A single-channel non-intrusive C_{50} estimator correlated with speech recognition performance

Pablo Peso Parada, *Student Member, IEEE,* Dushyant Sharma, *Member, IEEE,* Jose Lainez, Daniel Barreda, Toon van Waterschoot, *Member, IEEE,* and Patrick A. Naylor, *Senior Member, IEEE*

***Abstract*—Several intrusive measures of reverberation can be computed from measured and simulated room impulse responses, over the full frequency band or for each individual mel-frequency subband. It is initially shown that full-band *clarity index* C_{50} is the most correlated measure on average with reverberant speech recognition performance. This corroborates previous findings but now for the dataset to be used in this study. We extend the previous findings to show that C_{50} also exhibits the highest mutual information on average. Motivated by these extended findings, a non-intrusive room acoustic (NIRA) estimation method is proposed to estimate C_{50} from only the reverberant speech signal. The NIRA method is a data-driven approach based on computing a number of features from the speech signal and it employs these features to train a model used to perform the estimation. The choice of features and learning techniques are explored in this work using an evaluation set which comprises approximately 100000 different reverberant signals (around 93 hours of speech) including reverberation from measured and simulated room impulse responses. The feature importance of each feature with respect to the estimation of the target C_{50} is analysed following two different approaches. In both cases the newly chosen set of features shows high importance for the target. The best C_{50} estimator provides a root mean square deviation around 3 dB on average for all reverberant test environments.**

***Index Terms*—Room acoustic parameter estimation, reverberant speech recognition, reverberation**

I. Introduction

In enclosed acoustic spaces such as rooms, sound emitted from a source propagates through the air and reflects off the walls and different objects in the room creating the effect known as reverberation. The energy associated with the reflected waves determines the reverberation level in the room and is often quantified relative to the energy at the receiver due to direct path propagation. Reverberation is known to degrade automatic speech recognition (ASR) performance and it is therefore highly valuable to be able to quantify the relation between room acoustic effects and ASR performance.

The acoustic characteristics of a given enclosure, source and receiver geometry can be represented using a room impulse response (RIR), which depends on the room properties as well as the position of the source and receiver. The reverberant sound $y(n)$ measured at a receiver in the room can be modelled as the convolution of the RIR $h(m)$, assumed time-invariant, and the source signal in the room $s(n)$ so that for each time index n

$$y(n) = \sum_{m=0}^{M-1} h(m)s(n-m) \tag{1}$$

where M is the effective length of $h(m)$.

Several room acoustic parameters derived from the RIR have been proposed in the literature [1] [2] in order to measure the level of reverberation. The reverberation time T_{60} is a widely used metric that characterizes the room acoustics properties. Alternative parameters, such as the direct-to-reverberant ratio (DRR) [1], the *definition* D_{50} [1] or the *clarity index* C_{50} [1], provide further measures describing the reverberation level in a signal.

These room acoustic parameters are employed for a wide range of tasks. For example, in [3] a non-linear mapping of T_{60}, DRR and room spectral variance is proposed to estimate the human perception of the reverberation disturbance in speech signals. Kuttruff [2] suggests that D_{50} can be used as an indicator of the speech intelligibility in reverberant environments. Several room acoustic parameters have been employed to predict the ASR performance for reverberant speech. In [4] a new metric derived from D_{50} is proposed as an estimator of the ASR performance. Tsilfidis et al. [5] present a correlation analysis of several room acoustic parameters (T_{60}, C_{50}, D_{50} ...) showing that C_{50} is the most correlated parameter with ASR performance, reaching the same conclusion as in [6]. In [7] the ASR performance was investigated as a function of early reflection duration. An analysis of the impact of the RIR shape on the ASR performance [8] concludes that the first 50 ms of the RIR barely affect the ASR performance and therefore D_{50} could be used to predict the word accuracy rate. Additionally, several room acoustic parameters have been applied in different dereverberation methods to suppress the reverberation in the signal. C_{50} is used in [9] [10] and T_{60} in [11][12] to select the ASR acoustic model that better represents the reverberant conditions of the input utterance. In [13] T_{60} is used to add to the current hidden Markov model state the contribution of previous states by applying a piece-wise energy decay curve that is separated in early reflections and late reverberation contributions. The T_{60} information is

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.

P. Peso Parada, Jose Lainez and Daniel Barreda are with Nuance Communications, Inc., Marlow SL7 2AF, U.K. (e-mail:pablo.peso@nuance.com; jose.lainez@nuance.com; daniel.barreda@nuance.com).

Dushyant Sharma is with Nuance Communications, Inc., Sunnyvale, CA 94085, U.S. (e-mail: dushyant.sharma@nuance.com).

T. van Waterschoot is with the ESAT-STADIUS, Stadius Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, 3001 Leuven, Belgium, and also with ESAT-ETC, Advise Lab, KU Leuven, 2440 Geel, Belgium (e-mail: toon.vanwaterschoot@esat.kuleuven.be).

P. A. Naylor is with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: p.naylor@imperial.ac.uk).

also applied in [14] to suppress late reverberation through a wavelet packet tree decomposition. From these examples, it is clear that knowledge of estimation of room acoustic parameters can be beneficially exploited in the processing of reverberant signals.

In most real applications, the room impulse response is unknown and the only available information is the reverberant speech signal. Consequently the room acoustic parameters need to be estimated non-intrusively from this signal rather than directly from the RIR. Several methods have been proposed to estimate T_{60} non-intrusively. The method of [15] estimates the decay rate from a statistical model of the sound decay by using the *maximum likelihood* (ML) approach and then uses this decay rate to find the ML estimate for T_{60}. The T60 estimator [16] is based on spectral decay distributions. In this case the signal is analysed with a mel-frequency filter bank in order to compute the decay rate by applying a least-square linear fit to the time-frequency log magnitude bins. Variance of the negative gradients in the distribution of decay rates is then mapped to T_{60} with a polynomial function. A method to compute the reverberation time in the modulation domain is proposed in [17], exploiting the fact that low modulation frequency energy (below 20 Hz) is only slightly 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) whose features are the ratio of the average of low modulation frequency energy to different averages of high modulation frequency energy. The overall ratio is then mapped to estimate the DRR. Two methods to estimate T_{60} or C_{80} from speech and music signals are proposed in [18]. The first method exploits the power spectral density (PSD), which is estimated as the sum of the Hilbert envelopes computed per frequency band. The second method employs a ML approach to estimate the decay curve of the cleanest section in the signal and then averages the partial estimation to create the final estimate. In [19] a multilayer perceptron is built with spectro-temporal modulation features extracted from a 2D-Gabor filter bank in order to estimate the type of room that created the reverberant signal. Although room acoustic parameters can be also estimated from multichannel recordings, such as T_{60} [20] or DRR [21], this paper focuses on the problem of single-channel room acoustic parameter estimation.

In this work we provide evidence using different set-ups, ASR engine and newly measured RIRs, that corroborates previous evidence that C_{50} is the most correlated parameter to ASR performance. Furthermore, we include new features and a learning algorithm that provides a per-frame C_{50} estimate. These new features and the learning method were not proposed in previous work [6][22]. Additionally, the performance is tested over an extensive database including newly measured RIRs and different noise conditions and the relative importance of the different features is analysed. The proposed configuration of the data-driven method outperforms previous C_{50} estimators reported in [22] providing estimates highly correlated with ASR performance.

The remainder of the article is organized as follows. Section II presents the motivation to estimate full-band C_{50}. We describe the data-driven approach proposed to estimate this room acoustic parameter in Section III. Section IV introduces the evaluation metrics and the database utilized to evaluate different aspects of the estimator performance shown in Section V. Finally, conclusions are drawn in Section VI.

II. Parameters and evaluation

Before addressing the task of non-intrusive estimation of room acoustic parameters, an analysis of intrusive room acoustic parameters is first performed to investigate the relationship of various room acoustic parameters with ASR performance and thus find the parameter most correlated with ASR performance.

A. Room acoustic parameters

The reverberation time 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] and it is computed following [23]. An alternative measurement is the DRR calculated as [24]

$$\mathrm{DRR} = 10\log_{10}\left(\frac{E_d}{\left(\sum_{m=0}^{M-1} h^2(m)\right) - E_d}\right)\mathrm{dB}, \qquad (2)$$

where E_d is the direct path energy computed by convolving a sinc function with the RIR around the direct path sample n_d, given by

$$E_d = \max_{\sigma} \sum_{m=-\eta}^{\eta} \left(\mathrm{sinc}(\pi(m+\sigma))h(m+n_d)\right)^2, \qquad (3)$$

where η=8 is the number of sinc sidelobes and σ=[-1:1] is the offset considered to find the maximum energy.

Similarly, the C_{50} and D_{50} can be formulated as follows

$$\mathrm{C}_{\tau} = 10\log_{10}\left(\frac{\sum_{m=0}^{N_\tau} h^2(m)}{\sum_{m=N_\tau+1}^{M-1} h^2(m)}\right)\mathrm{dB}, \qquad (4)$$

$$\mathrm{D}_{\tau} = 10\log_{10}\left(\frac{\sum_{m=0}^{N_\tau} h^2(m)}{\sum_{m=0}^{M-1} h^2(m)}\right)\mathrm{dB}, \qquad (5)$$

where τ = 50 ms in this case and N_τ represents the number of samples in the RIR $h(m)$ from the beginning to τ ms after the reception of the direct path. Additionally, the *centre time* (Ts) is a measure of reverberation that represents the centre of gravity of the squared RIR and it is computed as follows [2]

$$\mathrm{Ts} = \frac{\sum_{m=0}^{M-1} \frac{m}{fs}\, h^2(m)}{\sum_{m=0}^{M-1} h^2(m)} \quad \mathrm{s}, \qquad (6)$$

where fs is the sampling frequency.

