KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

| | |
|---|---|
| **Citation/Reference** | Pablo Peso Parada, Dushyant Sharma, Toon van Waterschoot, and Patrick A. Naylor<br>**Confidence measures for non-intrusive estimation of speech clarity index**<br>*J. Audio Eng. Soc.*, vol. 65, no. 1/2, pp. 90-99, Jan./Feb. 2017. |
| **Archived version** | Author manuscript: the content is identical to the content of the submitted paper, but without the final typesetting by the publisher |
| **Published version** | https://doi.org/10.17743/jaes.2016.0058 |
| **Journal homepage** | http://www.aes.org/journal/ |
| **Author contact** | toon.vanwaterschoot@esat.kuleuven.be<br>*+ 32 (0)16 321927* |
| **IR** | ftp://ftp.esat.kuleuven.be/pub/SISTA/vanwaterschoot/abstracts/16-193.html |

*(article begins on next page)*

Journal of the Audio Engineering Society
Vol. 65, No. 1/2, January/February 2017 (© 2017)
DOI: https://doi.org/10.17743/jaes.2016.0058



# Confidence Measures for Nonintrusive Estimation of Speech Clarity Index

**PABLO PESO PARADA**[1], **DUSHYANT SHARMA**[2],
(pablo.peso@nuance.com) (dushyant.sharma@nuance.com)

**TOON VAN WATERSCHOOT,**[3] *AES Associate Member*, **AND PATRICK A. NAYLOR,**[3] *Associate Member*
(toon.vanwaterschoot@esat.kuleuven.be) (p.naylor@imperial.ac.uk)

[1]*Nuance Communications Inc., Wethered House, Pound Lane, SL7 2AF Marlow, UK*
[2]*Dept. of Electrical Engineering (ESAT-STADIUS/ETC), KU Leuven, Kasteelpark Arenberg, 3001 Leuven, Belgium*
[3]*Dept. of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ UK*

We present a confidence measure based on prediction intervals for a non-intrusive method to estimate the clarity index ($C_{50}$) of reverberant speech. The method employed to estimate $C_{50}$ is a data driven approach. Multiple features are extracted from a reverberant speech signal and these are then used to train a bidirectional long short-term memory model that maps from the feature space into the target $C_{50}$ value. The prediction intervals are derived from the standard deviation of the per frame $C_{50}$ estimates and the confidence measures are obtained by normalizing these prediction intervals. These confidence measures are shown to provide a high correlation coefficient with the absolute $C_{50}$ estimation errors, i.e., $-0.95$ when applying conditional averaging. The performance of the prediction intervals and confidence measure are shown to be consistent among different noisy reverberant environments.

## 0 INTRODUCTION

Sound propagates from a source to a receiver placed in a room following multiple paths due to reflections from walls or objects in the room. This multipath propagation produces a reverberant sound, the characteristics of which depend on the room properties and the positions of both source and receiver. The reverberation time ($T_{60}$) indicates an acoustic property of a room and is independent of the positions of both source and receiver in a diffuse reverberant field. $T_{60}$ is defined as the time needed for the sound pressure level in the room to drop 60 dB after the acoustic excitation ceases [1]. Different objective measurements such as Direct-to-Reverberation Ratio (DRR) [2] or clarity index ($C_{50}$) [2] may be employed to characterize the level of reverberation associated with a particular scenario. The computation of these measures requires an estimation of the Room Impulse Response (RIR). However, in many real scenarios this information is unavailable and these measures need to be non-intrusively estimated from the reverberant signal.

Several methods have been proposed for non-intrusive estimation of $T_{60}$ [3–5]. Kendrick et al. [6] compare two methods to estimate from speech and music signals different room acoustic parameters, specially $T_{60}$ and $C_{80}$. The first method finds the cleanest sections of free decays in the signal to estimate with a maximum-likelihood (ML) approach the decay curve and averages this estimation to obtain the final estimator. The second algorithm uses an artificial neural network with 40 features extracted by sampling the power spectral density estimation of the sum of the Hilbert envelopes computed for certain frequency bands. This method was adapted in [7] to compute $C_{50}$ instead of $C_{80}$ for comparison purposes. Although room acoustic parameters can also be estimated from multichannel recordings, such as $T_{60}$ [8] or DRR [9], or per frequency bin [10], this paper focuses on the problem of single-channel full-band $C_{50}$ estimation due to the high correlation achieved with speech recognition performance compared to other measures of reverberation [11].

Estimates of room acoustics parameters have a number of applications, for example, to estimate the perceived quality [12] or intelligibility [2] of reverberant recordings, to de-reverberate speech signals [1] or to perform reverberant speech recognition [13]. In addition, information about the confidence of the estimator can be important in many situations in order to quantify the risk of applying the estimate in an application.

In this work we use $C_{50}$ to characterize reverberation in the signal because it was shown to be highly correlated

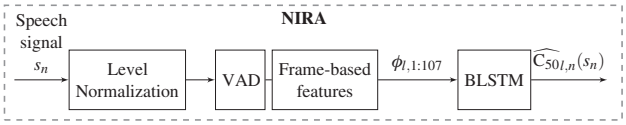



Fig. 1. The NIRA method.

with the speech recognition performance compared to other measures of reverberation [7, 14]. The $C_{50}$ parameter is computed as follows,

$$C_{50} = 10\log_{10}\left(\frac{\sum_{n=0}^{N_{50}} h^2(n)}{\sum_{n=N_{50}+1}^{\infty} h^2(n)}\right) \text{dB}, \quad (1)$$

where $h(n)$ is the RIR sampled at 8 kHz and $N_{50}$ is an integer number of samples corresponding to 50 milliseconds after the time arrival of the direct path.

This paper now extends an earlier version of the work presented in [15] including an analysis of the $C_{50}$ estimator per $C_{50}$ ranges and proposing a confidence measure based on prediction intervals. This technique was tested on a single-channel database created with simulated and real RIRs. Unlike the prediction intervals, confidence measures provide a bounded metric from 0 to 1 indicating how confident is the $C_{50}$ estimate.

