KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Pablo Peso Parada, Dushyant Sharma, Patrick A. Naylor, and Toon van Waterschoot **Analysis of prediction intervals for non-intrusive estimation of speech clarity index** in *Proc. AES 60th Int. Conf. Dereverberation and Reverberation of Audio, Music, and Speech*, Leuven, Belgium, Feb. 2016.
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	http://www.aes.org/e-lib/browse.cfm?elib=18077
Conference homepage	http://www.aes.org/conferences/60
Author contact	toon.vanwaterschoot@esat.kuleuven.be *+ 32 (0)16 321927*
IR	ftp://ftp.esat.kuleuven.be/pub/SISTA/vanwaterschoot/abstracts/15-139.html

(article begins on next page)

Analysis of prediction intervals for non-intrusive estimation of speech clarity index

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

[1]*Nuance Communications Inc., Wethered House, Pound Lane, SL7 2AF Marlow, UK*

[2]*Dept. of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ UK*

[3]*Dept. of Electrical Engineering (ESAT-STADIUS/ETC), KU Leuven, Kasteelpark Arenberg, 3001 Leuven, Belgium*

Correspondence should be addressed to Pablo Peso Parada (pablo.peso@nuance.com)

ABSTRACT

We present an analysis of prediction intervals for a non-intrusive method to estimate the clarity index (C_{50}). The method employed to estimate C_{50} is a data driven approach that extracts multiple features from a reverberant speech signal which are then used to train a bidirectional long-short term memory model which maps 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. This approach was shown to provide a coverage probability of 80%, i.e. 80% of times the ground truth lies between the estimated intervals, where the interval bounds are computed by using 5.6 times the standard deviation of the per-frame estimates. This accuracy is shown to be consistent with other noisy reverberant environments.

1. INTRODUCTION

Sound propagates from the source to the 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 which depends on the room characteristics and the positions of both source and receiver. The reverberation time (T_{60}) indicates the acoustic properties of a room and it is independent of the positions of both source and receiver in a diffuse reverberant field. Different objective measurements such as Direct-to-Reverberation Ratio (DRR) [1] or clarity index (C_{50}) [1] may be employed to take into account this position dependency. The computation of these measures requires an estimation of the Room Impulse Response (RIR), however in many real scenarios this information remains unavailable and these measures need to be non-intrusively estimated from the reverberant signal.

Several methods have been proposed to blindly estimate T_{60} [2–4]. Kendrick et al. [5] compare two methods to estimate from speech and music signals different room acoustic parameters, mainly T_{60} and C_{80}. The first 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 [6] 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} [7] or DRR [8], or per frequency bin [9], this paper focuses on the problem of single-channel full-band C_{50} estimation.

Estimates of room acoustics parameters has a number of applications, for example, to estimate the perceived quality [10] or intelligibility [1] of reverberant recordings, to de-reverberate speech signals [11] or to perform reverberant speech recognition [12]. In addition, information about the accuracy 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 with the speech recognition performance compared to other measures of reverberation [6, 13].

The key contributions of this paper are to propose a

non-intrusive room acoustic (NIRA) parameter estimation method, i.e. the only available information to compute the acoustic parameter is the reverberant signal, to estimate C_{50} based on extracting a number of per-frame features from the reverberant speech and to compute prediction intervals. This method differs from the method presented previously [6] in two ways. First, only the frame-based features are used along with new features based on modulation domain representation and deep scatter spectrum transformation. Second, the Classification And Regression Tree (CART) is replaced by a recurrent neural network which models the relationship between these features and the room acoustics parameter. This technique was tested on a single-channel database created with simulated and real RIRs. The computation of prediction intervals from the per-frame C_{50} estimates provides additional information about the estimates.

The remainder of the paper is organized as follows. Section 2 and 3 describe the methods proposed in this work to estimate C_{50} and its prediction intervals respectively. In Section 4 the metrics used to evaluate the methods are introduced and results are presented in Section 5. Finally, in Section 6 the conclusions of this contribution are drawn.

