MEASURING, MODELLING AND PREDICTING PERCEIVED REVERBERATION

Hamza A. Javed1, Benjamin Cauchi2,4, Simon Doclo2,3,4, Patrick A. Naylor1 and Stefan Goetze2,4

1 Imperial College London, Dept. of Electrical and Electronic Engineering, London, UK
2 Fraunhofer IDMT, Project Group Hearing, Speech and Audio Technology, Oldenburg, Germany
3 University of Oldenburg, Dept. of Medical Physics and Acoustics, Oldenburg, Germany
4 Cluster of Excellence Hearing4all, Oldenburg, Germany

ABSTRACT

This paper investigates the relationship between the perceived level of reverberation and parameters measured from the room impulse response (RIR), as well as the design of an instrumental measure that predicts this perceived level. We first present the results of an experimental listening test conducted to assess the level of perceived reverberation in speech captured by a single microphone, before analysing the gathered data to assess the influence of parameters such as the reverberation time ($T_{60}$) or the direct-to-reverberant ratio (DRR). Secondly, we use the results of this analysis to improve the signal based reverberation decay tail ($R_{DT}$) measure, previously proposed by the authors to predict the perceived level of reverberation. The accuracy of the proposed measure is evaluated in terms of correlation with the subjective scores and compared to the performance of predictors using parameters extracted from the RIR. Results show that the proposed modifications to the $R_{DT}$ does improve its accuracy. Though still slightly outperformed by measures based on parameters of the RIR, we believe the proposed measure to be useful in scenarios in which the RIR or its parameters are unknown.

Index Terms— Perceived reverberation, experimental listening tests, instrumental measures

1. INTRODUCTION

Reverberation, broadly defined as the multipath propagation of sound in an enclosed space, has been shown to degrade both the intelligibility and perceived quality of speech. The effects of reverberation are particularly prominent when the desired signal is captured from a distant microphone [1]. With a growing number of speech communication applications facilitating distant-talking speech input, developing measures capable of quantifying the level of perceived reverberation in captured speech is an increasingly important task. Such measures would be particularly useful in evaluating dereverberation algorithms, bypassing the need to conduct expensive and time consuming listening tests [2].

Despite the growing interest in reverberation measures that correlate with human perception, to date there is no comprehensive consensus within the research community on the use of a single measure. In contrast, more consensus is found on other related issues such as, for example, with the use of the perceptual evaluation of speech quality (PESQ) to evaluate overall speech quality [3]. As a result, it is common practice to use quality measures not specifically designed to evaluate reverberation perception. For example, many proposed dereverberation techniques have their performance evaluated by the apparent reduction they are able to achieve in objective acoustic parameters, e.g. the reverberation time ($T_{60}$). By describing the acoustic space a sound is produced in, these metrics clearly provide insight into reverberation perception. However, considered on their own they may only be weakly correlated with perception [4].

The relationship between acoustic parameters and perceived reverberation has been explored most extensively in the context of architectural acoustics and concert hall designs, e.g. in [5]. Whilst valuable insights can be gained from these studies, they usually concern acoustic settings with high levels of reverberation that are not typically encountered in day-to-day scenarios. On the other hand, the parametric models proposed in [6, 7, 8, 9] may be more applicable for predicting the perceived effect of reverberation on speech. In particular the measure proposed in [9], derived from extensive experimental testing, was shown to strongly correlate with perception. However, the measures described above require knowledge of parameters extracted from the room impulse response (RIR) between the sound source and sensor.

Knowledge of the RIR or of its parameters is often not available and signal based measures, computed either only from the reverberant signal or with an additional clean reference signal are more easily applicable. For this purpose, the reverberation decay tail ($R_{DT}$) measure, proposed in [10], aims at predicting the perceived level of reverberation from a (noisy) reverberant speech signal and the corresponding reference signal. Aiming at improving the accuracy of the measure, we proposed a modified version of the $R_{DT}$ [11]. However this extension, denoted $R_{DTX}$ in the remainder of this paper, did not consistently outperform the original $R_{DT}$ [2]. Therefore, further improvements and validations are needed.