The motivation of this work is to estimate the measure of reverberation that is most correlated with the ASR performance. We therefore analyse T_{60}, Ts, DRR, C_τ and D_τ over a range of τ.

B. Evaluation metrics

The ASR performance is measured as the phoneme error rate (PER)

$$\mathrm{PER} = \frac{D+I+S}{N_{phn}} \qquad (7)$$

where N_{phn} is the total number of phonemes recognized, D is the number of deletions, S is the number of substitutions and I the number of insertions. The performance is measured per phoneme to avoid possible influences of the language model or dictionary rules and therefore be able to measure more accurately the impact of reverberation on the acoustic modelling of ASR. For this purpose a context-dependent GMM-HMM phoneme recognizer was employed based on Kaldi [25] following the TIMIT recipe 's5'. The feature vector includes mel-frequency cepstral coefficients with delta and delta-delta features computed from non-reverberant utterances of the TIMIT training set.

In addition to PER, we include the Perceptual Evaluation of Speech Quality (PESQ) in the evaluation as a commonly used baseline that is helpful to obtain a quantitative insight into the nature of the test data. PESQ [26] is an intrusive objective method to estimate the speech quality. The reference signal used in the PESQ calculation is the original anechoic clean speech.

Two different metrics are used to evaluate the relevance of different measures to ASR performance. The first is the absolute value of the Pearson correlation coefficient computed as

$$\rho = \left| \frac{\sum_{u=1}^{U} (y_u - \overline{y})(x_u - \overline{x})}{\sqrt{\sum\limits_{u=1}^{U} (y_u - \overline{y})^2 \sum\limits_{u=1}^{U} (x_u - \overline{x})^2}} \right| \tag{8}$$

where $\overline{x}$ is the average of the PER scores x_u per utterance, $\overline{y}$ is the average of a particular measure of reverberation y_u computed for each utterance, and U is the total number of utterances included. Additionally, we also use the mutual information between these variables computed as [27]

$$I(X,Y) = \int_X \int_Y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}\, dx\, dy \tag{9}$$

where $p(x)$ and $p(y)$ are the marginal probability density functions of X and Y respectively and $p(x,y)$ is the joint probability density function of X and Y. In (9) $I(X,Y)$ measures the amount of information shared between both random variables, where the variables in this case are PER scores and the values of a particular measure of reverberation.

C. Evaluation data

The data used to compute ρ and $I(X,Y)$ for the different measures of reverberation is taken from two sets described in Section IV-B. The first set is extracted from the training set presented in Section IV-B by selecting only the reverberant utterance without noise giving a total of 6144 utterances (5.55 hours). The second set uses the RealInf set from the evaluation set presented in Section IV-B which comprises 3960 reverberant utterances (3.70 hours) obtained with measured impulse responses.

D. Correlation of room acoustic parameters with ASR performance

We first review in this Section the correlation of different room acoustic parameters with PER, as well as with PESQ for comparison. Our aim is to corroborate the findings of [6] in the case of our specific test data, and then extend the previous findings to include mutual information analysis and also room acoustic parameters computed from each individual mel-frequency subband of the RIR.

1) Full frequency-band room acoustic parameters: Table I displays the correlation coefficients obtained with simulated impulse responses. It shows that the most correlated measure with PER is C_{50}, which is in accordance with the results obtained in [5]. Additionally C_{50} is seen again to be the most correlated with PESQ. Figure 1 shows the correlation of C_τ and D_τ where C_τ from τ approximately 20 ms to 50 ms achieves the highest correlation coefficients for PESQ and PER and D_τ shows its highest correlation coefficients with smaller τ. Similar results are obtained with real RIRs which are given in Table II and in Fig. 2.

	T_{60}	DRR	Ts	D_{50}	C_{50}
PER	0.70	0.68	0.73	0.73	**0.85**
PESQ	0.75	0.75	0.78	0.78	**0.91**

TABLE I
CORRELATION COMPARISON OF PER AND PESQ WITH DIFFERENT ACOUSTIC PARAMETERS FOR SIMULATED IMPULSE RESPONSES. THE MAXIMUM VALUES ARE BOLD.

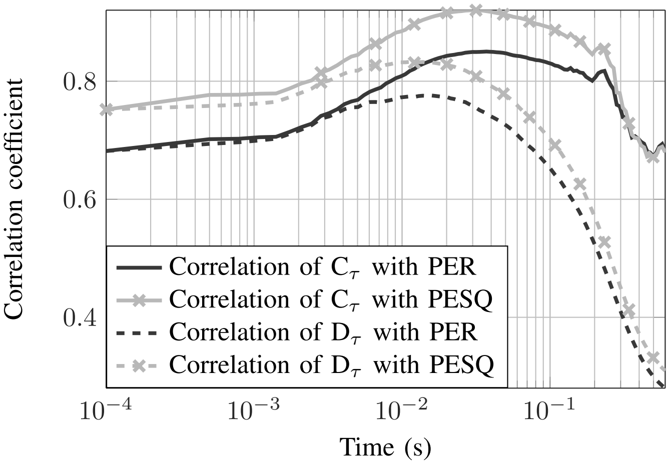


Fig. 1. PER and PESQ correlation coefficients obtained with C_τ and D_τ for τ between 0.1 ms and 600 ms using simulated RIRs.

	T_{60}	DRR	Ts	D_{50}	C_{50}
PER	0.75	0.37	0.72	0.60	**0.85**
PESQ	0.79	0.42	0.78	0.66	**0.94**

TABLE II
CORRELATION COMPARISON OF PER AND PESQ WITH DIFFERENT ACOUSTIC PARAMETERS FOR REAL MEASURED IMPULSE RESPONSES. THE MAXIMUM VALUES ARE BOLD.

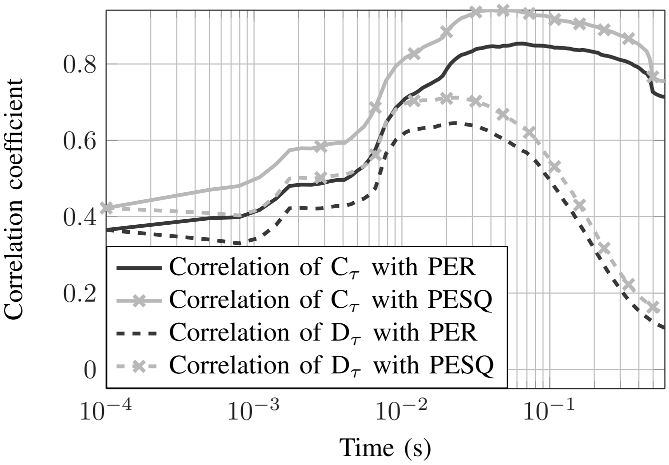


Fig. 2. PER and PESQ correlation coefficients obtained with C_τ and D_τ for τ between 0.1 ms and 600 ms using real RIRs.

Table III gives the magnitude of the mutual information between the measure of reverberation and PER and PESQ. It shows that Ts provides the highest mutual information value with PER and PESQ, closely followed by the C_{50}. DRR is seen to be the measure that shares the least information with PER and PESQ.

	T_{60}	DRR	Ts	D_{50}	C_{50}
PER	0.59	0.40	**0.66**	0.58	0.64
PESQ	0.79	0.66	**0.96**	0.81	0.95

TABLE III
MUTUAL INFORMATION COMPARISON OF PER AND PESQ WITH DIFFERENT ACOUSTIC PARAMETERS FOR SIMULATED IMPULSE RESPONSES. THE MAXIMUM VALUES ARE BOLD.

Figure 3 shows the magnitude of mutual information achieved for C_τ and D_τ for a range of τ from 0.1 ms to 600 ms. It shows in all cases higher values with C_τ than with D_τ. The highest value of the mutual information of C_τ with PER is at approximately τ = 50 ms. Regarding the mutual information of C_τ with PESQ, the highest values are around τ = 30 ms. On the other hand, mutual information of D_τ with PER and PESQ shows lower values compared to C_τ with the highest values obtained towards lower τ values.

Table IV shows the mutual information magnitude of several measures of reverberation with the ASR performance (PER) and PESQ obtained on reverberant data generated with real measured impulse responses. Despite Ts showing high mutual information in some cases, C_{50} is the measure of reverberation that provides the highest values on average over the two datasets.

Figure 4 shows the mutual information of C_τ and D_τ with PER and PESQ respectively. All the figures presented in this Section lead to the same conclusions: C_τ provides higher correlation and mutual information values than D_τ and the highest values of C_τ are in the range centred at τ = 50 ms.

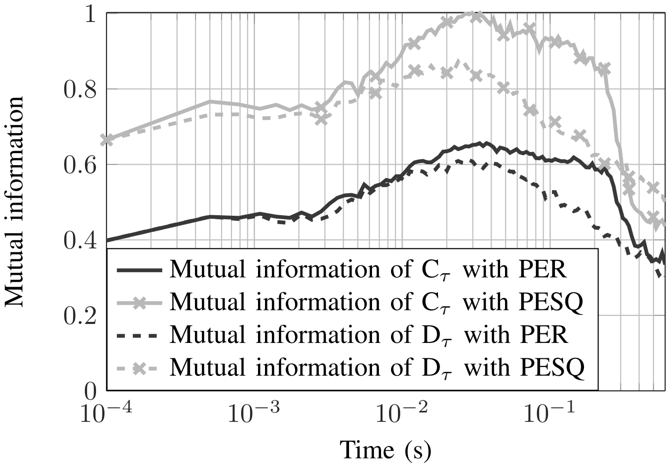


Fig. 3. PER and PESQ mutual information magnitude obtained with C_τ and D_τ for τ between 0.1 ms and 600 ms using simulated RIRs.