2. NIRA METHOD

The method shown in Fig. 1 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) [14] is employed to discard non-speech frames. The following 107 frame-based features are computed from the reverberant signal based on [6] 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 [15] 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 [16] 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 [17].

These features are used to train a Bidirectional Long-Short Term Memory (BLSTM) [18] 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 correlation present in reverberation with its feedback connections and memory cells. Alternative learning algorithms such as CART, 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. Since we assume that the room acoustic properties remain unchanged within each utterance, the $\widehat{C_{50_n}}(s_n)$ estimate of the nth utterance s_n is computed as the mean of the per-frame estimates $\widehat{C_{50_{l,n}}}(s_n)$ for that utterance:

$$\widehat{C_{50_n}}(s_n) = \frac{1}{L}\sum_{l=1}^{L}\widehat{C_{50_{l,n}}}(s_n) \quad \text{dB}, \tag{1}$$

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/

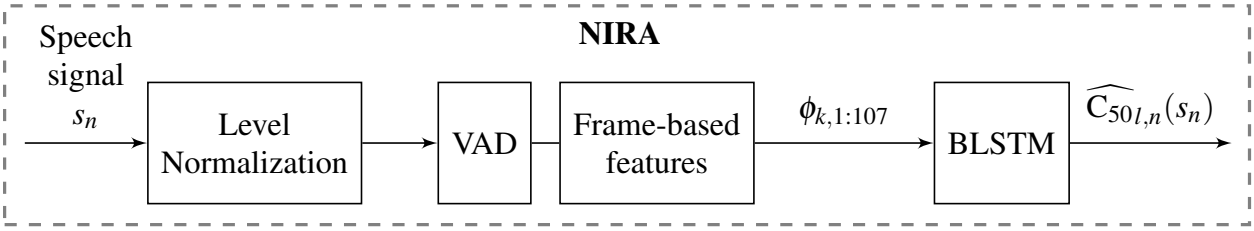


Fig. 1: The NIRA method.

3. PREDICTION INTERVALS

The NIRA method provides a $\widehat{\mathrm{C}_{50n}}(s_n)$ estimation for the nth utterance s_n using (1), however 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 $\widehat{\mathrm{C}_{50}}_{l,n}(s_n)$ estimate.

The ground truth C_{50n} computed directly from the RIR differs from the $\mathrm{C}_{50n}(s_n)$ that can be observed in the recording. This difference is caused by several factors. One of these is due to the spectrum of the speech signal s_n which is not flat and therefore only some frequencies of the RIR are excited in the recording. Therefore $\mathrm{C}_{50}(s_n)$ will differ by $\varepsilon(s_n)$ from C_{50n} and can be written as follows,

$$\mathrm{C}_{50n} = \mathrm{C}_{50n}(s_n) + \varepsilon_n(s_n) \quad \text{dB}, \tag{2}$$

which can be rewritten as,

$$\mathrm{C}_{50n} - \widehat{\mathrm{C}_{50n}}(s_n) = (\mathrm{C}_{50n}(s_n) - \widehat{\mathrm{C}_{50n}}(s_n)) + \varepsilon_n(s_n) \text{ dB}. \tag{3}$$

While confidence intervals address the differences between $\mathrm{C}_{50n}(s_n)$ and $\widehat{\mathrm{C}_{50n}}(s_n)$, prediction intervals deal with the left-hand side of (3), which is related to $P(\mathrm{C}_{50n}|\widehat{\mathrm{C}_{50}}(s_n))$ [19]. We are interested in this last parameter since the estimator is evaluated using the ground truth C_{50n}.

There are two sources of uncertainty σ_v and σ_m in (3). The former variance σ_v is due to data limitations, e.g. the spectrum is not flat, and the latter variance σ_m is due to model limitations causing estimation errors. Considering these errors to be statistically uncorrelated, we can write

$$\sigma_{TOTAL} = \sigma_v + \sigma_m \quad \text{dB}. \tag{4}$$

Confidence intervals are based on the uncertainty σ_m, however prediction intervals are based on σ_{TOTAL} [20].

