NON-INTRUSIVE ESTIMATION OF THE LEVEL OF REVERBERATION IN SPEECH

Pablo Peso Parada[⋆]*, Dushyant Sharma*[⋆]*, Patrick A. Naylor*[†]

[⋆]Nuance Communications Inc. Marlow, UK
[†]Department of Electrical and Electronic Engineering, Imperial College London, UK
{pablo.peso, dushyant.sharma}@nuance.com, p.naylor@imperial.ac.uk

ABSTRACT

We show corroborating evidence that, among a set of common acoustic parameters, the clarity index C_{50} provides a measure of reverberation that is well correlated with speech recognition accuracy. We also present a data driven method for non-intrusive C_{50} parameter estimation from a single channel speech signal. The method extracts a number of features from the speech signal and uses a binary regression tree, trained on appropriate training data, to estimate the C_{50}. Evaluation is carried out using speech utterances convolved with real and simulated room impulse responses, and additive babble noise. The new method outperforms a baseline approach in our evaluation.

Index Terms— C_{50} estimation, speech recognition.

1. INTRODUCTION

Sound propagation in enclosed spaces may follow multiple paths from the source to the receiver due to reflections from surfaces in the room, in addition to direct path propagation. These reflections create a reverberant sound which varies with the acoustic characteristics of the room and positions of source and receiver. Whereas the reverberation time (T_{60}) is widely used to characterize the acoustic properties of a room, this measure is independent of the source-receiver distance. Since the level of reverberation in a signal is sensitive to the source and sensor positions as well as the room characteristics, it is often desirable to estimate measures that take account of these factors. Two such measures are clarity index (C_{50}) and direct-to-reverberation ratio (DRR)) [1].

Room acoustic properties can be determined from the Room Impulse Response (RIR). However, in many real situations the RIR is unavailable so any acoustic parameters must be estimated non-intrusively from the reverberant signal.

Room acoustic parameters can be used to estimate the perceived quality [2] or intelligibility [3] of reverberant recordings, and also to predict speech recognition performance [4] [5] [6]. Furthermore, a wide range of de-reverberation algorithms employ room acoustic parameter information to suppress reverberation [1] [6] [7] [8] [9]. Hence, the measurements related to room acoustics have a key role in many situations and therefore it is of strong importance to find an accurate method to estimate these parameters from speech signals.

A baseline method for clarity index is [10]. The authors propose two methods to estimate different room acoustic parameters from speech and music signals. The first one uses an artificial neural network with 40 features extracted by sampling the power spectrum density (PSD) estimation of the sum of the Hilbert envelopes computed for certain frequency bands. The second method finds the cleanest sections of free decays in the signal to estimate with ML approach the decay curve and average this estimation to obtain the final estimator. We have chosen the first approach as the baseline method due to its higher performance for speech signals [10]. Its original form measures C_{80} but it has been modified in the current paper to measure C_{50} in order to be compatible with our evaluation. Falk and Chan [11] proposed a method to compute DRR in the modulation domain. The algorithm is based on the observation that low modulation frequency energy (below 20Hz) is barely affected by the reverberation level whilst high modulation frequency energy increases with the reverberation level. The overall ratio can be linearly mapped to estimate DRR parameter. Additionally, a similar idea is applied to estimate T_{60}. Many more methods are available for T_{60} estimation including [12] based on the decay rate from a statistical model of the sound decay or [13] based on spectral decay distributions.

In this paper we propose a Non-Intrusive Room Acoustic (NIRA) estimation method to estimate the room acoustic parameters based on a Classification And Regression Tree (CART) which is created with a set of features computed from the reverberant signals. The paper is organised as follows: Section 2 presents the motivation of this work. In Section 3 the method proposed to estimate room acoustic parameters is presented. The evaluation database and the results obtained are detailed in Section 4 and 5 respectively. Finally, in Section 6 the conclusions are drawn.

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° ITN-GA-2012-316969.

