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 **Reverberant speech recognition exploiting clarity index estimation** *EURASIP J. Adv. Signal Process.*, vol. 2015, Article no. 54, 12 p., 2015.
Archived version	Final publisher's version / pdf
Published version	http://dx.doi.org/10.1186/s13634-015-0237-7
Journal homepage	http://asp.eurasipjournals.com
Author contact	toon.vanwaterschoot@esat.kuleuven.be *+ 32 (0)16 321927*
IR	ftp://ftp.esat.kuleuven.be/pub/SISTA/vanwaterschoot/abstracts/15-39.html

(article begins on next page)

Parada *et al. EURASIP Journal on Advances in Signal Processing* (2015) 2015:54
DOI 10.1186/s13634-015-0237-7

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Reverberant speech recognition exploiting *clarity index* estimation

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

Abstract

We present single-channel approaches to robust automatic speech recognition (ASR) in reverberant environments based on non-intrusive estimation of the *clarity index* (C_{50}). Our best performing method includes the estimated value of C_{50} in the ASR feature vector and also uses C_{50} to select the most suitable ASR acoustic model according to the reverberation level. We evaluate our method on the REVERB Challenge database employing two different C_{50} estimators and show that our method outperforms the best baseline of the challenge achieved without unsupervised acoustic model adaptation, i.e. using multi-condition hidden Markov models (HMMs). Our approach achieves a 22.4 % relative word error rate reduction in comparison to the best baseline of the challenge.

Keywords: Reverberant speech recognition; C_{50}; HLDA; Acoustic model selection

1 Introduction

Automatic speech recognition (ASR) is increasingly being used as a tool for a wide range of applications in diverse acoustic conditions (e.g. health care transcriptions, automatic translation, voicemail-to-text, and voice interface for command and control). Of particular importance is distant speech recognition, where the user can interact with a device placed at some distance from the user. Distant speech recognition is essential for natural and comfortable human-machine voice interfaces such as used in, for example, the automotive sector and smartphone mobile applications.

1.1 Signal model

In a distant-talking scenario, reverberation causes a significant degradation in ASR performance. A reverberant sound is created in enclosed spaces by reflections from surfaces which create a multipath sound propagation from the source to the receiver. This effect varies with the acoustic properties of the room and the source-receiver distance, and it is characterized by the room impulse response (RIR). The reverberant signal $y(n)$ can be modelled as the convolution between the RIR $h(m)$ and the source signal $s(n)$ as follows:

$$y(n) = \sum_{m=0}^{m=\infty} h(m)s(n-m). \quad (1)$$

Typical RIRs can be divided into three different parts: the direct path, the early reflections corresponding to the first 50 ms after the direct path, and the late reverberation corresponding to reflections that are delayed more than 50 ms after the direct path. Early reflections cause spectral colouration of the signal, whereas late reverberation causes temporal smearing and characteristic ringing echoes of the signal [1].

1.2 Room acoustic measures

Several acoustic measures have been proposed that estimate the reverberation level present in a signal [2] by using the RIR $h(m)$ or the source $s(n)$ and received signal $y(n)$, but in many applications, the only information available is the received signal $y(n)$. Recently, methods have been proposed to estimate room acoustic measures from reverberant signals such as the reverberation time (T_{60}) [3–5] which characterizes the room acoustic properties. However, alternative measures have been shown to be more correlated with ASR performance such as *clarity index* (C_{50}) which is the ratio of the energy in the early reflections over the energy in late reflections [6], defined as

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

*Correspondence: pablo.peso@nuance.com
[1] Nuance Communications, Inc., Wethered House, Pound Lane, Marlow SL7 2AF, UK
Full list of author information is available at the end of the article

Springer

© 2015 Parada et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

where N_τ is an integer number of samples corresponding to 50 ms after the time arrival of the direct path. Such measures have been shown to predict ASR performance with significant reliability [7, 8] compared to other measures of reverberation. Moreover, different values of N_τ have been investigated in [7] showing that the number of samples N_τ corresponding to the range from 50 to 100 ms after the direct path provides the highest correlation values with the ASR performance.

1.3 Distant-talking ASR

ASR techniques robust to reverberation can be divided into two main groups [9–11]: front-end-based and back-end-based. The former approach suppresses the reverberation in the feature domain; therefore, the processing is performed after feature extraction. Li et al. [12] propose to train a joint sparse transformation to estimate the clean feature vector from the reverberant feature vector. In [13], a model of the noise is estimated from observed data by considering the late reverberation as additive noise, and then the feature vector is enhanced by applying vector Taylor series. A feature transformation based on discriminative training criterion inspired on Maximum Mutual Information is suggested in [14]. Additional features related to the amount of diffuse noise in each frequency bin and frame are employed in [15] to improve deep neural network-based ASR accuracy in noisy and reverberant environments. Yoshioka and Gales [16] present several front-end approaches such feature transformation or feature set expansion that are tailored to deep neural network acoustic models employed for distant-talking recognition.

The latter approach, back-end-based, modifies the acoustic models or the observation probability estimate to suppress the reverberation effect. Sehr et al. [17] suggest to adapt the output probability density function of the clean speech acoustic model to the reverberant condition in the decoding stage. Selection of different acoustic models trained for specific reverberant conditions using an estimation of T_{60} is proposed in [18]. In [19], an attenuation of late reverberation is proposed such as [20] to build several reverberant acoustic models which are selected using ground truth T_{60}. The RIR attenuation parameters are tuned to provide the highest recognition rate on a reverberant test set created with measured RIR. The early-to-late reverberation ratio, considering the first 110 ms of the RIR as part of the early reverberation, is used in [21] instead of T_{60} to select between different reverberant acoustic models. In [22], the likelihood scores of the ASR acoustic models based on Gaussian mixture models are maximized to select the optimum acoustic model. An adaptation of multiple reverberant acoustic models trained with different T_{60} values is proposed in [23]. The mean vector of the optimal adapted model is estimated in a maximum-likelihood sense from the reverberant models. The idea in [24] is to add to the current state the contribution of previous acoustic model states using a piece-wise energy decay curve which considers the early reflections and late reverberation as different contributions.

In addition to front-end-based and back-end-based approaches, signal-based methods are intended to dereverberate the acoustic signal in the time domain, before being processed by the ASR feature extraction module [2]. In [25], a complementary Wiener filter is proposed to compute suitable spectral gains which are applied to the reverberant signal to suppress late reverberation. In [26], a denoising autoencoder is used to clean a window of spectral frames and then overlapping frames are averaged and transformed to the feature space. All these three approaches may be combined to create complex robust systems [27, 28].

Additionally, ASR techniques robust to reverberation can be also classified according to the number of microphones used to capture the signal such as single-channel methods [13, 20, 26, 29] or multi-channel techniques [12, 27, 30, 31].