	T_{60}	DRR	Ts	D_{50}	C_{50}
PER	**0.67**	0.36	0.60	0.50	0.66
PESQ	1.07	0.78	0.99	0.88	**1.15**

TABLE IV
MUTUAL INFORMATION COMPARISON OF PER AND PESQ WITH DIFFERENT ACOUSTIC PARAMETERS FOR REAL MEASURED IMPULSE RESPONSES. THE MAXIMUM VALUES ARE BOLD.

2) Mel-frequency subbands room acoustic parameters: In ASR, the input acoustic signal is commonly processed to extract the mel-frequency cepstral coefficients [28]. In this section we compute parameters using the same mel-frequency filter bank applied in the ASR [25] in order to investigate whether room acoustic parameters per mel-frequency subband provide higher correlation and mutual information values than the full-band counterpart.

Figures 5 and 6 show the correlation and mutual information values for different acoustic parameters computed per mel-frequency subband for simulated and real impulse responses respectively. On average, C_{50} provides again the highest values, especially at high frequencies. However, these values are lower than (in certain cases approximately equal to) the

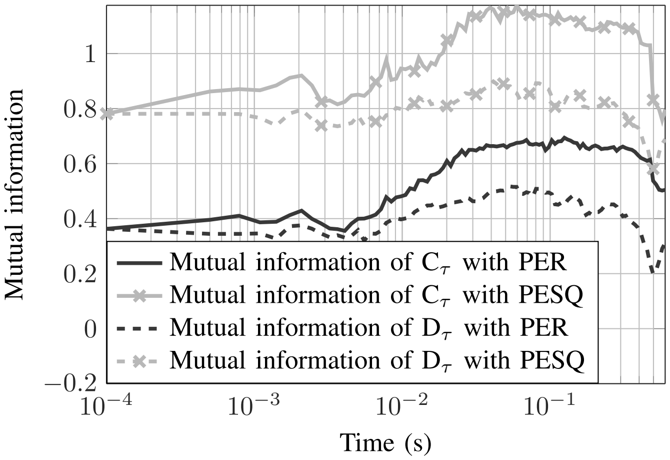


Fig. 4. PER and PESQ mutual information magnitude obtained with C_τ and D_τ for τ between 0.1 ms and 600 ms using real RIRs.

C_{50} full-band correlation and mutual information. Thus, C_{50} computed from the full-band impulse response is the most correlated room acoustic parameter with ASR performance and provides on average the highest mutual information value with ASR performance. Motivated by this finding, in Section III we propose a method to estimate full-band C_{50} non-intrusively using only the reverberant speech signal.

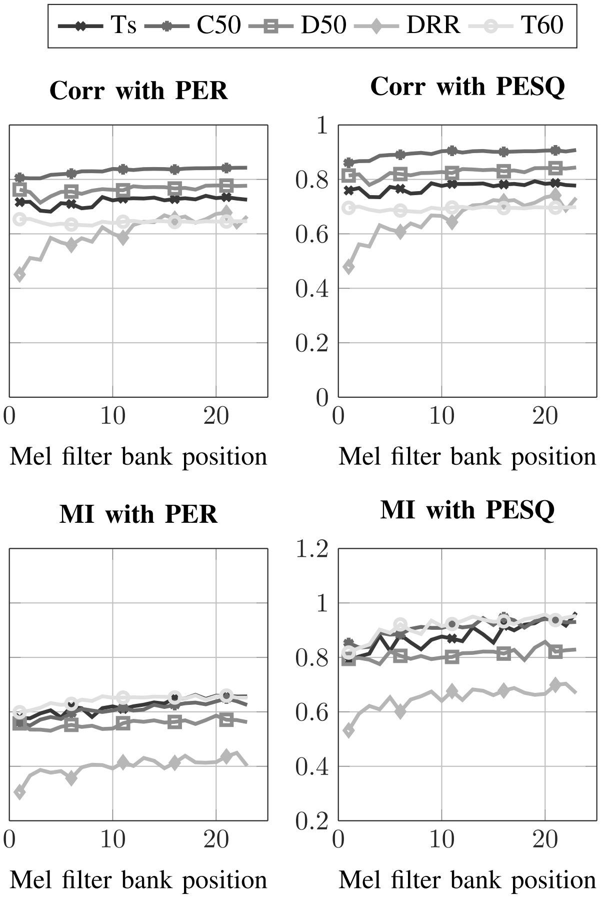


Fig. 5. PER and PESQ correlation coefficients (top) and mutual information values (bottom) obtained with five measures of reverberation computed per mel-frequency subband using simulated RIRs.

III. Method description

The proposed method to estimate C_{50} is a data-driven approach which computes 409 features per utterance from a single-channel speech signal at a sampling rate of 8 kHz. Figure 7 presents the general block diagram of the NIRA method. The features are used to build a model from which the C_{50} value will be estimated.

A. Feature extraction

Features derived from modulation domain (MD) [29] and from deep scattering spectrum (DSS) transformation [30] are now proposed.

The modulation domain provides information about the spectral envelopes of the signal. Speech is dominated by modulation frequencies from 2 Hz to 8 Hz [31]. However, the reverberation effect boosts higher modulation frequencies [17] in the speech signal. Motivated by this fact, modulation domain features are extracted by first selecting the frequency-band with the highest energy in the average modulation domain representation and then computing the first four central moments of this frequency band and its two adjacent modulation frequency bands.

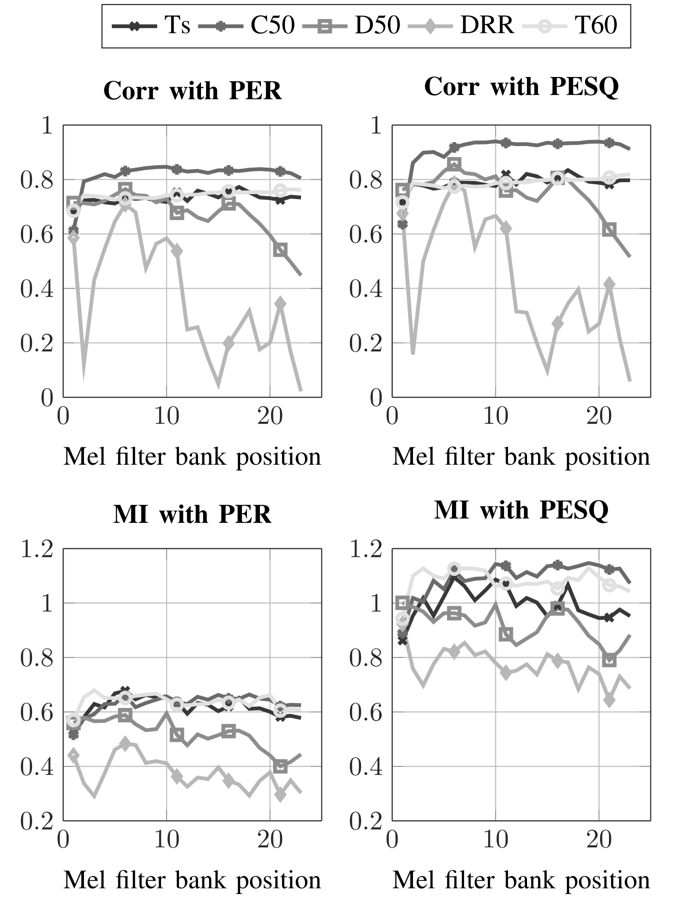


Fig. 6. PER and PESQ correlation coefficients (top) and mutual information values (bottom) obtained with five measures of reverberation computed per mel-frequency subband using real RIRs.

Deep scattering spectrum features are extracted from a scattering transformation applied to the signal [30]. This wavelet transformation is particularly interesting due to its locally translation invariant representation and its stability to time-warping deformations. The transformation comprises a cascade of wavelet decomposition and modulus operators. The MFCCs are approximately equal to the first-order scattering coefficients whereas second-order coefficients characterize transient observations (e.g., onsets or amplitude modulation) [30]. Since MFCCs are already included in the feature set and reverberation effect causes distortions in transient periods, features are extracted only from this second wavelet layer. The DSS features are computed by employing one wavelet per octave in both layers normalized by the first-order coefficients and an average window size of 20 ms with 50% overlap.

In addition to these features, we employ the utterance-based and frame-based features proposed in [6].

Utterance-based features are computed from Long-Term Average Speech Spectrum (LTASS) deviation by mapping it into 16 bins with equal bandwidth as well as the slope of the unwrapped Hilbert phase of the input signal.

Frame-based features comprise the following parameters:

- Line Spectrum Frequency (LSF) features computed by mapping the first 10 LPC coefficients to the LSF representation.
- Zero-crossing rate (ZCR).
- Speech variance.
- Pitch period estimated with the PEFAC algorithm [32].
- Estimation of the importance-weighted Signal-to-Noise Ratio (iSNR) in units of dB [33].
- Variance and dynamic range of the Hilbert envelope.
- 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 the LTASS power spectrum magnitudes.
- 12th order mean- and variance-normalized MFCCs computed from the fast Fourier transform as well as the rate of change of the per-frame features.

The rate of change for all short-time features, excluding the 12th order MFCCs, is also computed.