For the sake of clarity, we describe in the following bullet points four different $C_{50}$ notations employed throughout this paper to refer to different approaches of computing $C_{50}$:

- $C_{50n}$ is the $C_{50}$ computed directly from the RIR used to create the reverberant signal $s_n$ employing Eq. (1).
- $C_{50n}(s_n)$ is the hypothesized $C_{50}$ in the reverberant signal $s_n$. This is effectively the $C_{50n}$ with the speech spectrum bias due to spectral coloration of $s_n$ and it is employed to described the prediction intervals in Sec. 2.
- $\widehat{C50}_{l,n}(s_n)$ is the $C_{50}$ estimated at frame $l$ from the reverberant signal $s_n$ which is obtained with the non-intrusive room acoustic (NIRA) method described in Sec. 1.
- $\widehat{C50}_{n}(s_n)$ is the $C_{50}$ estimated per utterance from the reverberant signal $s_n$. This estimate is computed as the average of the per frame estimates $\widehat{C50}_{l,n}(s_n)$ for the given reverberant signal $s_n$.

The remainder of the paper is organized as follows. Secs. 1, 2, and 3 introduce the methods proposed in this work to estimate $C_{50}$, the prediction intervals, and the confidence measures respectively. In Sec. 4 the metrics used to evaluate the methods are described and results are presented in Sec. 5. Finally, in Sec. 6 the conclusions of this study are drawn.

## 1 NON-INTRUSIVE ROOM ACOUSTIC METHOD

Fig. 1 outlines the NIRA method that has recently been detailed in [11]. This method computes a set of frame-based features from a signal sampled at 8 kHz using a window size of 20 ms and a frame increment of 10 ms. A Voice Activity Detector (VAD) [16] is employed to discard non-speech frames. The following 107 frame-based features are computed from the reverberant signal based on [7] in addition to novel modulation domain and deep scatter features:

- Line Spectrum Frequency (LSF) features computed by mapping the first 10 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 [17] 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 [18] 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 extracted from a scattering transformation applied to the signal [19].

These features are used to train a Bidirectional Long-Short Term Memory (BLSTM) [20] model to provide an estimate of $C_{50}$ every 10 ms. The main motivation for using a BLSTM is that it can model long temporal correlations present in reverberation with its feedback connections and memory cells. Alternative learning algorithms such as classification and regression tree, linear regression or deep belief neural network have been investigated for use in $C_{50}$ estimation, however BLSTM showed a better performance in our experiments even against a $C_{50}$ estimator baseline [11]. Since we assume that the room acoustic properties remain unchanged within each utterance, the $\widehat{C50}_{n}(s_n)$ estimate of the $n$th utterance $s_n$ is computed as the mean of the per frame estimates $\widehat{C50}_{l,n}(s_n)$ for that utterance:

$$\widehat{C50}_{n}(s_n) = \frac{1}{L}\sum_{l=1}^{L} \widehat{C50}_{l,n}(s_n) \quad \text{dB}, \quad (2)$$

where $L$ is the number of frames.

Different architectures of the BLSTM[1] are explored with one to four layers including 64, 128, and 256 neurons per layer and minibatch (i.e., number of utterances considered to update the weights of the neural network) size of 25, 50, 100, and 200 samples.

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

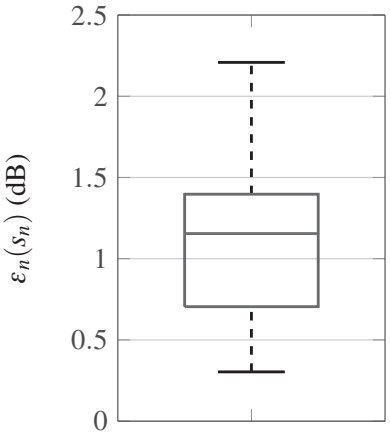



Fig. 2. Boxplot of the $\varepsilon_n(s_n)$ obtained with different utterance $s_n$ using the same RIR.

## 2 PREDICTION INTERVALS

A prediction interval comprises an upper and lower bounds quantifying the likely uncertainty in the estimation. Wide prediction intervals indicate high level of uncertainty [21].

The NIRA method outputs $\widehat{C_{50n}}(s_n)$, an estimation of the $C_{50n}$ value for the $n$th utterance $s_n$ using Eq. (2). As described in the introduction, it is also important to know the accuracy of each estimate. In this section we describe a method to compute the prediction interval (PI) of each utterance employing the per frame $C_{50}$ estimate, $\widehat{C_{50l,n}}(s_n)$.

In practice, the ground truth $C_{50n}$ estimate, computed directly from the RIR may differ from the hypothesized $C_{50n}(s_n)$ in the reverberant signal. This difference is caused by several factors. One of these is due to the spectrum of the speech signal $s_n$ which is colored and therefore only some frequencies of the RIR are excited in the reverberant signal. Therefore $C_{50}(s_n)$ will differ by the error $\varepsilon_n(s_n)$ from $C_{50n}$ and can be written as follows,

$$C_{50n} = C_{50n}(s_n) + \epsilon_n(s_n) \quad \text{dB}. \tag{3}$$

Fig. 2 shows a boxplot of the error $\varepsilon_n(s_n)$ obtained for 24 different TIMIT utterances and employing one recorded RIR from the MARDY database [22]. This error is computed as the average of the per frame errors where $C_{50n}(s_n)$ per frame is derived from the RIR achieved after convolving the given speech frame with the original recorded RIR. The frame size employed is 20 ms with a 50% overlap and the frames containing silence are discarded. The magnitude of the error $\varepsilon_n(s_n)$ obtained is on average 1.13 dB.