The method now proposed is a hybrid approach based on front-end-based and back-end-based single-channel techniques. The C_{50} estimate is employed to select different acoustic models (back-end approach) which are trained on feature vectors appended to include the C_{50} value (front-end approach). The resulting appended feature vector is then reduced in dimension to match the original dimensionality by applying heteroscedastic linear discriminant analysis (HLDA) [32]. The technique was tested within the ASR task of the REVERB Challenge [33] which was launched by the IEEE Audio and Acoustic Signal Processing Technical Committee in order to compare ASR performance on a common data set of reverberant speech. This paper now extends an earlier version of the work presented in [34] including an improved C_{50} estimator, which provides estimates per frame, and a performance comparison of the new system with the previous method.

The remainder of this paper is organized as follows: Section 2 introduces the C_{50} estimators employed in this work. In Section 3, the training and test data from the REVERB Challenge is analysed. Section 4 describes the methods proposed, and Section 5 discusses the comparative performance of the these techniques. Finally, in Section 6, the conclusions are drawn.

2 C_{50} estimator

Two different single-channel C_{50} estimators are employed in this work: non-intrusive room acoustic estimation using classification and regression trees (NIRA-CART) and non-intrusive room acoustic estimation using

bidirectional long-short term memory (NIRA-BLSTM). In this work, we use C_{50} to characterize reverberation in the signal instead of T_{60} as in [18] because this last measure is independent of the source-receiver distance which is a key factor in the speech degradation. Moreover, C_{50} was shown to be highly correlated with the ASR performance compared to other measures of reverberation [7, 8] which makes it suitable for this purpose.

2.1 NIRA-CART

This method in [7] computes a set of features from the signal which can be divided into long-term features and frame-based features. The former features are taken from long-term average speech spectrum (LTASS) deviation by mapping it into 16 bins with equal bandwidth and additionally from the slope of the unwrapped Hilbert transformation. The latter features are created with pitch period, importance weighted signal-to-noise ratio (iSNR), zero-crossing rate, variance and dynamic range of Hilbert envelope and speech variance. In addition, spectral centroid, spectral dynamics and spectral flatness of the power spectrum of long-term deviation (PLD) are included in the feature vector as well as 12th-order mel-frequency cepstral coefficients (MFCCs) with delta and delta-delta and line spectrum frequency (LSF) features computed by mapping the first 10 linear predictive coding (LPC) coefficients to LSF representation.

The first-order numerical difference is used to compute the rate of change for all frame-based features, excluding the MFCCs. The complete feature vector is created by adding to the long-term features the mean, variance, skewness and kurtosis of all frame-based features and therefore creating a 313-element vector. Finally, a CART regression tree [35] is built to estimate C_{50}. The CART uses the complete feature vector, and it is trained on the training set from the REVERB Challenge.

2.2 NIRA-BLSTM

The feature configuration of this method (P Peso Parada, D Sharma, J Lainez, D Barreda, P A Naylor, T van Waterschoot, A single-channel non-intrusive C50 estimator with application to reverberant speech recognition, submitted) is based on computing the frame-based features of NIRA-CART and including in addition 12 features extracted from the modulation domain. Consequently, the per frame feature vector comprises a total of 94 features. Moreover, rather than building a CART model to estimate C_{50}, a particular recurrent neural network architecture called BLSTM [36] is trained with these features to provide an estimation every 10 ms. Since REVERB Challenge data assumes that the room acoustic properties remain unchanged within each utterance, only the temporal average for each utterance of all per frame estimations is considered.

2.3 Wide-band feature set extension

In [7] and (P Peso Parada, D Sharma, J Lainez, D Barreda, P A Naylor, T van Waterschoot, A single-channel non-intrusive C50 estimator with application to reverberant speech recognition, submitted), these estimators were originally proposed to operate on speech signals sampled with a sampling frequency of 8 kHz. Therefore, an adaptation of the features has been developed here in order to process wider bandwidth signals. For speech signals sampled at 16 kHz, 10 LPC coefficients and their corresponding LSFs are not sufficient to characterize the speech [37]. For wide-band speech therefore, the order of the LPC is increased to 20. Hence, the feature vector for NIRA-CART comprises 393 elements and 106 features per frame for NIRA-BLSTM.

3 Analysis of the challenge data

The database provided in REVERB Challenge comprises three different sets of eight-channel recordings: training set, development set and evaluation set. Real data recorded in a reverberant room and simulated data created by convolving non-reverberant utterances with measured RIRs are included in the development set and evaluation set, whereas the training set only comprises simulated data. This section analyses the RIRs of different data sets in terms of C_{50} inasmuch as this is a key aspect in the design of the algorithms proposed in this work.

Figure 1 shows the histogram of C_{50} values for the 24 training RIRs including all channels of each response. As seen in Fig. 1, the RIR training set covers a wide range of C_{50} spanning approximately 25 dB. These RIRs are used to create the data set employed to train our C_{50} estimator [7] by convolving these RIRs with speech signals from the training set which, for the REVERB Challenge, was formed from the WSJCAM0 training set [38].

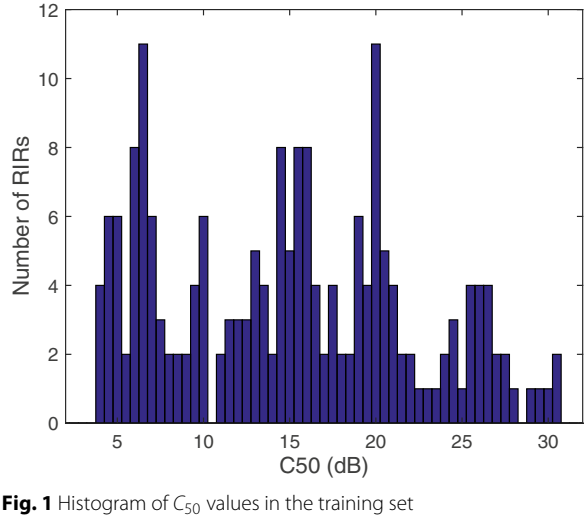


Fig. 1 Histogram of C_{50} values in the training set

Table 1 presents the measured C_{50} of the RIRs included in the development and evaluation sets of simulated data. It shows a significant difference between the small room recordings (Room1) which are less reverberant ($T_{60} = 0.25$), and the medium and large room recordings (Room2 and Room3, respectively) which have higher reverberation times ($T_{60} = 0.5$ and $T_{60} = 0.7$, respectively). Furthermore, the two distances of the speaker from the microphone, that is, *near* = 50 cm and *far* = 200 cm from, show a constant C_{50} difference of 8 to 10 dB.

Real recordings are captured in a reverberant meeting room from two different distances: near (≈100 cm) and far (≈250 cm). The development and evaluation sets of these recordings are not analysed in terms of measured C_{50} since the RIRs of these sets are unavailable.

3.1 C_{50} estimator performance

The evaluation metric used to compare the C_{50} estimator performance is the root-mean-square deviation (RMSD) given as

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\widehat{C_{50,n}} - C_{50,n})^2}\ \text{dB}, \tag{3}$$

where N is the total number of measured ground truth values $C_{50,n}$ and estimated scores $\widehat{C_{50,n}}$ considered to compute the RMSD.