2. ROOM ACOUSTIC PARAMETERS

In this section we provide evidence to support the usefulness of C_{50} estimation by showing that C_{50} is the best correlated to phoneme recognition performance of a set of common acoustic parameters. A number of parameters have been presented in the literature to characterize the effect of reverberation [1]. We measure the recognition performance using phoneme accuracy rate ($\mathcal{A}_{ph}$) to avoid the influence of language model or dictionary rules in the results and thus characterize the acoustic distortions more specifically. The $\mathcal{A}_{ph}$ is defined as

$$\mathcal{A}_{ph} = \frac{N - D - I - S}{N} = \frac{H - I}{N}, \tag{1}$$

where N represents the total number of phones recognized, H is the number of correctly recognized phones, D is the number of deletions, S is the number of substitutions and I the number of insertions.

2.1. Phone recognition configuration

A standard phone recognizer was implemented using HTK [14]. The TIMIT database [15] was used to train the models following the phone folding proposed in [16] and excluding the 2 dialect sentences (SA). The testing database contains 168 TIMIT sentences equally distributed per dialect convolved with 140 RIRs simulated with the image method [17].

2.2. Room acoustic parameters evaluation

The acoustic parameters evaluated in this experiment are T_{60}, DRR, center time (Ts) [3], D_τ computed as [1]

$$D_\tau = 10\log_{10}\left(\frac{\sum_{n=0}^{N_\tau} h^2(n)}{\sum_{n=0}^{\infty} h^2(n)}\right) \text{dB}, \tag{2}$$

and C_τ defined as

$$C_\tau = 10\log_{10}\left(\frac{\sum_{n=0}^{N_\tau} h^2(n)}{\sum_{n=N_\tau+1}^{\infty} h^2(n)}\right) \text{dB}, \tag{3}$$

where τ represents a variable ranging from 0.1 ms to 1 s, N_τ is an integer number of samples corresponding to τ seconds and $h(n)$ is the RIR.

2.3. Results

Table 1 shows the absolute Pearson correlation coefficients obtained for each measurement. As expected from reference [6], it can be seen that T_{60} is not well correlated with $\mathcal{A}_{ph}$. One reason for this behaviour is the independence of this measure with the source-receiver distance which is a key factor in the degradation of the clean speech. Figure 1 plots the correlation of D_τ with 10^{-1} ms $\leq \tau \leq 10^3$ ms. The correlation of this parameter tends to decrease when τ increases. At $\tau = 50$ ms, this is D_{50}, the correlation coefficient is low. Figure 1 also shows the correlation of $\mathcal{A}_{ph}$ with C_τ, where it can be seen that the maximum correlation is in the range 50 ms$\leq \tau \leq$100 ms. The correlation coefficient for C_{50} is the highest for the set evaluated. These results are consistent with [6]. Results are also given for the correlation of each of the parameters with the speech quality score PESQ [18].
We can conclude that C_{50} is the most strongly correlated acoustic parameter in our tests (0.80 and 0.96 correlation with $\mathcal{A}_{ph}$ and PESQ respectively) and therefore, in the following, we propose a method to estimate C_{50}.

	T_{60}	DRR	Ts	D_{50}	C_{50}
$\mathcal{A}_{\mathbf{ph}}$	0.6449	0.6937	0.7868	0.642	**0.8044**
PESQ	0.6997	0.8352	0.9549	0.715	**0.9571**

Table 1. Correlation comparison of $\mathcal{A}_{ph}$ and PESQ with different acoustic parameters.

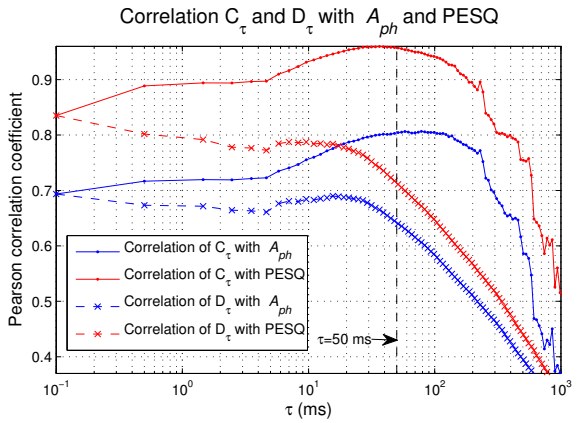

Fig. 1. Chart of $\mathcal{A}_{ph}$ and PESQ correlation coefficients obtained with C_τ and D_τ for τ between 0.1 ms and 1 s.

