Using Acoustic Parameters for Intelligibility Prediction of Reverberant Speech

Ahmed Alghamdi*, Wai-Yip Chan*, Daniel Fogerty†
*Department of Electrical and Computer Engineering, Queen's University, Kingston, Canada,
{12aa89, chan}@queensu.ca
†Department of Communication Sciences and Disorders, University of South Carolina, Columbia, USA,
fogerty@sc.edu

***Abstract*—This work addresses the problem of predicting the subjective intelligibility of reverberant speech. Using new subjective listening test data, we evaluate the performance of three objective intelligibility measures that can be computed from the room impulse response. The measures are found to correlate well with the word and phoneme recognition rates of reverberant speech. In particular, one of the examined measures more readily relates to spatial dimensions and hence may be more convenient for environment-tracking acoustic interface applications.**

***Index Terms*—Reverberant speech, intelligibility measures, room acoustics, room impulse response.**

I. Introduction

Modern speech communication systems are deployed in highly variable and adverse conditions involving distortions caused by background noise, reverberation, packet loss, denoising and so forth. Therefore, it is important to measure the impact of these degradations on human perception of speech in order to design, monitor, and improve the performance of speech-centric applications.

Human perception of speech has many attributes depending on the application settings. However, speech quality and intelligibility are the key subjective measures of speech perception [1]. Quality is related to the pleasantness and naturalness of speech whereas intelligibility is the proportion of correctly identified units of speech (e.g. phonemes, syllables, words, and sentences).

The relationship between subjective quality and intelligibility is not fully clear. There exists some correlation between these attributes where good quality speech generally has high intelligibility, and vice versa. However, there are cases where poor quality speech provides high intelligibility. In the remainder of this paper, we will focus entirely on the intelligibility of speech corrupted with reverberation.

Reverberation occurs when speech propagates in enclosed spaces such as rooms and halls. The reverberant sound consists of the direct sound and multiple distorted versions caused by reflections off the objects and surfaces. Moderate levels of reverberation might have a pleasing effect. However, excessive reverberation reduces speech intelligibility. Therefore, we need to quantify the impact of reverberation on speech perception to ensure appropriate operation of speech applications in acoustic environments.

The ground truth to assess speech intelligibility is to carry out subjective tests where human listeners judge the intelligibility of reverberant speech. However, formal listening tests are costly and time consuming as well as impractical for applications requiring fast and cheap evaluation. Consequently, objective methods have been proposed to replace subjective tests.

Objective measures to predict intelligibility of reverberant speech can be divided into two categories [2]: signal-based, computed from reverberant speech and acoustic-based, computed from room acoustic parameters. The two approaches are complementary as they offer different advantages. For instance, an acoustic-based measure is useful for architectural acoustic design and planning and real-time operation of acoustic interfaces.

Additionally, acoustic-based estimation of intelligibility isolates the reverberant component of degradation when speech is degraded by multiple means besides reverberation, such as additive noise, non-ideal speaker, transmission impairment, etc. The model for this component can be combined with models for other degradation components to construct a comprehensive model.

ISBN 978-90-827970-1-5 © EURASIP 2018

In enclosed environments reverberation depends on several factors such as the room geometry, reflectivity of its surfaces, and positions of the source and receiver. The acoustic characteristics of a given enclosure can be modelled with a linear time invariant system where the reverberant sound at the receiver can be computed as the source sound convolved with the room impulse response (RIR). The RIR will vary as the positions of the source and receiver and surrounding objects change.

Many acoustic parameters computed from the RIR have been proposed to quantify the reverberation level. These include the reverberation time T_{60}, the direct-to-reverberant ratio (DRR), the clarity C_τ index, the definition D_τ index, and several other parameters [2]. Researchers have investigated the use of acoustic parameters in estimating the human perception of reverberant speech as well as the performance of automatic speech recognition (ASR) systems in reverberant environments.

