14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

AN EVALUATION MEASURE FOR REVERBERANT SPEECH USING DECAY TAIL MODELLING

Jimi Y.C. Wen, and Patrick A. Naylor

Department of Electrical and Electronic Engineering, Imperial College London
South Kensington Campus, SW7 2AZ, UK
email: {yung-chuan.wen,p.naylor}@imperial.ac.uk

ABSTRACT

An objective measure for the perceived effect of reverberation is an essential tool for research into dereverberation algorithms or acoustic space modeling. There are two different effects that contribute to the total perceived reverberation, colouration and the reverberation tail effect. This paper presents an objective reverberation decay tail measure R_{DT}, and measures perceptually the reverberation tail effect directly from speech. R_{DT} uses the perceptually weighted Bark Spectral Distortion (BSD) in conjunction with an end-point search algorithm and decay tail modelling. The measure was evaluated against T_{60} and BSD for colourless and constant colouration reverberation impulse responses.

1. INTRODUCTION

Reverberation is caused by the multi-path propagation of acoustic signals from source to microphones. Reverberant speech can be described as sounding distant with noticeable echo and colouration and these effects generally increase with increasing distance from source to microphone for a given reverberant room. Reverberation also usually increases with increasing room size for a given distance from source to microphone. Reverberation has a neglible effect in telephony applications with traditional handsets. However, in hands-free systems, reverberation affects the quality and intelligibility of speech and is a significant problem for telecommunications, speech recognition applications [1], and hearing aids.

Dereverberation is the process of forming an estimate of the original source from one or more observations of the reverberant signal. Several dereverberation algorithms have been proposed and can be considered in two categories: (i) speech enhancement processing and (ii) blind channel estimation/inversion algorithms [1]. There is a need to have a reliable objective measure for the perceived reverberation, which can allow immediate and reliable estimation of the perceptual significance of a particular algorithm's processing. Reliable quantitative measurement of the level of reverberation in a speech signal is particularly difficult and a unanimously accepted methodology has yet to emerge [1]. It has been observed that most of the current objective speech quality measures which give good prediction of overall speech quality do not give good prediction for the perceived reverberation.

The Bark Spectral Distortion (BSD) [2] is an often used objective measure for speech quality. Assuming the only distortion present is reverberation, the BSD measures the reverberation as a perceptually weighted spectral difference of the original and reverberant signal. This difference consists of two parts, the colouration effect, and the transient reverberation decay tail effect. The colouration effect is due to the early room reflections [3], and the reverberation tail effect is due to the (late) decay tail of the acoustic room impulse response. The early room reflection is usually defined as the impulse response from the direct path to the next 50 ms to 80 ms, and the late reverberation or decay tail is defined as the remaining later taps of the impulse response.

The aim of this paper is to develop an objective measure called R_{DT} of the perceived effect of the reverberant decay tail from the speech signals, without the need for knowledge or estimation of the impulse response. We note that the currently accepted objective measures of the reverberation decay tail effect such as reverberation time (T_{60}), Clarity index (C_{50}) [4] and time centre of gravity (TS), are not independent of the colouration. The dependency is worse for input-output type of measures, such as BSD, Cepstral Distance (CD), and Log Likelihood Ratio (LLR) [5]. Our approach is to consider colouration and the reverberation decay tail effect separately since their perceptual effects are very different. In this paper our focus is on the measurement of the reverberation tail effect whilst the measurement of the colouration effect is the subject of our future work. The measure should be independent of the colouration and of the the signal. Previous work on estimation of reverberation time [6], [7] and [8] is related to our work but does not consider perceptual significance nor separation of colouration and decay tail. Whereas T_{60} is a measure of the overall impulse response, the aim of R_{DT} is to characterize specifically the late reverberation. In the special case when the colouration due to the early reflection is insignificant, then T_{60} and R_{DT} will be equivalent measures.

The outline of the paper is as follows. In Section 2 we introduce parametric decay tail modeling, and formulate the perceived tail measure R_{DT}. Section 3 describes a simple end-point search algorithm and constrained averaging to estimate decay tail model parameters from real speech data. The results of implementation are discussed in Section 4, and we conclude in the Section 5.