A voice activity detector (VAD) is applied to the power-normalized input signal to extract all the features employing only active speech segments. This VAD uses the P.56 method [34].

Table V summarizes all the features. The complete feature vector is created by appending to the long-term features the mean (μ), variance (σ^2), skewness (s) and kurtosis (k) of all short-time features and thereby creating the final vector with 409 features.

Description	**Feature**	**ΔFeature**
LSFs	$\phi_{1:10}$	$\phi_{11:20}$
ZCR, 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, centroid and dynamics of PLD	$\phi_{33:35}$	$\phi_{36:38}$
MFCCs with delta and delta-delta	$\phi_{39:74}$	-
LTASS	$\phi_{75:90}$	-
Unwrapped Hilbert phase	ϕ_{91}	-
MD	$\phi_{92:103}$	-
DSS	$\phi_{104:124}$	-

TABLE V

NIRA FEATURES: $\phi_{1:74-104:124}$ ARE FRAME-BASED FEATURES COMPUTED FRAME BY FRAME, WHOSE STATISTICS ARE USED IN THE LEARNING ALGORITHM, AND $\phi_{75:103}$ ARE UTTERANCE-BASED FEATURES CALCULATED OVER THE ENTIRE UTTERANCE. ΔFEATURE REPRESENTS THE RATE OF CHANGE OF THE FEATURE.

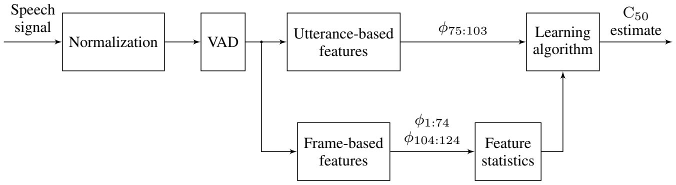


Fig. 7. The NIRA method.

The feature configuration described above is used to estimate C$_{50}$ per utterance. Additionally, we propose in this work a C$_{50}$ estimated per frame which employs a different feature configuration. This configuration is based on computing features $\phi_{1:74}$ with a 20 ms window size with 50% overlap and computing $\phi_{92:103}$ per frame instead of averaging over all per-frame modulation domain representations for each utterance. A wider window size with the same overlap is used for the modulation domain features, 256 ms window size, and pitch estimation, 90 ms window size, to preserve higher frequency resolution. The remaining utterance-based features are excluded (i.e. $\phi_{75:91}$).

B. Learning algorithms

The learning algorithms employed to build the NIRA models, designed to estimate C$_{50}$ with the features presented in Section III-A, are now presented.

1) Classification And Regression Trees (CART): Classification And Regression Trees [35] offer a non-parametric methodology to build binary trees. These trees split the data recursively into smaller partitions in order to find the best fit. The training process involves three main steps: tree building, stopping tree building and pruning the tree.

The predicted output is obtained according to the leaf reached after having recursively traversed the tree in depth, deciding the branch to follow at each node based on one or more input feature values. We use CART in a regression mode rather than a classification mode since our target is to estimate a room acoustic parameter within a continuous range.

2) Linear regression (LR): The estimate $\widehat{\mathrm{C}_{50,\mathrm{u}}}$ is computed using linear regression [36] as

$$\widehat{\mathrm{C}_{50,\mathrm{u}}} = \sum_{j=1}^{J} \theta_j \Phi_{j,u} + \theta_0, \tag{10}$$

where $\mathbf{\Phi_u} = [\Phi_{1,u}, \ldots, \Phi_{J,u}]^T$ represents the length-J observed variables (i.e. feature vector) for the uth utterance and $\boldsymbol{\theta} = [\theta_0, \ldots, \theta_J]$ is a vector comprised of $J+1$ linear regression coefficients.

The optimal coefficient vector $\boldsymbol{\theta}$ to model the target C$_{50,\mathrm{u}}$ is obtained by minimizing the sum of squared errors according to the cost function

$$\min_{\boldsymbol{\theta}} \frac{1}{2U} \sum_{u=1}^{U} \left(\left(\sum_{j=1}^{J} \theta_j \Phi_{j,u} + \theta_0 \right) - \mathrm{C}_{50,u} \right)^2 + \lambda \sum_{j=1}^{J} \theta_j^2, \tag{11}$$

where λ is the regularization parameter and U represents the total number of utterances. This minimization problem is solved by applying the gradient descent algorithm [37]. Additionally, an L2 regularization term is included in the cost function to avoid complex and overfitted models.

3) Deep belief neural network (DBN): A deep belief network structure allows complex non-linear models to learn how to fit the input data to the target C_{50} values. The discriminative training of these networks is applied to a stack of generative pretrained layers. This generative training attempts to learn the structure of the input data in an unsupervised manner by setting the output values to the input values at each layer. Pretrained networks reduce overfitting and discriminative training effort [38]. Sparse autoencoders [39] are used to pretrain each layer that aim to find optimal weights with the backpropagation algorithm subject to sparsity constraints. This sparsity constraint facilitates the task of finding dependencies in the input data. Additionally, dropout [40] is applied in the fine-tuning by randomly removing units of the network at each training step to prevent overfitting. The fine-tune training is carried out with stochastic gradient descent and adaptive momentum [41].

Whereas the DBN is widely used for classification tasks, in this work the output layer uses a linear regression on the final hidden layer of neurons in order to estimate a continuous value for C_{50}.

4) Bidirectional long short-term memory (BLSTM): Recurrent neural networks (RNN) have been applied in different tasks [42] [43] [44] [45]. This type of neural network can be seen as a neural network with at least one feedback connection, hence the output of the activation function is employed to compute the output in the next time step. This configuration provides memory capabilities in the RNN which enables it to learn sequences such as temporal correlations. In addition to the forward propagation, bidirectional RNNs also exploit future context information by processing the data in time reverse direction. The principal drawback of conventional RNNs is the vanishing gradient problem during learning [46] which is overcome by introducing the Long Short-Term Memory (LSTM) cells [47] in the network. LSTM is better at modelling long-term dependencies and it can be combined with a bidirectional RNN to form a bidirectional LSTM.

We employ this structure to build a model [48], which provides a C_{50} estimation per frame, motivated by the bidirectional long-term dependency capabilities of the BLSTM which can potentially represent temporal smearing effects of reverberation.

IV. Experimental setup

Experiments have been performed to assess different C_{50} estimators considered in this work. Section IV-A defines the evaluation parameters while Section IV-B introduces the database employed to evaluate the methods. Section IV-C describes the trained neural network topology finally employed for each model.

A. Evaluation metrics

The root mean square deviation (RMSD) of C_{50} is computed using

$$E_u = \widehat{\mathrm{C}_{50,u}} - \mathrm{C}_{50,u}\ \mathrm{dB},$$

$$\mathrm{RMSD} = \sqrt{\frac{1}{U}\sum_{u=1}^{U}(E_u)^2}\ \mathrm{dB}, \tag{12}$$

where $\widehat{\mathrm{C}_{50,u}}$ and $\mathrm{C}_{50,u}$, both measured in dB, correspond to the estimated and ground truth C_{50} value of the uth utterance respectively, and U is the total number of utterances.

In addition, the mean (μ_E) and standard deviation (σ_E) of the estimation error are also included in the analysis to provide further information about the C_{50} estimation error, and are computed as

$$\mu_E = \frac{1}{U}\sum_{u=1}^{U}E_u\ \mathrm{dB}, \tag{13}$$

$$\sigma_E = \sqrt{\frac{1}{U}\sum_{u=1}^{U}(E_u-\mu_E)^2}\ \mathrm{dB}. \tag{14}$$

Moreover, the Pearson correlation coefficient ρ and the mutual information $I(X,Y)$ are employed in the analysis to measure the linear relationship between the estimated and ground truth values. They are computed following (8) and (9), where $x_u = \widehat{\mathrm{C}_{50,u}}$, $y_u = \mathrm{C}_{50,u}$ and X and Y represent the estimated and ground truth C_{50} respectively.

B. Data sets

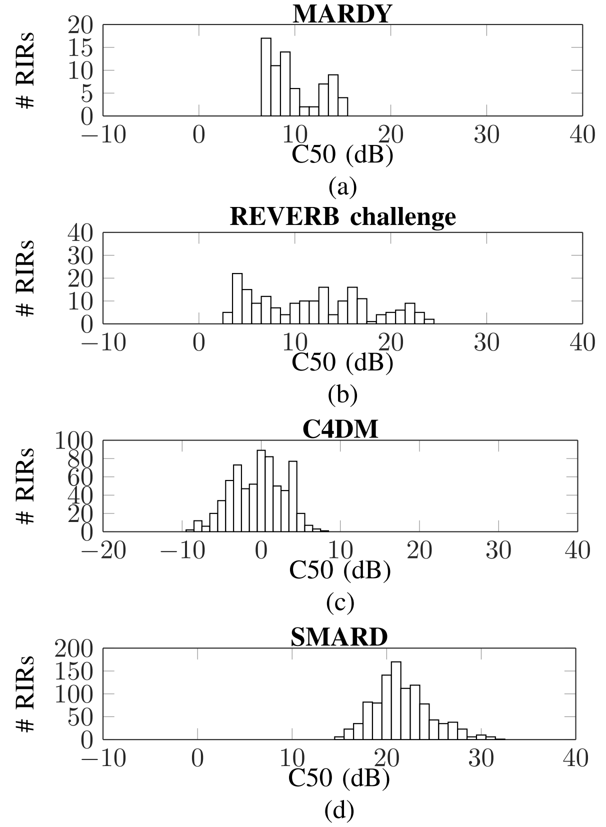
Three data sets are employed. The training set is used to train the methods which are tuned with the development set, whereas the evaluation set is used only to evaluate the methods. The utterances, RIRs and noise signals are different for each set and are all sampled at 8 kHz.

