EVALUATING THE NON-INTRUSIVE ROOM ACOUSTICS ALGORITHM WITH THE ACE CHALLENGE

Pablo Peso Parada[1*], Dushyant Sharma[1], Toon van Waterschoot[2], Patrick A. Naylor[3]

[1] Voicemail-To-Text Research, Nuance Communications Inc., Marlow, UK
[2] Dept. of Electrical Engineering (ESAT-STADIUS/ETC), KU Leuven, Belgium
[3] Department of Electrical and Electronic Engineering, Imperial College London, UK
{pablo.peso, dushyant.sharma}@nuance.com
toon.vanwaterschoot@esat.kuleuven.be,p.naylor@imperial.ac.uk

ABSTRACT

We present a single channel data driven method for non-intrusive estimation of full-band reverberation time and full-band direct-to-reverberant ratio. The method extracts a number of features from reverberant speech and builds a model using a recurrent neural network to estimate the reverberant acoustic parameters. We explore three configurations by including different data and also by combining the recurrent neural network estimates using a support vector machine. Our best method to estimate DRR provides a Root Mean Square Deviation (RMSD) of 3.84 dB and a RMSD of 43.19 % for T_{60} estimation.

***Index Terms*—** Reverberant speech, DRR estimation, T_{60} estimation

1. INTRODUCTION

Sound propagation from the source to the receiver placed in a room may follow multiple paths due to reflections from walls or objects in the enclosed space. This multipath propagation creates a reverberant sound which depends on the characteristics of the room and positions of both source and receiver. The reverberation time (T_{60}) characterizes the acoustic properties of an enclosed space and it is theoretically independent of the source-receiver distance. Alternative objective measurements such as Direct-to-Reverberation Ratio (DRR) or clarity index (C_{50}) [1] may be employed to take into account this dimension. The calculation of these measures of reverberation require an estimation of the Room Impulse Response (RIR), however in many real situations this information remains unavailable and these measures need to be non-intrusively estimated from the reverberant signal.

Several methods have been proposed to blindly estimate T_{60}. The method proposed by Löllmann et al. [2] estimates the decay rate from a statistical model of the sound decay using the *maximum likelihood* (ML) approach and then from this decay rate the method finds the ML estimate for T_{60}. The Eaton et al. [3] T_{60} estimator is based on spectral decay distributions. In this case the signal is filtered with uniform Mel-spaced filters and from the output of this filter bank the decay rate is computed by applying a least-square linear fit to the time-frequency log magnitude bins. The variance of the negative gradients in the distribution of decay rates is then mapped to T_{60} with a polynomial function. Falk and Chan [4] proposed a method to compute the reverberation time in the modulation domain. The algorithm is based on the idea 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 estimator is created with a Support Vector Regressor (SVR) and with the ratio of the average of low modulation frequency energy to different averages of high modulation frequency energy as the input features. In addition, the overall ratio can be mapped to estimate directly the DRR parameter. Kendrick et al. [5] compare two methods to estimate from speech and music signals different room acoustic parameters, mainly T_{60} and C_{80}. The first one uses an artificial neural network with 40 features extracted by sampling the power spectrum density 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. Although room acoustic parameters can be also estimated from multichannel recordings, such as T_{60} [6] or DRR [7], or per frequency bin [8], this paper focuses on the problem of single-channel full-band room acoustic parameter estimation.

These measures of reverberation have been applied to estimate the perceived quality [9] or intelligibility [10] of reverberant recordings. These were also shown to predict speech recognition performance [11] [12] [13] [14]. In addition to these applications, several de-reverberation algorithms use measures of reverberation to suppress reverberation in speech [1] [13][15] [16] [17], and so it is important to develop methods that estimate these measures directly from the reverberant signal.

We propose a non-intrusive (NIRA) method to estimate the room acoustic parameters based on extracting a number of per-frame features from the reverberant speech. A recurrent neural network is then employed to model the relationship between these features and the room acoustics parameters, i.e. DRR and T_{60}. This technique was tested on the single-channel configuration of the ACE challenge [18] organized by the IEEE Audio and Acoustic Signal Processing Technical Committee to compare different approaches to estimate DRR and T_{60}.