Eq. (3) can be rewritten as

$$C_{50n} - \widehat{C_{50n}}(s_n) = (C_{50n}(s_n) - \widehat{C_{50n}}(s_n)) + \epsilon_n(s_n) \quad \text{dB}. \tag{4}$$

Prediction intervals deal with Eq. (4), which is related to $P(C_{50n}|\widehat{C_{50n}}(s_n))$ [21]. There are two sources of uncertainty in Eq. (4): first the standard deviation $\sigma_v$ due to data limitations, e.g., the spectrum is colored, which is associated to the error $\varepsilon_n(s_n)$, and second the standard deviation $\sigma_m$ due to model limitations causing estimation errors associated with the magnitude $C_{50n}(s_n) - \widehat{C_{50n}}(s_n)$. Considering these errors to be statistically uncorrelated, we can write

$$\sigma_{TOTAL} = \sigma_v + \sigma_m \quad \text{dB}. \tag{5}$$

Prediction intervals are based on $\sigma_{TOTAL}$ [23]. In this paper we propose to estimate $\sigma_{TOTAL}$ for the $n$th utterance ($\sigma_{TOTAL,n}$) from the $L$ per frame estimates $\widehat{C_{50l,n}}(s_n)$ as

$$\sigma_{TOTAL,n} = \sqrt{\frac{1}{L}\sum_{l=1}^{L}(\widehat{C_{50l,n}}(s_n) - \widehat{C_{50n}}(s_n))^2} \quad \text{dB}.$$

The lower ($L_n$) and upper ($U_n$) bounds of the PI for the $n$th utterance are computed as

$$L_n = \widehat{C_{50n}}(s_n) - k \cdot \sigma_{TOTAL,n} \quad \text{dB}, \tag{6}$$

$$U_n = \widehat{C_{50n}}(s_n) + k \cdot \sigma_{TOTAL,n} \quad \text{dB}, \tag{7}$$

where $k$ is a tuning parameter that defines the width of the intervals. Fig. 3 shows different PIs computed for a given $\widehat{C_{50l,n}}(s_n)$ using several values of $k$.

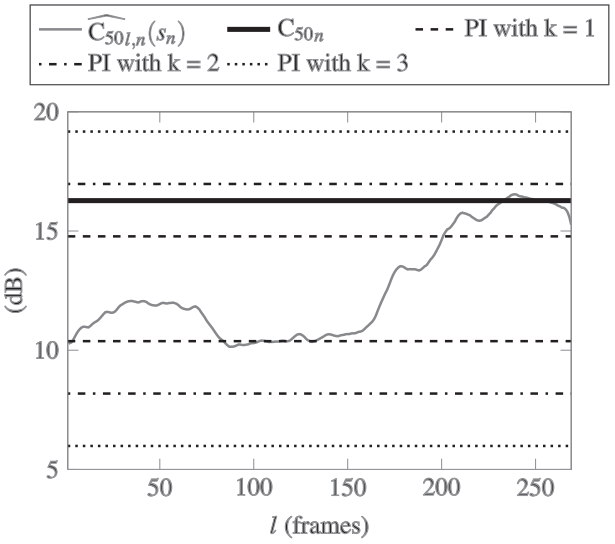



Fig. 3. Different PIs depending on $k$ for a given utterance of the development set.

## 3 CONFIDENCE MEASURE

Prediction intervals provide an indication of the accuracy of the estimates, however their values are unbounded and they are related to a $C_{50}$ range. In this section a confidence measure that provides an accuracy metric bounded from 0 to 1 is proposed. This confidence measure is based on normalizing the prediction interval width, $U_n - L_n$, by the total $C_{50}$ range observed in the training dataset, $R_{tr}$. Thus, the confidence measure $CM_n$ for the $n$th utterance is computed as

$$CM_n = \max\left\{1 - \frac{U_n - L_n}{R_{tr}}, 0\right\}, \tag{8}$$

where $U_n - L_n = 2 \cdot k \cdot \sigma_{TOTAL,n}$. The normalized width $\frac{U_n - L_n}{R_{tr}}$ is subtracted from 1 in order to make the confidence measure directly proportional to the accuracy level and the

maximization is carried out to ensure that the lower bound of the confidence measure is 0.

## 4 EXPERIMENTAL SETUP

### 4.1 Database

Three data sets are employed: the training set used to train the $C_{50}$ estimation method; the development set used to find the optimal model which minimizes the estimation error in this set and to determine the factor $k$ that is used to compute the prediction intervals in Eq. (6) and Eq. (7); and the evaluation set used only to evaluate the $C_{50}$ estimator and the prediction intervals. All the utterances, RIRs, and noise signals are different for each set and are sampled at 8 kHz.

#### 4.1.1 Training Set

A total of 32 utterances are selected randomly from the training set of the TIMIT [24] database ensuring that 2 male and 2 female speakers are included for each dialect and excluding "SA sentences." These speech signals are convolved with 192 simulated RIRs generated with the randomized image method [25] which are carefully selected to create a set of RIRs with a uniformly distributed $C_{50}$ in the interval [–3,28] dB. White noise and babble noise from the NOISEX corpus [26] are added to the reverberant speech at SNRs of 0 dB to 30 dB in steps of 5 dB where the speech power is computed using P.56 [16].

#### 4.1.2 Development Set

This set is created following the training set configuration using 16 utterances and 64 RIRs. Neither of the speech signals nor RIRs of this set are included in the training set.

#### 4.1.3 Evaluation Set

One utterance of each TIMIT core set speaker is included, excluding SA sentences, resulting in 24 sentences. Babble noise and white noise are also added in the evaluation set at 6 different SNR levels: 2 dB, 7 dB, 12 dB, 17 dB, 22 dB, 27 dB. Both simulated and real (measured) RIRs are included in this set employing four different databases to build the real RIR set: MARDY [22]; REVERB challenge [27]; C4DM RIR [28]; and SMARD [29]. The same selection procedure applied to simulated RIRs is employed in this case to build a set of RIRs with a uniform distribution of $C_{50}$ in the interval [–3,28] dB.

This evaluation set is divided into 26 subsets as outlined in Table 1 that are evaluated independently to assess the performance of $C_{50}$ estimation and the prediction intervals and to provide insights into the impact of different types and levels of noise on the final performance.