2. PARAMETRIC DECAY TAIL MODELING

Bark Spectral Distortion is an often used measure of speech quality, as it takes into account auditory frequency warping, critical-band integration, amplitude sensitivity variations with frequency and subjective loudness [2]. However BSD does not measure specifically the two attributes of reverberation, colouration and decay tail. Colouration is a frequency distortion which is due to the stronger early reflections. Colouration could cause a sound to be 'boxy', 'thin', 'bright', and so on. The reverberation decay tail causes a 'distant' and 'echo-ey' sound quality. Current objective measures which operate on impulse responses such as the T_{60}, C_{50}, TS, and on input-output signals such BSD, CD, LLR measure a combination of the two effects discussed above. A measure of the latter kind which only requires the reverberant and reference signals would be more practical because the room impulse response (or how this response is modified by a dereverberation algorithm) may not be available or may be difficult to estimate. We propose a parametric decay tail modeling, similar to those of [6] and [9], which has the advantage of estimating the decay tail without needing to estimate the full impulse response and measuring the reverberation more independently of the colouration effect.

2.1 Decay tail model

Habets [9] employed a simplified time-domain model for a Room Impulse Response (RIR)

$$h(t) = b(t)Ae^{-\lambda t}, \qquad (1)$$

where $b(t)$ is white zero-mean Gaussian stationary noise, A is a scaling factor and λ is the rate of decay which is linked to the reverberation time. In this paper we aim to estimate the decay tail model parameters from Bark Spectral (BS) difference of the reverberant and reference speech in each BS bin. The unit for the Bark Spectral difference in each BS bin is perceptual loudness in sones. We choose to estimate the decay tail parameters from the BS difference so as to eliminate from our measurement any decay tail in the reference speech, due perhaps to natural vocal tract resonances. We assume that the two signals can be time aligned and can be normalized in energy in the BS domain.

In this Section we define end-points as instant of time at which the speech energy falls abruptly and we define flat-regions as periods of time immediately following an end-point during which there are no significant increases in energy due to speech onset. We assume first that we can find end-points in the energy of each BS bin. Examples of such signal end-points can easily be found in simple signals such as an impulse (Fig. 1(a)) and a finite duration WGN sequence (Fig. 1(b)). Decay responses obtained after reverberation are shown in Fig. 1(c) and (d) for inputs (a) and (b) respectively. Fig. 1(e)-(j) shows the BS difference decay curves for three arbitrary BS bins. In line with model of (1), we model these decay curves starting from frame n for BS bin k using

$$d(n,k) = A_k e^{-\lambda_k n}, \qquad n = 1,2,...,I+J \qquad (2)$$

The A_k and λ_k are obtained using sum of squares error minimization curve fitting (Fig. 2(c)) to the decay curve from the BS difference as shown in Fig. 2(a). Here, n starts from 1 instead of 0 because we estimate our tail decay from the BS *difference*, and the peaks at the origin approximately cancel. The term corresponding to $n = 0$ is defined to be the direct path frame which is estimated from the end-point in the clean reference signal (Fig. 2(b)). The results in Fig. 1 employ $(I+J = 10)$. In real speech the decay tail following an end-point may not have enough time to decay before the signal energy increases again due to a new utterance starting. This problem will be addressed in Section 3.

Using Taylor expansion we can rewrite the first two terms of our decay tail model for k-th BS bin as

