Development of a clarity parameter using a time-varying loudness model (L)

Doheon Lee,[1,a)] Jasper van Dorp Schuitman,[2] Xiaojun Qiu,[1] and Ian Burnett[1]
[1]*Faculty of Engineering and Information Technology, University of Technology Sydney, New South Wales 2007, Australia*
[2]*Sound Intelligence, Amersfoort, The Netherlands*

(Received 10 August 2017; revised 26 April 2018; accepted 16 May 2018; published online 8 June 2018)

The perceived sound clarity is often estimated with the clarity index, which is calculated on the basis of physical acoustic measures that can correlate weakly to the way humans perceive sound for certain test conditions. Therefore, this study proposes a clarity parameter based on a binaural room impulse response processed with a time-varying loudness model. The proposed parameter is validated by calculating the correlation coefficient with subject responses collected from previous listening experiments. Results show that the parameter outperforms the clarity index in most of the tested conditions, but its performance is less robust than parameter for clarity (P_{CLA}).
 https://doi.org/10.1121/1.5040480

[JB] Pages: 3455–3459

I. INTRODUCTION

The sound clarity is defined as a degree to which every musical detail in a music piece or each syllable in speech can be heard, and has been considered as one of the most important room acoustic attributes. ISO 3382-1 (2009) recommends using the clarity index, definition, and centre time for the estimation of the sound clarity (hereafter, referred to as the ISO clarity parameters). Among those, the clarity index (Reichardt *et al.*, 1975; ISO 3382-1, 2009) with an early time limit of 50 ms (hereafter, referred to as C_{50}) and 80 ms (hereafter, referred to as C_{80}) have been most commonly used for speech clarity and music clarity, respectively. In order to obtain a better estimation of the perceived clarity, a new clarity parameter is proposed in this study, which is calculated by using the output of a time-varying loudness model (TVL) (Glasberg and Moore, 2002) from binaural room impulse responses (BRIRs).

C_{50} and C_{80} are defined as a ratio of the early sound energy to the late sound energy, and this energy ratio is measured on the basis of the sound pressure level (SPL) in octave bands of a room impulse response (RIR). As the SPL considers very little of the transformation from sound to perception, it does not agree well with the subjective sound strength (hereafter, loudness) in many acoustic scenarios (Fastl and Zwicker, 2007). For this reason, the measured early-to-late SPL ratio of a RIR is not always consistent with the early-to-late loudness ratio of a RIR, leading to the discrepancy between the clarity index and the perceived sound clarity.

To address this problem, van Dorp Schuitman *et al.* (2013) proposed a clarity parameter calculated using an auditory model based on the model by Breebaart (Breebaart, 2001; Breebaart *et al.*, 2001), i.e., parameter for clarity (P_{CLA}). For obtaining P_{CLA}, a running signal such as music and speech is processed with the auditory model, after which a level ratio of two model outputs (i.e., of the direct and reverberant stream) is calculated. In this way, P_{CLA} incorporates many complexities of the auditory system. Results of multiple listening experiments showed that P_{CLA} provides a closer match to the perceived clarity than C_{50} and C_{80} in various listening conditions (van Dorp Schuitman *et al.*, 2013). Similarly, Griesinger (2010) proposed localizability (LOC) by comparing the number of nerve firings resulting from the onset of direct sound to those resulting from reflections. LOC is a measure for the perceived engagement that is closely related to the perceived clarity.

Unlike that in calculating P_{CLA}, in this study a BRIR is used for the derivation of the new clarity parameter. To do this, a BRIR is processed with the TVL, and a ratio of the early-to-late loudness of a BRIR is calculated in the similar way to that for the clarity index. The TVL calculates loudness of an input sound in the following way. An finite impulse response (FIR) filter simulates the combined effect of the middle and outer ear transfer functions. Then, six Hanning windows are applied and six parallel fast Fourier transforms are executed to calculate spectral magnitudes for frequencies from 20 Hz to 15 kHz. From the short-term spectrum at intervals of 1 ms, an excitation pattern is derived and transformed to a specific loudness pattern. The total area under the specific loudness pattern is the instantaneous loudness, which is an intervening variable that is not consciously perceivable (Glasberg and Moore, 2002). For this reason, the short-term loudness output of the TVL, which models the perceived loudness at any instant, is used for the loudness analysis of a BRIR in this study. This short-term loudness is calculated from the instantaneous loudness by executing a set of functions for the auditory temporal integration.

