AN EXTENDED REVERBERATION DECAY TAIL METRIC AS A MEASURE OF PERCEIVED LATE REVERBERATION

Hamza A. Javed and Patrick A. Naylor

Dept. of Electrical and Electronic Engineering, Imperial College London, UK

ABSTRACT

In this paper the development and evaluation of an extended Reverberation Decay Tail (R$_{DT}$) metric is described. The signal-based metric predicts the perceived impact of reverberation on speech, by identifying and characterising energy decay characteristics in the signal Bark spectrum. In comparison with a previous metric, the new metric is extended to operate on wideband speech and incorporates an improved perceptual model and decay curve detection scheme. Furthermore, contributions of this work include experimental testing and validation of the metric on reverberant speech. The tests conducted show positive correlation with objective measures such as C$_{50}$ as well as with subjective listening test scores. Potential applications of the measure include use as an evaluation tool for dereverberation research.

***Index Terms*—** R$_{DT}$, late reverberation, perceptual reverberation, speech quality measure.

1. INTRODUCTION

In many real-world settings, speech signals captured by a microphone frequently suffer from the effects of reverberation. Broadly defined as the multipath propagation of sound in an enclosure, reverberation is known to degrade the overall quality, and intelligibility, of captured speech signals [1]. This degradation subsequently adversely affects the performance of a wide range of speech processing technologies, from hearing aids to speech recognition systems [2][3]. With a growing number of devices accepting speech input from distant microphones up to and beyond the critical distance [4], the development of dereverberation algorithms is currently an important and practically well motivated task.

However, despite the increased interest in dereverberation research over the last few years, objective measures capable of estimating perceived levels of reverberation have received far less attention. Whilst numerous speech and audio quality metrics exist, even popular measures such as the Bark Spectral Distortion (BSD) [5], Perceptual Evaluation of Speech Quality (PESQ) [6] and Perceptual Objective Listening Quality Assessment (POLQA) [7], are limited to giving indications correlated with overall speech quality. In many situations, such measures, originally designed to evaluate the quality of network transmitted speech, have been shown to be poor to average predictors of perceived reverberation distortion [8][9].

A reverberation specific quality measure would be an important development, as it would allow the effectiveness of dereverberation algorithms to be tested quickly and consistently, avoiding the need for time-consuming and labor intensive listening tests. Such a measure would therefore be a valuable tool for dereverberation research, whilst also being able to provide useful acoustic information about the enclosure in which the speech signals propagate.

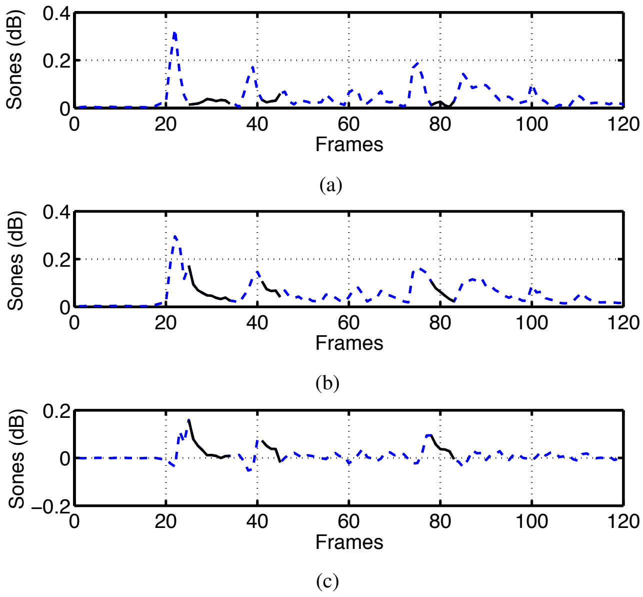
The aim of this work is to develop such a reverberation measure, by extending the Reverberation Decay Tail (R$_{DT}$) metric proposed in [10]. Novel contributions include extension of the measure to wideband operation, the incorporation of an improved perceptual model, and an improved decay curve detection scheme. Furthermore, an evaluation of the new metric has also been conducted on large speech sets. Objective evaluation as well as listening tests have been performed to assess and validate the measure's accuracy as a predictor of the perceived level of reverberation.

Having provided the motivation and context for this work, the remainder of this paper is organised as follows. In Section 2, the R$_{DT}$ measure is reviewed, whilst proposed extensions and improvements in the measure are presented in Section 3. The experiments conducted to test the extended measure are then described in Section 4, with key results highlighted. Finally, conclusions are drawn in Section 5.

