is again thought to be due to the real room impulse responses having greater reflection density than the simulated ones, and so it is easier to fit the simplified room model to the data. For high clarity values, where underestimation occurs, then, the reverberance of the notes is causing an overestimation of both the early and late sound energies, hence, resulting in a lower clarity than expected. For low clarity values, the MLE method struggles to fit the detailed effects of strong reflections in the early sound field, and so produces a consistent overestimation.

IV. DISCUSSIONS

Music and speech are obvious choices for measurement using naturalistic signals in performance spaces, and each has their own particular advantages and disadvantages.

Narrated speech has the advantage of containing plenty of gaps between the signal excitations, and so is rich in clean decay phases. This offers many opportunities for averaging estimations when using the MLE method and more accurate estimation with the envelope spectrum method, as there will be many decay phases where late reflections are not masked by subsequent utterances. However, speech excitation has limited bandwidth and lacks energy in low frequencies, say, for the 250 Hz octave band and below, and consequently, estimations are less accurate in lower octave bands,[4] and consequently, music is needed for these bandwidths. The music tracks used in this study lacked high frequency excitation above 2 kHz, which means that estimation accuracy is affected there also.

Table I compares the standard error of each parameter estimate for speech and music by using both the MLE and envelope spectrum methods. Table I indicates that the MLE method more accurately predicts the late decay properties than the envelope spectrum method, while the envelope spectrum method more accurately predicts the early decay behavior than the MLE method.

The ideal music signal is one that has plenty of short transient sounds, and whose excitation is even across the octave band, in other words, a piece of music which breaks traditional rules of western harmony and gives the same weight to all notes in the chromatic scale. Alternatively, averaging over different pieces of music can be effective, especially if the pieces are in a variety of musical keys. Another approach is to use pieces with lots of untuned percussion. The best music also has large gaps between the notes so the decay phrases are prominent.

Table II indicates how many regions of decay with at least 25 dB of dynamic range each of the anechoic signals contain. This can indicate how accurate the ML estimates are using a particular signal. Table II also indicates the percentage of silence within each signal; this combined with the spectral variance previously detailed in Fig. 3 is a good indicator as to the suitability of the signal for parameter estimation using the envelope spectrum method.

Neither the envelope spectrum method nor the MLE technique offers a single foolproof method for estimating room acoustic parameters from naturalistic signals. The MLE technique is appealing because it is not empirical and so is potentially more robust. Furthermore, it is truly blind and does not require explicit knowledge of the anechoic excita-

TABLE II. Average number of decay phases per minute exhibiting at least 25 dB of decay, and percentage of relative quiet in each anechoic signal. Percentage of relative quiet is calculated by computing the percentage of (non-overlapping) 0.05s length windows in each signal with energy 40 dB less than the maximum energy.

Signal	Average number of decay phases/minute	Amount of relative quiet in signal (%)
Music track 1	4.2	0.19
Music track 2	4.3	2.2
Music track 3	4.8	0.86
Music track 4	12	6.4
Music track 5	3.2	0.53
Music track 6	5.0	0.17
Speech	37	25

tion signal. However, the MLE method has problems accurately estimating the early sound field because it cannot compensate for the inherent reverberance of music notes and speech utterances. As the calculation of reverberation time by definition avoids this problematic region of decay, the MLE method is most successful in estimating reverberation times.

Currently, the inaccuracies of geometric room acoustic models used to provide training data for the envelope spectrum method limit the accuracy that can be obtained. Hopefully, as geometric room acoustic models improve, this problem will be resolved. The envelope spectrum method is not blind but requires knowledge of the test signal during training. This limits the applicability of the method. For example, with live orchestras, the accuracy will be compromised.