1) Training set: Speech signals from the TIMIT [49] database are employed to build the training data set. A total of 32 utterances are selected randomly from the training set ensuring that 2 different male and 2 female speakers are included for each dialect and excluding 'SA sentences'. The reverberant speech is created by convolving these speech utterances with simulated room impulse responses. These are generated randomly using the randomized image method [50] and then 192 RIRs are carefully selected to obtain 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 [51] are added to the reverberant speech at SNRs of 0 dB to 30 dB in steps of 5 dB.

2) Development set: The development set is created following the training set configuration using 16 utterances and 64 RIRs. None of the speech signals nor RIRs of this set are included in the training set.

3) Evaluation set: In the evaluation set, one utterance of each TIMIT core set speaker is included resulting in 24 sentences. The SA sentences are excluded. Babble noise and white noise are also included 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. Four different databases are considered to build the real room impulse response set: MARDY [52]; REVERB challenge [53]; C4DM RIR [54]; and SMARD [55]. Only recordings from the B-format microphone taken in the Great Hall are considered within the C4DM RIR database due to an artefact in the other C4DM recordings at the 125 Hz octave band. 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 range from -3 dB to 28 dB.

Accordingly, this evaluation set covers a wide range of reverberant scenarios from large rooms such as the Great Hall of the C4DM RIR database to medium rooms with low reverberation as in the SMARD database. Figure 8 illustrates the C_{50} distribution of each of the RIR data sets.



(a)

(b)

(c)

(d)

Fig. 8. Distribution of C_{50} in real measured RIR databases: (a) MARDY database [52] ; (b) RIRs collected from the training set of the REVERB challenge database [53]; (c) B-format microphone recording from the Great Hall of the C4DM database [54]; (d) SMARD database [55].

The average duration of simulated and real RIRs is 2 seconds and 1.17 seconds respectively, i.e. M in (1) is on average 16000 for simulated RIRs and 9360 for real RIRs.

This evaluation set is divided into 26 subsets which are evaluated independently to assess the performance of the methods for each specific situation as outlined in Table VI.

C. Learning algorithm topologies

The DBN architecture is selected using genetic algorithms [56] which find the topology that minimizes the estimation error in the development set. Two different DBN models are trained for comparison purposes employing features $\phi_{1:91}$, containing the features proposed in [6], and features $\phi_{1:124}$, adding to the previous feature vector $\phi_{1:91}$ the features proposed in this work. The main motivation for this splitting is to measure the improvement in performance obtained by including the new features proposed in this work $\phi_{92:124}$.

RIR type	**Noise type**	**SNR level**	**Name**
Simulated	none	∞	SimInf
	Babble / White	2	Sim2BA / Sim2WN
		7	Sim7BA / Sim7WN
		12	Sim12BA / Sim12WN
		17	Sim17BA / Sim12WN
		22	Sim22BA / Sim22WN
		27	Sim27BA / Sim27WN
Real	none	∞	RealInf
	Babble / White	2	Real2BA / Real2WN
		7	Real7BA / Real2WN
		12	Real12BA / Real12WN
		17	Real17BA / Real17WN
		22	Real22BA / Real22WN
		27	Real27BA / Real27WN

TABLE VI
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).

The topology selected in the DBN model is a two layer neural network with 75 and 79 neurons in the first and second layer respectively, whereas the latter model comprises a first layer of 160 neurons and a second layer of 110 neurons.

The BLSTM model trained with $\phi_{1:91}$ includes 3 layers of 256 neurons in each layer and the model trained with $\phi_{1:124}$ comprises 4 layers of 64 neurons in each layer.

V. PERFORMANCE EVALUATION

In this Section the methods presented in Section III are evaluated. Firstly, in Section V-A an analysis of the importance of the features with respect to the target C_{50} is presented. Two measures of feature importance are used to find the value of the feature to estimate C_{50}. The proposed C_{50} estimators are evaluated in Section V-B. A baseline method [18] provides a comparison. The baseline method originally estimates C_{80} based on PSD of the reverberant microphone signal. However here it has been adapted to estimate C_{50} by modifying the target values in the learning process. Finally, the correlation and mutual information values of the C_{50} estimates with ASR performance are compared in Section V-C to an upper bound on the performance, obtained using ground truth C_{50} values.

A. Feature importance

The importance to the C_{50} estimator of each of the features presented in Table V is now analysed. Numerous methods have been proposed in the literature to compute the feature importance [57] [58]. We employ two different methods to rank the features according to their importance: CART [35] and Regressional ReliefF method (RReliefF) [59].

The first approach relies on the CART learning algorithm presented in Section III-B1. This decision tree method attempts to find the feature to split the data set at each node that provides the best discrimination between a set of targets. Once the tree is built, the importance is computed as a function of the purity reduction due to the split at each node. Since CART is employed to estimate C_{50}, we also use the already trained model for feature importance purposes.

The RReliefF [59] method computes the importance of the features based on the capability to differentiate target values that are close together. The importance is defined as a function of three different terms:

- Probability of different values of the feature given the nearest observations.
- Probability of different target values given the nearest observations.
- Probability of different target values given different feature values and the nearest observations.

We use this method because it provides an importance ranking of the features. Additionally, this method is faster than wrapper methods [60] and it is not targeted to any specific learning algorithm.

Table VII shows the 10 most important features for each method using the features proposed in previous work $\phi_{1:91}$ [6]. The ranking of feature importance estimated in each case is different, however there are some common features: ϕ_{29}, ϕ_{52}, ϕ_{64}, ϕ_{65}, ϕ_{66}. The results also suggest that the MFCC features are highly important for C_{50} estimation.

RANK	CART	RReliefF
1	$\sigma^2\phi_{54}$	$\sigma^2\phi_{64}$
2	$\sigma^2\phi_{63}$	$s\phi_{26}$
3	$\mu\phi_{29}$	$\mu\phi_{29}$
4	$\sigma^2\phi_{52}$	$\sigma^2\phi_{66}$
5	$\sigma^2\phi_{64}$	$\mu\phi_{30}$
6	$\sigma^2\phi_{66}$	$k\phi_{26}$
7	$\sigma^2\phi_{28}$	$s\phi_{22}$
8	$s\phi_{52}$	$\sigma^2\phi_{67}$
9	$\sigma^2\phi_{38}$	$\sigma^2\phi_{65}$
10	$\sigma^2\phi_{65}$	$\sigma^2\phi_{52}$

TABLE VII
RANKED FEATURE IMPORTANCE EMPLOYING CART AND RRELIEFF WITH THE FEATURE SET $\phi_{1:91}$ EXTRACTED FROM THE TRAINING SET. THE VARIANCE, MEAN, SKEWNESS AND KURTOSIS OF THE PER-FRAME FEATURES ARE REPRESENTED WITH σ^2, μ, s AND k RESPECTIVELY.

Table VIII shows the top 10 important features for the full feature set, including now the newly proposed MD and DSS features to the previous existing feature set presented in [6] (i.e. $\phi_{1:91}$). CART and RReliefF show some common features to be highly important: ϕ_{98} and ϕ_{64}. In both cases, some of the new features (i.e. features within $\phi_{92:124}$) are present, in particular MD features appear 8 times in the top 10 for RReliefF. Looking further in the RReliefF ranking, DSS features appear 19 times in the first 100 features, which indicates that these features are also important. Additionally, it should be mentioned that CART only uses 46 features after pruning, of which 11 are DSS features and 2 are MD features. These results highlight the suitability of these new features for the estimation of C_{50}.

RANK	CART	RReliefF
1	$\sigma^2\phi_{54}$	ϕ_{101}
2	$\sigma^2\phi_{63}$	ϕ_{100}
3	ϕ_{98}	ϕ_{103}
4	$\mu\phi_{29}$	ϕ_{93}
5	$\sigma^2\phi_{64}$	$\sigma^2\phi_{64}$
6	$\sigma^2\phi_{66}$	ϕ_{92}
7	$\sigma^2\phi_{28}$	$s\phi_{26}$
8	$\sigma^2\phi_{38}$	ϕ_{99}
9	$\sigma^2\phi_{118}$	ϕ_{95}
10	$\sigma^2\phi_{55}$	ϕ_{98}

TABLE VIII
RANKED FEATURE IMPORTANCE EMPLOYING CART AND RRELIEFF WITH THE FEATURE SET $\phi_{1:124}$ EXTRACTED FROM THE TRAINING SET. THE VARIANCE, MEAN, SKEWNESS AND KURTOSIS OF THE PER-FRAME FEATURES ARE REPRESENTED WITH σ^2, μ, s AND k RESPECTIVELY.

B. C_{50} *estimators*

Figure 9 shows a comparison of the estimators' performance with regards to RMSD for all evaluation sets. In this first analysis only features ϕ_{1-91} are included in the feature vector. It is important to note that the BLSTM provides an estimation per frame, hence for comparison purposes only the average of all the frame estimations per utterance is taken into account. Figure 9 suggests that the estimation accuracy is lower with babble noise compared to the same RIRs with white noise, and estimation accuracy is better in lower levels of noise as expected. The best estimations are achieved with BLSTM, whereas the baseline provides the worst RMSD values.