In this paper, we investigate a human listeners ability to perceive different levels of reverberation in speech captured by a single microphone. To this end, experimental listening tests were conducted in which participants scored realistic reverberant speech. Motivated by practical considerations, the test signals selected for assessment were produced in acoustic environments that speech processing technologies could typically be expected to operate in. Secondly, reverberation measures designed to estimate perceived level of reverberation were evaluated on the listening test data and their level of correlation with subjective scores was assessed. Finally, a novel weighting scheme is proposed and integrated into the computation of the $R_{DT}$ and $R_{DTX}$ measures. We show that the resulting weighted $R_{DT}$ ($wR_{DT}$) and weighted $R_{DTX}$ ($wR_{DTX}$) predict the perceived level of reverberation more accurately than their unweighted counterparts.

Having provided the motivation and context for this work, the remainder of this paper is organised as follows. In Section 2, reverberation measures evaluated in this work are summarised. The experimental listening test conducted is described in Section 3. The results and analysis of the data gathered is presented in Section 4, with concluding remarks made in Section 5.
2. REVERBERATION MEASURES

Reverberant speech can be modelled as the convolution between an acoustic channel and the speech signal in question. The channel, or RIR, thus characterises the effect of the room on the sound produced, for a given source-listener or source-microphone placement. For the study of perceptual reverberation, a RIR can be viewed as being made up of early and late reflections. The former are defined as reflections typically arriving within 50 ms of the direct path, and are actually considered beneficial [12]. The remaining reflections make up the reverberant tail of the RIR. They are responsible for the sound ringing on and are, by contrast, deemed detrimental to perceptual quality. The level of perceived reverberation, whilst also dependent on the source signal itself [7], is therefore strongly influenced by the RIR envelope profile.

2.1. Room Impulse Response Based Measures

The structure of a room impulse response can be mathematically described using well known acoustic metrics. These are primarily based on the rate of energy decay, e.g. the $T_{60}$, or on energy ratios of different portions of the RIR, e.g. the clarity-index ($C_{50}$) or the direct-to-reverberant ratio (DRR) [13]. Numerous reverberation measures proposed in the literature, attempt to quantify the relative impact of RIR based acoustic parameters on perception.

Perhaps the first attempt to develop a measure based on a parametric model of perceived reverberation, was by Allen [6]. The subjective preference of reverberant speech was reported to be a function of the reverberation time and room spectral variance (RSV). Referring to as Allen’s score in this work, it is mathematically expressed as

$$A_s = A_{max} - \sigma^2 T_{60},$$

(1)

where $A_{max}$ is chosen arbitrarily to represent the maximum preference score possible and $\sigma^2$ denotes the RSV.

More recently, Paulus et. al. proposed a provisional model derived from listening test data in which users scored artificially generated reverberant signals [7]. The measure, referred hereafter as Paulus’ score, models the impact of changes in $T_{60}$ and DRR on overall perception. Namely,

$$\Delta P_s = 19.2 \Delta T_{60} - 2.35 \Delta DRR,$$

(2)

indicating the relatively higher importance of the $T_{60}$ in reverberation perception, which has been observed in other works [14].

In [4], three main acoustic features were identified as playing a key role in predicting perceived reverberation. In the work that followed [8, 9], the QAreverb measure was proposed

$$Q = -\left(\frac{\lambda_{avg}}{\lambda_{avg}}\right)\sigma^2 \left(\frac{T_{60}}{DRR}\right)^\gamma,$$

(3)

where $\sigma^2$ denotes the RSV and exponents $\alpha$, $\beta$ and $\gamma$ are introduced to model the relative contribution of the three acoustic parameters. Initial experiments showed that by incorporating the DRR into Allen’s score ($\alpha = 1$, $\beta = 1$, $\gamma = 0.3$), the resulting QAreverb measure, $Q_{\text{avg}}$, correlated best with subjective ratings [8]. However, more extensive experiments presented in [9], demonstrated that the RSV could be discarded altogether, with optimal weights instead of $\alpha = 0.6$, $\beta = 0$ and $\gamma = 0.15$. To distinguish itself from it’s predecessor, the QAreverb measure computed using these weights is hereafter referred to as $Q_{\text{avg}}$.