V. CONCLUSIONS

This paper has examined two methods for extracting room acoustic parameters from music and speech. The motivation is to enable in-use measurements of spaces without using artificial test signals. The first method uses a machine learning approach combined with a envelope spectrum preprocessor. Previously, this method had been used for speech; this paper details adaptations necessary to make this method work for music. The second method uses a MLE applied to decay phases at the end of speech utterances or musical notes. The two methods are compared and contrasted for common monaural measures used in performance space design. The MLE method is best for estimating reverberation time. For center time, clarity, and EDT, neither method is better. The indications are that when geometric room acoustic models become more accurate, enabling better quality training data for the machine learning algorithm, then the envelope spectrum method will be most successful for parameters needing accurate estimation of early decays. Both speech and music have a role in enabling naturalistic measurement. Speech is a better signal at mid- and high frequencies because it naturally has more pauses and has more predictable low frequency statistics. However, music is needed to gain parameters at lower frequency bands, where speech has insufficient power for excitation.

ACKNOWLEDGMENTS

Thanks go to Henrik Möller for supplying the real room impulse responses. The authors would like to acknowledge the support of the Engineering and Physical Sciences Research Council, UK for funding this project.

[1]F. F. Li and T. J. Cox, "Speech transmission index from running speech: A neural network approach," J. Acoust. Soc. Am. **113**, 1999–2008 (2003).
[2]F. F. Li and T. J. Cox, "A neural network for blind identification of speech transmission index," Proc. IEEE **2**, 757–760 (2003).
[3]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. Acoust. Soc. Am. **114**, 2877–2892 (2003).
[4]P. Kendrick, F. F. Li, T. J. Cox, Y. Zhang, and J. A. Chambers, "Blind Estimation of Reverberation Parameters for Non-Diffuse Rooms," Acta. Acust. Acust. **93**, 760–770 (2007).
[5]P. Kendrick, T. J. Cox, F. F. Li, Y. Zhang, and J. A. Chambers, "Blind estimation of clarity, centre time and deutlichkeit from speech and music signals," Proceedings of the 19th ICA, Madrid (2007), Paper No. RBA-07-013.
[6]International Standard, "Acoustic—Measurement of the reverberation time of rooms with reference to other acoustical parameters," Report No. EN ISO 3382:2000.
[7]H. P. Seraphim, "Investigations on the difference limen of exponentially decaying bandlimited noise pulses," Acustica **8**, 280–284 (1958).
[8]L. Cremer and H. A. Müller, *Principles and Applications of Room Acoustics*, translated by T. Schultz (Applied Science, London/New York, 1982), Vol. **1**.
[9]T. J. Cox, W. J. Davies, and Y. W. Lam, "The sensitivity of listeners to early sound field changes in auditoria," Acustica **79**, 27–41 (1993).
[10]CATT-ACOUSTIC v8.0c, room acoustic modeling software, http://www.catt.se/ Last viewed 21/04/2008
[11]T. Lahti, A. Ruusuvuori, and H. Moller, "The acoustic conditions in Finnish concert spaces-Preliminary results," Proceedings of the Audio Engineering Society 110th convention (2001), p. 5356.
[12]Denon Anechoic Orchestral Music Recording, CD PG-6006, 1995.
[13]Music for Archimedes. CD B&O 101, 1992.
[14]M. T. Hagan and M. Menhaj, "Training feedforward networks with the Marquardt algorithm," IEEE Trans. Neural Netw. **5**, 989–993 (1994).
[15]T. Houtgast and H. J. M. Steeneken, "Envelope spectrum and intelligibility of speech in enclosures," Proceedings of the IEEE-AFCRL Speech Conference (1972), pp. 392–95.
[16]H. Kuttuff Kuttruff, *Room Acoustics*, (Spon, London, 2000).
[17]P. C. Mahalanobis, "On the generalized distance in statistics," Proc. Nat. Inst. Sri. India A **2**, 49–55 (1936).
[18]J. Aldrich, "R.A. Fisher and the making of maximum likelihood 1912-1922," Stat. Sci. **12**, 162–176 (1997).