The bias (μ_E) and standard deviation (σ_E) of the estimation errors for each set are shown in Fig. 10. CART provides a low-biased estimator. However, due to its high variance the estimation accuracy is degraded. BLSTM achieves the lowest standard deviations for all sets, while the baseline provides the worst bias and standard deviation of the estimation error.

Figure 11 plots the improvement in RMSD achieved by including the additional features proposed in this work (ϕ_{92-124}). This improvement is measured as

$$\Delta\mathrm{RMSD} = \mathrm{RMSD}_{\phi_{1-91}} - \mathrm{RMSD}_{\phi_{1-124}}, \tag{15}$$

where the subscripts indicate the set of features considered to build the estimators. Figure 11 shows that all estimators improve when using the new features (i.e. ϕ_{91-124}). The highest improvement is achieved with DBN, which is about 0.4 dB on average across all sets. Despite this fact, the best overall performance is achieved with BLSTM, approximately RMSD = 3.3 dB on average.

Figure 12 summarizes the reduction of the bias ($\Delta\mu_E$) and standard deviation ($\Delta\sigma_E$) of the estimation error. These are quantified as follows

$$\begin{aligned} \Delta\mu_E &= \mu_{E_{\phi_{1-91}}} - \mu_{E_{\phi_{1-124}}}, \\ \Delta\sigma_E &= \sigma_{E_{\phi_{1-91}}} - \sigma_{E_{\phi_{1-124}}}, \end{aligned} \tag{16}$$

where the lowest subscripts indicate the range of features considered to perform the estimations. BLSTM shows a significant reduction of the bias while the standard deviation is

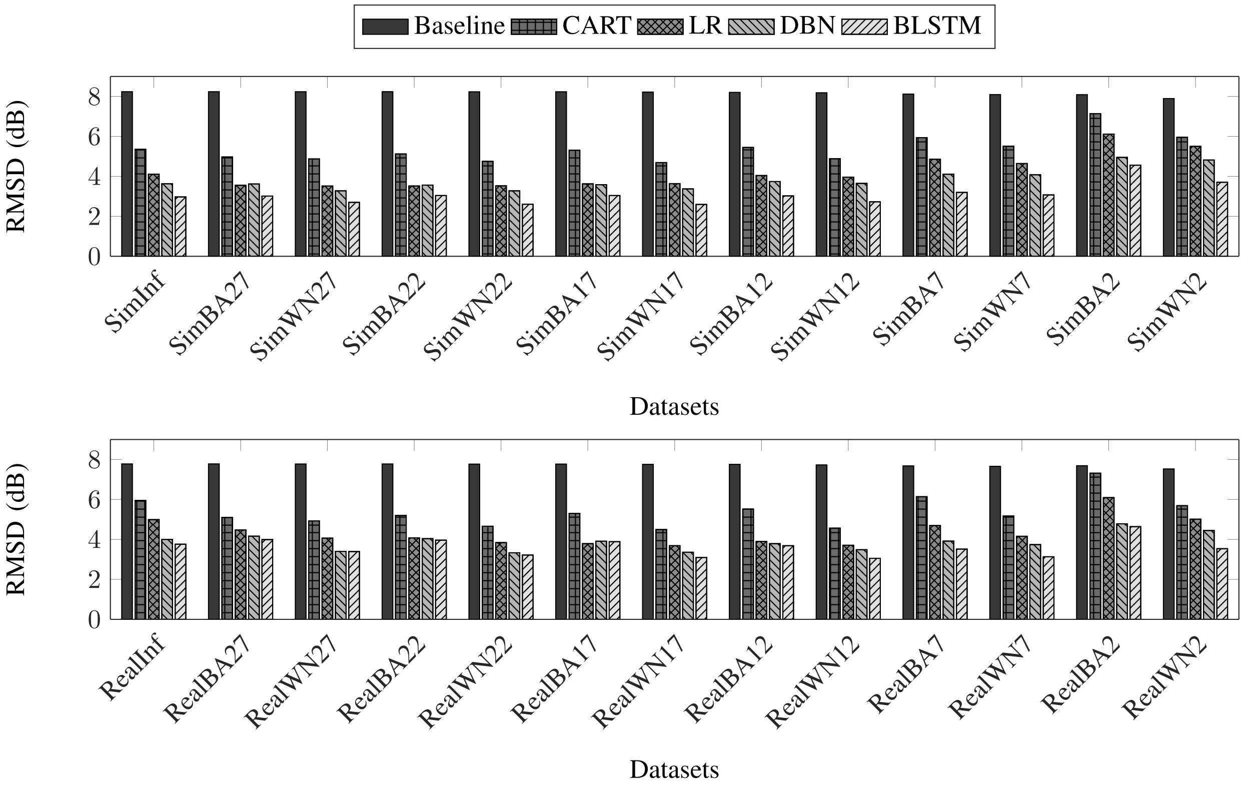


Fig. 9. RMSD obtained for different room impulse responses (simulated and real) including different noise types (WN: white, BA: babble).

increased. On the contrary, all methods except BLSTM achieve a significant reduction of the standard deviation but their bias is increased.

Figure 13 shows the ground truth C_{50} and the estimated C_{50} for the BLSTM based method that achieves the lowest RMSD on average and the baseline method. Only two different sets are shown for the sake of clarity: SimInf and SimBA2 which provide approximately the worst and best performance in terms of RMSD for the BLSTM.

From an application point of view, the minimum number of frames required to provide a C_{50} estimate relatively close to the estimate achieved when using the entire utterance is relevant in order to reduce the computational cost of the estimate and the latency in real-time applications. For this purpose we analyse the per-frame performance of the best C_{50} estimator presented previously (i.e. BLSTM). Figure 14 illustrates the effect of the number of frames employed to estimate C_{50} on the final RMSD. This performance curve converges to the RMSD value of this estimator, plotted in dashed line in Fig. 14, when approximately 180 frames are considered. Taking into account that the window size and increment are 20 ms and 10 ms respectively, approximately 1.9 seconds are required to achieve the same performance as with the full utterance.

Additionally, Fig. 15 presents the RMSD average per frame k obtained with the same estimator when employing n frames available for the estimation. Note that the RMSD of the frames decreases when the number n of frames included increases. The main reason is because BLSTM applies backward propagation (as well as forward propagation) to provide an estimation, therefore the performance depends not only on previous frames but on future frames as well. Figure 15 indicates that, even from the first frame, a low C_{50} estimate deviation is achieved using 180 frames which is similar to the RMSD obtained with the entire utterance information and estimation errors are higher in the last frames.

C. Correlation and mutual information of the C_{50} *estimates with PER*

In Section II we have shown that ground truth C_{50} values provide high correlation and mutual information values with ASR performance. The correlation and mutual information of the estimated C_{50} values with ASR performance is summarized in Table IX. This shows that the C_{50} estimates provide a high correlation value which is comparable to the value obtained with the ground truth C_{50} values. Furthermore, the use of C_{50} within the context of speech recognition has been investigated and the results documented in [9] [10].

Metric	GT	Baseline	CART	LR	DBN	BLSTM
ρ	0.85	0.56	0.77	0.84	0.85	0.85
$I(X,Y)$	0.66	0.36	0.67	0.67	0.69	0.73

TABLE IX
CORRELATION (ρ) AND MUTUAL INFORMATION ($I(X,Y)$) VALUES OF THE GROUND TRUTH C_{50} (GT) AND THE ESTIMATED C_{50} (BASELINE, CART, LR, DBN AND BLSTM) WITH PER FOR REALINF EVALUATION SET.

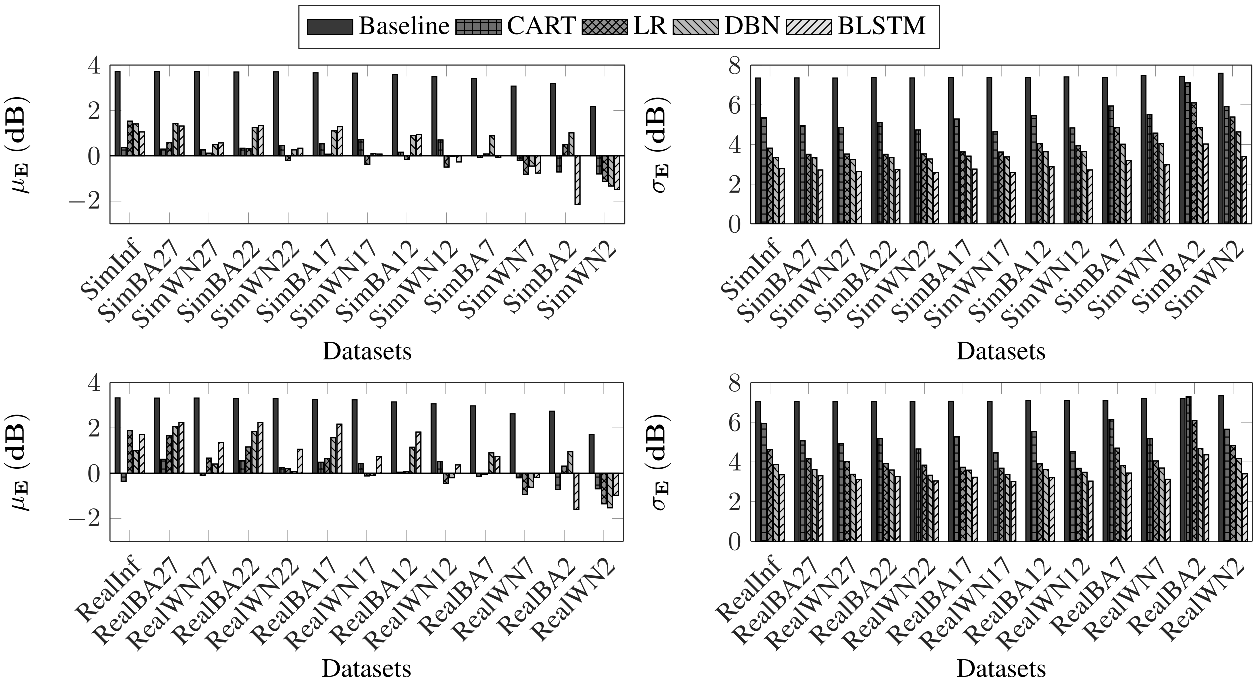


Fig. 10. Mean and standard deviation of the estimation error obtained for different room impulse responses (simulated and real) including different noise types (WN: white, BA: babble).