3. NIRA METHOD

3.1. Feature extraction

It begins with a short time segmentation of the signal into 20 ms non-overlapping frames using a Hanning window. Short-term features ($\phi_{1:73}$ in Table 2) are extracted per active speech frame. The mean (μ), variance (σ^2), skewness (s) and kurtosis (k) of $\phi_{1:73}$ are computed per utterance and appended to the long-term features ($\phi_{74:90}$ in Table 2).
We include a novel feature in this vector based on the Hilbert phase computed as

$$\theta_H(t) = \tan^{-1}(s_i(t)/s_r(t)), \tag{4}$$

where $s_r(t)$ represents the signal to be analysed and $s_i(t)$ its Hilbert transform defined as

$$s_i(t) = \mathcal{H}(s_r(t)) = \frac{1}{\pi t}\int_{-\infty}^{+\infty} \frac{s_r(\tau)}{t-\tau} d\tau. \tag{5}$$

Description	Feature	Rate of change of feature
Line Spectrum Frequency (LSF) computed from 10 first Linear Prediction Coding coefficients	$\phi_{1:10}$	$\phi_{11:20}$
Zero crossing rate, Speech variance, Pitch period and iSNR	$\phi_{21:24}$	$\phi_{25:28}$
Variance and dynamic range of Hilbert envelope	$\phi_{29:30}$	$\phi_{31:32}$
Spectral flatness and spectral centroid of the Power Spectrum of long term Deviation (PLD)	$\phi_{33:34}$	$\phi_{35:36}$
Spectral dynamics of the Power Spectrum of long term Deviation (PLD)	ϕ_{37}	-
12th order Mel-Frequency Cepstral Coefficients with delta and delta-delta	$\phi_{38:73}$	-
16 frequency bins of the Long Term Average Speech Spectrum (LTASS) deviation	$\phi_{74:89}$	-
Unwrapped Hilbert phase	ϕ_{90}	-

Table 2. NIRA features: $\phi_{1:73}$ are short-term features computed by frame, whose statistics are used in the CART, and $\phi_{74:90}$ are long-term features calculated over the entire utterance.

This parameter was shown to be a key factor for sound localization [19]. It is know that reverberant environments have the effect of diffusing the sound source [20], hence Hilbert phase can provide useful and relevant information indicative of the reverberation level. Figure 2 shows the behaviour of the unwrapped Hilbert phase for three different reverberant conditions applied to the same speech file. The slope of this phase increases with the reverberation level and thus it can be used for estimating this room acoustic parameter.

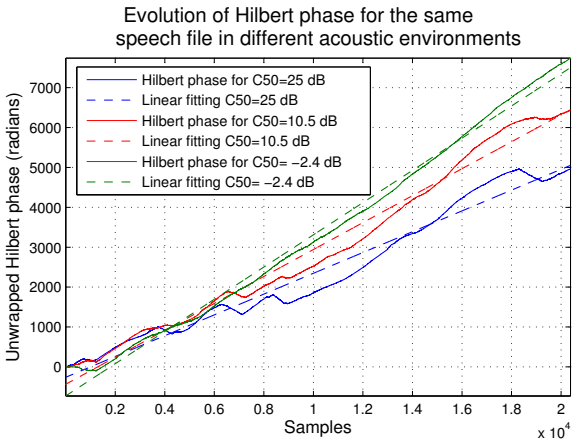

Fig. 2. Unwrapped Hilbert phase and linear regressions.

3.2. CART classifier

A CART regression tree [21] is used to model the 309 element feature vectors. The CART was trained with the database described in Section 4.1 to output the C_{50} estimate as a continuous variable. The advantage of using the CART approach is its fast output prediction and its human readable structure.

4. PERFORMANCE EVALUATION

Here we present the database and evaluation metrics used to compare NIRA with the method of [10], which is based on statistical machine learning using envelope spectrum features and serves the role as a baseline approach from the state-of-the-art, against which our method can be evaluated.

4.1. Database

The database was created to evaluate the C_{50} estimators comprising a training set used to build the different models; and an independent testing set.