$$A_k e^{-\lambda_k n} \cong A_k - A_k \lambda_k n. \qquad (3)$$

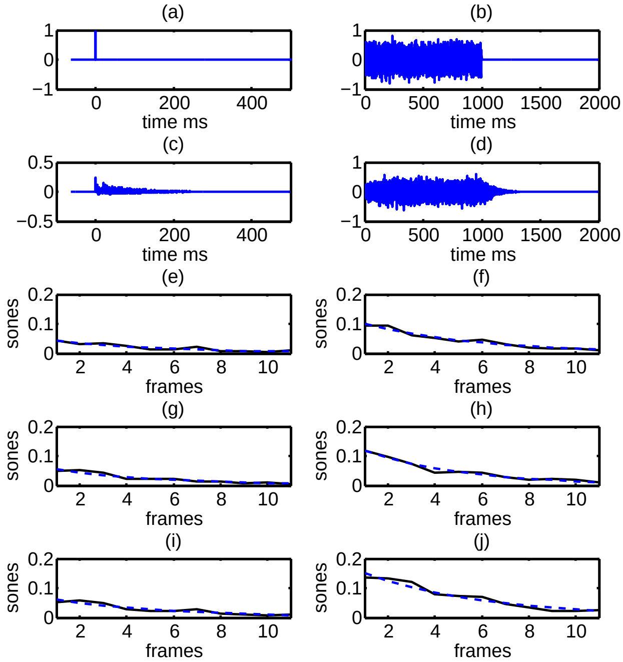

Figure 1: (a)impulse, (b)white noise, (c) impulse response, (d) reverberant noise, (e)(g)(i) fitted decay tail curves for impulse response for different BS bins (50,100,150), (f)(h)(j) fitted decay tail curves for reverberant noise for different BS bins (50,100,150). Dash lines indicate the fitted curve and solid lines indicate the actual values.

The decay tail model parameters are computed for all end-points as an average across all BS bins in the form of

$$A_{avg} e^{-\lambda_{avg} n} \cong A_{avg} - A_{avg} \lambda_{avg} n. \qquad (4)$$

Hence, the model parameters can be rewritten

$$A_{avg} = \frac{\sum_{k=1}^{M} A_k}{M} \qquad (5)$$

$$\lambda_{avg} = \frac{\sum_{k=1}^{M} A_k \lambda_k}{M A_{avg}}. \qquad (6)$$

2.2 Formulation of the R_{DT} Measure

For a particular end-point in M BS bins the R_{DT} includes the following three terms; A_{avg} represents the average absolute decay tail energy for each BS bin and is obtained from (5), λ_{avg} represents the average decay tail rate for each BS bin and is obtained from (6), D_{avg} represents the average direct path energy for each BS bin and is obtained from

$$D_{avg} = \frac{\sum_{k=1}^{M} D_k}{M}, \qquad (7)$$

where D_k is the direct path energy estimate in the k-th BS bin. The direct path term D_{avg} is estimated from the BS of the clean reference signal. R_{DT} is defined as the ratio of the amplitude and decay rate of the exponential decays normalized to the amplitude of the direct component and is written

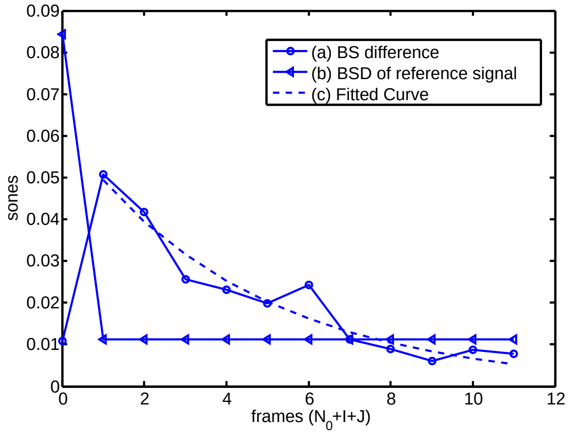

Figure 2: (a) Reverberation decay tail estimation from the BS difference, (b) the direct path estimate for a particular decay tail estimate from the BS of the reference signals, (c) Fitted curve using sum squares error minimization

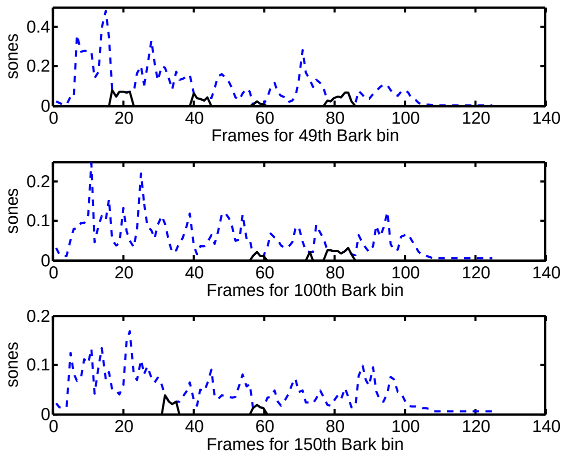

Figure 3: Results of end-points/flat-regions search algorithm. Dash lines indicates loudness level, and solid lines indicates detected flat-regions.

$$R_{DT} = \frac{A_{avg}}{\lambda_{avg} D_{avg}}. \tag{8}$$

This measure jointly characterizes the relative energy in the decay tail and the rate of decay. We note that other averaging schemes can be considered.