In this paper we propose to estimate σ_{TOTAL} for the nth utterance ($\sigma_{TOTAL,n}$) from the L per-frame estimates $\widehat{\mathrm{C}_{50}}_{l,n}(s_n)$ as

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

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

$$\mathrm{L}_n = \widehat{\mathrm{C}_{50n}}(s_n) - k \cdot \sigma_{TOTAL,n} \quad \text{dB} \tag{5}$$

$$\mathrm{U}_n = \widehat{\mathrm{C}_{50n}}(s_n) + k \cdot \sigma_{TOTAL,n} \quad \text{dB} \tag{6}$$

where k is a tuning parameter that defines the width of the intervals. Figure 2 shows different PIs computed for a given $\widehat{\mathrm{C}_{50}}_{l,n}(s_n)$ using several values of k.

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 (5) and (6); 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.

Training set

A total of 32 utterances are selected randomly from the training set of TIMIT [21] database ensuring that 2 different male and 2 female speakers are included for each dialect and excluding 'SA sentences'. These speech signals are convolved with 192 simulated RIRs generated

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 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).

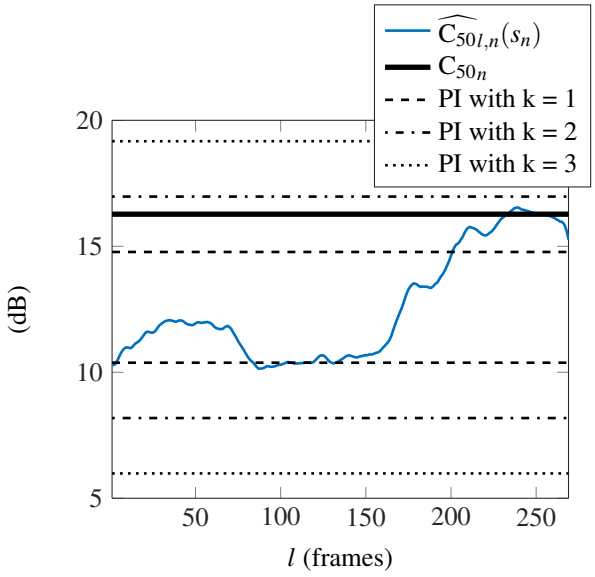


Fig. 2: Different PIs depending on k for a given utterance of the development set.

with the randomized image method [22] 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 [23] 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 [14].

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.

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 [24]; REVERB challenge [25]; C4DM RIR [26]; and SMARD [27]. 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 which are evaluated independently to assess the performance of C_{50} estimation and the prediction intervals, and to provide insights into impact of different types and levels of noise on the final performance.

4.2. Evaluation metrics

The C_{50} estimator method described in this paper is evaluated using box plots of the estimation error, which is computed for the nth utterance as $\widehat{C_{50_n}}(s_n) - C_{50_n}$.

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

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

where

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

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):

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

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

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 which was found to provide the lowest error of all explored architectures in the development set.

5.1. C_{50} estimation

Figure 3 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. 3 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. 3 with higher IQR and bias compared to the environments with white noise. Nevertheless, in all cases the bias is low.

5.2. Prediction intervals

Figure 4 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 behaviour 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. 4, at $k = 5.6$ with a NMPIW=30.20% in the development test. 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 $\widehat{C_{50_n}}(s_n)$ is in the middle ground truth range, i.e. $\widehat{C_{50_n}}(s_n) = 15.5$ dB, the prediction interval bounds are the same as the ground truth range limits when NMPIW=100%.