In [3], the authors proposed a speech quality measure based on a nonlinear mapping of three acoustic parameters. The resulting measure has high correlation with subjective quality scores. In [4], the authors showed that the clarity index C_{50} is highly correlated with phone recognition rate. In this paper, we will extend the work of [3] and [4] for the case of predicting subjective intelligibility of reverberant speech.

The remainder of this paper is organized as follows. Section II details the room acoustic parameters to be used in predicting subjective intelligibility of reverberant speech. Our newly acquired subjective intelligibility dataset is described in Section III. In Section IV, we report the experimental results. Finally, conclusions are provided in Section V.

II. Room Acoustic Parameters

The RIR sequence is often divided into two regions as illustrated in Fig. 1. The first region represents early reflections which roughly span the first 80 ms of the RIR. The second region extends over the reminder of the RIR and represents late reflections. The early reflections region includes mainly distinctive impulses of large amplitudes whereas the late reflections region has noise-like components with smaller amplitudes. The early reflections contribute to desirable perceptual attributes such as pleasantness and intelligibility where the late reflections are associated with annoying attributes of reverberation. The precise boundary between the two regions may be varied depending on specific perceptual attribute(s) of interest.

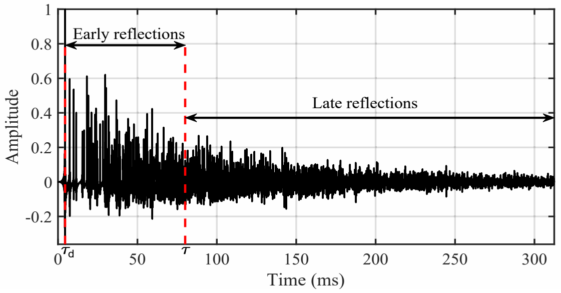


Fig. 1: An example of a normalized RIR sequence indicating the time instant of direct-path sound (τ_d) and the boundary between early and late reflections (τ).

The RIR sequence shown in Fig. 1 has a peak at time instant τ_d which is caused by the direct-path propagation of sound from the source to receiver. The position of the direct-path peak depends on the source-receiver distance and its relative amplitude depends on the reflectivity of objects and walls in the acoustic space. In the following subsections we will define several acoustic parameters and formulas that will be used in intelligibility prediction.

A. Reverberation Time (RT)

The RT, first proposed by Sabine [5], is a common measure of the impact of reverberation on a given sound signal. The RT, T_{60}, is defined as the time required for the reverberant energy to decay by 60 dB from its steady state level after the sound source is switched off. Longer RT time results in more reverberation effect. The value of T_{60} is approximately invariant across the room and it is determined by the reflectivity of surfaces and the room geometry.

There are several methods proposed in the literature to estimate T_{60} from a measured RIR $h(t)$ [6]–[8]. One technique uses the normalized energy decay curve (EDC) defined as [8]

$$\text{EDC}(t) = 10\log_{10}\left(\frac{\int_t^\infty h^2(\tau)d\tau}{\int_0^\infty h^2(\tau)d\tau}\right) \qquad (1)$$

The EDC has a maximum value of 0 dB. The EDC in the range from -5 dB down to a stop point, at least 10 dB above the noise floor, is approximated by a straight line [4], [6]. T_{60} is computed as the length of time needed for the linear approximation to decay from 0 to -60 dB.

B. Direct-to-Reverberant Ratio (DRR)

The DRR is intended to quantify factors due to the source-receiver distance and the reflection characteristics of the acoustic space. The DRR, R, is defined as the ratio of the direct-path energy E_d to the reverberant energy

E_r where E_d is the energy component of the RIR within a short time interval around the direct-pulse instant τ_d and E_r is the energy of the remaining part of the RIR. R is computed as [9]

$$R = \frac{E_d}{E_r} = \frac{\int_{\tau_d - 0.5ms}^{\tau_d + 1ms} h^2(\tau) d\tau}{\int_{\tau_d + 1ms}^{\infty} h^2(\tau) d\tau} \tag{2}$$