3. IMPLEMENTATION

We propose a search algorithm to find end-points in each BS bin of real speech. Data from multiple end-points in any given BS bin can be combined to improve the accuracy of the measure. Averaging (5)(6)(7) is performed across time-frames for any given bin and then across all BS bins to obtain a single value R_{DT}.

We define a function $\Delta\chi(n,l,k)$

$$\Delta\chi(n,l,k) = \chi(n,k) - \chi(n+l,k), \tag{9}$$

where χ is the BS of the speech signal, n is the time frame index, k is the BS bin index and l is a positive integer. We used a frame size of 32 ms. The instant of an end-point, n_0, is defined as a sample for which

$$\Delta\chi(n_0,1,k) > \delta_{max1} \tag{10}$$

is satisfied. For each end-point we measure I the number of frames for which

$$\Delta\chi(n_0+i,1,k) > \delta_{max2} \quad i = 0,1...,I \tag{11}$$

is satisfied. Flat-regions for which the BS ripple is within δ_{min} and below a threshold δ_t over J frames are found by searching

$$\begin{aligned} &\Delta\chi(n_0+i,j,k) < \delta_{min} \\ &\chi(n_0+i+j,k) < \delta_t. \quad j = 0,1...,J \end{aligned} \tag{12}$$

In our implementation, we restrict end-points/flat-regions to those which satisfy $J > 4$, and we do not include decays that are too short to obtain an R_{DT} measurement. In Fig. 3 we show the end-points/flat-regions result from our search algorithm (solid lines). We then fit the decay model (2) the with length $(I_{k,p} + J_{k,p})$ from the Bark Spectral difference of the reverberant and reference signals in the regions found from the search algorithm, where $(I_{k,p} + J_{k,p})$ is the length of the p-th end-points/flat-regions estimated in the k-th BS bin. Values of the δs were chosen experimentally as a percentage of the maximum loudness in each BS bins, with $(\delta_{max1} = 0.2)$, $(\delta_{max2} = 0.1)$, $(\delta_{min} = 0.1)$ and $(\delta_t = 0.2)$.

From the end-points we obtain sets of $A_{k,p}$, $\lambda_{k,p}$ and $D_{k,p}$. We ignore any outliers in these sets according to $0 < A \le A_{max}$ and $0 < \lambda \le \lambda_{max}$. Good values were found experimental as $A_{max} = 1$ $\lambda_{max} = 100$. Subsequently averaging is first performed over p and then k to obtain A_{avg}, λ_{avg} and the corresponding D_{avg} from which R_{DT} is obtained using (8).

4. RESULTS

To evaluate our measure, we set our impulse response using white noise enveloped by different rate of decay slopes, then compared our measure against T_{60} calculated from the generated impulse responses using Schroeder's method in [10]. Using white noise to generate the impulse response eliminates any colouration effect measured by the T_{60}, allowing the T_{60} to measure only the reverberation tail effect for the case of colourless reverberation.

Figure 4 compares the R_{DT} measure obtained from the speech signals with T_{60}s obtained from the room impulse response. Two speech test signals have been employed in addition to an impulse and noise stimulus. Room responses with T_{60} values ranging from 200 ms to 1800 ms have been generated using white noise with different decay slopes. The results show a strong correlation between R_{DT} and T_{60} under colourless reverberation in all cases, which means that the R_{DT} is a good indicator of the perceived reverberation decay tail effect.

Figure 5 shows the results of R_{DT} against T_{60} for two speech signals convolved with impulse responses of vary-

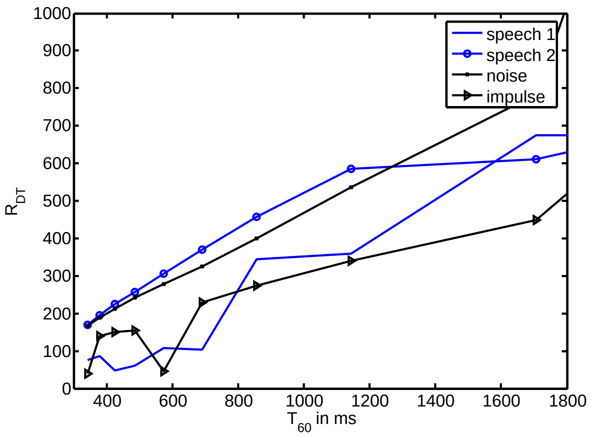

Figure 4: Comparison of R_{DT} against T_{60} of a sets of impulse/noise response and reverberant speech 1 and speech 2,