The training set is randomly split into a training subset (80 % of the data used to train the models) and evaluation subset (20 % of the recordings employed to evaluate the models) in order to provide insights into the performance of both C_{50} estimators. Additionally, the performance of the C_{50} estimators is also evaluated using the development set and evaluation set of the simulated data whose C_{50} measures are presented in Table 1. Table 2 summarizes the RMSD performance of each estimator evaluated in these data sets. NIRA-BLSTM achieves the lowest deviation in each data set, providing on average a RMSD 1.6 dB lower than that of NIRA-CART. Both estimators exhibit lower deviations on the evaluation subset of the training set (i.e. *training set - eval. subset*) because this reverberant subset is similar to the data used for training the C_{50} estimators.

4 Methods

In this section, we describe different configurations for reverberant speech recognition. The idea underpinning these methods is to exploit estimated C_{50} to improve robustness of ASR to reverberation. Section 4.1 introduces the front-end techniques, Section 4.2 describes the back-end methods, and finally, Section 4.3 presents the combination as outlined in Fig. 2.

4.1 C_{50} as a supplementary feature in ASR

In this approach, the estimated C_{50} of the utterance is included as an additional feature in the ASR feature vector. The baseline recognition system uses a feature vector with 13 mel-frequency cepstral coefficients, with the first and second derivatives of these coefficients followed by cepstral mean subtraction.

We now propose two alternative improved configurations. The first configuration proposed (*C50FV*) is to add C_{50} estimation directly to this feature vector. Therefore, the modified feature vector comprises 40 elements.

In the second configuration (*C50HLDA*), the feature vector dimension is reduced using linear discriminant analysis (LDA) [39]. This method projects the input feature vector $\mathbf{x}_k$ onto a new space $\mathbf{y}_k$ by applying a linear transformation $\mathbf{W}$ such that

$$\mathbf{y}_k = \mathbf{W}^T\mathbf{x}_k, \tag{4}$$

where $\mathbf{W}$ is an $p \times q$ matrix. This transformation in general retains the class discrimination in the transformed feature space. The transformation $\mathbf{W}$ is obtained by maximizing the ratio of the between-class scatter matrix $\mathbf{S}_B$ to the within-class scatter matrix $\mathbf{S}_W$, that is,

$$\breve{\mathbf{W}} = \arg\max_{\mathbf{W}} \frac{|\mathbf{W}^T\mathbf{S}_B\mathbf{W}|}{|\mathbf{W}^T\mathbf{S}_W\mathbf{W}|}. \tag{5}$$

The projection that maximizes (5) corresponds to $\breve{\mathbf{W}}$ whose columns are the eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$ with the q highest eigenvalues so that q is the dimension of the reduced feature space.

In this work, a model-based generalization of LDA [32] is used. In this case, the linear transformation is estimated from Gaussian models using the expectation-maximization algorithm. For these models, it is assumed that class distributions with equal mean and variance across all classes do not contain discriminant classification information.

In all configurations, the acoustic models are trained using the modified feature space.

Table 1 C_{50} measures of the RIRs included in the development set (dev. set) and evaluation set (eval. set) of the simulated data from the REVERB Challenge

		Room1		Room2		Room3	
		Near	Far	Near	Far	Near	Far
Dev. set	C_{50} (dB)	30.78	21.62	16.52	7	16.37	6.69
Eval. set	C_{50} (dB)	29.44	22.043	14.47	6.27	15.10	7.06

Table 2 RMSD of the C_{50} estimators tested in three different sets

Estimator	RMSD (dB)		
	Training set - eval. subset	Sim. data - dev. set	Sim. data - eval. set
NIRA-CART	1.86	3.60	3.16
NIRA-BLSTM	0.46	2	1.37

4.2 Model selection

The proposed back-end approach aims to select the optimal acoustic model $\breve{\mathcal{A}}$ such as

$$\breve{\mathcal{A}}(C_{50}) = \begin{cases} \mathcal{A}_1 & -\infty < C_{50} \leq \theta_1 \\ \mathcal{A}_2 & \theta_1 < C_{50} \leq \theta_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \mathcal{A}_J & \theta_{J-1} < C_{50} < \infty \end{cases} \quad (6)$$

where J represents the number of available acoustic models $\boldsymbol{\mathcal{A}} = \{\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_J\}$ and $\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots, \theta_{J-1}\}$ is the vector with the C_{50} threshold values sorted in ascending order.

4.2.1 Model switching between REVERB Challenge acoustic models

The first configuration (Clean&Multi cond.) is based on selecting between the two acoustic models provided in the challenge (clean-condition hidden Markov models (HMMs) and multi-condition HMMs) according to the level of C_{50} estimated from the input signal. In this case, $\mathcal{A}_1$ represents the multi-condition HMMs and $\mathcal{A}_2$ is the clean-condition HMMs. By empirical optimization over the development data set and considering the analysis carried out in Section 3, we choose the model switching threshold $\theta_1 = 23$ dB. Therefore, input speech signals with estimated C_{50} higher than 23 dB are recognized using clean-condition HMMs, whereas signals with C_{50} lower than this threshold are recognized using multi-condition HMMs.

4.2.2 Model switching using newly trained acoustic models

The second and subsequent configurations are now introduced based on training new reverberant acoustic models. The data set used to train the models is always the clean training set convolved with the training RIRs (Fig. 1). In order to include in the trained models $\boldsymbol{\mathcal{A}}$ all representative data of the acoustic units (i.e. triphones), all L clean training utterances are convolved with a subset of M training RIRs to create a reverberant acoustic model $\mathcal{A}_i$ such as

$$\mathbf{y}_l = \mathbf{H}_i(l \bmod M) * \mathbf{s}_l \quad l = 1, 2, \cdots, L \quad (7)$$

where $\mathbf{y}_l$ is the reverberant speech obtained with the clean utterance $\mathbf{s}_l$ and the RIR in the row $(l \mod M)$ of the matrix $\mathbf{H}_i$. This matrix contains the M RIRs with a C_{50} value that satisfies $\theta_{i-1} < C_{50} \leq \theta_i$.