The rational for using this method comes from the study by Lee and Cabrera (2010) and Lee *et al.* (2012). In these studies, the outputs of the dynamic loudness model (Chalupper and Fastl, 2002) and the TVL from a RIR provides a better match to the perceived decay of a RIR than the SPL decay of a RIR. As the masking largely affects the perceived clarity, accurate modelling of the perceived sound decay with this method can

[a)]Electronic mail: dosyd@hotmail.com

enable better estimation of the perceived clarity. It should be noted that only the TVL is used in this study, as it allows a binaural input and includes functions for modelling the binaural loudness perception (Moore and Glasberg, 2007). Therefore, instead of RIRs, BRIRs are used for the derivation of the new clarity parameter.

An issue here is that listening to a BRIR is different from listening to anechoic samples convolved with the same BRIR. Nevertheless, in previous studies (Lee *et al.*, 2012; Lee *et al.*, 2017), the reverberation parameters calculated from the loudness decay of a RIR outperformed significantly the conventional reverberation time and early decay time for various anechoic samples convolved with the same RIR. This is because the perceived decay of a RIR is a good match to the overall perceived reverberation decay of the convolution products (as a RIR is the only source of the reverberation of the convolution products). This finding supports the use of the loudness decay analysis of a RIR (or a BRIR) for the development of room acoustic parameters. As such a psychoacoustic analysis of a RIR is the system analysis, the proposed parameter might perform well for both for dynamic and stationary sounds played in the same system.

The proposed clarity parameter is validated by calculating the correlation coefficient with the subjective data collected from the previous listening experiments by van Dorp Schuitman *et al.* (2013), in which cello and speech samples were tested.

II. METHOD

A. Calculation of C_N

The proposed parameter is named C_N (as a subscript "N" stands for loudness), and it is calculated as follows: (1) L_{AFmax} (which is the maximum A-weighted SPL with a "fast" temporal integration, i.e., using a 125 ms time constant) of a BRIR is adjusted to a desired or measured value of L_{Aeq} (which is the power-average of the A-weighted SPL over a given time period) of music and speech; (2) the loudness of the level-adjusted BRIR is calculated using the TVL; and finally (3) C_N is calculated in a similar way as the clarity index, using the loudness of the level-adjusted RIR

$$C_N = 10 \cdot \log_{10}\left(\frac{\int_o^{t_e} N(t)dt}{\int_{t_e}^{\infty} N(t)dt}\right), \quad (1)$$

where $N(t)$ is the loudness of a level-adjusted RIR, t_e is the early time limit. As mentioned above, the short-term loudness output of the TVL is used for the calculation of C_N because it approximates the momentary loudness perception. As it is unclear which value of t_e is suitable for C_N, this study tested t_e from 10 to 100 ms with 10 ms intervals.

B. Experiments

van Dorp Schuitman *et al.* (2013) conducted four experiments in which subjects listened to four sets of binaural audio samples and rated them in terms of four acoustic qualities, namely, reverberance, clarity, apparent source width, and listener envelopment, on a range from "very low" to "very high." Only the responses for clarity are used in the present study. Each set of binaural audio samples represents different acoustic conditions as listed in Table I. For the samples, four sets of measured or simulated BRIRs were convolved with an anechoic solo cello recording and anechoic speech. The convolved speech and music samples have a length of 10 s. Note that C_{50} and C_{80} in Table I are calculated from the monaural RIRs (as recommended by ISO 3382-1, 2009), which were measured with the same source−receiver positions as the BRIRs.

The experiments were conducted with a double-blind task, following a so-called "mixed procedure" method proposed by Chevret and Parizet (2007), which is a mix between a paired comparison and a direct evaluation method. Using this method, the subjects were allowed to apply direct rating to the samples using a slider on the screen, and then the collected subject responses were sorted from the highest to lowest rating, allowing for paired comparisons by fine-tuning the ratings.