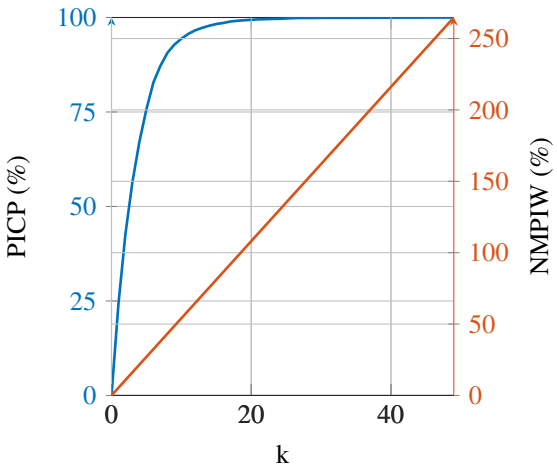


Fig. 4: Values of PICP and NMPIW depending on the tuning parameter k tested on the development set.

In order to assess the repeatability of this performance we use the evaluation set. Figure 5 shows the PICP and NMPIW 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 character-

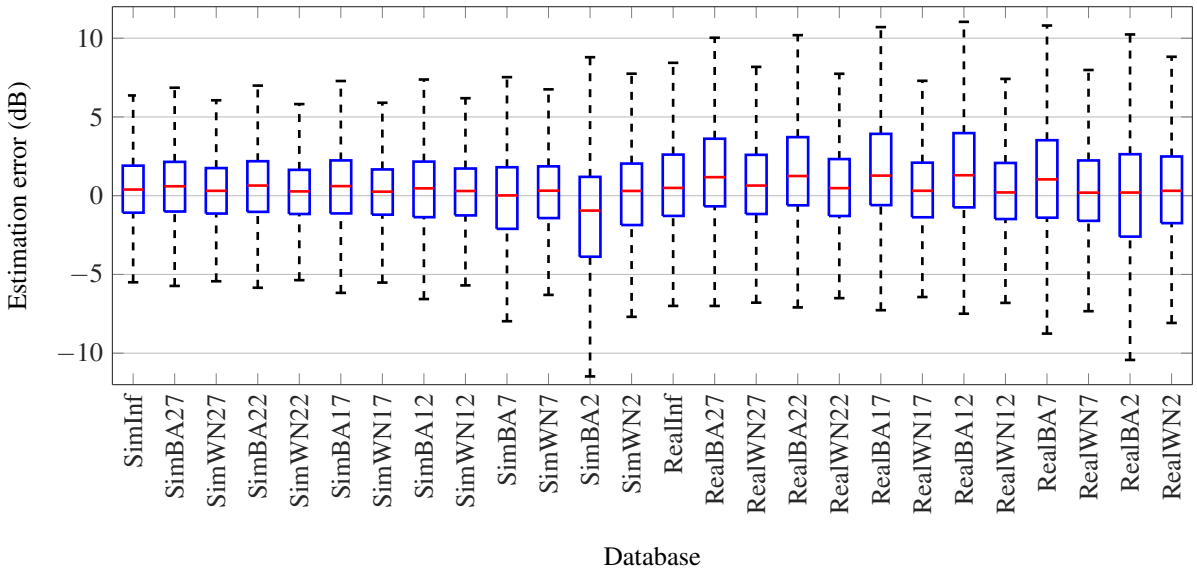


Fig. 3: Distribution of the C_{50} estimation errors for each database. The edges of the boxes indicate the lower and upper quartile range, while the horizontal lines inside the boxes represent the medians for each method. Moreover, the horizontal lines outside the boxes indicate the estimation error up to 1.5 times the interquartile range. Outliers are removed from this plot.

istics of the training set which only includes simulated RIRs.

The higher NMPIW values for the scenarios which 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, which is more prominent as the SNR decreases.

Overall, Figure 5 shows that the values of PICP and NMPIW for each evaluation subset are consistent with the PICP and NMPIW obtained in the development set.

6. CONCLUSIONS

We have presented a non-intrusive method for C_{50} estimation and an approach that can provide information about the accuracy of the estimation. 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 ± 4.69 dB from the estimation.

This method was validated with an evaluation set comprised of 26 subsets that includes different RIRs, noise levels and noise types. 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 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.

7. REFERENCES

[1] H. Kuttruff, *Room Acoustics*. London: Taylor & Francis, fifth ed., 2009.