Table 1. Subsets of the evaluation set regarding RIR type, noise type, and SNR level. In all cases, the same 24 utterances are convolved with 160 RIRs. Therefore each subset comprises 3840 files (approximately 3.6 hours).

| RIR type | Noise type | SNR level | Name |
|---|---|---|---|
| Simulated | None | – | SimInf |
| | Babble / White | 2 | SimBA2 / SimWN2 |
| | | 7 | SimBA7 / SimWN7 |
| | | 12 | SimBA12 / SimWN12 |
| | | 17 | SimBA17 / SimWN17 |
| | | 22 | SimBA22 / SimWN22 |
| | | 27 | SimBA27 / SimWN27 |
| Real | None | – | RealInf |
| | Babble / White | 2 | RealBA2 / RealWN2 |
| | | 7 | RealBA7 / RealWN7 |
| | | 12 | RealBA12 / RealWN12 |
| | | 17 | RealBA17 / RealWN17 |
| | | 22 | RealBA22 / RealWN22 |
| | | 27 | RealBA27 / RealWN27 |

### 4.2 Evaluation Metrics

The $C_{50}$ estimator method described in this paper is evaluated as described above and the results obtained are shown using box plots of the estimation error, which is computed for the $n$th utterance as $\widehat{C_{50n}}(s_n) - C_{50n}$. Additionally, the Root Mean Square Deviation (RMSD) is employed to compare the performance per $C_{50}$ band

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\widehat{C_{50n}}(s_n) - C_{50n}\right)^2}, \tag{9}$$

where $N$ is the total number of utterances.

The estimated prediction intervals are evaluated using the Prediction Interval Coverage Probability (PICP) and the Normalized Mean Prediction Interval Width (NMPIW) metrics [21]. The PICP measures the percentage of times the ground truth $C_{50n}$ lies between the upper and lower bound of the estimated PI:

$$\mathrm{PICP} = \frac{1}{N}\sum_{n=1}^{N} c_n \cdot 100 \quad \%,$$

where

$$c_n = \begin{cases} 1, & C_{50,n} \in [L_n, U_n], \\ 0, & C_{50,n} \notin [L_n, U_n]. \end{cases}$$

The PICP provides information about the accuracy of the PI, however additional information regarding the width of the PIs is needed since high PICP can be achieved by using wide intervals. The width of the intervals is measured using Normalized Mean Prediction Interval Width (NMPIW):

$$\mathrm{NMPIW} = \frac{\frac{1}{N}\sum_{n=1}^{N}(U_n - L_n)}{R} \cdot 100\%,$$

where $R$ represents the range of values covered by $C_{50n}$.

Finally, the confidence measure method is evaluated using the Pearson correlation between the absolute $C_{50}$ estimation error $|C_{50n} - \widehat{C_{50,n}}(s_n)|$ and the confidence values $CM_n$.

## 5 RESULTS

The results shown in the following subsections are based on a BLSTM with 4 layers of 64 nodes and using a mini-batch size of 50 samples. This configuration was found by

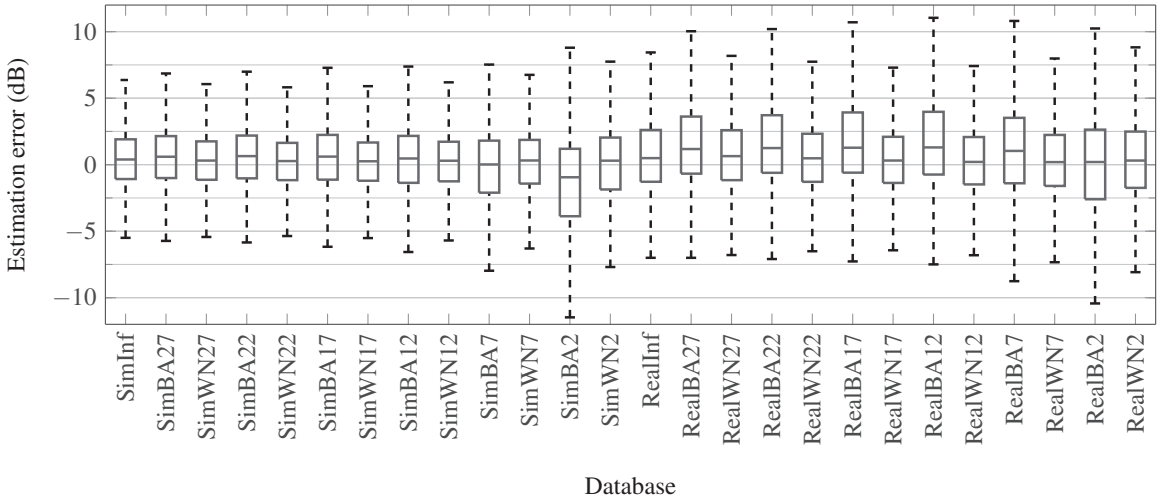



Fig. 4. Distribution of the $C_{50}$ estimation errors for each database in the evaluation set. The edges of the boxes indicate the lower and upper quartile range, while the horizontal lines inside the boxes represent the medians for each method. Moreover, the horizontal lines outside the boxes indicate the estimation error up to 1.5 times the interquartile range. Outliers are removed from this plot.

an empirical investigation to provide the lowest error of all explored architectures in the development set.

### 5.1 $C_{50}$ Estimation

Fig. 4 summarizes the performance of the $C_{50}$ estimator for all the evaluation subsets. Table 1 shows the relationship between each acronym displayed on the horizontal axis of Fig. 4 and the characteristics of RIRs and noise employed to create the subset. The first 13 whisker plots show on average less interquartile range (IQR) and bias compared to the remaining 13 in the right hand side of the figure. This is due to mismatch between the training set and development set, which are created only using simulated RIRs, and these last 13 subsets created with real RIRs.

Babble noise causes the highest estimation errors, which is reflected in Fig. 4 with higher IQR and bias compared to the environments with white noise. Nevertheless, in all cases the bias is low.

The average RMSD achieved on the evaluation set is 3.3 dB. This error is comparable to the minimum $C_{50}$ variation required to perceive a difference in reverberant speech in everyday situations shown in [30] to be in the region of 3 dB.