As shown in Table I, experiments 1 and 2 include "virtual" rooms, for which the BRIRs were simulated using an acoustic shoebox model (van Dorp Schuitman, 2011). The main difference between the two experiments is that for experiment 1 "realistic" rooms were chosen, whereas rooms for experiment 2 had more "non-realistic" properties in terms of dimensions, shape, and spatial distribution of absorption. For example, one of the rooms in experiment 2 has T_{20} of 1.75 s with side walls are completely absorbing.

The use of such unrealistic rooms was done in an attempt to decrease the correlation between acoustic parameters that is often present in real rooms. For example, highly reverberant rooms often show low clarity values and vice versa. In order to investigate if acoustic parameters model

TABLE I. An overview of experiments. Clarity index represents C_{80} for the cello samples and C_{50} for the speech samples.

Experiment	L_{Aeq} range (dB)	Clarity index (dB)	No. of rooms	Room type	Loudness normalized
Exp. 1: Cello	68.3 to 69.7	−9.10 to 64.13	9	Virtual (realistic)	Yes
Exp. 1: Speech	64.7 to 66.3	−10.16 to 50.91			
Exp. 2: Cello	68.8 to 69.9	−0.56 to 3.64	8	Virtual (unrealistic)	Yes
Exp. 2: Speech	65.2 to 66.2	−1.37 to 2.13			
Exp. 3: Cello	45.9 to 72.2	−15.88 to 45.78	10	Real	No
Exp. 3: Speech	42.8 to 71.2	−20.61 to 45.62			
Exp. 4: Cello	64.2 to 69.5	−15.88 to 45.78	10	Real	Yes
Exp. 4: Speech	64.9 to 70.2	−20.61 to 45.62			

the corresponding perceived attributes correctly, the parameters should be more or less independent wherever possible. In both experiments, the samples were normalized to the same estimated loudness using the Replaygain 1.0 algorithm (Robinson, 2001). This algorithm estimates the loudness by applying an equal loudness filter, followed by RMS energy calculations in 50 ms blocks. Finally, the 95% highest RMS value is picked as the overall loudness. In contrast, experiments 3 and 4 included real rooms. While experiments 3 and 4 used the same real rooms, only samples in experiment 3 retained their original loudness differences. Note that values of C_{50} and C_{80} are same in experiments 3 and 4 because they are not affected by the SPL.

Five subjects participated in experiments 1 and 2. They were working at the acoustics department at TU Delft with in-depth knowledge about the room acoustical parameters and had experience in assessing those parameters. Fifteen subjects participated in experiments 3 and 4. They consisted mostly of students with mixed musical experiences and preferences. All subjects reported normal hearing, and received instructions (including audio examples) explaining sound clarity before the start of the experiments. More details of the experiment method can be found in van Dorp Schuitman *et al.* (2013).

III. RESULTS

The performance of C_N is validated by calculating correlation coefficients between C_N and the subject responses, as in the study of van Dorp Schuitman *et al.* (2013). For example, the correlation coefficient in experiment 1 is calculated between values of C_N in the nine rooms and the subject responses collected from the same rooms. The correlation coefficient indicates the strength and direction of the linear relationship between two factors, and its value ranges from -1 to $+1$ (Privitera, 2015). Therefore, an ideal clarity parameter is supposed to yield a correlation coefficient of $r = 1$. As each subject may have rated "very low" and "very high" differently on the continuous scale, the subject responses were normalized according to ITU-R BS.1284-1 (ITU-R, 2003) to compensate for variations in interpretation of the scale

$$z_i = \frac{x_i - \bar{x}_i}{\sigma_i} \cdot \sigma + \bar{x}, \tag{2}$$

where z_i is the normalized results for subject i, x_i is the results for subject i, $\bar{x}_i$ is the mean result for this subject and σ_i is the standard deviation. $\bar{x}$ and σ are the mean and the standard deviation for all subjects, respectively.