ing length. Two cases are studied: the case of colourless reverberation, and the case in which the colouration is non-white, constant and independent of impulse response length. Constant colouration is generated using Allen and Berkley's method of images [11][12] with different T_{60} but of constant colouration caused by the strong distinct early reflection. This is achieved by fixing the room dimension but varying the reflectivity of the walls. In this way only the amplitude of the reflections varies which will have negligible perceptual effect on the spectrum of the impulse response. Figure 6 shows the BSD against the T_{60} for the same speech data. By comparing Fig. 5 and 6, it can be seen that the R_{DT} is less dependent on the signal than BSD. The correlation of R_{DT} and BSD is 0.92 and 0.87 respectively for the colourless case across all speech sets, while for the constant colouration case is 0.96 to 0.75. We can see that for the constant colouration case, there is a high loss in the performance of the BSD to measure the reverberation decay tail effect due to the interaction between the room colouration and the signal's timbre, while the performance of the R_{DT} is not negatively affected by colouration and signal.

Figures 7 and 8 compare the R_{DT} measure against the BSD for different colouration of the same reverberation time ($T_{60} = 1\ s$). The colouration is different for each test shown on the horizontal axis, therefore T_{60} would be expected to be influenced by the colouration. It can be seen that the R_{DT} matches better than BSD across the different speech of the same reverberation system, therefore showing the independence of R_{DT} to different signal timbre's interaction with different colouration. It can also be seen that the R_{DT} is less dependent on the room's colouration in general. The mean and standard deviation of the R_{DT} calculated over the 20 points of Fig. 7 is 539.8 and 95.7, which is equivalent to a coefficient of variation of 17.7% due to different colouration for a given T_{60}. This is much less than the values calculated from Fig. 8 with mean 0.337, standard deviation 0.112 and coefficient of variation 33.3% for BSD.

5. CONCLUSION

A measure for the perceived reverberation decay tail effect R_{DT} was presented in this paper. This measure is based on an exponential decay of the Bark Spectrum of the tail, averaged over all Bark Spectral bins and all speech end-points.

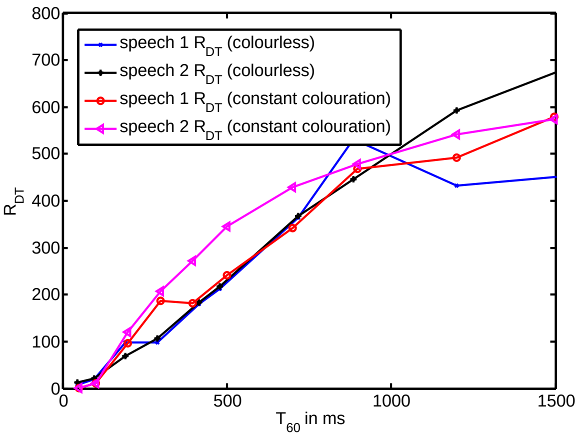

Figure 5: Comparison of R_{DT} against T_{60} on a set of colourless and constant colouration reverberant speech.

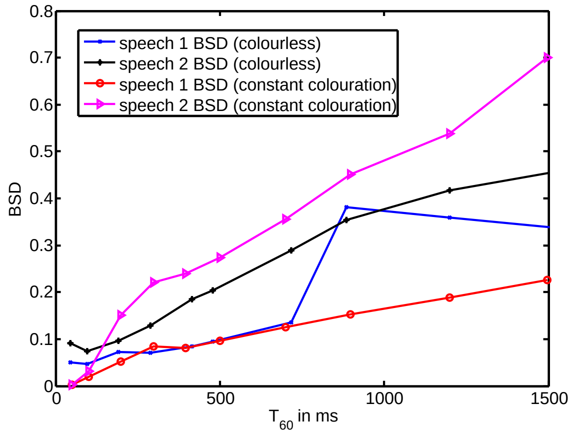

Figure 6: Comparison of BSD against T_{60} on a set of colourless and constant colouration reverberant speech.

