SINGLE-CHANNEL REVERBERANT SPEECH RECOGNITION USING C_{50} ESTIMATION

Pablo Peso Parada[⋆], *Dushyant Sharma*[⋆], *Patrick A. Naylor*[†], *Toon van Waterschoot*[‡]

[⋆]Nuance Communications Inc. Marlow, UK
[†]Dept. of Electrical and Electronic Engineering, Imperial College London, UK
[‡]Dept. of Electrical Engineering (ESAT-STADIUS), KU Leuven, Belgium
{pablo.peso, dushyant.sharma}@nuance.com,
p.naylor@imperial.ac.uk,toon.vanwaterschoot@esat.kuleuven.be

ABSTRACT

We present several single-channel approaches to robust speech recognition in reverberant environments based on single-channel estimation of C_{50}. Our best method includes this estimation in the feature vector as an additional parameter and also uses C_{50} to select the most suitable acoustic model according to the reverberation level. We evaluate our method on the REVERB challenge database and show that our method outperforms the best baseline of the challenge, reducing the word error rate by 5.7% (corresponding to 16.8% relative word error rate reduction).

Index Terms— 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, command automation, etc.). Of particular importance is distant speech recognition, where the user can interact with a device placed a short distance from the user. Such systems allow for more natural and comfortable interaction between the technology and the Human (e.g. hands free ASR systems in a car) which is crucial for increasing the acceptance of ASR among potential users.

In a distant-talking scenario, there is a significant degradation in ASR performance due to reverberation. The reverberant sound is created in enclosed spaces by reflections from surfaces which creates 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 can be modeled as the convolution between the RIR and the transmitted signal in the room.

RIRs can be divided in three different parts: direct path; early reflections (first 50 milliseconds after the direct path corresponding to spectral colouration); and late reverberation (reflections delayed more than 50 milliseconds causing temporal smearing of the signal [1]).

Several acoustic measures have been proposed to compute the reverberation level present in a signal by using the RIR or the reference and reverberant signal, but in many applications the only information available is the reverberant signal. Recently, some methods have been proposed to estimate room acoustic measures from reverberant signals such as the reverberation time (T_{60}) which characterizes the acoustic room properties. However, alternative measures have been shown to be more correlated with ASR performance such as C_{50} [2] which is the ratio of the energy in the early reflections over the energy in late reflections measured in dB. Such measures could be used to predict ASR performance or employed as a tuning parameter in de-reverberation algorithms.

ASR techniques robust to reverberation can be divided in two main groups [3][4]: front-end-based and back-end-based. The former approach suppresses the reverberation in the feature vector domain. Li et al. [5] propose to train a joint sparse transformation to estimate the clean feature vector from the reverberant feature vector. In [6] a model of the noise is estimated from observed data and considering the late reverberation as additive noise 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 [7]. The latter approach, back-end-based, modifies the acoustic models or the observation probability estimate to suppress the reverberation effect. Sehr et al. [8] 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 a estimation of T_{60} is proposed in [9]. The idea in [10] 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 reverbera-

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.

tion as different contributions. In addition to front-end-based and back-end-based approaches, signal-based methods are intended to de-reverberate the acoustic signal. In [11] a complementary Wiener filter is proposed to compute suitable spectral gains which are applied to the reverberant signal to suppress late reverberation. In [12] 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 [13].

Additionally, ASR techniques robust to reverberation can be also split according to the number of microphones used to capture the signal into single-channel [6] or multi-channel methods based on beamforming techniques [14].

The method proposed in this work is a hybrid approach based on front-end-based and back-end-based single-channel techniques. The idea is to estimate C_{50} [15] from the reverberant signal and use this estimation to select different acoustic models which were trained including C_{50} in the feature vector. The final feature vector size keeps the original dimensionality by applying HLDA [16]. The technique was tested within the ASR task of the REVERB challenge [17] which was launched by the IEEE to compare ASR performance on a common data set of reverberant speech.

The remainder of this paper is organized as follows: in Section 3 the challenge data is analysed. Section 4 describes the methods proposed and Section 5 discusses the performance of the these techniques. Finally, in Section 6 the conclusions are drawn.

2. C_{50} ESTIMATOR

This C_{50} estimator has recently been proposed in [15], therefore only an outline is provided here. This method 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 from the slope of the unwrapped Hilbert transformation. The latter group is 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 LPC coefficients to LSF representation.

For all frame-based features, excluding PLD spectral dynamics and the 12th order MFCCs, the rate of change is computed. 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 309 element vector. Finally, a CART regression tree [18] is built to estimate C_{50} using the complete feature vector.

3. ANALYSIS OF THE CHALLENGE DATA

The database provided in REVERB challenge comprises 3 different sets of 8-channel recordings: training, development set and evaluation set. This section analyses the RIRs of the training set and the reverberant recordings of development test in terms of C_{50} because this is a key aspect in the design of the algorithms proposed in this work. Evaluation test set is not analysed because this set must be only used to assess the algorithms.

Figure 1 shows the histogram of the 24 training RIRs according to C_{50} including all channels of each response. This acoustic parameter is computed as follows,

$$\mathrm{C}_{50} = 10\log_{10}\left(\frac{\sum_{n=0}^{N_{50}} h^2(n)}{\sum_{n=N_{50}+1}^{\infty} h^2(n)}\right)\mathrm{dB}, \qquad (1)$$

where h is the RIR and N_{50} is an integer number of samples corresponding to 50 milliseconds after the time arrival of the direct path.