The first approach is to create three reverberant acoustic models (*MS3*) according to the C_{50} values of the RIRs as shown in Fig. 3a. The threshold vector is set to $\boldsymbol{\theta} = \{10, 20\}$ dB, which was derived from the C_{50} estimations of the development set. The aim is to cluster the development set into three groups with similar ASR performance and train a model for each group. The most reverberant model $\mathcal{A}_1$ is trained with the RIRs that have C_{50} lower than 10 dB. The second acoustic model $\mathcal{A}_2$ is trained with

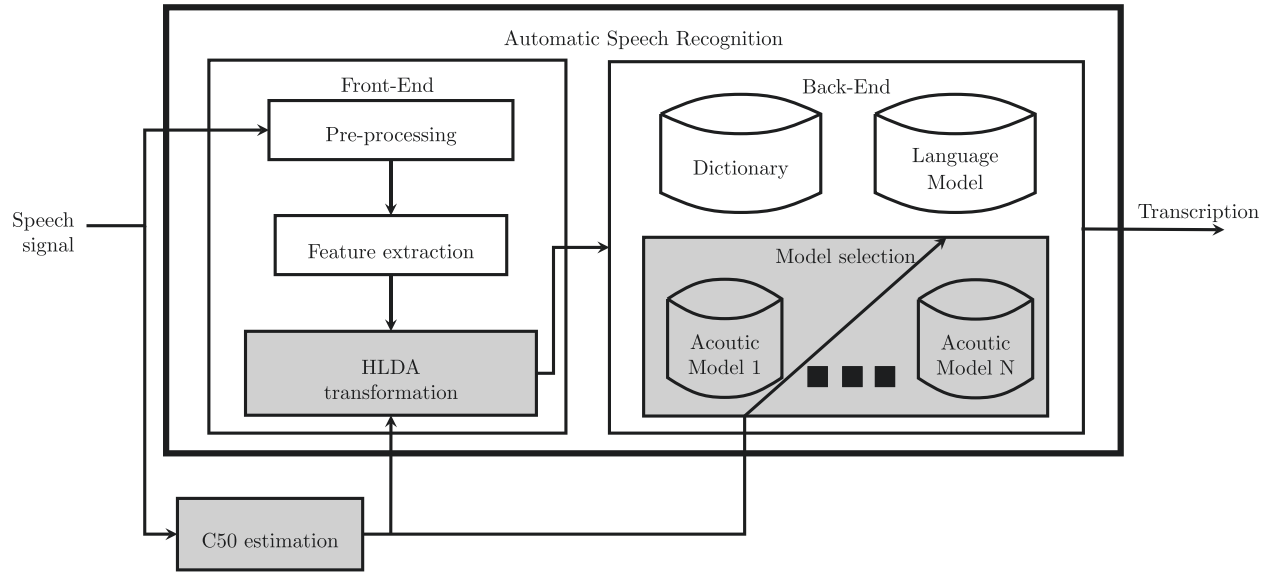


Fig. 2 Reverberant speech recognition using C_{50} estimation

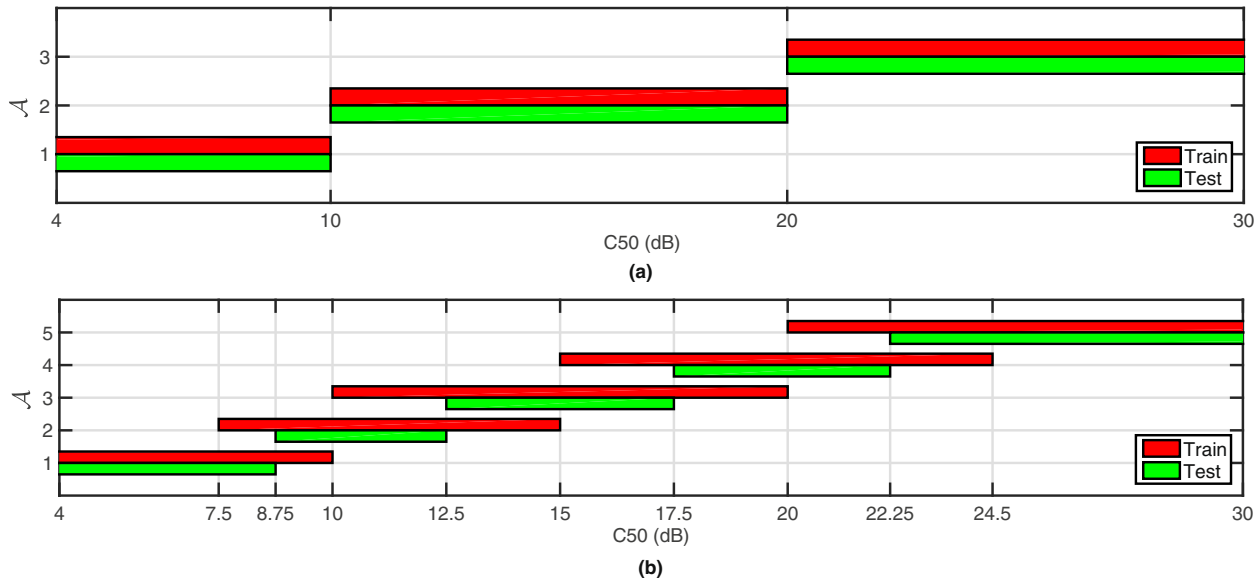


Fig. 3 Comparison of MS3 (**a**) and MS5 (**b**) configurations for training the acoustic (*red bars*) models and recognizing testing data (*green bars*) according to C_{50}. The difference is in the overlapping of the training data for MS5 configuration

RIRs that have C_{50} between 10 and 20 dB. Finally the third model $\mathcal{A}_3$, which represents the least reverberant conditions, is trained with those RIRs with a C_{50} higher than 20 dB.

The next configuration (*MS5*) includes the use of classes with overlapping ranges of C_{50} in order to build the acoustic models. For each class, the overlapping range of C_{50} used was approximately 50 % of the size of the neighbouring class. This configuration results in the same previous models (MS3) but adds two additional models spanning the transitional ranges of C_{50}. These two models provide a smoother transition between acoustic models. The acoustic model most representative of reverberation level estimated from the utterance is selected in the recognition phase. Figure 3b shows the construction of MS5 during training (red bars) and the thresholds used to select models in the recognition stage (green bars).

Additional configurations were tested by increasing the number of models trained: 8 overlapped acoustic models (*MS8*), 11 overlapped acoustic models (*MS11*), 14 overlapped acoustic models (*MS14*) and 18 overlapped acoustic models (*MS18*). These models are obtained by further dividing the original MS3 configuration. By increasing the number of models, the range of C_{50} of the training data of each model is decreased in terms of C_{50} which creates acoustic models more specific for each reverberant condition. Figure 4 shows the ranges of C_{50} used for MS11.

4.3 Model selection including C_{50} in the feature vector

This method combines the two approaches described above: C50HLDA and model selection. Figure 2 shows the block diagram of this method where green modules represent the modifications included to design this method. Firstly, C_{50} is estimated from the speech signal. The C_{50} estimate is then included in the feature vector before applying the HLDA transformation and also used to select the most suitable acoustic model.

All the tested configurations employ the C_{50} thresholds as described in Section 4.2 to create the data to train the acoustic models and select the appropriate acoustic model in the recognition stage. These configurations are referred as MS*N*+C50HLDA, where N represents the number of acoustic models created.