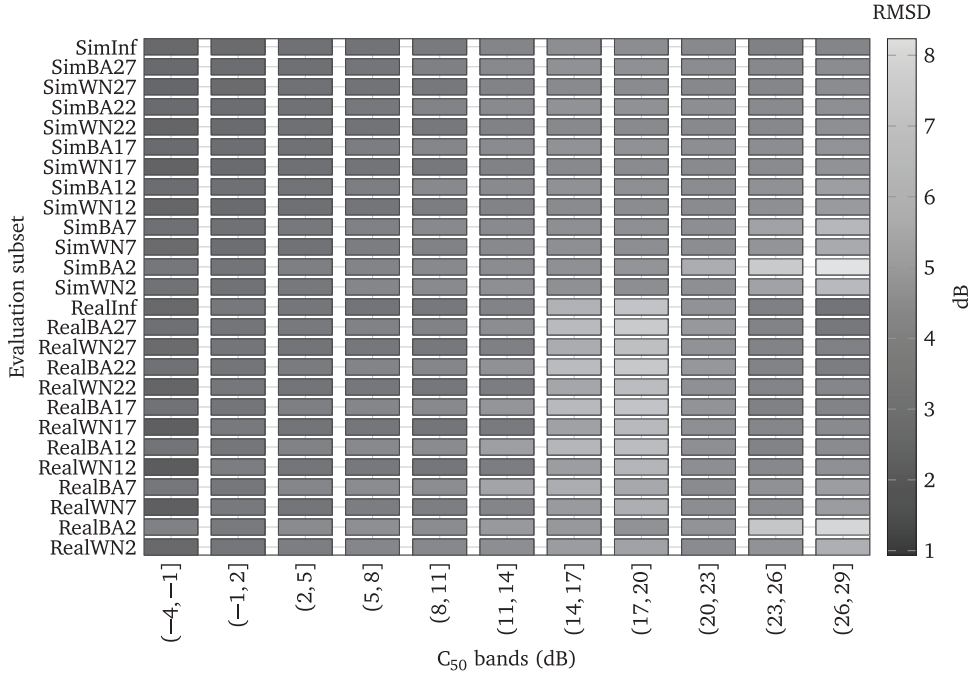



Fig. 5. Root mean square deviation of the $C_{50}$ estimator for the different evaluation subsets split in different bands according to the ground truth $C_{50}$.

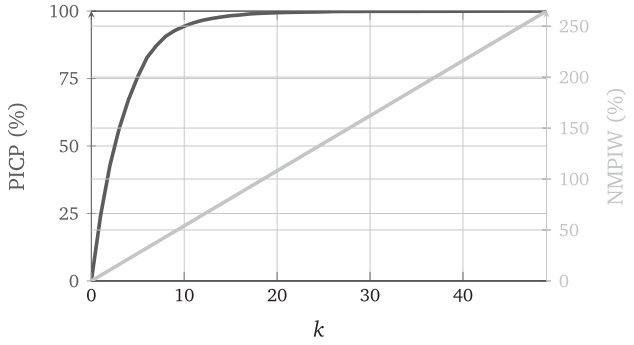



Fig. 6. Values of PICP and NMPIW depending on the tuning parameter $k$ tested on the development set.

In Fig. 5 the $C_{50}$ estimator performance is analyzed per $C_{50}$ bands. In this plot, the width of the $C_{50}$ bands is 3 dB. This figure indicates that NIRA provides more accurate estimations for low $C_{50}$ values, i.e., R1-R3, compared to high $C_{50}$ values, i.e., R7-R11. In general, the RMSD tends to increase as $C_{50}$ increases. Fig. 5 shows again that in general the accuracy achieved estimating real RIRs is lower than the accuracy obtained for simulated RIRs, even though this is not the case for high $C_{50}$ (R10–R11), and babble noise is again the most problematic condition to estimate $C_{50}$. Moreover, in environments with high SNR levels, NIRA tends to provide less accurate estimates in middle-high $C_{50}$ bands, i.e., R7 and R8, which is more prominent for real RIRs.

### 5.2 Prediction Intervals

Fig. 6 shows the PICP and the NMPIW for the development set using different values of $k$. The PICP rapidly increases with $k$ for low values of $k$, whereas NMPIW has a linear behavior in the whole range of $k$. Assuming we require a high PICP for a given application, e.g, PICP = 80%, this is achieved, as shown in Fig. 6, at $k = 5.6$ with a NMPIW = 30.20% in the development set. It is worth noting that NMPIW = 100% may not indicate that the prediction intervals cover the full ground truth range [–3,28] dB, it only suggests that the width of the intervals is comparable to the width of the ground truth $C_{50}$. In contrast, if the estimate $\widehat{C_{50n}}(s_n)$ is in the middle ground truth range, i.e., $\widehat{C_{50n}}(s_n) =$ 15.5 dB, the prediction interval bounds are the same as the ground truth range limits when NMPIW = 100%.

In order to assess the repeatability of this performance we use the evaluation set. Fig. 7 shows the PICP and NMPIW

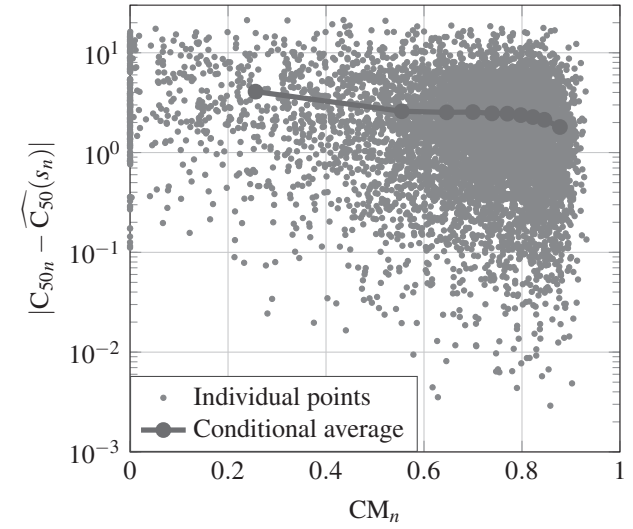



Fig. 8. Confidence measures obtained in the development test set.