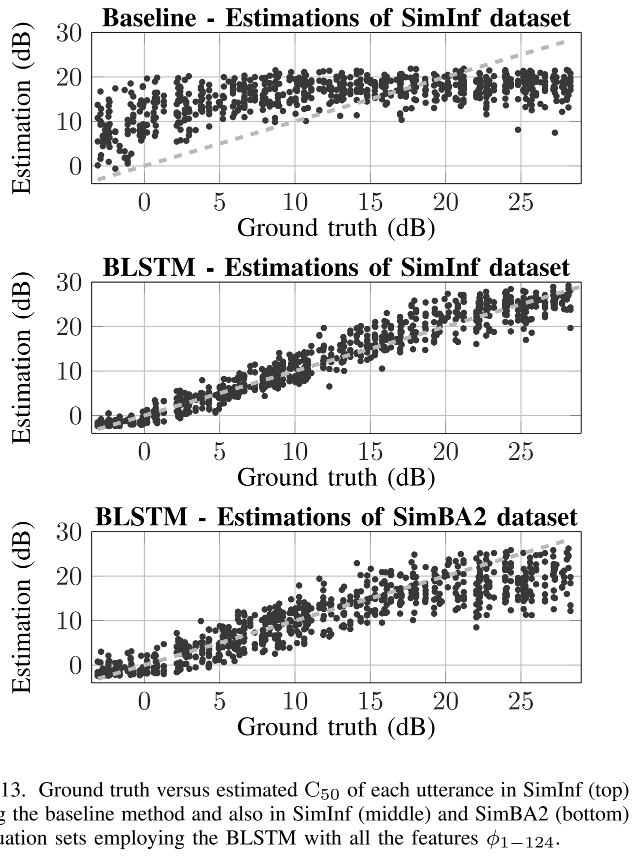


Fig. 13. Ground truth versus estimated C_{50} of each utterance in SimInf (top) using the baseline method and also in SimInf (middle) and SimBA2 (bottom) evaluation sets employing the BLSTM with all the features ϕ_{1-124}.

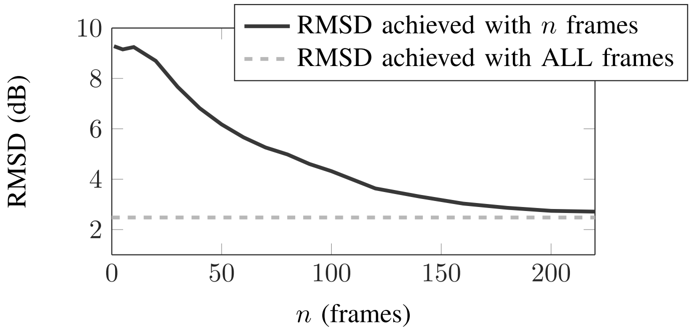


Fig. 14. RMSD achieved with BLSTM employing the n first frames of each utterance in SimInf evaluation set.

VI. Summary and conclusions

We have shown that the full frequency-band C_{50} is the most relevant measure of reverberation to predict phoneme recognition in terms of correlation and mutual information. Motivated by this finding, we have proposed a data-driven method (NIRA) to estimate C_{50} from the reverberant speech signal using a single microphone observation. New features based on modulation domain and deep scatter spectrum have been included in NIRA and have been shown to improve the performance of NIRA and to be highly ranked in terms of feature importance. Additionally, we have introduced recurrent neural networks in NIRA to model the time smearing effect of reverberation and provide an estimation per frame. This configuration has shown the best performance on average across all evaluation sets, which include measured impulse responses, achieving a root mean square deviation of 3.3 dB in C_{50} estimation. This deviation is similar to the minimum C_{50} variation necessary to perceive a change in reverberant speech in everyday situations stated on [61] to be in the region of 3 dB.

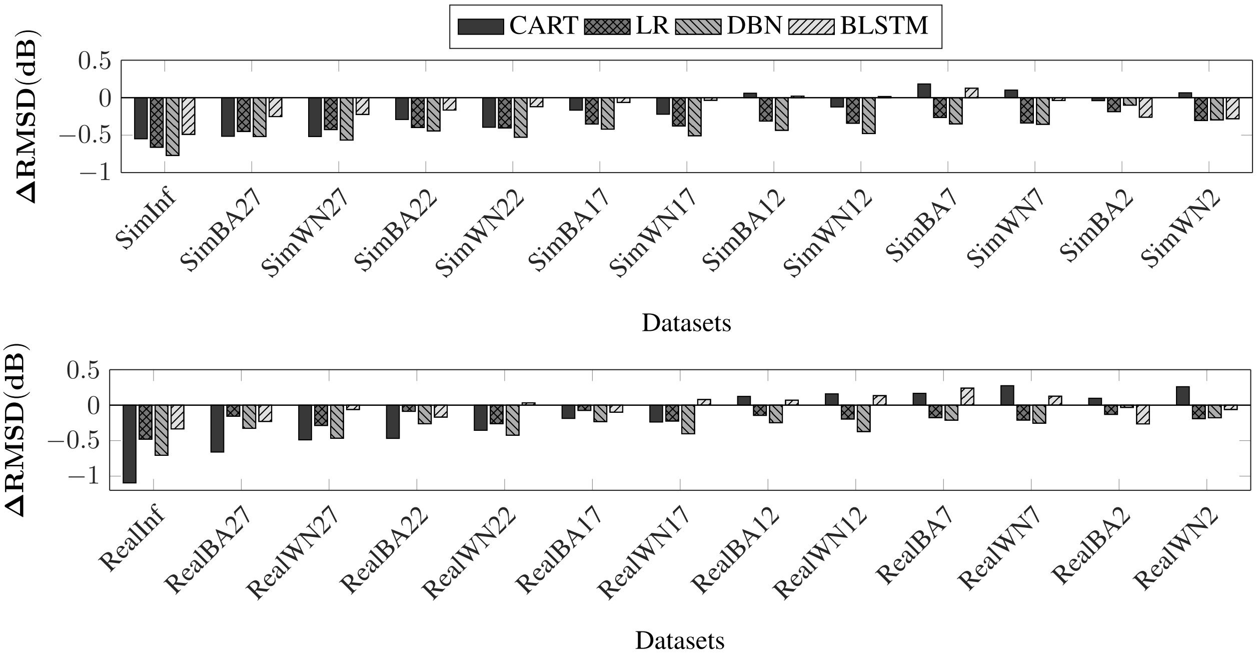


Fig. 11. RMSD improvement including new features (DSS and MD) for different room impulse responses (simulated and real) including different noise types (WN: white, BA: babble).

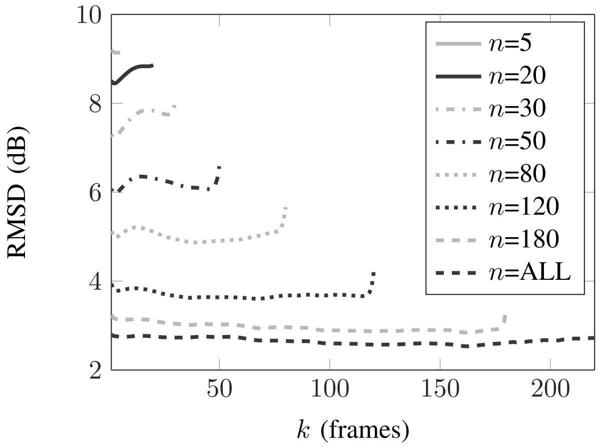


Fig. 15. RMSD per frame k achieved with BLSTM employing only the n first frames of each utterance in SimInf evaluation set to perform the estimation.

References

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

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

[3] 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)*, Vancouver, Canada, 2013, pp. 8169–8173.

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

[5] 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.

[6] 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)*, Florence, Italy, 2014, pp. 4718–4722.

[7] A. Brutti and M. Matassoni, "On the use of early-to-late reverberation ratio for asr in reverberant environments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 4638–4642.

[8] 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)*, Tel Aviv, Israel, 2010.

[9] P. Peso Parada, D. Sharma, P. A. Naylor, and T. van Waterschoot, "Single-channel reverberant speech recognition using C50 estimation," in *Proc. REVERB Challenge*, Florence, Italy, May 2014.

[10] P. Peso Parada, D. Sharma, P. A. Naylor, and T. v. Waterschoot, "Reverberant speech recognition exploiting clarity index estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, 2015.

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

[12] J. Liu and G.-Z. Yang, "Robust speech recognition in reverberant environments by using an optimal synthetic room impulse response model," *Speech Communication*, vol. 67, pp. 65–77, 2015.

[13] 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)*, Bucharest, Romania, 2012, pp. 365–369.

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

[15] 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)*, Tel Aviv, Israel, 2010, pp. 1–4.