The remainder of the paper is organized as follows. Section 2 describes the method proposed in this work. In Section 3 the metrics used to evaluate the methods are introduced and results obtained on the ACE Challenge database are detailed in Section 4. Finally, in Section 5 the conclusions of this contribution 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.

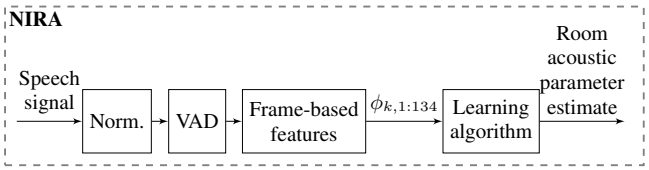

Figure 1: The NIRA method.

2. NIRA METHOD

The method shown in Fig. 1 computes a set of frame-based features from the signal using a window size of 20 ms and a 50% overlap. Non-speech frames are dropped out with a Voice Activity Detector (VAD) using P.56 method [19]. This estimator was originally proposed for estimating C_{50} from 8kHz speech signals in [20] and extended to 16kHz signals in [21]. In this work we have employed the latter configuration which estimates 134 frame-based features from the reverberant signal:

- Line Spectrum Frequency (LSF) features computed by mapping the first 20 linear prediction coefficients to the LSF representation and their rate of change.
- Zero-crossing rate and its rate of change.
- Speech variance and its rate of change.
- Pitch period estimated with the PEFAC algorithm [22] and its rate of change.
- Estimation of the importance-weighted Signal-to-Noise Ratio (iSNR) in units of dB and its rate of change.
- Variance and dynamic range of the Hilbert envelope and their rate of change.
- Three parameters extracted from the Power spectrum of the Long term Deviation (PLD): spectral centroid, spectral dynamics and spectral flatness. The PLD is calculated per frame using the log difference between the signal power spectrum and long term average speech spectrum. Their rate of change is also included.
- 12th order mean- and variance-normalized Mel-frequency cepstral coefficients computed from the fast Fourier transform with delta and delta-delta.
- Modulation domain features [23] derived from computing the first four central moments of the highest energy frequency band and its two adjacent modulation frequency bands.
- Deep scattering spectrum features are extracted from a scattering transformation applied to the signal [24].

These features are used to train a Bidirectional Long-Short Term Memory (BLSTM) [25] recurrent neural network to provide an estimate of DRR and T_{60} every 10 ms. The main motivation for using this architecture is that it can model temporal correlation such as reverberation due to its feedback connections. Alternative learning algorithms as classification and regression tree, linear regression or deep belief neural network have been investigated in the frame of C_{50} estimation however BLSTM showed a better performance [20]. Since ACE Challenge data assumes that the room acoustic properties remain unchanged within each utterance, only the temporal average for each utterance of all per frame estimations is considered.

Different architectures of the BLSTM[1] are explored with one to four layers including 64, 128 and 256 neurons per layer and a minibatch size of 25, 50, 100 and 200 samples. Three different configurations were explored using this framework which are described in the following subsections.

2.1. NIRAv1

This configuration is based on training the NIRA framework presented in Fig. 1 using only the ACE Challenge development database. All data from the different microphone configurations was split randomly into three parts: training set (*trainSet*), development set (*devSet*) and evaluation set (*evalSet*). The *trainSet* comprises 70% of the files in the ACE Challenge development database, whereas *devSet* and *evalSet* comprise 20 % and 10 % respectively. In this case, *trainSet* is used to train the model and *devSet* is employed to validate the model, then the selected model is the one that minimizes the estimation error in *devSet*.

2.2. NIRAv2

This configuration employs the NIRA framework shown in Fig. 1 trained on three different databases in order to introduce new data in the model which could generalize the model to a wider range of scenarios. In this case 60% of the files are extracted from the ACE Challenge development database, 20% of the files from the REVERB Challenge database and the remainder of the files are taken from a database created with TIMIT database [26] and real impulse responses from MARDY [27], SMARD [28], C4DM RIR [29] and REVERB Challenge [30] database. Similarly, *devSet* is created with the same proportions and from the same databases but the total number of files is 30% of *trainSet*.