The authors of [3] proposed the following objective measure to predict subjective quality of reverberant speech:

$$Q = -\frac{(T_{60})^{\alpha}}{R^{\gamma}} \tag{3}$$

This formula extends Berkley and Allen's [10] formula that models relative quality degradation as proportional to the product of T_{60} and the standard deviation of the room spectral response. Since the spectral deviation increases as DRR drops and saturates at around 5.6 dB when DRR reaches 0 dB [11], the original formula is applicable only for DRR greater than 0 dB, i.e., when the direct sound is stronger than the reflected sounds. The formula of [3] is applicable for both the direct sound-dominant and reflected sounds-dominant (DRR $<$ 0 dB) regions, with parameters α and γ provided to fit the quality data. To determine the best values of α and γ, [3] used exhaustive search in the interval [0,1] with steps of 0.05. For every combination of α and γ, the correlation with subjective quality scores is computed and the exponents that achieve maximum correlation are determined. In this paper, we will follow the same approach, but the correlation is computed with subjective intelligibility scores.

C. *Clarity and Definition Indexes*

The clarity and definition indexes are measures of the strength of early reflections. The clarity index is defined as the dB ratio of early to late energies [12],

$$C_{\tau} = 10\log_{10}\left(\frac{\int_{\tau_d}^{\tau} h^2(t) dt}{\int_{\tau}^{\infty} h^2(t) dt}\right) \tag{4}$$

and the definition index is defined as the dB ratio of early to total energies [12],

$$D_{\tau} = 10\log_{10}\left(\frac{\int_{\tau_d}^{\tau} h^2(t) dt}{\int_{\tau_d}^{\infty} h^2(t) dt}\right) \tag{5}$$

where τ is the extent of early reflections and τ_d is the position of the direct-path peak.

The authors of [4] showed that C_{50} is well correlated with the accuracy of a standard phone recognizer. Motivated by their conclusion, we will find the length of early reflections that makes C_{τ} and D_{τ} highly correlated with subjective intelligibility scores. Below, we describe a subjective intelligibility dataset we produced to validate the above measures.

III. Subjective Intelligibility Data

The subjective intelligibility database contains 12 reverberation conditions (RCs) with 4 values of T_{60} $\{$0.9, 1.2, 1.5, 2.1 seconds$\}$ and for each T_{60} there are 3 R levels $\{$0, -10, -20 dB$\}$. The RIRs were simulated using the image method with a 16 kHz sampling rate [13], [14]. Reverberant speech in each RC was generated by convolving clean utterances with the RIR.

The speech text was taken from the Northwestern University Auditory Test No. 6 (NU6) corpus [15]. Each RC had a total of 28 monosyllabic words in the carrier phrase "say the word..." spoken by a native male or female speaker (14 sentences each). The speech signals were sampled at 16 kHz. Fifteen normal-hearing English-speaking participants listened to the reverberant utterances in a sound attenuating booth over headphones. Listeners responded by typing the target word. Typed responses were automatically corrected for common typing and spelling errors and were phonetically transcribed from the CELEX database [16] and were also visually inspected. Responses were scored using both word and phoneme scoring. The intelligibility score assigned to each RC condition was obtained by averaging over the 15 listeners. Subjective intelligibility is expressed in two forms: word recognition rate (WRR) and phoneme recognition rate (PRR). Repeated measures ANOVA demonstrated significant ($p < .001$) main effects (T_{60}, R) and interactions ($T_{60} \times R$) for both WRR and PRR.