2.2. Signal Based Measures

Along with the RIR based measures described earlier, we evaluate the ability of the $R_D$ and of its extension to accurately predict the perceived level of reverberation. These intrusive measures use a reference signal to identify time periods which would be most affected by reverberation in the perceptually motivated Bark spectral domain, and characterises them in the corresponding sections of the reverberant test signal. The $R_D$ is then computed as the ratio between the average reverberant energy $A_{avg}$, relative to the direct path energy $D_{avg}$, and the reverberation rate of decay $\lambda_{avg}$, i.e.

$$R_D = \frac{A_{avg}}{\lambda_{avg} D_{avg}},$$

(4)


Despite being a perceptually motivated measure, previous studies have shown both the $R_D$ and $R_D^2$ correlate poorly in some cases with subjective assessment of reverberation [2, 15]. It is hypothesised that this is due to the equal weighting of each component in the calculation of the measure, which can be thought of as a ratio of the estimated $T_{60}$ and DRR. We therefore incorporate a weighting scheme into the computation of the measure,

$$wR_D = \left(\frac{1}{\lambda_{avg}}\right)^\alpha \left(\frac{A_{avg}}{D_{avg}}\right)^\gamma \left(\frac{T_{60}}{DRR}\right)^\gamma,$$

(5)

where the introduced exponents $\alpha$ and $\gamma$, better model the relative importance of the $T_{60}$ and DRR. Inspired from [9], the $wR_D$ measure takes $\alpha$ and $\gamma$ values of 0.6 and 0.15 respectively. A thorough investigation into the choice of these values, through experimental training and testing, will be the subject of future work.

Finally, we also consider the perceptual evaluation of speech quality (PESQ) score [3], which despite being originally developed for the evaluation of audio codec quality, is often used evaluate the performances of speech dereverberation algorithms [16, 17]. Furthermore, it has been shown to correlate reasonably well with subjective assessment of reverberation [18].

3. EXPERIMENT

This section describes the listening test conducted; the reverberant test conditions considered for assessment are summarised, and the experimental setup detailed.

3.1. Test Signals

The audio signals presented for assessment $y_{i,j}(n)$, were generated by convolving anechoic speech $s_i(n)$ with recorded RIRs $h_j(n)$ of length $L_h$ and adding ambient noise $v_j(n)$ for each discrete time index $n$, i.e.

$$y_{i,j}(n) = \sum_{l=0}^{L_h-1} s_i(n-l) h_j(l) + v_j(n),$$

(6)

where subscripts $i$ and $j$ correspond to the utterance and test condition number respectively. All signals were sampled at 16 kHz and the generated test signals have durations ranging from 11 s to 17 s.

For all listeners, a total of $M = 10$ RIRs were used to generate reverberant test signals. The RIRs were selected from either the SMARD corpus [19] ($j = 1$), or from the ACE Challenge [20]
3.2. Listening Test

To assess the perceived reverberation in the test signals presented, a multi-stimuli test with hidden reference and anchor (MUSHRA) was used. The frameworks, which refer to a multi-stimuli test with hidden reference and anchor (MUSHRA), was previously proposed by the authors and considered to be an effective method for perceptual evaluation of reverberation [22]. It differs from the standard MUSHRA scheme in reference selection to be an equaliser evaluated for reverberation and anchor signals. Furthermore, the scoring scale (0-100) is inverted such that a more reverberant signal is attributed with a lower score, which is more intuitive to the listener.

A total of 28 self-reported normal hearing listeners, aged 24 to 60, participated in each session of the multi-stimuli test. All participants conducted the test in a soundproof booth and listened to the diotic signals reproduced using an audio interface and closed-back headphones (Sennheiser HD800).

The response of the audio chain was accounted for with appropriate equalisation performed in order to obtain a flat field response. Test signals were normalised from one to nine in English. All test signals had a signal to noise ratio (SNR) of 25 dB, which was computed using the early reverberant energy according to [21]. Test signals were normalised, and signal energy was limited to 65 SPL.

In order to produce a wide range of reverberant conditions for assessment, RIR recordings obtained from an office, meeting room, lecture theatre and building lobby were selected. All of which are far configurations respectively. Furthermore, the scoring scale (0-100) is inverted such that a more reverberant signal is attributed with a lower score, which is more intuitive to the listener.