5 Results and discussion

Methods described in Section 4 were tested using NIRA-CART and NIRA-BLSTM to estimate C_{50}, and we compare the performance of each method in terms of the word error rate (WER) obtained using the REVERB Challenge ASR task [33]. The ASR evaluation tool is based on the hidden Markov model tool kit (HTK) provided by the REVERB Challenge. It uses mel-frequency cepstral coefficient (MFCC) features including delta and delta-delta coefficients and tied-state HMM acoustic models with 10 Gaussian components per state for clean-condition models and 12 Gaussian components per state for multi-condition models.

Table 3 shows the average WER achieved with the non-reverberant recordings (Clean), simulated reverberant recordings (Sim.) and real reverberant recordings (Real) of the REVERB Challenge evaluation test set including the average of all subsets in the last column, while

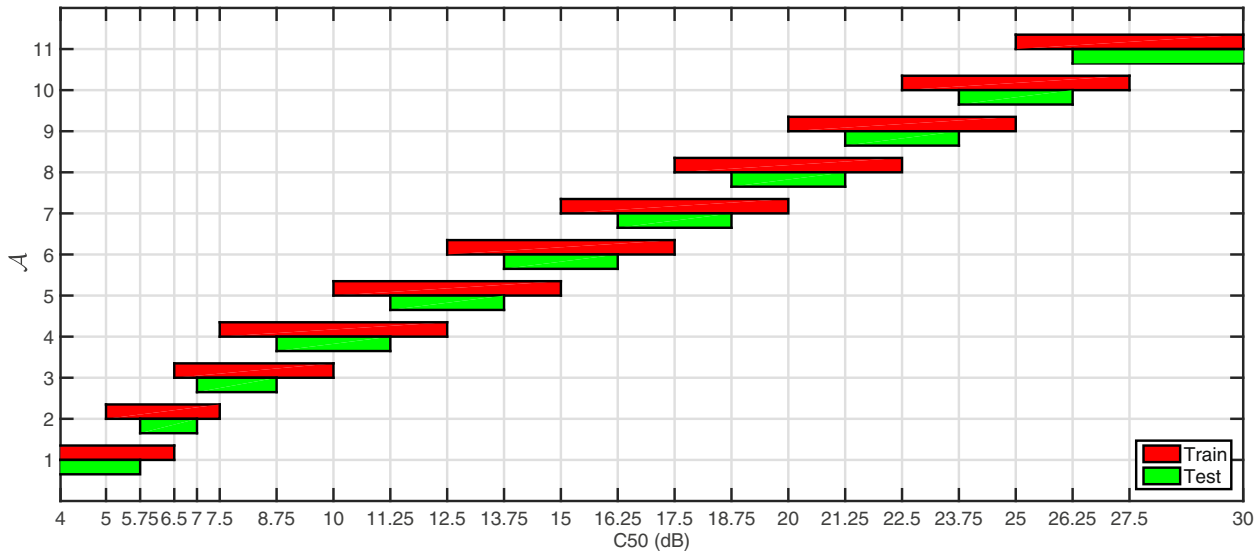


Fig. 4 MS11 configurations for training the acoustic models (*red bars*) overlapping of the training data and recognizing testing data (*green bars*) according to C_{50}

Tables 4 and 5 show with more detail these results for each scenario. Moreover, Fig. 5 summarizes these results, displaying the average WER for the development test set and evaluation test set.

Baseline methods are also tested in order to compare the performance. The baseline methods consist of decoding the data using the two acoustic models provided in the REVERB Challenge: the acoustic model trained with non-reverberant data (Clean-cond.) and the acoustic model trained with reverberant data (Multi-cond.). The performance of these baselines is shown in the first two rows of Tables 3, 4 and 5. Clean-cond. models provide a better performance in non-reverberant environments, whereas Multi-cond. models provide a significant reduction in WER for reverberant environments.

5.1 C_{50} as a new feature

The C50FV method provides a similar performance compared with the baselines. This outcome is due to the fact that we are using a diagonal covariance matrix to build the acoustic model. Therefore, this feature only provides information regarding the probability of observing the acoustic unit in this reverberant environment not taking into account possible dependences with the MFCC.

On the other hand, the last method described in Section 4.1 (C50HLDA) outperforms on average the WER obtained with the baselines. The main reason for this result is the use of the discriminative transformation matrix to combine the feature space. Regarding the C_{50} estimator employed, NIRA-BLSTM provides similar WER to that obtained with NIRA-CART for this configuration. This small performance difference suggests that C50HLDA does not strongly depend on the accuracy of the estimations. Furthermore, the averaged WER obtained by applying HLDA to the feature space without the C_{50} feature is 32.20 %. This result supports the previous suggestion about the dependence of C_{50} estimation accuracy upon C50HLDA performance and moreover indicates that the improvement achieved with C50HLDA is mainly due to the HLDA transformation.

5.2 Model selection

Tables 3, 4 and 5 also display the performance obtained with the methods described in Section 4.2 based on model selection. First, they show that a considerable WER reduction of the baseline is achieved by employing the two acoustic models provided by REVERB Challenge and exploiting our estimate of C_{50} to select the most appropriate model for each utterance between them (i.e. Clean&Multi cond.). Further improvement is achieved by training more reverberant models. The MS3 configuration employs three reverberant models (Fig. 3a) and the performance in reverberant conditions is improved in most of the situations, but on average, the error rate has been increased with respect to the Clean&Multi cond. mainly due to the poor performance in clean environments. The performance of this configuration is slightly improved, from WER = 30.82 % to WER = 30.35 % in the evaluation set, by overlapping the training data to build the acoustic models (MS5). Increasing the number of models trained using overlapping ranges of C_{50} (i.e. MS8, MS11, MS14 and MS18) results in further WER reductions.

Table 3 WER (%) averages obtained in evaluation data set

	Clean Avg.	Sim. Avg.	Real Avg.	Avg.
Clean-cond.	12.21	52.22	89.17	48.04
Multi-cond.	30.13	29.50	56.94	34.67
NIRA-CART				
Clean&Multi cond.	13.51	29.29	56.94	30.02
C50FV	28.65	29.72	56.84	34.37
C50HLDA	25.52	27.78	55.00	32.12
MS3	22.17	27.22	54.57	30.82
MS3+C50HLDA	19.90	25.24	52.51	28.75
MS5	22.32	26.35	54.38	30.35
MS5+C50HLDA	20.07	24.80	52.65	28.58
MS8	21.57	26.10	53.17	29.80
MS8+C50HLDA	19.69	24.08	51.04	27.79
MS11	21.10	26.04	56.62	30.26
MS11+C50HLDA	19.83	24.24	53.13	28.30
MS14	21.34	25.97	55.13	30.02
MS14+C50HLDA	19.38	23.75	52.31	27.76
MS18	21.96	25.97	55.85	30.32
MS18+C50HLDA	20.73	23.95	53.12	28.38
NIRA-BLSTM				
Clean&Multi cond.	*12.35*	29.06	56.94	29.58
C50FV	28.75	29.65	56.91	34.37
C50HLDA	25.86	27.75	54.56	32.12
MS3	20.67	26.79	53.44	29.98
MS3+C50HLDA	18.70	24.57	52.56	28.07
MS5	21.33	26.24	53.87	29.93
MS5+C50HLDA	19.42	24.46	52.07	28.11
MS8	19.97	25.51	53.14	29.03
MS8+C50HLDA	18.58	23.61	*50.96*	27.22
MS11	18.64	25.40	54.73	28.90
MS11+C50HLDA	17.76	23.47	51.92	27.09
MS14	18.99	25.09	54.31	28.75
MS14+C50HLDA	17.50	*23.17*	52.15	*26.90*
MS18	18.40	25.08	56.00	28.89
MS18+C50HLDA	16.96	23.30	52.64	26.91