IV. Results

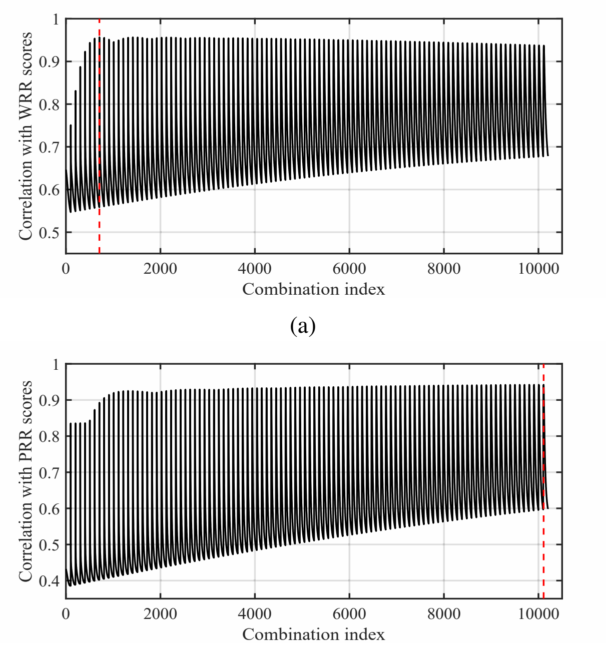
In the following subsections, we find the correlation-maximizing parameters α, γ, and τ by exhaustive search over a grid of values.

A. *Finding α and γ*

To find the best values of α and γ we vary each of these parameters in the interval [0,1] with steps of 0.01 and we compute the correlation between the objective scores computed using Eq. (3) and the subjective scores expressed in terms of WRR and PRR. The best values of the exponents are the combination that achieves highest correlation.

Fig. 2 shows plots of the correlation coefficients obtained from the 101^2 indexed combinations of α and γ values. Fig. 3 shows scatter plots for the chosen exponents along with a linear fit. We achieve a correlation of 0.957 with WRR scores using $\alpha = 0.07$ and $\gamma = 0.01$.

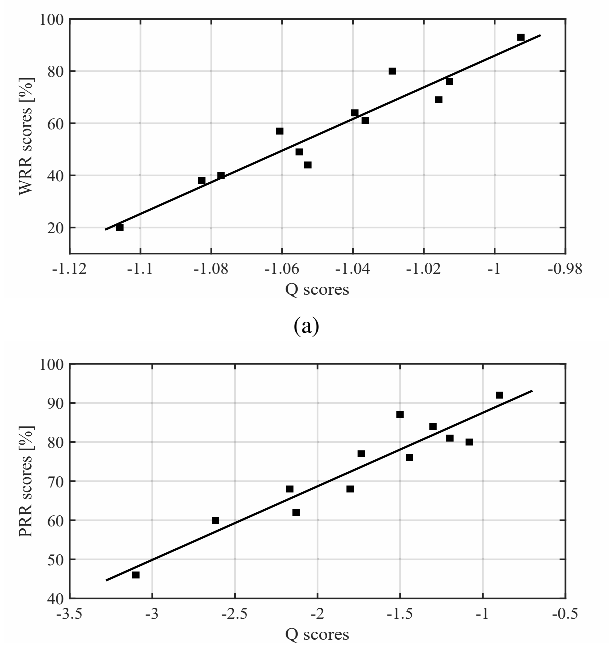
ISBN 978-90-827970-1-5 © EURASIP 2018



(a)

(b)

Fig. 2: Correlation coefficient versus combination index for (a) WRR scores and (b) PRR scores. The vertical dotted line shows the location of maximum correlation.



(a)

(b)

Fig. 3: Scatter plots between objective and subjective scores for the chosen exponents. (a) WRR scores and (b) PRR scores.

For the case of PRR scores, we achieve a correlation of 0.942 using $\alpha = 1$ and $\gamma = 0.08$. From WRR to PRR, α increases by a factor of 14 while γ increases by a factor of 8. This suggests that PRR is numerically more sensitive to T_{60} than WRR. The shorter duration of phonemes makes them more vulnerable to increasing time-domain smearing as T_{60} increases.

Fig. 2 also shows that many combinations of α and γ achieve nearly maximum correlation. This can be explained as follows. Let c be a positive number. There are many values of c such that if we replace the above correlation-maximizing α and γ with $c\alpha$ and $c\gamma$ in Eq. (3), respectively, the correlation is still close to the maximum.