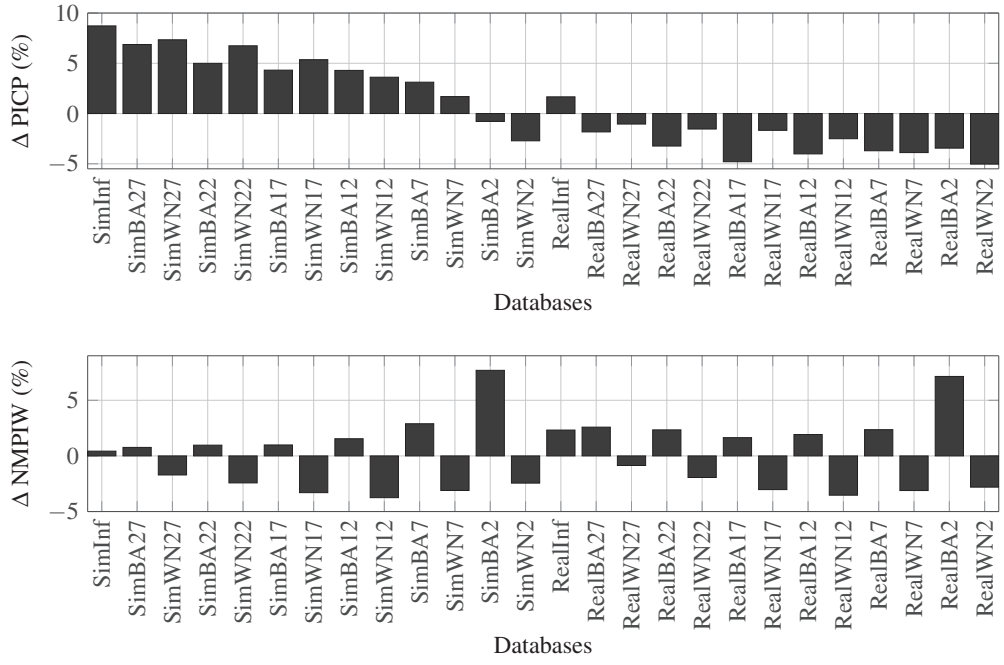



Fig. 7. Difference between the PICP and NMPIW achieved in the different evaluation subsets and the PICP and NMPIW obtained for the development set using $k = 5.6$, which provides a PICP=80% in the development set.

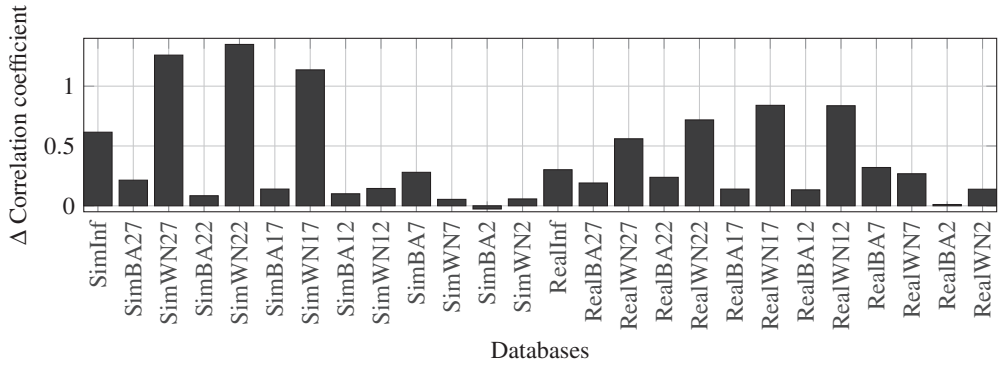



Fig. 9. Difference between the correlation coefficients achieved in the individual evaluation subsets and those achieved in the development set. These correlation coefficients are obtained by conditional averaging the absolute estimation errors and the confidence measures obtained.

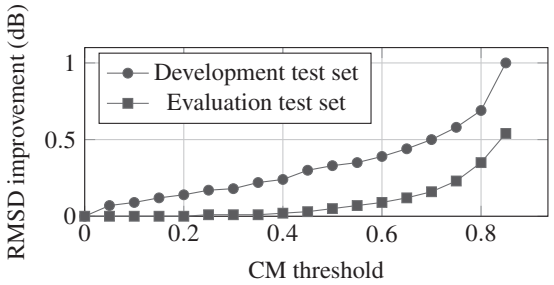



Fig. 10. RMSD improvement achieved for difference confidence measure thresholds.

absolute differences between development set and each evaluation subset for $k = 5.6$. It shows higher PICP for many of the subsets that include simulated RIRs with limited variations of NMPIW and slightly lower PICP for many of the subsets that include real RIRs. The main reason for this difference is due to the characteristics of the training set that only includes simulated RIRs.

The higher NMPIW values for the scenarios that include babble noise indicate that this type of noise creates higher variations in the per frame estimate $\widehat{C_{50l,n}}(s_n)$ compared to the other subsets.

Overall, Fig. 7 shows that the values of PICP and NMPIW for each evaluation subset are consistent with the PICP and NMPIW obtained in the development set.

### 5.3 Confidence Measure

As it was shown in the confidence measure formulation Eq. (8), this metric depends on the value of $k$. In this section all the results are obtained with the same value of $k$ applied in the prediction intervals example, i.e., $k = 5.6$.

Fig. 8 shows the confidence values obtained in the development set. In addition to the individual confidence values, the conditional average extracted from these points is plotted. The conditional average is computed by averaging the absolute estimation error differences with similar confidence measure values. This average is carried out over confidence measure bands obtained after sorting all the points by confidence value and then splitting these points in 10 bands that comprise the same number of elements.

The correlation achieved with the conditional averaging for development set is –0.95, which indicates that the confidence measure decreases as the estimation error increases. The correlation differences between each evaluation subset and the development set are plotted in Fig. 9 in order to assess the repeatability of these results. On average, this difference is 0.39. However, there are three cases (SimWN27, SimWN22, and SimWN17) that show low positive correlation coefficients. This is due to the fact that there are few low confidence points in these evaluation subsets, causing a noisy average for the lowest confidence measure band. In fact, the correlation coefficients are $-0.59$, $-0.61$, and $-0.30$ against the original coefficients 0.31, 0.40, and 0.18 when the averaged point corresponding to the lowest confidence measure band is ignored in SimWN27, SimWN22, and SimWN17 respectively.