2.3. NIRAv3

This configuration follows the structure shown in Fig. 2. It is based on training 4 different BLSTM models using different data: NIRAv1; NIRA_{α} using the whole REVERB Challenge development set; NIRA_{β} and NIRA_{γ} employing real and simulated RIRs respectively convolved with TIMIT database. The real RIRs are taken from MARDY, SMARD, C4MD and REVERB Challenge database, while the simulated RIRs are created with the randomized image method [31]. These 4 estimators are combined by averaging the per-frame estimations of each utterance and by training a SVR model [32] with the 4-dimensional estimate vector obtained from the individual estimators. The training data for this SVR is *devSet* from NIRAv1 and *evalSet* is used for validation purposes.

3. PERFORMANCE EVALUATION

All methods described in this paper are evaluated on the Acoustic Characterization of Environments (ACE) challenge [18]. This challenge provides a common framework where different approaches of estimating DRR and T_{60} can be directly compared. In addition to the box plots provided by the challenge to compare the different approaches, the algorithms are compared in this paper in terms of Root Mean Square Deviation (RMSD). This metric is computed for

[1] http://sourceforge.net/projects/currennt/

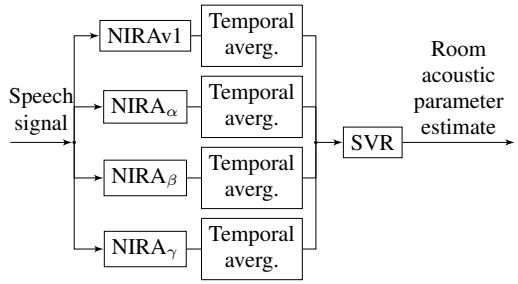

Figure 2: The NIRAv3 method for DRR and T_{60} estimation.

the DRR estimators as

$$\mathrm{RMSD}_{\mathrm{DRR}} = \sqrt{\frac{\sum_{n=1}^{N}(\widehat{\mathrm{DRR}}_{\mathrm{n}} - \mathrm{DRR}_{\mathrm{n}})^2}{N}}\ \mathrm{dB}, \qquad (1)$$

where $\mathrm{DRR}_{\mathrm{n}}$ and $\widehat{\mathrm{DRR}}_{\mathrm{n}}$ are the ground truth and the estimated DRR respectively of the n-th utterance and N is the total number of utterances.

On the other hand, the RMSD of the T_{60} estimators is calculated as

$$\mathrm{RMSD}_{\mathrm{T}_{60}} = \sqrt{\frac{\sum_{n=1}^{N}(100 \cdot (\widehat{\mathrm{T}_{60_{\mathrm{n}}}} - \mathrm{T}_{60_{\mathrm{n}}})/\mathrm{T}_{60_{\mathrm{n}}})^2}{N}}\ \%, \qquad (2)$$

where $\mathrm{T}_{60_{\mathrm{n}}}$ and $\widehat{\mathrm{T}_{60_{\mathrm{n}}}}$ are the ground truth and the estimated T_{60} respectively.

4. RESULTS

The evaluation results for the different approaches are shown in this section. These approaches are tested on two datasets: *evalSet* described in Section 2.1 and the ACE Challenge evaluation set.

4.1. Performance in *evalSet*

Table 1 shows the performance of the three approaches in terms of RMSD on the *evalSet* dataset introduced in Section 2.1. NIRAv1 and NIRAv3 show the best performance for DRR estimation and NIRAv2 the highest estimation error deviation. Figure 3 displays the box plot for the same dataset. NIRAv2 shows a wider interquartile range (IQR) and a negative bias which explains the higher RMSD value compared to the other two methods. Regarding T_{60} estimation, Tab. 1 indicates that the best approach is NIRAv1, whereas NIRAv3 provides the lowest performance mainly due to the bias and the wide IQR displayed in Fig. 4.