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

[3] J. Eaton, N. D. Gaubitch, and P. A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *Proc. IEEE International Confer-*

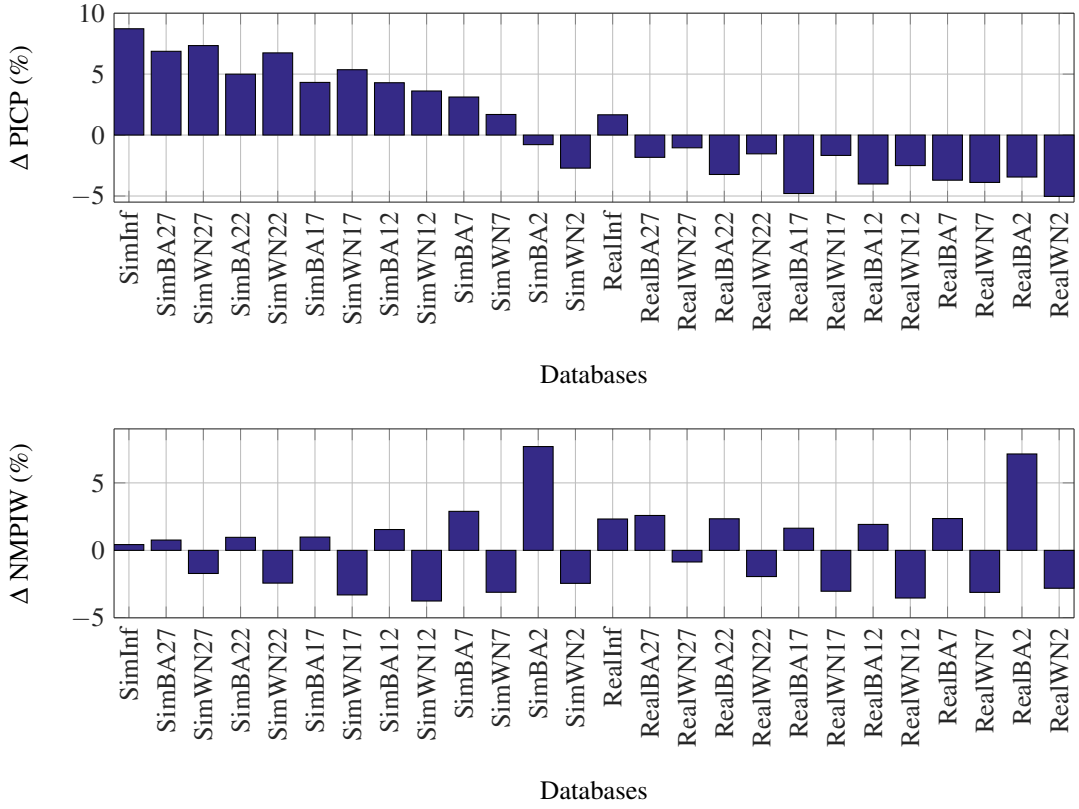


Fig. 5: Difference between the PICP and NMPIW achieved in the development set and the PICP and NMPIW obtained for the different evaluation subsets using $k = 5.6$ which provides a PICP=80% in the development set.

ence on Acoustics, Speech and Signal Processing (ICASSP), pp. 161–165, 2013.

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

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

[6] 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)*, pp. 4718–4722, 2014.

[7] 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)*, pp. 5187–5191, May 2014.

[8] 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)*, pp. 4713–4717, May 2014.

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

[10] 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)*, pp. 8169–8173, 2013.

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

[12] 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.

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

[14] "Objective measurement of active speech level," *ITU-T Recommendation P.56*, March 1993.

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

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

[17] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, pp. 4114–4128, Aug 2014.

[18] 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.

[19] 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.

[20] 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.

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

[22] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, pp. 774–786, April 2015.

[23] 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.

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

[25] 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)*, pp. 1–4, 2013.

[26] 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)*, pp. 165–168, March 2010.

[27] 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)*, pp. 40–44, Sep. 2014.