[16] 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)*, Vancouver, Canada, 2013, pp. 161–165.

[17] 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.

[18] P. Kendrick, T. J. Cox, F. F. Li, Y. Zhang, and J. A. Chambers, "Monaural

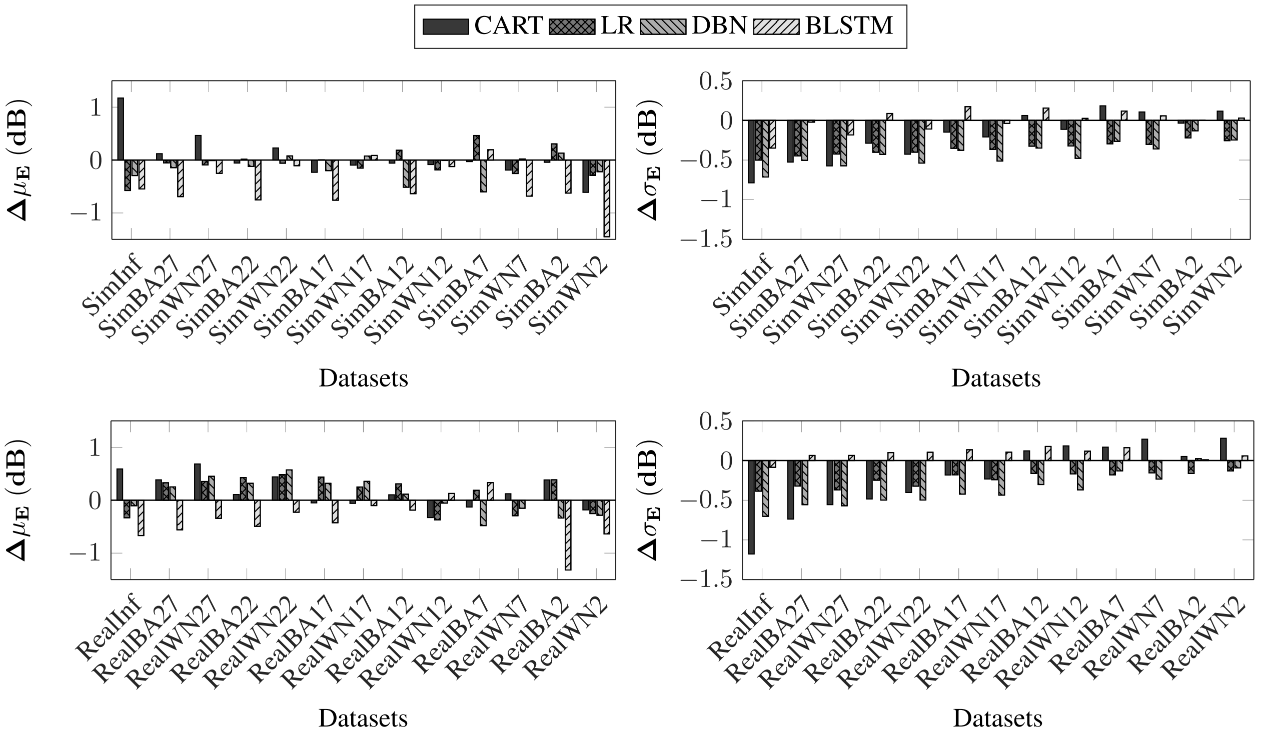


Fig. 12. Increment of the absolute mean and standard deviation of the estimation error including new features (DSS and MD) for different room impulse responses (simulated and real) including different noise types (WN: white, BA: babble).

room acoustic parameters from music and speech," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 278–287, 2008.

[19] F. Xiong, S. Goetze, and B. T. Meyer, "Estimating room acoustic parameters for speech recognizer adaptation and combination in reverberant environments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 5522–5526.

[20] 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)*, Florence, Italy, May 2014, pp. 5187–5191.

[21] 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)*, Florence, Italy, May 2014, pp. 4713–4717.

[22] P. Peso Parada, D. Sharma, J. Lainez, D. Barreda, P. A. Naylor, and T. van Waterschoot, "A quantitative comparison of blind C50 estimators," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Juan les Pins, France, 2014, pp. 298–302.

[23] M. Karjalainen, P. Ansalo, A. Mäkivirta, T. Peltonen, and V. Välimäki, "Estimation of modal decay parameters from noisy response measurements," *Journal of the Audio Engineering Society*, vol. 50, no. 11, pp. 867–878, 2002.

[24] R. Stanton and M. Brookes, "Speech dereverberation in the STFT domain," Imperial College London, Tech. Rep., June 2013. [Online]. Available: http://arxiv.org/pdf/1509.07411v1.pdf

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Hawaii, USA, 2011, pp. 1–4.

[26] 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*, International Telecommunications Union (ITU-T) Recommendation P.862, Feb. 2001.

[27] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, p. 066138, Jun 2004.

[28] 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, UK: Cambridge University Engineering Department, 2006.

[29] 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)*, Florence, Italy, May 2014, pp. 7024–7028.

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

[31] H. Hermansky, "Modulation spectrum in speech processing," in *Signal Analysis and Prediction*. Springer, 1998, pp. 395–406.

[32] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *Proc. of the 19th European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, 2011, pp. 451–455.

[33] D. Sharma, G. Hilkhuysen, N. D. Gaubitch, P. A. Naylor, M. Brookes, and M. Huckvale, "Data driven method for non-intrusive speech intelligibility estimation," in *Proc. of the 18th European Signal Processing Conference (EUSIPCO)*, Aalborg, Denmark, 2010, pp. 1899–1903.

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

[35] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Florida: CRC Press, 1984.

[36] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.

[37] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: John Wiley & Sons, 2001.

[38] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[39] A. Ng, "Sparse autoencoder," *Stanford University, CS294A Lecture notes*, 2011.

[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[41] T. K. Leen and G. B. Orr, "Optimal stochastic search and adaptive mo-

mentum," in *Proc. Advances in neural information processing systems (NIPS)*, Denver, USA, 1994, pp. 477–484.

[42] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 6645–6649.

[43] F. Weninger, S. Watanabe, J. Le Roux, J. R. Hershey, Y. Tachioka, J. Geiger, B. Schuller, and G. Rigoll, "The MERL/MELCO/TUM system for the REVERB Challenge using Deep Recurrent Neural Network Feature Enhancement," in *Proc. REVERB challenge*, Florence, Italy, 2014.

[44] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 1045–1048.

[45] Z. Zhang, J. Pinto, C. Plahl, B. Schuller, and D. Willett, "Channel mapping using bidirectional long short-term memory for dereverberation in hands-free voice controlled devices," *IEEE Transactions on Consumer Electronics*, vol. 60, no. 3, pp. 525–533, Aug 2014.

[46] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[47] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[48] 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.

[49] 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.

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

[51] 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.

[52] 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)*, Paris, France, 2006.

[53] 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)*, New Paltz, USA, 2013, pp. 1–4.

[54] 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)*, Dallas, USA, March 2010, pp. 165–168.

[55] 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)*, Juan les Pins, France, Sep. 2014, pp. 40–44.

[56] A. Blanco, M. Delgado, and M. Pegalajar, "A genetic algorithm to obtain the optimal recurrent neural network," *International Journal of Approximate Reasoning*, vol. 23, no. 1, pp. 67 – 83, 2000.

[57] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 12, pp. 245–271, 1997.

[58] M. A. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 6, pp. 1437–1447, 2003.

[59] M. Robnik-Šikonja and I. Kononenko, "An adaptation of Relief for attribute estimation in regression," in *Machine Learning: Proceedings of the Fourteenth International Conference (ICML)*, Nashville, USA, 1997, pp. 296–304.

[60] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[61] 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.

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, and M.Sc.Res degree in Signal Theory and Communications from the University of Vigo in 2013. 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 is currently an ESR Marie Curie fellow at Nuance Communications, Inc., United Kingdom, since April 2013. 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, 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.

Jose Lainez received his M.Sc. degree in Telecommunication Engineering from the University of Zaragoza, Spain, in 2006, and M.Sc.Res degree in Information Technology and Communications on Mobile Networks from the University of Zaragoza in 2008. Between 2006 and October 2012 he was as research engineer on the Speech processing group of the University of Zaragoza contributing to basic and applied research in a broad variety of fields: Automatic Speech Recognition (ASR), Distributed Speech Recognition (DSR), automatic subtitle generation, and pitch and formants detection. He is currently a Senior Research Scientist at Nuance Communications, Inc., Marlow, United Kingdom, since October 2012, carrying out applied research in ASR, language model adaptation, confidence estimation, and signal processing.

Daniel Barreda received his M.Sc. degree in Telecommunication Engineering from the Polytechnic University of Catalonia and the European Masters in Language and Speech (EMLS) in 2004, after doing an internship in LIMSI (France) working on Speech Understanding. Between November 2005 and July 2006 he did an internship in Toshiba R& D (Japan), focusing his work on Robust Speech Recognition. He is a senior research scientist at Nuance Communications, United Kingdom, where he has been working since 2006 on speaker diarisation, language identification, natural language processing, and more recently on machine learning.

Toon van Waterschoot (S'04, M'12) received the MSc degree (2001) and the PhD 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 General 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 (M'89, SM'07) received his BEng degree in Electronic and Electrical Engineering from the University of Sheffield, U.K., in 1986 and the PhD. degree from Imperial College, London, U.K., 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 multi-channel 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 (EURASIP) and formerly an associate editor of IEEE Signal Processing Letters and IEEE Transactions on Audio Speech and Language Processing.