Method	$\mathbf{RMSD_{DRR}}$ **(dB)**	$\mathbf{RMSD_{T_{60}}}$ **(%)**
NRIAv1	0.64	3.18
NRIAv2	0.92	3.66
NRIAv3	0.63	7.15

Table 1: RMSD of the three approaches to estimate DRR and T_{60} using *evalSet* dataset.

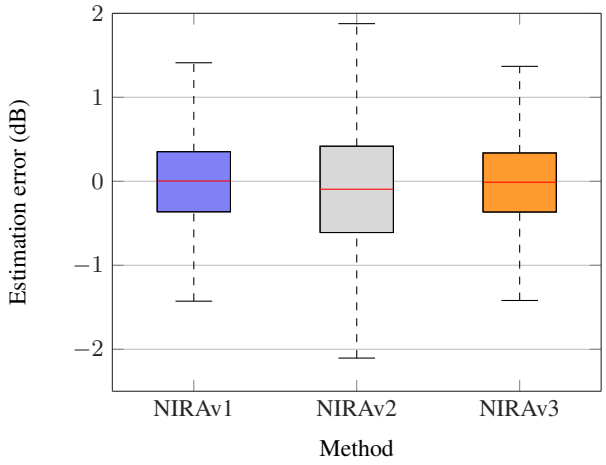

Figure 3: Distribution of the DRR estimation errors for each method using *evalSet*. The edges of the boxes indicate the lower and upper quartile range, while the horizontal lines inside the boxes represent the medians for each method. Moreover, the horizontal lines outside the boxes indicate the estimation error up to 1.5 times the interquartile range.

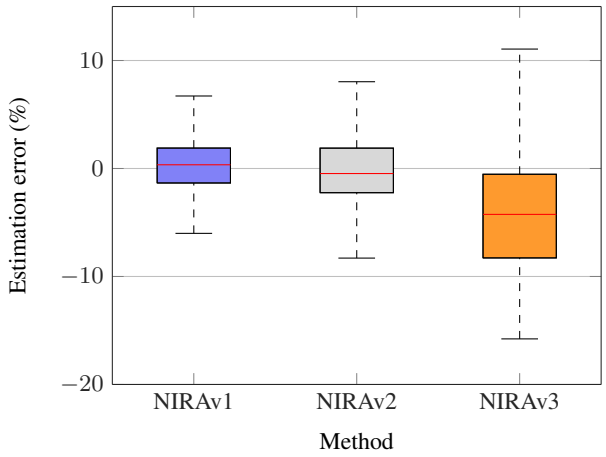

Figure 4: Distribution of the T_{60} estimation errors for each method using *evalSet*.

4.2. Performance in ACE Challenge Evaluation set

Table 2 shows the performance of the three approaches on the ACE Challenge evaluation dataset. NIRAv3 and NIRAv1 still provide the best performance when estimating DRR and T_{60} respectively on this dataset, however the deviations are considerably increased.

Figure 5 shows the distribution of the DRR estimation error for each method. The three methods present similar distributions, however NIRAv3 is less biased which is in accordance with the results displayed in Tab. 2. Figure 6 shows the box plot for each method proposed to estimate T_{60}. NIRAv3 presents the higher interquartile range and NIRAv1 the least biased estimation, which is reflected in the deviation shown in Tab. 2.

An analysis of the performance of the best approaches to estimate DRR and T_{60} is shown in Fig. 5 and 6 respectively for each

Method	$\mathbf{RMSD_{DRR}}$ **(dB)**	$\mathbf{RMSD_{T_{60}}}$ **(%)**
NRIAv1	3.87	43.19
NRIAv2	3.85	44.80
NRIAv3	3.84	44.18

Table 2: RMSD of the three approaches to estimate DRR and T_{60} using ACE Challenge evaluation set.