The training set of reverberant speech is created with 32 clean recordings extracted from training partition of TIMIT [15]. The clean speech is convolved with 140 RIRs generated with the image method [17]. Figure 3 displays the distribution of the RIRs according to their C_{50}. Babble noise is added at different SNRs from 0 dB to 30 dB in steps of 5 dB as well as SNR = ∞ dB.

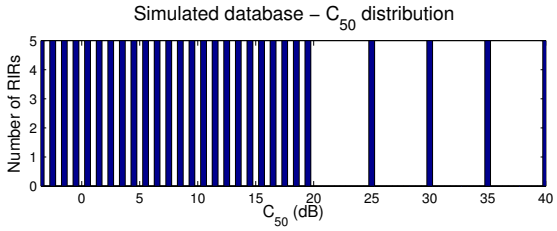

Fig. 3. C_{50} distribution for simulated RIRs.

The test set is formed of 32 clean speech files from TIMIT convolved with a group of 140 RIRs created with image method. The C_{50} distribution of the impulse responses follows also the previous distribution (represented in Fig. 3) but are totally independent of the training set. Real impulse responses, obtained from MARDY database [22], are also included for testing. The distribution of these RIRs in terms of C_{50} is shown in Fig. 4. Furthermore, babble noise is also included in the test set at two arbitrarily chosen SNR levels, in this case is 3 dB and 17 dB. In total, 35.57 hours of reverberant data were used for training the CART model and 86.65 hours for testing.

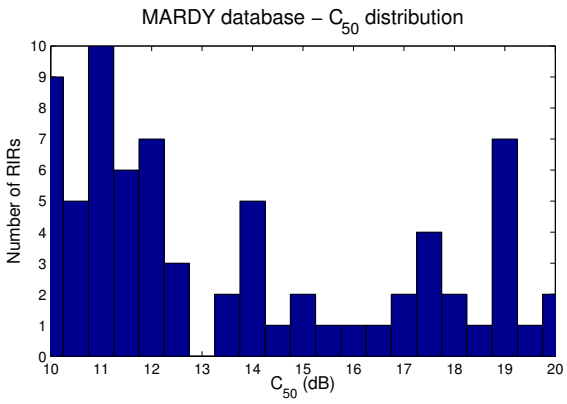

Fig. 4. C_{50} distribution for measured RIRs.

4.2. Figures of merit

The evaluation is performed using two performance figures: the Pearson coefficient computed as

$$\rho = \frac{\sum_{n=1}^{N}(y_n - \overline{y})(x_n - \overline{x})}{\sqrt{\sum_{n=1}^{N}(y_n - \overline{y})^2 \sum_{n=1}^{N}(x_n - \overline{x})^2}}; \qquad (6)$$

and the root mean square deviation calculated as

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{n=1}^{N}(y_n - x_n)^2}{N}}\ \mathrm{dB}, \qquad (7)$$

where $\overline{x}$ is the average of N ground truth scores x_i, and $\overline{y}$ represents the mean of N estimated scores y_i.

5. RESULTS

In this section we compare our C_{50} estimator with the method of [10] in terms of ρ and RMSD. These results are summarized in Table 3 for simulated ('Sim.') and Real RIRs.

		ρ		RMSD (dB)	
RIR	**SNR**	**NIRA**	**Baseline**	**NIRA**	**Baseline**
Sim.	3	0.70	0.43	8.74	11.23
	17	0.83	0.51	6.32	10.41
	∞	0.85	0.52	6.07	10.47
Real	3	0.44	0.18	9.85	9.53
	17	0.62	0.20	7.6	9.00
	∞	0.57	0.20	5.52	9.05

Table 3. Performance comparison of the C_{50} estimator proposed against the baseline including babble noise.

5.1. CART feature importance

Feature selection is automatically performed as an intrinsic step of the training processes of the CART model. For the database employed, 118 of the 309 available features were selected for the trained model. The 10 most important features are presented in order in the following vector ψ,

$$\psi = \{\mu\phi_{29},\ \mu\phi_{37},\ s\phi_{51}\ ,\sigma^2\phi_{63}\ ,\phi_{78}\ ,\cdots \\ \sigma^2\phi_{33}\ ,\mu\phi_{24}\ ,\mu\phi_{1}\ ,\phi_{90}\ ,\sigma^2\phi_{65}\}. \qquad (8)$$

5.2. Simulated room impulse responses

The first three rows in Table 3 show the evaluation metrics obtained with the simulated RIRs from which it can be seen that NIRA outperforms the envelope spectrum in every environment in terms of correlation and deviation. It can also be seen that both methods are robust to babble noise at the arbitrarily chosen test condition of SNR=17 dB. Nevertheless, at a high noise level of SNR=3 dB the performance of NIRA is reduced approximately by 2 dB RMSD.