2. REVIEW

The effects of reverberation on an audio signal can be classified into two perceptually distinct phenomena [1], arising from the structure of a room impulse response (RIR). The first effect, known as colouration, is the result of early reflections corresponding to reverberation effects of the acoustic environment approximately 50 - 80 ms after the direct path arrival at the microphone. Due to the small time difference of arrival between the direct sound and early reflections, the

The authors would like to thank Prof. Dr.-Ing Jürgen Herre of International Audio Labs, Erlangen, for providing experimentally assessed listening test data for this work.



(a)

(b)

(c)

Fig. 1: Identified flat regions highlighted in the energy envelopes for (a) the clean speech signal, (b) the reverberant signal and (c) the Bark spectral difference. Dashed and solid lines indicate loudness and detected flat regions, in the 29th Bark bin.

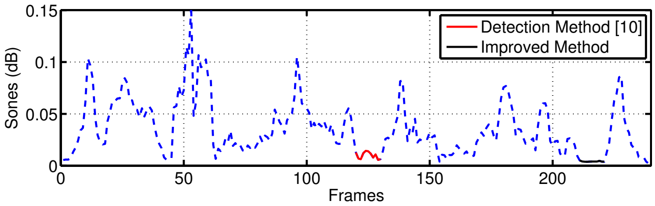


Fig. 2: Identified flat regions highlighted in the energy envelope of a clean signal, using the method of [10] and the improved scheme proposed here. Dashed lines indicate loudness, solid line detected flat region, in the 50th Bark bin.

human auditory system does not normally resolve these signal components individually, but instead perceives them as a single sound. Early reflections typically modify the spectral characteristics of signals, in such a way that they are often described using subjective terms such as ‘boomy’, ‘boxy’ and so on.

Late reverberation, or the decay tail effect, is the second perceived effect of reverberation. It is attributed to the effects of mutipath room acoustics propagation, arriving at the microphone later than the early reflections. Although weaker in amplitude compared to early reflections, these later effects are represented by noise-like and more densely packed RIR coefficients, and are perceptually significant, causing a captured audio signal to ring-on. In contrast to early reflections, which have been shown to have potential to increase speech intelligibility [11], late reverberation is generally responsible for reducing overall quality and intelligibility of speech, particularly in combination with noise [12][13].

The R_{DT} measure, first introduced in [10], aims to predict the perceived level of late reverberation in speech. It operates intrusively using as inputs the clean reference signal and the test reverberant signal. Working under the premise that a room impulse response can be modelled as exponentially decaying, zero-mean, white Gaussian noise [14], this is achieved by identifying and characterising energy decays introduced by reverberation, in the signal Bark spectra. Computing the measure involves three main steps, namely (i) Bark spectrum calculation; (ii) end-point and flat region detection; and (iii) parametric decay curve fitting.

The first step in the computation of R_{DT} requires mapping the clean and reverberant speech signals to the Bark spectrum domain, frame by frame [10]. The Bark domain models the peripheral auditory system by generating an indication of perceived loudness from a weighted frequency spectral analysis. It is therefore well suited to assessing the perceptual impact of late reverberation.

In the R_{DT} implementation of [10], the Bark spectra of a clean signal and its reverberant counterpart is calculated using the approach described in [5], in which several key processing steps carried out by the human ear are emulated in the spectra computation. By having both signal spectra available, the perceived impact of late reverberation can be assessed by detecting regions where such distortions would be most audible across time frames and Bark bins, and quantifying such distortions in these regions.

The identification is performed on the clean Bark spectra, where abrupt decreases in energy followed by plateaus several frames long, are detected. These are known as end-points and flat regions respectively. They are identified in each bin, and are determined from the energy envelope of the reverberant signal by gradient and absolute thresholds. These thresholds are defined as percentage values of the global peak in the energy envelope of that bin. Figure 1a illustrates an example of flat region detection in an arbitrary Bark bin of an example clean speech signal.

In the time aligned and energy normalised reverberant Bark spectra, these corresponding regions exhibit the effect of reverberation; the spreading of energy across frames, as shown in Fig. 1b. The final step in the calculation of R_{DT} involves characterising the energy decays in these regions in the Bark spectral difference of the reverberant and clean signal, shown in Fig. 1c. The difference is used rather than the reverberant spectrum, so as to discount any natural decays present in the clean speech signal.