The confidence measures can be used to select the most accurate estimates by rejecting estimations below a confidence measure threshold. Fig. 10 shows the improvement obtained with different confidence measure thresholds. The maximum value of this threshold is set to 0.85 due to the limited number of estimates with a confidence higher than 0.9, as it is shown in Fig. 8. The RMSD is reduced as the confidence measure threshold is increased. As a result of tuning the parameter $k$ on the development set, the RMSD improvement is higher on this set than on the evaluation set. The highest RMSD improvement on the evaluation set, i.e., 0.54 dB which leads to a 16% relative improvement, is achieved with a confidence measure threshold of 0.85. This indicates that the confidence measure can be successfully employed to reject inaccurate estimations.

## 6 CONCLUSIONS

We have analyzed the performance of NIRA method, a non-intrusive method for $C_{50}$ estimation, and proposed an approach that can provide information about the accuracy of the $C_{50}$ values obtain from NIRA.

We have shown that the prediction intervals, which provide an upper and lower bound of the estimate, can be derived from the standard deviation of the per frame estimations. On average, in 80% of the per utterance estimates, the ground truth is between the prediction intervals in the development set when computing these intervals as 5.6 times of the standard deviation. The mean width of these intervals is 30% of the ground truth range, therefore on average the intervals are placed $\pm 4.69$ dB from the estimation. Moreover, a confidence measure is derived from these prediction intervals which has shown high correlation coefficients with the $C_{50}$ estimation error.

These methods were validated with an evaluation set comprised of 26 subsets that includes different RIRs, noise levels, and noise types. The NIRA method provided an interquartile range of the $C_{50}$ estimation error of 3.67 dB with a median of 0.47 dB. Regarding the prediction intervals and confidence measure, the results in this evaluation set using $k = 5.6$ showed similar performance to the results obtained in the development set with the same value of $k$, which suggests the results, i.e., prediction intervals and confidence measure performance, are consistent over different databases and therefore repeatable.

The $C_{50}$ estimator showed less accurate estimations for high levels of babble noise, which was reflected in higher mean width of the prediction intervals.

In addition, the procedure proposed in this paper for deriving $C_{50}$ prediction intervals and confidence measures could as well be applied to other room acoustic parameter estimations (e.g., $T_{60}$ or DRR [31]).

## 7 ACKNOWLEDGMENTS

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.

## 8 REFERENCES

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

[2] H. Kuttruff, *Room Acoustics* , Fifth Ed. (London: Taylor & Francis, 2009).

[3] H. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An Improved Algorithm for Blind Reverberation Time Estimation," *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)* (Tel Aviv, Israel), pp. 1–4 (2010).

[4] J. Eaton, N. D. Gaubitch, and P. A. Naylor, "Noise-Robust Reverberation Time Estimation Using Spectral Decay Distributions with Reduced Computational Cost," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Vancouver, Canada), pp. 161–165 (2013).

[5] 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).

[6] P. Kendrick, T. J. Cox, F. F. Li, Y. Zhang, and J. A. Chambers, "Monaural Room Acoustic Parameters from Music and Speech," *J. Acous. Soc. Am.*, vol. 124, no. 1, pp. 278–287 (2008).

[7] P. Peso Parada, D. Sharma, and P. A. Naylor, "Non-Intrusive Estimation of the Level of Reverberation in Speech," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Florence, Italy), pp. 4718–4722 (2014 May).

[8] B. Dumortier and E. Vincent, "Blind RT60 Estimation Robust across Room Sizes and Source Distances," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Florence, Italy), pp. 5187–5191 (2014 May).

[9] E. Georganti, J. Mourjopoulos, and S. van de Par, "Room Statistics and Direct-to-Reverberant Ratio Estimation from Dual-Channel Signals," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Florence, Italy), pp. 4713–4717 (2014 May).

[10] 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," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brisbane, Australia) (2015).

[11] P. Peso Parada, D. Sharma, J. Lainez, D. Barreda, T. van Waterschoot, and P. A. Naylor, "A Single-Channel Non-Intrusive C50 Estimator Correlated with Speech Recognition Performance," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, pp. 719–732 (2016 Apr.).

[12] 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," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Vancouver, Canada), pp. 8169–8173 (2013).

[13] P. Peso Parada, D. Sharma, P. A. Naylor, and T. van Waterschoot, "Reverberant Speech Recognition Exploiting Clarity Index Estimation," *EURASIA J. Advances in Signal Processing*, vol. 2015, no. 1 (2015).

[14] 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).

[15] P. Peso Parada, D. Sharma, P. A. Naylor, and T. van Waterschoot, "Analysis of Prediction Intervals for Non-Intrusive Estimation of Speech Clarity Index," presented at the *AES 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)* (2016 Jan.), conference paper 5-2.

[16] "Objective Measurement of Active Speech Level," *ITU-T Recommendation P.56* (1993 Mar.).

[17] S. Gonzalez and M. Brookes, "A Pitch Estimation Filter Robust to High Levels of Noise (PEFAC)," *Proc. of the 19th European Signal Processing Conference (EUSIPCO)*(Barcelona, Spain), pp. 451–455 (2011).

[18] Y. Wang and M. Brookes, "Speech Enhancement Using a Modulation Domain Kalman Filter Post-Processor with a Gaussian Mixture Noise Model," *Proc. IEEE International Conference on Acoustics, Speech and Signal Pro-*

*cessing (ICASSP)* (Florence, Italy), pp. 7024–7028 (2014 May).

[19] J. Andén and S. Mallat, "Deep Scattering Spectrum," *IEEE Transactions on Signal Processing*, vol. 62, pp. 4114–4128 (2014 Aug.).