The first two rows correspond to the baseline methods, and the remainder are the methods proposed in this work. Best performance results in each column are shown in italics

Table 4 WER (%) obtained with the non-reverberant part of the evaluation data set

	Clean		
	R1	R2	R3
Clean-cond.	12.83	12.20	11.62
Multi-cond.	30.29	30.00	30.10
NIRA-CART			
Clean&Multi cond.	13.98	13.76	12.81
C50FV	28.87	28.80	28.29
C50HLDA	25.84	24.97	25.76
MS3	22.31	21.64	22.59
MS3+C50HLDA	19.91	19.87	19.95
MS5	22.72	21.39	22.86
MS5+C50HLDA	20.18	19.57	20.47
MS8	21.94	20.69	22.11
MS8+C50HLDA	20.62	19.07	19.38
MS11	21.70	20.04	21.58
MS11+C50HLDA	20.67	19.76	19.06
MS14	21.57	20.63	21.84
MS14+C50HLDA	19.77	19.07	19.31
MS18	22.26	21.13	22.52
MS18+C50HLDA	21.47	20.31	20.41
NIRA-BLSTM			
Clean&Multi cond.	*12.98*	*12.32*	*11.76*
C50FV	28.80	29.02	28.44
C50HLDA	26.45	25.28	25.87
MS3	20.89	20.13	21.02
MS3+C50HLDA	18.84	18.40	18.87
MS5	21.62	20.68	21.70
MS5+C50HLDA	19.16	19.15	19.96
MS8	20.35	19.39	20.20
MS8+C50HLDA	19.04	18.33	18.39
MS11	19.03	18.04	18.87
MS11+C50HLDA	18.26	17.80	17.23
MS14	19.37	18.75	18.87
MS14+C50HLDA	17.55	17.74	17.23
MS18	18.50	18.20	18.51
MS18+C50HLDA	17.38	16.82	16.69

The first two rows correspond to the baseline methods, and the remainder are the methods proposed in this work. R1, R2 and R3 represent room numbers 1, 2 and 3, respectively. Best performance results in each column are shown in italics

For these experiments, the best performance is obtained with MS8 using the NIRA-CART C_{50} estimator (WER = 29.8 %), whereas NIRA-BLSTM provides the lowest WER with MS14 (WER = 28.7 %). This is due to the fact that NIRA-BLSTM achieves more accurate C_{50} estimations than NIRA-CART; hence, it is able to select acoustic models trained with a narrower, and therefore better matched, C_{50} range.

Table 5 WER (%) obtained with the reverberant part of the evaluation data set

	Sim.						Real	
	R1		R2		R3		R1	
	Near	Far	Near	Far	Near	Far	Near	Far
Clean-cond.	17.91	25.67	42.85	83.70	54.22	89.08	90.19	88.15
Multi-cond.	20.60	21.09	23.70	38.72	28.08	44.86	58.45	55.44
NIRA-CART								
Clean&Multi cond.	18.67	21.59	23.83	38.72	28.15	44.86	58.45	55.44
C50FV	20.62	20.74	23.12	39.14	28.19	46.61	58.19	55.50
C50HLDA	18.38	19.99	21.34	37.03	27.44	42.55	55.92	54.09
MS3	18.08	19.82	21.92	35.94	27.35	40.25	55.64	53.51
MS3+C50HLDA	17.16	19.40	20.60	32.67	25.37	36.32	53.53	51.49
MS5	16.32	18.52	20.49	36.34	25.85	40.62	55.35	53.41
MS5+C50HLDA	16.44	17.93	19.91	32.51	24.45	37.62	53.66	51.65
MS8	16.72	19.32	20.79	34.02	26.50	39.31	53.24	53.11
MS8+C50HLDA	15.72	18.26	19.79	30.76	24.16	35.85	*52.06*	50.03
MS11	16.50	18.99	21.14	34.75	25.85	39.09	57.87	55.37
MS11+C50HLDA	16.10	17.79	19.95	31.58	23.90	36.21	54.77	51.49
MS14	16.50	19.06	21.37	34.64	24.83	39.50	55.61	54.66
MS14+C50HLDA	15.88	17.93	19.73	30.78	*22.39*	35.86	52.67	51.96
MS18	16.25	19.13	21.19	34.96	24.94	39.40	56.50	55.20
MS18+C50HLDA	15.64	18.23	19.79	31.15	22.83	36.15	53.78	52.46
NIRA-BLSTM								
Clean&Multi cond.	18.01	21.08	23.70	38.72	28.08	44.86	58.45	55.44
C50FV	20.52	20.50	23.07	39.25	28.07	46.58	58.70	55.13
C50HLDA	18.40	19.69	21.31	37.16	27.18	42.83	55.35	53.78
MS3	16.93	18.87	21.79	35.99	27.25	39.98	54.52	52.36
MS3+C50HLDA	15.96	18.18	20.05	32.38	25.00	35.93	53.37	51.76
MS5	16.01	18.25	20.50	36.32	26.02	40.42	54.90	52.84
MS5+C50HLDA	16.16	17.15	19.50	32.67	24.07	37.25	53.40	50.74
MS8	16.01	18.50	20.13	34.12	25.39	39.00	53.66	52.63
MS8+C50HLDA	15.66	17.01	19.50	*30.65*	23.22	35.68	52.22	*49.70*
MS11	15.88	17.94	20.07	34.75	24.96	38.87	55.76	53.71
MS11+C50HLDA	14.79	16.79	18.85	31.50	22.85	36.09	53.27	50.57
MS14	15.66	17.42	20.08	34.23	24.25	38.95	55.41	53.21
MS14+C50HLDA	*14.35*	17.40	*18.48*	30.95	22.61	*35.32*	52.48	51.82
MS18	15.06	17.52	19.86	34.14	24.83	39.14	57.11	54.90
MS18+C50HLDA	14.81	*16.66*	18.93	31.02	22.49	35.93	54.07	51.22

The first two rows correspond to the baseline methods, and the remainder are the methods proposed in this work. R1, R2 and R3 represent room numbers 1, 2 and 3, respectively. Best performance results in each column are shown in italics

5.3 Model selection including C_{50} in the feature vector

The performance of the full system presented in Fig. 2 is now discussed. A significant improvement is observed by combining both methods; the WER is decreased by approximately 2 % absolute with respect to the error achieved using only model selection. NIRA-CART offers the best performance with MS8+C50HLDA (WER = 27.8 %) and NIRA-BLSTM with MS14+C50HLDA (WER