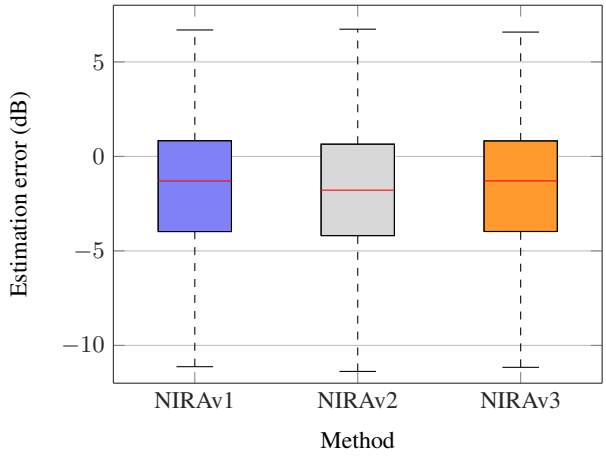

Figure 5: Distribution of the DRR estimation errors for each method using ACE Challenge evaluation dataset.

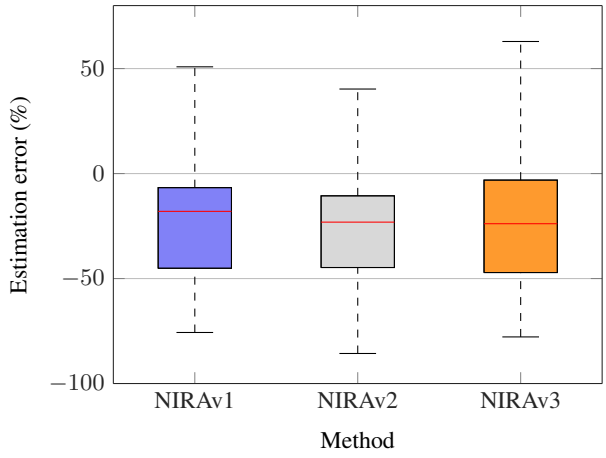

Figure 6: Distribution of the T_{60} estimation errors for each method using ACE Challenge evaluation dataset.

noise condition. These figures suggest that babble noise provides the lowest RMSD for DRR estimation whereas fan noise in the recordings brings higher DRR estimation errors. On the contrary, fan noise provides the lowest T_{60} deviation and babble noise brings the highest T_{60} estimation errors.

5. CONCLUSION

We have presented in this paper three data-driven approaches to estimate full-band DRR and T_{60} from single-channel reverberant speech. These approaches are based on training a BLSTM with

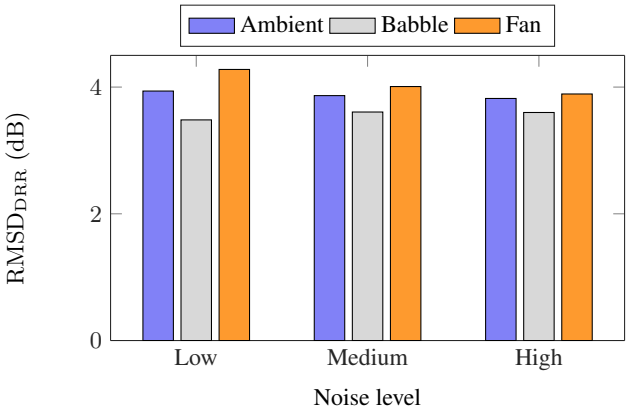

Figure 7: Performance of NIRAv3 estimating DRR on the ACE Challenge evaluation dataset for different noise conditions.

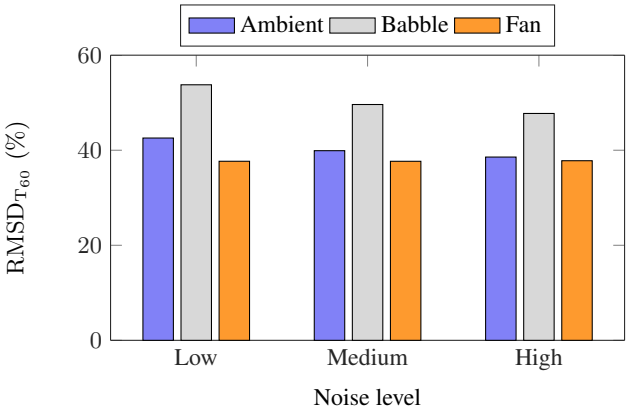

Figure 8: Performance of NIRAv1 estimating T_{60} on the ACE Challenge evaluation dataset for different noise conditions.

different datasets. Additionally, we explored the combination of these networks trained with different datasets by employing a SVR. The best DRR estimation performance was achieved with NIRAv3, $\mathrm{RMSD_{DRR}} = 3.84$ dB with IQR = 4.79 dB and median of -1.3 dB. This is based on training with different databases several BLSTMs and combining their individual time averaged estimations with a SVR. On the other hand, NIRAv1 provides the best T_{60} estimation performance, $\mathrm{RMSD_{T_{60}}} = 43.19$ % with IQR = 44 % and median of -23.88 %. This configuration is based on training a BLSTM employing only the ACE Challenge development dataset.