The correlation coefficients between the normalized subject responses (hereafter, subject responses) and each of C_N, C_{50}, and C_{80} are shown in Fig. 1. All the correlation coefficient values in Fig. 1 are statistically significant ($p < 0.05$), except C_N and C_{50} in experiment 2 for the speech samples. As shown in Fig. 1, C_N outperforms C_{80} for all the tested early time limits in the experiments with the cello samples. For the speech samples, C_N performs better or similarly to C_{50} in experiments 1, 3, and 4 for all the tested early time limits.

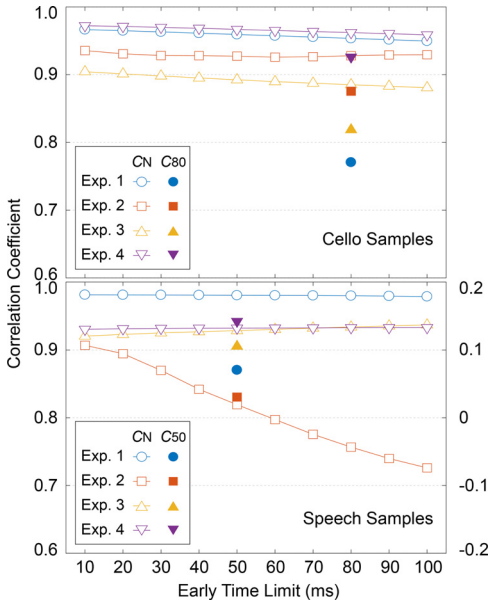

FIG. 1. (Color online) The correlation coefficients between the subjective responses and each of C_N, C_{80}, and C_{50}. The right y-axis in the lower figure is for experiment 2. The x axis is the early time limit t_e in Eq. (1).

In Table II, C_N at the early time limit of 50 ms (C_{N50}) and 80 ms (C_{N80}) are compared with C_{50}, C_{80}, and P_{CLA} (which is a psychoacoustic clarity parameter proposed by van Dorp Schuitman *et al.*, 2013). Note that C_N with other early time limits would also yield similar results (see Fig. 1), but 80 and 50 ms are chosen for this comparison as C_{80} and C_{50} are also based on them. In the table, the highest correlation in each experiment is underlined. The comparisons reveal that C_{N80} is the most accurate parameters for the cello samples. For the speech samples, P_{CLA} exhibits the most robust performance while C_{50} and C_{N50} are inaccurate in the unrealistic rooms (i.e., experiment 2). The table also shows that the outperformance of C_N over the clarity index is greater for the cello samples (i.e., C_{N80} vs C_{80}) than for the speech samples (i.e., C_{N50} vs C_{50}).

IV. DISCUSSION

This study proposed a loudness-based clarity parameter, named C_N. To validate the performance of C_N, the correlation coefficients between C_N and the subject responses

TABLE II. The correlation coefficients between the subject responses and each of C_{50}, C_{80}, C_{N80}, C_{N50}, and P_{CLA}. C_{N80} and C_{N50} are C_N with the early time limit of 80 and 50 ms, respectively. Statistically insignificant correlations are parenthesized ($p > 0.05$). The highest correlation in each experiment is underlined.

	Cello			Speech		
	C_{80}	P_{CLA}	C_{N80}	C_{50}	P_{CLA}	C_{N50}
Exp. 1	0.77	0.94	0.95	0.87	0.90	0.98
Exp. 2	0.87	0.83	0.93	(0.03)	0.82	(0.02)
Exp. 3	0.82	0.79	0.89	0.91	0.82	0.93
Exp. 4	0.93	0.96	0.96	0.94	0.87	0.93

collected from the listening experiments by van Dorp Schuitman *et al*. (2013) are calculated. The results show that C_N outperforms the clarity index and P_{CLA} in most of the realistic acoustic conditions, and that the performance of P_{CLA} is the most robust in the unrealistic acoustic conditions. The performance of C_N is scarcely affected by the evaluation range (see Fig. 1). Note that the correlations in experiment 2 for the speech samples are not statistically significant ($p > 0.05$).