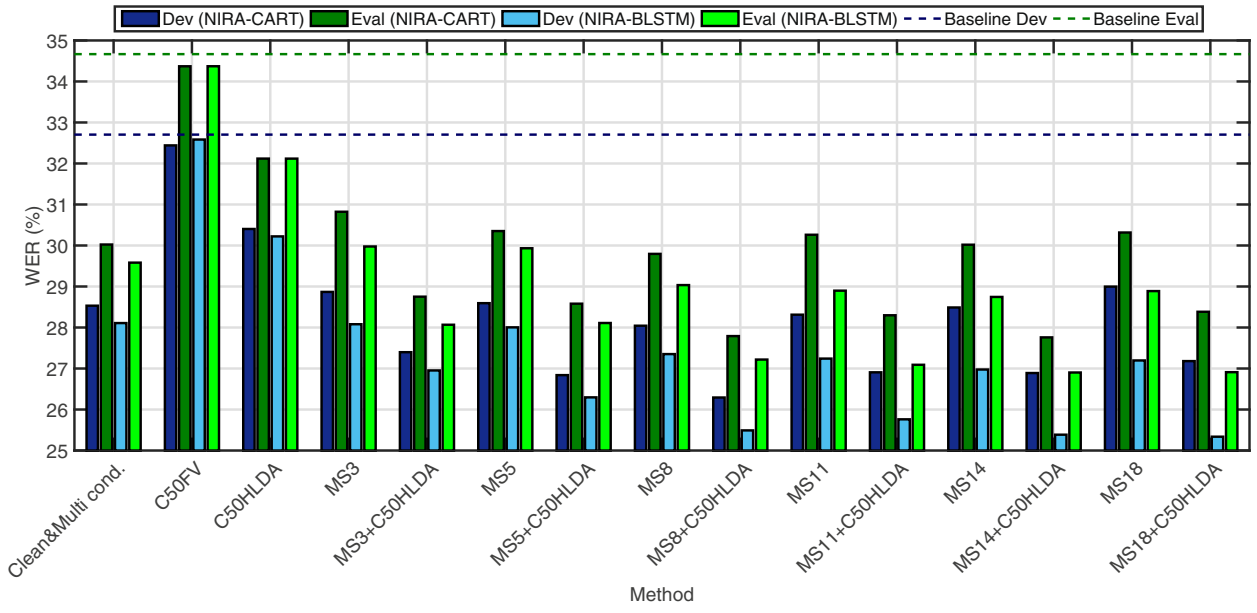


Fig. 5 Comparison of the ASR performance of several methods (*bars*) against the baselines (*dotted lines*) for the development test set (*blue*) and evaluation test set (*green*) using both C_{50} estimators (NIRA-CART and NIRA-BLSTM)

= 26.9 %), which outperform the best baseline method (Multi-cond.) by 6.9 % and 7.8 %, respectively, in the evaluation set.

Tables 3, 4 and 5 highlight in italics the lowest WER obtained in each data set. The best performance in reverberant conditions is achieved with this full system (i.e. MS*N*+C50HLDA); however, Clean&Multi cond. shows the best performance in the non-reverberant condition. This is mainly because all the data used to train MS*N*+C50HLDA is reverberant data, while Clean&Multi cond. uses reverberant and clean data to train the acoustic models. Therefore, MS*N*+C50HLDA could be further improved including a clean acoustic model to recognize the non-reverberant data.

All these reverberant speech recognition approaches were investigated in the previous work [34] using NIRA-CART. Figure 5 shows that using a more accurate C_{50} estimator, i.e. NIRA-BLSTM, the WER is further reduced.

The method proposed in Fig. 2 may potentially be complementary to some other reverberation-robust speech recognition methods, such as applying speaker adaptation, acoustic model adaptation or preprocessing schemes (e.g. beamforming) [40]. For example, performing an unsupervised acoustic model adaptation using constrained maximum likelihood linear regression (CMLLR) with the best method proposed in this work (MS14+C50HLDA using NIRA-BLSTM), the average WER is further reduced to 24.34 %, that is, a relative WERR of 9.88 % with respect to the best baseline of the REVERB Challenge using CMLLR.

6 Conclusions

Various approaches for single-channel reverberant speech recognition using clarity index (C_{50}) estimation have been presented. One approach investigated was to include C_{50} estimated from two different estimators (NIRA-CART and NIRA-BLSTM) as an additional feature in the ASR system and apply a dimensionality reduction technique (i.e. HLDA) to match the original feature vector dimension. This approach helped to improve the ASR performance of the best baseline by a relative word error rate reduction (WERR) of 7.35 % for NIRA-CART and NIRA-BLSTM. This improvement was shown to be in a significant part due to the HLDA transformation. Another approach was to use the C_{50} information to perform acoustic model selection, which in turn gave a relative WERR of 14.04 % with NIRA-CART and 17.07 % with NIRA-BLSTM. The best performance was achieved by combining both approaches and using NIRA-BLSTM, leading to a relative WERR of 22.41 % (7.77 % absolute WERR). It is worth noting that only data from the REVERB Challenge data sets was used to train all the models employed in the system (including the C_{50} estimator); furthermore, the method presented is complementary to other techniques such as CMLLR, and an example combination was shown to improve further the best performance, increasing the relative WERR to 29.8 %.

As expected, more accurate C_{50} estimations lead to a further reduction in the final WER. In the two algorithms exploited in this study, NIRA-BLSTM is more accurate than NIRA-CART by 1.6 dB RMSD, which results in a

relative WERR of 3.24 %. These results clearly indicate that C_{50} can be successfully used for reverberant speech recognition tasks and the accuracy in the C_{50} estimation is crucial.

DNN-based ASR can capture various characteristics of reverberant speech in different reverberant environments; therefore, future work will address the usefulness of the proposed method in such systems.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
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 and from the Research Foundation Flanders.