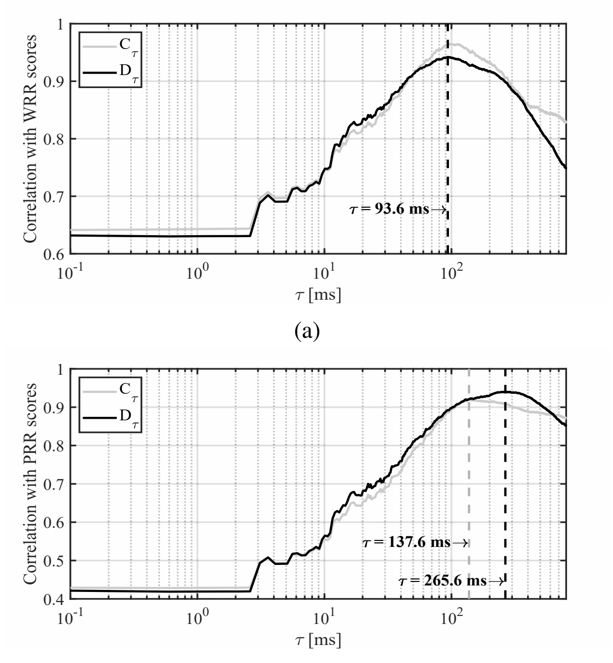
For the case of reverberant speech quality prediction, the highest correlation obtained in [3] is 0.942 with $\alpha = 0.6$ and $\gamma = 0.15$. The larger $\alpha - \gamma$ ratio for WRR and PRR than the subjective quality in [3] suggests that for sufficiently reverberant speech, intelligibility degradation is more sensitive to T_{60} than quality degradation.

B. Finding τ

We find the time length of early reflections that is best correlated with subjective intelligibility by varying τ in the range 0.1 ms to 800 ms with 0.5 ms steps and examine the correlation between subjective scores and C_τ and D_τ, separately.

Fig. 4 shows plots of correlation obtained with C_τ and D_τ as a function of the time index τ. Fig. 5 shows scatter plots for the acoustic indexes yielding highest correlation and their linear fitting. When predicting WRR scores the maximum correlation for both C_τ and D_τ is obtained at $\tau = 93.6$ ms where the correlation is 0.965 for C_τ and 0.942 for D_τ. Regarding PRR scores, C_τ provides maximum correlation of 0.920 obtained at $\tau = 137.6$ ms and D_τ provides maximum correlation of 0.941 obtained at $\tau = 265.6$ ms. The larger τ values needed for PRR are akin to the need to rebalance the exponent weights α and γ on RT and DRR, as we saw earlier.

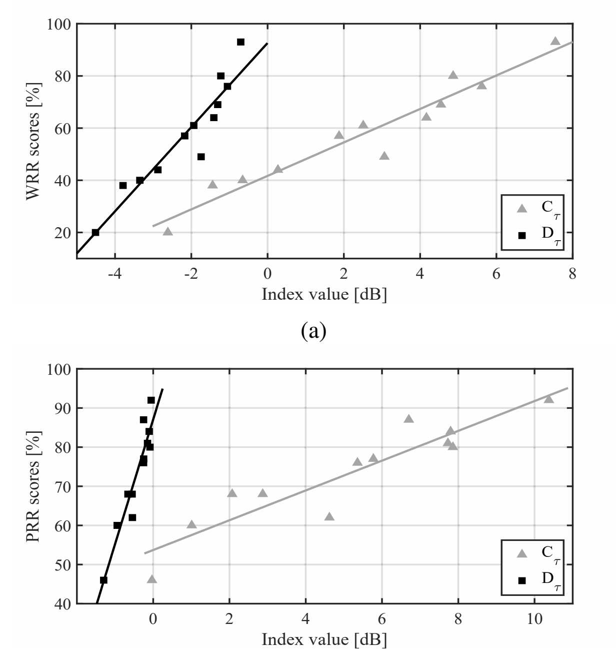
The maximum correlation with WRR and PRR using RIR parameter C_τ or D_τ is about the same as obtained using the Q score (Eq. (3)). This corroborates the general understanding that intelligibility is a function of the contrast between early and late reflection energies. T_{60} and R are more directly relatable to the physical acoustic setting. The Q formula (Eq. (3)) may be more convenient for sensor-rich human computer interfaces that track their environments.