The motivation for proposing C_{N} is that the research from Lee and Cabrera (2010) shows that the loudness decay of a RIR (calculated using the TVL) provides a closer match to the perceived sound decay of a RIR. The most striking feature of the loudness decay of a RIR is its level dependency (e.g., a slow decay for an increased SPL of a RIR, and vice versa). While the SPL influence on the perceived clarity has not been clearly defined yet, a negative relationship between the SPL and the perceived clarity is likely to occur because prolonged reverberance due to an SPL increase (as reported in Lee *et al*., 2012) can mask more subsequent sounds.

Partly for this reason, C_{80} combined with the A-weighted G (i.e., a parameter for sound strength specified in ISO 3382-1, 2009) is correlated better with subject responses than the sole use of C_{80} (Soulodre and Bradley, 1995). The SPL influence on the perceived clarity is also observed in results of experiments 3 and 4, in which the same samples are played at different SPLs (see Table I). As shown in Table II, for the cello samples, the outperformance of $C_{80\mathrm{N}}$ (which is the clarity index calculated from the loudness decay function of a BRIR with the early time limit of 80 ms) over C_{80} is greater in experiment 3 (where the spread in the SPL is large) than in experiment 4 (where the spread in the SPL is small). For the speech samples, $C_{50\mathrm{N}}$ (which is the clarity index calculated from the loudness decay function of a BRIR with the early time limit of 50 ms) is less correlated with the perceived clarity than C_{50} in experiment 4.

Comparing C_{N} and P_{CLA} is of interest, because they are based on different psychoacoustic models and calculation methods. C_N uses the output of the TVL from a BRIR, and P_{CLA} uses the output of an auditory model from a running signal. That way, the latter considers directly the acoustic properties of samples such as the spectral distribution and temporal envelope. Furthermore, P_{CLA} is calculated from a ratio of direct-to-reverberant sound, while C_N is calculated from a ratio of early-to-late sound. As shown in Table II, C_N (and also the clarity index) does not correlate with the subject responses in experiment 2 for the speech samples. This is partly because the BRIRs in experiment 2 have atypical reflection distribution, due to the unrealistic acoustic conditions. Therefore, unlike in most realistic conditions, their ratios of early-to-late sound are independent from reverberation provided by the same BRIRs, which leads to a situation where a high C_N value is yielded in a high reverberant condition. Because reverberation substantially degrades the perceived clarity, C_N does not correctly estimate the perceived clarity in such unrealistic acoustic conditions. However, the same result is not observed in experiment 2 for the cello samples. This is partly due to quasi-stationary ("legato") passages in the cello samples, which mask a substantial part of reverberant sounds and therefore the effect of atypical reflection distributions of the BRIRs can be mitigated.

For the calculation of C_N, 10 times the logarithm to base 10 (the common logarithm) is applied to the loudness ratio of a BRIR [see Eq. (1)]. When C_{N80} and C_{N50} are calculated without the logarithm, they yield lower correlation coefficients than C_{80} and C_{50} in all the tested conditions, except in experiment 2 for the cello samples. Furthermore, without being multiplied by 10, the scale of C_{N} becomes very small. For example, in experiment 1 for the cello samples, the range of C_{N80} without being multiplied by 10 is only from -1.18 to 0.48, while the range of C_{80} is from -9.10 to 64.13 dB (note that the rooms in experiment 1 have the reverberation time from 0.01 to 6.92 s).

In Table II, C_{N80} and C_{N50} are compared with C_{80} and C_{50}, respectively, but this study does not strongly suggest the early time limit of 80 and 50 ms for C_{N}, because C_{N} is similarly correlated with the subject responses for all the tested early time limits. To generalize the use of such early time limits, the performance of C_{N} needs to be investigated extensively with various music and speech samples. In future work, it would be interesting to calculate C_N with a monaural RIR (as the TVL also allows a monaural input) and to investigate the performance of C_{N} when a different loudness model is used.

V. CONCLUSION

This paper proposes a loudness clarity parameter C_N based on the BRIR processed with the TVL for better estimation of the perceived sound clarity than the clarity index (i.e., C_{50} and C_{80}) specified in ISO 3382-1 (2009). The results show that C_{N} is correlated better with the perceived clarity than the clarity index in most of the tested acoustic conditions, and the outperformance of C_N over the clarity index is greater for the cello samples than for the speech samples. This outperformance of C_N is greater when the spread in the SPL between samples is larger, and the early time limit scarcely affects the accuracy of C_N. However, the performance of C_N is not robust in the tested unrealistic acoustic conditions. These results provide a basis for future research into the use of loudness modelling for the estimation of the perceived sound clarity.