Moreover, the performance of these approaches was tested with 10 % of the ACE Challenge development files, not previously used in the training process, i.e. *evalSet*. The best performance of DRR and T_{60} was obtained with NIRAv3 and NIRAv1 respectively, as it occurs on ACE Challenge evaluation dataset. However, the deviations were considerably lower, $\mathrm{RMSD_{DRR}} = 0.63$ dB with IQR = 0.7 dB and median of -0.01 dB for DRR estimation and $\mathrm{RMSD_{T_{60}}} = 3.18$ % with IQR = 3.23 % and median of 0.34 % for T_{60} estimation, which may indicate an overfitting problem in the training process.

6. REFERENCES

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

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

[3] 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, pp. 161–165.

[4] 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.

[5] 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.

[6] B. Dumortier and E. Vincent, "Blind RT60 estimation robust across room sizes and source distances," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 5187–5191.

[7] E. Georganti, J. Mourjopoulos, and S. van de Par, "Room statistics and direct-to-reverberant ratio estimation from dual-channel signals," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4713–4717.

[8] C. S. J. Doire, M. Brookes, P. A. Naylor, D. Betts, C. M. Hicks, M. A. Dmour, and S. H. Jensen, "Single-channel blind estimation of reverberation parameters," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[9] J. M. F. del Vallado, A. A. de Lima, T. d. 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.

[10] H. Kuttruff, *Room Acoustics*, 5th ed. London: Taylor & Francis, 2009.

[11] 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.

[12] 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. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, 2010.

[13] 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.

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

[15] R. Gomez and T. Kawahara, "Dereverberation based on wavelet packet filtering for robust automatic speech recognition," in *Proc. INTERSPEECH*, 2012, pp. 1243–1246.

[16] 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.

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

[18] J. Eaton, A. H. Moore, N. D. Gaubitch, and P. A. Naylor, "The ACE Challenge - corpus description and performance evaluation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

[19] ITU-T, *Objective Measurement of Active Speech Level*, International Telecommunications Union (ITU-T) Recommendation P.56, Mar. 1993.

[20] P. Peso Parada, D. Sharma, J. Lainez, D. Barreda, P. A. Naylor, and T. van Waterschoot, "A single-channel non-intrusive C50 estimator with application to reverberant speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, 2015, submitted for publication.

[21] P. Peso Parada, D. Sharma, P. A. Naylor, and T. van Waterschoot, "Reverberant speech recognition exploiting clarity index estimation," *EURASIP Journal on Advances in Signal Processing*, 2015, accepted for publication.

[22] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2011, pp. 451–455.

[23] Y. Wang and M. Brookes, "Speech enhancement using a modulation domain Kalman filter post-processor with a Gaussian mixture noise model," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 7024–7028.

[24] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, Aug 2014.

[25] F. Weninger, J. Bergmann, and B. Schuller, "Introducing CURRENNT–the Munich open-source CUDA RecurREnt Neural Network Toolkit," *Journal of Machine Learning Research*, vol. 15, 2014.

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

[27] J. Wen, N. D. Gaubitch, E. A. P. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, 2006.

[28] J. K. Nielsen, J. R. Jensen, S. H. Jensen, and M. G. Christensen, "The single- and multichannel audio recordings database (SMARD)," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sep. 2014, pp. 40–44.

[29] R. Stewart and M. Sandler, "Database of omnidirectional and B-format room impulse responses," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2010, pp. 165–168.

[30] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4.

[31] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 4, pp. 774–786, April 2015.

[32] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.