(a)

(b)

Fig. 4: Correlation coefficients versus values of τ. (a) WRR scores and (b) PRR scores.



(a)

(b)

Fig. 5: Scatter plots between objective and subjective scores for the chosen values. (a) WRR scores and (b) PRR scores.

V. Conclusion

We have used newly gathered subjective intelligibility test data to optimize and validate several acoustic parameter-based measures for estimating the intelligibility of reverberated speech signals. The measures showed good correlation with the test scores. Further testing will be performed to characterize the relation of subjective intelligibility with other perceptual attributes.

References

[1] K. Kondo, *Subjective Quality Measurement of Speech: Its Evaluation, Estimation and Applications*. Springer, 2012.

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

[3] J. Del Vallado, A. A. de Lima, T. d. M. Prego, and S. L. Netto, “Feature analysis for the reverberation perception in speech signals,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 8169–8173.

[4] P. P. Parada, D. Sharma, and P. A. Naylor, “Non-intrusive estimation of the level of reverberation in speech,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 4718–4722.

[5] H. Kuttruff, *Room Acoustics*. CRC Press, 2016.

[6] M. Schroeder, “New method of measuring reverberation time,” *J. Acoust. Soc. Amer.*, vol. 37, pp. 409–412, 1965.

[7] P. Antsalo, A. Makivirta, V. Valimaki, T. Peltonen, and M. Karjalainen, “Estimation of modal decay parameters from noisy response measurements,” in *Audio Engineering Society Convention 110*. Audio Engineering Society, 2001.

[8] A. Lundeby, T. E. Vigran, H. Bietz, and M. Vorländer, “Uncertainties of measurements in room acoustics,” *Acustica,*, vol. 81, pp. 344–355, 1995.

[9] A. De Lima, T. Prego, S. Netto, B. Lee, A. Said, R. Schafer, T. Kalker, and M. Fozunbal, “On the quality-assessment of reverberated speech,” *Speech Communication*, vol. 54, pp. 393–401, 2012.

[10] D. Berkley and J. Allen, “Normal listening in typical rooms: The physical and psychophysical correlates of reverberation,” in *Acoustical Factors Affecting Hearing Aid Performance*, 2nd ed., G. Studebaker and I. Hochberg, Eds. Allyn and Bacon Boston, 1993, pp. 3–14.

[11] J. Jetzt, “Critical distance measurement of rooms from the sound energy spectral response,” *J. Acoust. Soc. Amer.*, vol. 65, pp. 1204–1211, 1979.

[12] A. De Lima, T. Prego, S. Netto, B. Lee, A. Said, R. Schafer, T. Kalker, and M. Fozunbal, “Feature analysis for quality assessment of reverberated speech,” in *Multimedia Signal Processing (MMSP), 2009 IEEE International Workshop on*, 2009, pp. 1–5.

[13] J. Allen and D. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[14] E. Habets, “Room impulse response (RIR) generator,” github.com/ehabets/RIR-Generator.

[15] T. Tillman and R. Carhart, “An expanded test for speech discrimination utilizing CNC monosyllabic words: Northwestern University Auditory Test No. 6,” USAF School of Aerospace Medicine, Tech. Rep., 1966.

[16] R. Baayen, R. Piepenbrock, and L. Gulikers, “The CELEX lexical database:Linguistic Data Consortium,” *University of Pennsylvania*, 1995.