Breebaart, D. J. (**2001**). "Modelling binaural signal detection," Ph.D. thesis, Eindhoven University of Technology.

Breebaart, D. J., van de Par, S., and Kohlrausch, A. (**2001**). "Binaural processing model based on contralateral inhibition. I. Model structure," J. Acoust. Soc. Am. **110**, 1074–1088.

Chalupper, J., and Fastl, H. (**2002**). "Dynamid loudness model (DLM) for normal and hearing-impaired listeners," Act. Acust. Acust. **88**, 378–386.

Chevret, P., and Parizet, E. (**2007**). "An efficient alternative to the paired comparison method for the subjective evaluation of a large set of sounds," in *Proceedings of the 19th International Congress on Acoustics*, Madrid, Spain.

Fastl, H., and Zwicker, E. (**2007**). *Psychoacoustics: Facts and Models*, 3rd ed. (Springer, Berlin), Chaps. 3, 4, 6, and 8.

Glasberg, B. R., and Moore, B. C. J. (**2002**). "A model of loudness applicable to time-varying sounds," J. Audio. Eng. Soc. **50**, 331–342.

Griesinger, D. (**2010**). "Phase coherence as a measure of acoustic quality, part one: The neural mechanism," in *Proceedings of the 20th International Congress on Acoustics*, Sydney, Australia (CD-ROM).

ISO 3382-1. (**2009**). "Acoustics—Measurement of room acoustic parameters—Part 1: Performance spaces" (International Organization for Standardization, Geneva, Switzerland).
ITU-R BS. 1284-1. (**2003**). "General methods for the subjective assessment of sound quality," ITU Radiocommunication Sector, Geneva Switzerland.
Lee, D., and Cabrera, D. (**2010**). "Effect of listening level and background noise on the subjective decay rate of room impulse responses: Using time-varying loudness to model reverberance," Appl. Acoust. **71**, 801–811.
Lee, D., Cabrera, D., and Martens, W. L. (**2012**). "The effect of loudness on the reverberance of music: Reverberance prediction using loudness models," J. Acoust. Soc. Am. **131**, 1194–1205.
Lee, D., van Dorp Schuitman, J., Cabrera, D., Qiu, X., and Burnett, I. (**2017**). "Comparison of psychoacoustic-based reverberance parameters," J. Acoust. Soc. Am. **142**, 1832–1840.
Moore, B. C. J., and Glasberg, B. R. (**2007**). "Modeling binaural loudness," J. Acoust. Soc. Am. **121**, 1604–1612.
Privitera, G. J. (**2015**). *Statistics for the Behavioural Sciences*, 2nd ed. (Sage, Los Angeles, CA), Chap. 15.
Reichardt, W., Abdel Alim, O., and Schmidt, W. (**1975**). "Definition und Meßgrundlage eines objektiven Maßes zur Ermittlung der Grenze zwischen brauchbarer und unbrauchbarer Durchsichtigkeit bei Musikdarbietung" ("Definition and foundation of a measure to assess the border between usable and non-usable transparency of musical performances"), Acustica **32**, 126–137.
Robinson, D. (**2001**). Replaygain—A proposed standard. http://wiki.hydrogenaud.io/index.php?title=ReplayGain_specification (Last viewed 20/07/2017).
Soulodre, G. A., and Bradley, J. S. (**1995**). "Subjective evaluation of new room acoustic measures," J. Acoust. Soc. Am. **98**, 294–301.
van Dorp Schuitman, J. (**2011**). "Auditory modelling for assessing room acoustics," Ph.D. thesis, Delft University of Technology, Delft, Netherlands.
van Dorp Schuitman, J., de Vries, D., and Lindau, A. (**2013**). "Deriving content-specific measures of room acoustic perception using a binaural, nonlinear auditory model," J. Acoust. Soc. Am. **133**, 1572–1585.