Each session started with spoken instruction (read from a prepared script), and a training phase which allowed the participant to familiarize themselves with the test environment. All participants conducted the test in a soundproof booth and listened to the diotic signals reproduced using an audio interface and closed-back headphones (Sennheiser HD800).
4. RESULTS & ANALYSIS

The mean scores obtained per listener for each test condition are depicted in Fig. 1. Despite several closely spaced medians that indicate the difficulty in discriminating between certain adjacent test conditions, a general trend can be observed in the data. Furthermore, both the anchor and reference have been correctly identified suggesting the usefulness of these signals in removing unreliable assessors as part of a post-screening criterion. Scores attributed to the reference and anchor are hereafter excluded from the remaining analysis.

In order to examine the perceived pairwise differences between conditions, the results of a Wilcoxon rank-sum test are depicted in Fig. 2. The plot illustrates that the first three test conditions (B₁, B₂ and C₁), which correspond to rooms with similarly low reverberation times, are not distinguishable with statistical significance by assessors despite their varied DRR values. This supports the notion that the perceived level of reverberation is more influenced by the T₆₀ than the DRR. Interestingly, the only other pair that was not perceived as significantly different were test conditions C₂ and D₁. This case indicates that a speech signal captured in a room with a higher T₆₀ can be perceived as equivalently reverberant to a different room with lower T₆₀, if the source-microphone placement is such that the DRR is considerably higher. These results show that to effectively predict the level of perceived reverberation, acoustic parameters such as the T₆₀ and DRR must be considered in combination.

To gain further insight into the parameters most influential towards perception, the Pearson correlation coefficients between participant scores and the considered RIR characteristics is computed and depicted in the left panel of Fig. 3. The results demonstrate that the T₆₀ gives the highest correlation with perceived reverberation for the chosen speech samples, whilst the DRR and RSV are by contrast less strongly correlated. It should be noted that in a previous work [2], the reverse was shown to be true. However, the difference can be attributed to the smaller number and range of test conditions previously used to compute the correlations.

The performance of reverberation measures detailed in Section 2, are also evaluated to determine how well they correlate with subjective assessment. We can see from the central panel of Fig. 3 that of the RIR based measures, Paulus’ score performs comparatively poorly (|rₚₚₚ| = 0.46). This could be due to the fact that it was modelled from data obtained in a relatively small listening test, in which users were asked to score artificially generated reverberant signals. By contrast, the other measures exhibited stronger correlations. In particular the QAreverb metric, with weights proposed in [9] and denoted here as Qₐₓ, gave the strongest correlation of all.

Correlations between signal based reverberation measures and participant scores are shown in the right panel of Fig. 3. The RₓDT measure, weighted or unweighted, performs better than both RₓDTₓ and wRₓDTₓ. Perhaps indicating that narrowband frequency operation contains more cues related to perceived reverberation. The better performance of RₓDT compared to RₓDTₓ is consistent with the results presented in [2]. However, the weighting proposed in this paper does improve correlation in the case of both RₓDT compared to wRₓDT and RₓDTₓ compared to wRₓDTₓ. Therefore, of all the considered signal based measures, wRₓDT appears to be the best predictor of the perceived level of reverberation with a Pearson correlation coefficient of 0.80. The wRₓDT is only outperformed by Qₐₓ, whose computation requires knowledge of the T₆₀ and of the DRR. Finally, relatively weak correlations are obtained from the PESQ measure, with a Pearson correlation coefficient of 0.41 highlighting its limitations as a predictor of the level of perceived reverberation.

5. CONCLUSION

The results of an experimental listening test to assess the perceived level of reverberation in speech captured by a single microphone were presented. The relation between the subjective scores and RIR characteristics showed that the T₆₀ has a larger influence than the DRR on reverberation perception, but that both need to be considered in combination to predict the perceived level of reverberation. Of the considered instrumental measures, the QAreverb of [9] showed the highest correlation with participant scores, with a Pearson correlation coefficient of 0.85. However, the proposed wRₓDT measure also performed well. Modified to weigh individual features more heavily, the signal-based measure has the advantage of not requiring knowledge of the RIR or of how this RIR would be modified by a dereverberation algorithm, making it a useful evaluation tool for dereverberation research.

6. REFERENCES