An implementation of an end-points/flat-regions detection algorithm and constrained averaging process for application to real speech has been presented. Results comparing R_{DT} against T_{60} were better than the original BSD measure, in the sense of independence of different perceptual effects. The aim of our measure is not to estimate the T_{60}, rather the perceptual characteristics similar to those of measured by T_{60} but independent of any colouration in the reverberation impulse response. The advantage of R_{DT} over measures like BSD and CD is the ability to separate the specific perception of the reverberation decay tail effect from colouration. In addition the R_{DT} measure does not require the impulse response of the reverberation system.

Future work will be to develop a measure which just measures the colouration of the reverberant speech independently to the reverberant tail effect. In conjunction with the R_{DT} measure discussed in this paper, it would then be possible to characterize and quantify the perceived colouration and reverberation tail effect separately for an acoustic space.

REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2005.

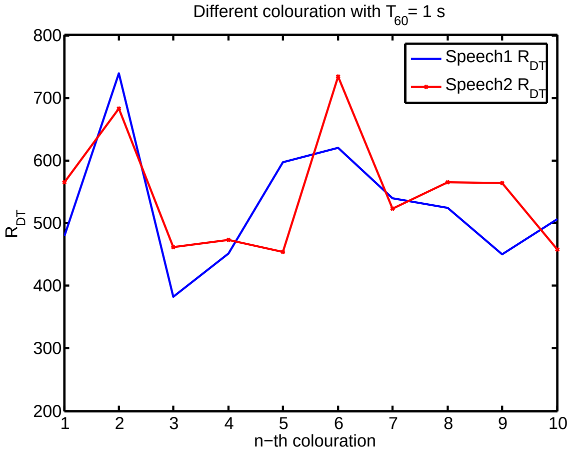

Figure 7: Comparison of R_{DT} against different colouration but of T_{60}=1 s on 2 sets of different colouration reverberant speech.

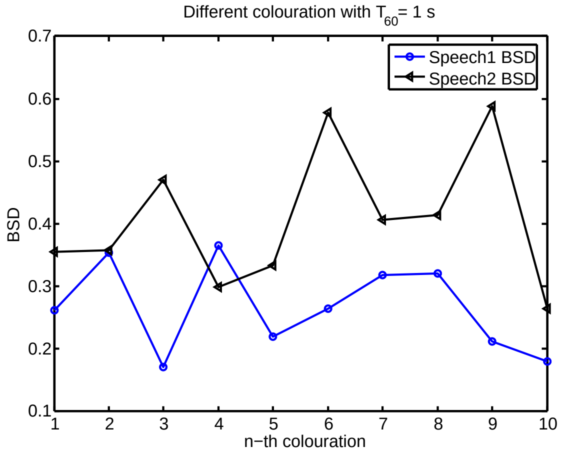

Figure 8: Comparison of BSD against different colouration but of T_{60}=1 s on 2 sets of different colouration reverberant speech.

[2] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Journal on Selected Areas in Communications,*, vol. 10, no. 5, pp. 819 – 829, 1992.

[3] M. R. Schroeder and B. F. Logan, "Colorless artificial reverberation," *IRE Transactions on Audio*, no. 6, pp. 209 – 214, 1961.

[4] H. Kuttruff, *Room Acoustics*, 4th ed. Taylor & Francis, Oct. 2000.

[5] A. J. Gray and J. Markel, "Distance measures for speech processing," *IEEE Trans. Speech and Audio Process.*, vol. 24, no. 5, pp. 381–390, Oct. 1976.

[6] R. Ratnam, D. L. Jones, B. C. Wheeler, J. William D. O'Brien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *J. Acoust. Soc. Amer.*, vol. 114, no. 5, pp. 2877–2892, 2003.

[7] T. J. Cox, F. Li, and P. Darlington, "Extracting room reverberation time from speech using artificial neural networks," *J. Audio Eng. Soc.*, vol. 49, no. 4, pp. 219–230, 2001.

[8] S. Vesa and A. Harma, "Automatic estimation of reverberation time from binaural signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, 2005, pp. 281 – 284.

[9] E. A. P. Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, 2005, pp. 173–176.

[10] M. R. Schroeder, "A new method of measuring reverberation time," *J. Acoust. Soc. Amer.*, vol. 37, pp. 409–412, 1965.

[11] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[12] P. M. Petereson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Amer.*, vol. 80, no. 5, pp. 1527–1529, 1986.