5.3. Real room impulse responses

This set of RIRs creates the most challenging test because the real impulse response are not included in the training stage; the CART model was trained using only simulated RIRs.
The last three rows of Table 3 show the performance of both estimators for this case. NIRA outperforms the envelope spectrum method in every condition tested except for SNR=3 dB in terms of RMSD where both methods provide similar performance. It is worth noting that in this test, the range of C_{50} values is narrower compared to the simulated RIRs, which may cause lower deviations.

6. CONCLUSION

We have presented results that confirm other indications in the literature that C_{50} has the highest correlation with phone recognition rate compared to other room acoustic measurements (T_{60}, DRR, Ts, D_τ and C_τ).
Motivated by this finding, we have presented a non-intrusive C_{50} estimator (NIRA) based on training a CART with multiple features. A new feature was also proposed, employing the Hilbert phase and it was found to be one of the 10 most important features for the CART algorithm. This C_{50} estimator was compared with a baseline C_{50} estimator algorithm from the literature. Both methods were tested on a database comprising clean and noisy reverberant speech obtained with real and simulated RIRs. The best performance was achieved with NIRA which outperforms the baseline implemented for low SNR levels in by least 2.6 dB RMSD.

7. REFERENCES

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

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

[3] H. Kuttruff, *Room Acoustics*, Taylor & Francis, London, fifth edition, 2009.

[4] T. Fukumori, M. Morise, and T. Nishiura, "Performance estimation of reverberant speech recognition based on reverberant criteria RSR-D_n with acoustic parameters," in *Proc. INTERSPEECH*, 2010, pp. 562–565.

[5] A. Sehr, E. A. P. Habets, R. Maas, and W. Kellermann, "Towards a better understanding of the effect of reverberation on speech recognition performance," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010.

[6] A. Tsilfidis, I. Mporas, J. Mourjopoulos, and N. Fakotakis, "Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing," *Computer Speech & Language*, vol. 27, no. 1, pp. 380–395, 2013.

[7] R. Gomez and T. Kawahara, "Dereverberation based on wavelet packet filtering for robust automatic speech recognition," in *Proc. INTERSPEECH*, 2012.

[8] L. Couvreur, C. Ris, and C. Couvreur, "Model-based blind estimation of reverberation time: application to robust ASR in reverberant environments.," in *Proc. INTERSPEECH*, 2001, pp. 2635–2638.

[9] A.W. Mohammed, M. Matassoni, H. Maganti, and M. Omologo, "Acoustic model adaptation using piecewise energy decay curve for reverberant environments," in *Proc. of the 20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 365–369.

[10] P. Kendrick, T. J. Cox, F. F. Li, Y. Zhang, and J. A. Chambers, "Monaural room acoustic parameters from music and speech," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 278–287, 2008.

[11] T. H. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 978–989, 2010.

[12] H. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010, pp. 1–4.

[13] J. Eaton, N. D. Gaubitch, and P. A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[14] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*, Cambridge University Engineering Department, Cambridge, UK, 2006.

[15] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," Technical report, National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Dec. 1988.

[16] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[17] J. B. Allen and D. A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[18] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Feb. 2001.

[19] Z. M. Smith, B. Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, no. 6876, pp. 87–90, March 2002.

[20] B. Blesser and L.-R. Salter, *Spaces speak, are you listening? : experiencing aural architecture*, Cambridge, Mass. ; London : MIT Press, first edition, 2009.

[21] L. Olshen, Breiman J. H., Friedman R. A., and Charles J. Stone, "Classification and regression trees," *CRC Press*, 1984.

[22] J. Wen, N. D. Gaubitch, E. A. P. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," Paris, France, Sept. 2006.