The identified energy decays in each Bark bin difference are modelled as negative exponentials, that is $Ae^{-\lambda n}$. Using a sum of squares error minimisation criterion, curves are fitted to each decay, and parameters of the exponential model esti-

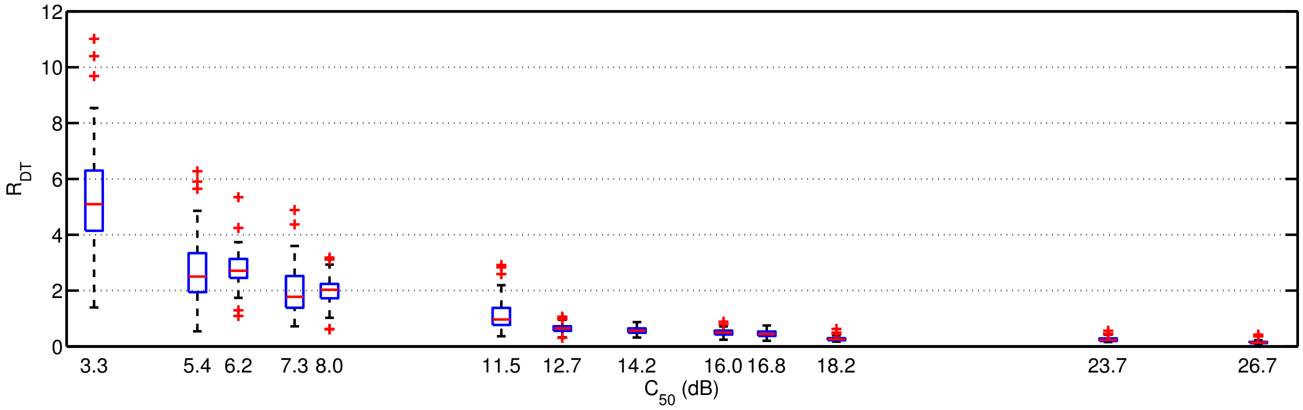


Fig. 3: Boxplots illustrating R$_{DT}$ value distribution for 100 TIMIT speech signals that have undergone varying levels of reverberation distortion (convolution with measured room impulse responses).

mated. The reverberant energies, represented by A, are then averaged by the number of detected curves, M, in each Bark bin, k_b, expressed mathematically as

$$\bar{A}_{k_b} = \frac{\sum_{p=1}^{M} A_{k_b}(p)}{M} \,. \tag{1}$$

A weighted average of the reverberation rate of decay is also computed in each Bark bin, $\bar{\lambda}_{k_b}$, expressed as

$$\bar{\lambda}_{k_b} = \frac{\sum_{p=1}^{M} A_{k_b}(p)\lambda_{k_b}(p)}{\bar{A}_{k_b} M} \,. \tag{2}$$

In addition to estimating the aforementioned exponential parameters, the energy of the clean speech signal preceding each flat region, is also recorded and averaged. The average direct path energy, $\bar{D}_{k_b}$, is written

$$\bar{D}_{k_b} = \frac{\sum_{p=1}^{M} D_{k_b}(p)}{M} \,. \tag{3}$$

Each parameter is then averaged across the total number of Bark bins, K_b. By characterising the identified reverberant energy decays with the above parameters, the R$_{DT}$ measure is defined as the ratio of the average reverberant energy to the average rate of decay, normalised to the average direct path component such that

$$R_{DT} = \frac{K_b \sum_{k_b=1}^{K_b} \bar{A}_{k_b}}{\sum_{k_b=1}^{K_b} \bar{\lambda}_{k_b} \sum_{k_b=1}^{K_b} \bar{D}_{k_b}} \,. \tag{4}$$

The measure indicates that for large average reverberation tail energies, relative to the direct sound energy, the perceived level of reverberation will be greater. Similarly, the effects of reverberation will be more noticeable for slower average decay rates.

3. EXTENDED R$_{DT}$

The R$_{DT}$ algorithm proposed in [10] employed a basic perceptual model in the calculation of a Bark spectrum, and was restricted to operating on narrowband speech (sampled at 8kHz). In this work the measure has been extended to wideband operation (24kHz), incorporates a more realistic perceptual model and, in addition, includes now an improved decay curve detection scheme.

3.1. Perceptual Model

In order to develop an R$_{DT}$ algorithm capable of operating on wideband speech, an increased number of critical band filters, centred at higher frequencies, were employed. In comparison to the implementation of [10], a trade-off between the Bark spectra resolution (determined by the number of critical band filters used), computational costs and overall performance was studied experimentally and a design was chosen using critical band filters positioned every 0.25 Bark.

Additionally equal loudness weightings, as specified in the revised ISO 226 standard [15], were incorporated into the Bark spectrum calculation used by the extended R$_{DT}$. In doing so, the auditory system's sensitivity to different frequencies could be modelled more accurately than with the high pass filter used in [10].