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

Received: 20 February 2015 Accepted: 5 June 2015
Published online: 01 July 2015

References

1. TH Falk, W-Y Chan, Temporal dynamics for blind measurement of room acoustical parameters. IEEE Trans. Instrum. Meas. **59**(4), 978–989 (2010)
2. PA Naylor, ND Gaubitch (eds.), *Speech Dereverberation* (Springer, London, 2010)
3. H Löllmann, E Yilmaz, M Jeub, P Vary, in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*. An improved algorithm for blind reverberation time estimation (Tel Aviv, Israel, 2010), pp. 1–4
4. J Eaton, ND Gaubitch, PA Naylor, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost (Vancouver, Canada, 2013), pp. 161–165
5. ND Gaubitch, HW Löllmann, M Jeub, TH Falk, PA Naylor, P Vary, M Brookes, in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*. Performance comparison of algorithms for blind reverberation time estimation from speech (Aachen, Germany, 2012), pp. 1–4
6. H Kuttruff, *Room Acoustics*, 5th edn. (Taylor & Francis, London, 2009)
7. P Peso Parada, D Sharma, PA Naylor, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Non-intrusive estimation of the level of reverberation in speech (Florence, Italy, 2014), pp. 4718–4722
8. A Tsilfidis, I Mporas, J Mourjopoulos, N Fakotakis, Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing. Comput. Speech Lang. **27**(1), 380–395 (2013)
9. T Yoshioka, A Sehr, M Delcroix, K Kinoshita, R Maas, T Nakatani, W Kellermann, Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition. IEEE Signal Process. Mag. **29**(6), 114–126 (2012)
10. R Haeb-Umbach, A Krueger, *Reverberant Speech Recognition*. (John Wiley & Sons, Chichester, 2012), pp. 251–281
11. M Wölfel, J McDonough, *Distant Speech Recognition*. (John Wiley & Sons, Chichester, 2009)
12. W Li, L Wang, F Zhou, Q Liao, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Joint sparse representation based cepstral-domain dereverberation for distant-talking speech recognition (Vancouver, Canada, 2013), pp. 7117–7120
13. T Yoshioka, T Nakatani, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Noise model transfer using affine transformation with application to large vocabulary reverberant speech recognition (Vancouver, Canada, 2013), pp. 7058–7062
14. Y Tachioka, S Watanabe, JR Hershey, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Effectiveness of discriminative training and feature transformation for reverberated and noisy speech (Vancouver, Canada, 2013), pp. 6935–6939
15. A Schwarz, C Huemmer, R Maas, W Kellermann, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Spatial diffuseness features for DNN-based speech recognition in noisy and reverberant environments (Brisbane, Australia, 2015)
16. T Yoshioka, MJF Gales, Environmentally robust ASR front-end for deep neural network acoustic models. Comput. Speech Lang. **31**(1), 65–86 (2015)
17. A Sehr, R Maas, W Kellermann, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Model-based dereverberation in the logmelspec domain for robust distant-talking speech recognition (Dallas, USA, 2010), pp. 4298–4301
18. L Couvreur, C Couvreur, Blind model selection for automatic speech recognition in reverberant environments. J. VLSI Signal Process. Syst. Signal Image Video Technol. **36**(2-3), 189–203 (2004)
19. J Liu, G-Z Yang, Robust speech recognition in reverberant environments by using an optimal synthetic room impulse response model. Speech Comm. **67**(0), 65–77 (2015)
20. A Sehr, R Maas, W Kellermann, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Frame-wise HMM adaptation using state-dependent reverberation estimates (Prague, Czech Republic, 2011), pp. 5484–5487
21. M Matassoni, A Brutti, P Svaizer, in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*. Acoustic modeling based on early-to-late reverberation ratio for robust ASR (Nice, France, 2014), pp. 263–267
22. L Wang, Y Kishi, A Kai, in *Chinese Conference on Pattern Recognition (CCPR)*. Distant speaker recognition based on the automatic selection of reverberant environments using GMMs, (2009), pp. 1–5
23. SM Ban, HS Kim, Instantaneous model adaptation method for reverberant speech recognition. Electron. Lett. **51**(6), 528–530 (2015)
24. AW Mohammed, M Matassoni, H Maganti, M Omologo, in *Proc. of the 20th European Signal Processing Conference (EUSIPCO)*. Acoustic model adaptation using piece-wise energy decay curve for reverberant environments (Bucharest, Romania, 2012), pp. 365–369
25. K Kondo, Y Takahashi, T Komatsu, T Nishino, K Takeda, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Computationally efficient single channel dereverberation based on complementary Wiener filter (Vancouver, Canada, 2013), pp. 7452–7456
26. T Ishii, H Komiyama, T Shinozaki, Y Horiuchi, S Kuroiwa, in *Proc. INTERSPEECH*. Reverberant speech recognition based on denoising autoencoder (Lyon, France, 2013), pp. 3512–3516
27. M Delcroix, T Yoshioka, A Ogawa, Y Kubo, M Fujimoto, N Ito, K Kinoshita, M Espi, T Hori, T Nakatani, *et al*, in *Proc. REVERB Workshop*. Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge (Florence, Italy, 2014)
28. F Weninger, S Watanabe, J Le Roux, JR Hershey, Y Tachioka, J Geiger, B Schuller, G Rigoll, in *Proc. REVERB Challenge*. The MERL/MELCO/TUM system for the REVERB Challenge using deep recurrent neural network feature enhancement, Florence, Italy, (2014)
29. EAP Habets, ND Gaubitch, PA Naylor, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Temporal selective dereverberation of noisy speech using one microphone (Las Vegas, USA, 2008), pp. 4577–4580
30. ML Seltzer, RM Stern, Subband likelihoodmaximizing beamforming for speech recognition in reverberant environments. IEEE Trans. Audio Speech Lang. Process. **14**, 2109–2121 (2006)
31. ND Gaubitch, PA Naylor, in *Proc. IEEE Intl. Conf. Digital Signal Processing (DSP)*. Spatiotemporal averaging method for enhancement of reverberant speech (Cardiff, UK, 2007), pp. 607–610
32. N Kumar, AG Andreou, Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. Speech Comm. **26**(4), 283–297 (1998)
33. K Kinoshita, M Delcroix, T Yoshioka, T Nakatani, E Habets, R Haeb-Umbach, V Leutnant, A Sehr, W Kellermann, R Maas, S Gannot, B Raj, in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. The REVERB challenge: a common evaluation framework for

dereverberation and recognition of reverberant speech (New York, USA, 2013), pp. 1–4
34. P Peso Parada, D Sharma, PA Naylor, T van Waterschoot, in *Proc. REVERB Workshop*. Single-channel reverberant speech recognition using C50 estimation (Florence, Italy, 2014)
35. L Breiman, J Friedman, CJ Stone, RA Olshen, *Classification and Regression Trees*. (CRC Press, Florida, USA, 1984)
36. F Weninger, J Bergmann, B Schuller, Introducing CURRENNT–the Munich open-source CUDA RecurREnt neural network toolkit. J. Mach. Learn. Res. **16**, 547–551 (2015)
37. LR Rabiner, RW Schafer, *Digital Processing of Speech Signals*. (Prentice Hall, Englewood Cliffs, USA, 1978)
38. T Robinson, J Fransen, D Pye, J Foote, S Renals, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition, vol. 1 (Detroit, USA, 1995), pp. 81–84
39. RO Duda, PE Hart, DG Stork, *Pattern Classification*, 2nd edn. (John Wiley and Sons, New York, 2001)
40. D Schmid, P Thuene, D Kolossa, G Enzner, in *Proc. of Speech Communication; 10. ITG Symposium*. Dereverberation preprocessing and training data adjustments for robust speech recognition in reverberant environments (Braunschweig, Germany, 2012), pp. 1–4

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Immediate publication on acceptance
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com