[20] F. Weninger, J. Bergmann, and B. Schuller, "Introducing CURRENNT–The Munich Open-Source CUDA RecurREnt Neural Network Toolkit," *J. Machine Learning Research*, vol. 16, pp. 547-551 (2015 Mar.).

[21] A. Khosravi, S. Nahavandi, D. Creighton, and A. F. Atiya, "Comprehensive Review of Neural Network-Based Prediction Intervals and New Advances," *IEEE Transactions on Neural Networks*, vol. 22, no. 9, pp. 1341–1356 (2011).

[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," *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)* (Paris, France) (2006).

[23] G. Papadopoulos, P. J. Edwards, and A. F. Murray, "Confidence Estimation Methods for Neural Networks: A Practical Comparison," *IEEE Transactions on Neural Networks*, vol. 12, no. 6, pp. 1278–1287 (2001).

[24] 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, MD (1988 Dec.).

[25] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the Modeling of Rectangular Geometries in Room Acoustic Simulations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 774–786 (2015 Apr.).

[26] A. Varga and H. J. Steeneken, "Assessment for Automatic Speech Recognition: II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251 (1993).

[27] 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," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY, USA), pp. 1–4 (2013).

[28] R. Stewart and M. Sandler, "Database of Omnidirectional and B-Format Room Impulse Responses," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Dallas, TX, USA), pp. 165–168 (2010 Mar.).

[29] J. K. Nielsen, J. R. Jensen, S. H. Jensen, and M. G. Christensen, "The Single- and Multichannel Audio Recordings Database (SMARD)," *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)* (Juan les Pins, France), pp. 40–44 (2014 Sep.).

[30] J. Bradley, R. Reich, and S. Norcross, "A Just Noticeable Difference in $C_{50}$ for Speech," *Applied Acoustics*, vol. 58, no. 2, pp. 99–108 (1999).

[31] P. Peso Parada, D. Sharma, T. van Waterschoot, and P. A. Naylor, "Evaluating the Non-Intrusive Room Acoustics Algorithm with the ACE Challenge," *ACE Challenge Workshop, a satellite event of IEEE-WASPAA 2015* (New Paltz, NY, USA) (2015 Oct.).

## THE AUTHORS


Pablo Peso Parada


Dushyant Sharma


Toon van Waterschoot


Patrick A. Naylor

Pablo Peso Parada received his B.Sc. and M.Sc. degrees in telecommunication engineering from the University of Vigo, Spain, in 2008 and 2011 respectively, the M.Sc.Res degree in signal theory and communications from the University of Vigo in 2013 and the Ph.D. degree from Imperial College London, UK, in 2016. Between July 2010 and September 2011, he was a research fellow at University of Vigo focusing on speech recognition confidence measures. He was a research intern at Sony, Germany, from June 2012 to November 2012, focusing on signal processing for brain-computer interface. Between December 2012 and March 2013 he was a research engineer at Voice INTER connect, Germany, working on embedded speech recognition. Pablo was an ESR Marie Curie fellow at Nuance Communications, Inc., UK, from April 2013 to June 2016. His research interests include automatic speech recognition (ASR), particularly distant speech recognition, diarization, and room acoustic parameter estimation.

•

Dushyant Sharma was born in New Delhi, India, in 1983. He received the M.Eng. degree in information systems engineering from Imperial College London in 2007 and then continued as a research student at the Centre for Law Enforcement Audio Research (CLEAR) at Imperial College from 2008 and received the Ph.D. degree in speech signal processing in 2012. He is a senior research scientist at Nuance Communications, Sunnyvale, CA, USA, working on non-intrusive speech signal characterization (quality and intelligibility assessment, CODEC identification, and room acoustic parameter estimation), and environment aware algorithms for automatic speech recognition.

•

Toon van Waterschoot received the M.Sc. degree (2001) and the Ph.D. degree (2009) in electrical engineering, both from KU Leuven, Belgium. He is currently a tenure-track Assistant Professor at KU Leuven, Belgium. He has previously held teaching and research positions with the Antwerp Maritime Academy, Belgium (2002), the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT), Belgium (2003–2007), KU Leuven, Belgium (2008–2009), Delft University of Technology, The Netherlands (2010–2011), and the Research Foundation - Flanders (FWO), Belgium (2011–2014). Since 2005 he has been a Visiting Lecturer at the Advanced Learning and Research Institute of the University of Lugano (Universit della Svizzera italiana), Switzerland. His research interests are in acoustic signal enhancement, acoustic modeling, audio analysis, and audio reproduction. He has been the Scientific Coordinator of the FP7-PEOPLE Marie Curie Initial Training Network on Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS). Dr. van Waterschoot has been serving as an Associate Editor for the *Journal of the Audio Engineering Society* (AES) and for the *EURASIP Journal on Audio, Music, and Speech Processing*, and as a Guest Editor for *Signal Processing*. He has been a Nominated Officer for the European Association for Signal Processing (EURASIP) and a member of the IEEE Audio and Acoustic Signal Processing Technical Committee (AASP-TC). He has been serving as an Area Chair for Speech Processing at the European Signal Processing Conference (EUSIPCO 2010, 2013–2015), and as Chair of the 60th AES Conference in Leuven, Belgium, 2016. He is a member of the AES, the Acoustical Society of America, EURASIP, and IEEE.

•

Patrick A. Naylor received his B.Eng. degree in electronic and electrical engineering from the University of Sheffield, UK, in 1986 and the Ph.D. degree from Imperial College, London, UK, in 1990. Since 1990 he has been a member of academic staff in the Department of Electrical and Electronic Engineering at Imperial College London. His research interests are in the areas of speech, audio, and acoustic signal processing. He has worked in particular on adaptive signal processing for dereverberation, blind multichannel system identification and equalization, acoustic echo control, speech quality estimation and classification, single and multichannel speech enhancement, and speech production modelling with particular focus on the analysis of the voice source signal. In addition to his academic research, he enjoys several fruitful links with industry in the UK, USA, and in mainland Europe. He is the Chair of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, a director of the European Association for Signal Processing (EURASIA), and formerly an associate editor of *IEEE Signal Processing Letters* and *IEEE Transactions on Audio Speech and Language Processing*.