3.2. Decay Curve Detection

As described in Section 2, the approach in [10] determines end-points and flat regions by gradient and absolute thresholds, which are in turn defined as fractions of a global peak in each Bark bin energy envelope. Such a detection scheme is susceptible to errors. Most notably, a spurious spike of energy due perhaps to impulsive noise, in even one segment of the captured audio will result in errors in flat region detection. Moreover, even in the absence of erroneous spikes, the scheme of [10] often misses valid flat regions whilst also occasionally causing false alarms in the flat region detection.

To overcome these limitations, the detection scheme proposed in this work uses local rather than global peaks to define the gradient and absolute thresholds. This is achieved by working on overlapping signal blocks, rather than the entire signal. A block size of duration 3s was found experimentally to be effective, after a range from 0.5s to 5s was tested. Using this approach more accurate flat region detection is achieved. An example case is illustrated in Fig. 2 in which the method of [10] recognises the frames shown in red as a flat region despite the significant ripple present, whilst the improved detection scheme does not. On the other hand, the improved scheme identifies a valid flat region missed by [10], because the energy drop preceding it was not seen as sufficiently large.

4. SIMULATIONS AND RESULTS

Two main experiments were conducted to test the effectiveness of the extended R$_{DT}$ measure. The first experiment assessed the performance of the metric against the early to late reverberation ratio, or clarity index, C$_{50}$. Whilst the second was a listening test. The setups and results obtained from these experiments are described in more detail below.

4.1. Objective Test

In order to assess the effectiveness of the extended R$_{DT}$, the measure was tested on a range of simulated reverberant speech. The speech signals used were from the TIMIT database [16]. Three sentences spoken by the same speaker were concatenated with a 1 second inter-sentence pause, to produce audio signals approximately 10-15 seconds long. Both male and female speakers across all dialect regions were used, and a speech set consisting of 100 signals was employed for testing purposes. Reverberant speech was simulated by convolving the TIMIT signals with measured room impulse responses, from the Aachen Impulse Response database [17]. The database contains impulse responses from a variety of acoustic environments, ranging from studio booths to cathedral halls, thus allowing a wide range of reverberation conditions to be simulated.

The results of the R$_{DT}$ metric on a reverberant speech set, as described experimentally above, is shown in Fig. 3. The figure illustrates the relationship between R$_{DT}$ and C$_{50}$. From inspection, an exponential function could be used to map between the two measures. Whilst the mean R$_{DT}$ score increases with decreasing C$_{50}$, the boxplots indicate that the variance does so as well. This becomes increasingly significant for small clarity index values. However, for a given speech signal that is convolved with room impulse responses of decreasing C$_{50}$, the R$_{DT}$ tends to decrease monotonically.

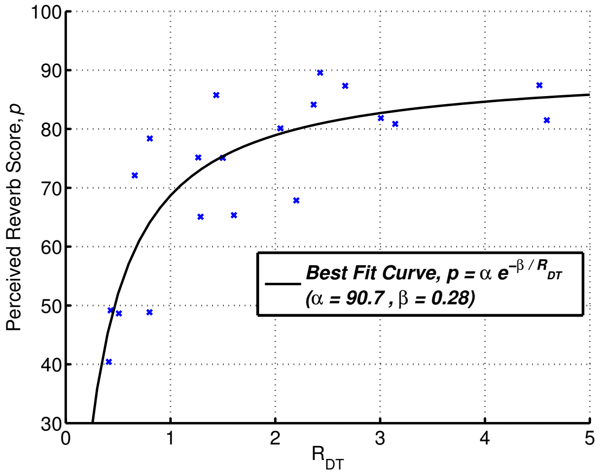


Fig. 4: Scatter plot of mean listening test scores for different reverberation conditions, against R$_{DT}$ scores. Fitted curve, p, modelled as $p = \alpha\, e^{(-\beta/R_{DT})}$.

4.2. Listening Test

By comparing the performance of R$_{DT}$ on reverberant speech, subjectively rated by listeners, we can determine how strongly correlated it is with human perception of reverberation. To this end listening test assessed speech data was used from [18]. In the subjective tests conducted speech signals, with varying levels of applied reverberation, were presented to sets of 12-14 listening subjects. These listeners were then asked to score the perceived level of reverberation out of 100, where a higher score indicates increasing reverberation.

Analysing the results of the listening tests showed that despite being provided with anchor signals, there is high variability in how humans score perceived reverberation. In our work we investigated the relationship between average listening scores for a given reverberant signal and the corresponding objective R$_{DT}$ score.

The results of this analysis is shown in Fig. 4. The relationship between listening test scores, p, and R$_{DT}$ was modelled parametrically as $p = \alpha\, e^{(-\beta/R_{DT})}$, giving a Pearson correlation coefficient of 0.84. Motivation for this particular model was provided by the negative exponential relationship observed between R$_{DT}$ and C$_{50}$. Parameters α and β were determined by fitting a curve through the data using a minimum sum of squares error criterion.

The produced curve allows R$_{DT}$ values to be mapped to an equivalent subjective listening score, as defined in [18], with a residual square error of 7.8 points. By comparison, the approach in [10] tested in the same way showed greater variability and spread for the same mathematical model, giving a Pearson correlation coefficient of 0.67 and a residual square error of 11.2 points.

5. CONCLUSIONS

A signal-based, extended R_{DT} metric incorporating an improved perceptual model and detection scheme was proposed and experimentally tested in this work. The metric predicts perceived reverberation by characterising reverberation decays in the perceptual Bark spectrum of a signal. Results show the extended metric correlates well with objective measures such as the clarity index, as well as with subjective listening test scores. In comparison with a previous metric [10], the extended R_{DT} is more positively correlated with subjective tests. A mathematical model that maps R_{DT} values onto subjective reverberation scores was also developed, making the metric a useful evaluation tool for dereverberation research.

6. REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, Springer, 2010.

[2] D. A. Berkley, *Acoustical factors affecting hearing aid performance*, chapter Normal Listeners in Typical Rooms - Reverberation Perception, Simulation, and Reduction, pp. 3–24, University Park Press, Baltimore, 1980.

[3] J. Pearson, Q. Lin, C. Che, D.-S. Yuk, L. Jin, B. deVries, and J. Flanagan, "Robust distant-talking speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, May 1996, vol. 1, pp. 21–24 vol. 1.

[4] H. Kuttruff, *Room Acoustics*, Taylor & Francis, London, fourth edition, 2000.

[5] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Sel. Areas Commun.*, vol. 10, no. 5, pp. 819–829, 1992.

[6] ITU-T P.862, "Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech coders," Tech. Rep., ITU-T, 2001.

[7] ITU-T, "Perceptual objective listening quality assessment: An advanced objective perceptual method for end-to-end listening speech quality evaluation of fixed, mobile, and IP-based networks and speech codecs covering narrowband, wideband, and super-wideband signals," Standard P.863, Jan. 2011.

[8] K. Kokkinakis and P. C. Loizou, "Evaluation of objective measures for quality assessment of reverberant speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 2420–2423.

[9] S. Goetze, A. Warzybok, I. Kodrasi, J.O. Jungmann, B. Cauchi, J. Rennies, E.A.P. Habets, A. Mertins, T. Gerkmann, S. Doclo, and B. Kollmeier, "A study on speech quality and speech intelligibility measures for quality assessment of single-channel dereverberation algorithms a study on speech quality and speech intelligibility measures for quality assessment of single-channel dereverberation algorithms," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014.

[10] J. Y. C. Wen and P. A. Naylor, "An evaluation measure for reverberant speech using tail decay modeling," in *Proc. European Signal Process. Conference*, 2006.

[11] G. A. Soulodre, N. Popplewell, and J. S. Bradley, "Combined effects of early reflections and background noise on speech intelligibility," *Journal of Sound and Vibration*, vol. 135, no. 1, pp. 123–133, 1989.

[12] J. P. A. Lochner and J. F. Burger, "The influence of reflections on auditorium acoustics," *Journal of Sound and Vibration*, vol. 1, no. 4, pp. 426–454, 1964.

[13] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *The Journal of the Acoustical Society of America*, vol. 21, no. 6, pp. 577–580, 1949.

[14] J. D. Polack, *La transmission de l'énergie sonore dans les salles.*, Ph.D. thesis, Université du Maine, Le Mans, 1988.

[15] ISO 226:2003 Acoustics, "Normal equal-loudness level contours," Tech. Rep. 2nd Edition, International Organization for Standardization (ISO), 2003.

[16] J. S. Garofolo, "Timit acoustic-phonetic continuous speech corpus," Tech. Rep., Linguistic Data Consortium, 1993.

[17] M. Jeub, M. Schafer, and P.Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Digital Signal Processing, 2009 16th International Conference on*. IEEE, 2009, pp. 1–5.

[18] J. Paulus, C. Uhle, and J. Herre, "Perceived level of late reverberation in speech and music," in *Audio Engineering Society Convention 130*, May 2011.