The training RIRs cover a wide range of C_{50}, approximately 25dB. These RIRs are used to create the data set employed to train our C_{50} estimator [15] by convolving these RIRs with the clean training set (i.e. WSJCAM0 training set [19]).

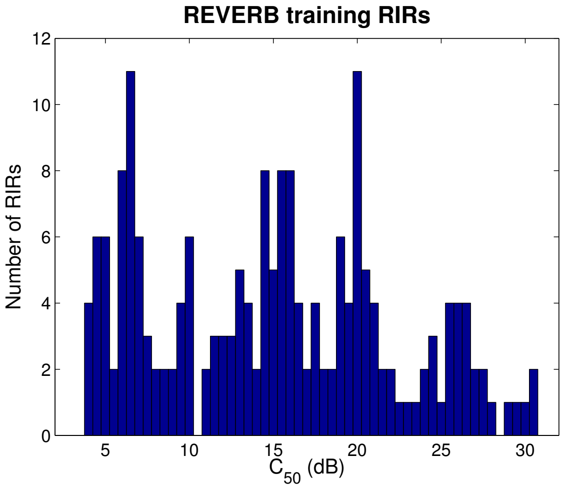

Fig. 1. Ground truth C_{50} value of the training RIRs.

Figure 2 displays the histogram for each reverberant condition (clean, near and far) according to the C_{50} estimated with our model. The first histogram represents the distribution of clean recordings according to the C_{50} estimated. This distribution is located at high C_{50} values indicating very low levels of reverberation. These signals are recorded in a five by five meters room with approximately the same recording configuration [19] for all speaker however some specific speakers have a lower estimated C_{50} (centered at approximately 19 dB). The second plot displays the histogram of those recordings with speaker placed near (50 cm) to microphone array. It shows a significant difference between the small room recordings (Room1) which are less reverberant, and the medium and large room recordings (Room2 and Room3 respectively) which have a higher reverberation level. At the bottom of Figure 2 is represented the distribution of speech signals with the speaker far (200 cm) from the microphone. In this case, the estimated C_{50} for all recordings have been dramatically decreased. All these C_{50} estimations are in accordance with the baseline results for ASR task (Table 3 in [17]): recordings with low C_{50} result in high word error rate while signals with high C_{50} perform considerably better.

Figure 3 shows the distribution of the real recordings captured in a reverberant meeting room for two different distances: near (∼=100 cm) and far (∼=250 cm). It shows that both configurations are similar in terms of C_{50} which agrees with the ASR performance (both have a similar word error rate).

The performance of the C_{50} estimator can not be tested in this development test because the RIRs of this set were unknown.

4. METHODS

In this section we describe different configurations for reverberant speech recognition. The idea underneath these methods is to exploit the C_{50} estimation to build an ASR robust to reverberation.

4.1. C_{50} as a new feature

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

The first configuration proposed (C_{50}FV) is to add C_{50} estimation directly to this feature vector. Therefore the modified feature vector comprises 40 elements.

The second configuration (C_{50}PCA) aims to decrease the dimensionality of the previous 40 element feature vector by employing principal component analysis decomposition (PCA). This technique is based on finding the eigenvectors of the scatter matrix $\mathbf{S}$

$$\mathbf{S} = \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t, \qquad (2)$$

where $\mathbf{x}_k$ represents the feature vector of the frame k, n the total number of frames and $\mathbf{m}$ is the sample mean. The data is projected onto the eigenvector space and only the N eigenvectors with the highest eigenvalues are kept to build the new feature space. In this case N is set to 39. This transformation reduces the dimensionality by keeping the dimensions with the highest variance (high eigenvalues), so PCA may not improve the discrimination between classes.

A third configuration (C_{50}HLDA) is tested based on reducing the feature vector dimension using linear discriminant analysis. This method projects the data in a new space by applying a linear transformation. Unlike PCA, this transformation aims to retain the class-discrimination in the transformed feature space. The linear function applied to data is computed by maximizing the ratio of between-class scatter to within-class scatter matrix. In this work a model-based generalization of linear discriminant analysis [16] is used. In this case the linear transformation is estimated from Gaussian models using expectation-maximization algorithm.

In all these configurations, the acoustic models are retrained since the feature extraction module is modified.

4.2. Model selection

This back-end approach is based on selecting the optimal acoustic model according to the level of reverberation present. In this work we use C_{50} to measure the amount of reverberation in the signal instead of T_{60} as in [9] because this last parameter measures the room acoustic properties. Moreover C_{50} was shown to be highly correlated with the ASR performance [15][2] which makes it suitable for this purpose.

The first configuration (Clean&Multi cond.) is based on selecting between the two acoustic models provided in the challenge (clean-condition HMMs and multi-condition HMMs) according to the level of C_{50} estimated from the signal. After performing some experiments and looking at the analysis carried out in section 3, we set the threshold to determine which acoustic model is used in the decoder to C_{50}=24.9 dB. This threshold provides the best separation between clean and reverberant signals in the development test set. Recordings with estimated C_{50} higher than 24.9 dB are recognized by applying clean-condition HMMs whereas recordings with C_{50} lower than this threshold are decoded employing multi-condition HMMs.

Following configurations are 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 (Figure 1). It is worth noting at this point that all utterances must be convolved with the subset of training RIRs to create each of the reverberant models, otherwise representative data of the acoustic units may be not included in the training. The first approach is to create three reverberant models (MS3) according to the C_{50} values of the RIRs. Using Figure 2 and Figure 3 the two thresholds are set to C_{50}=10 dB and C_{50}=20 dB. The aim is to cluster the development test set in three groups with similar ASR performance and train a

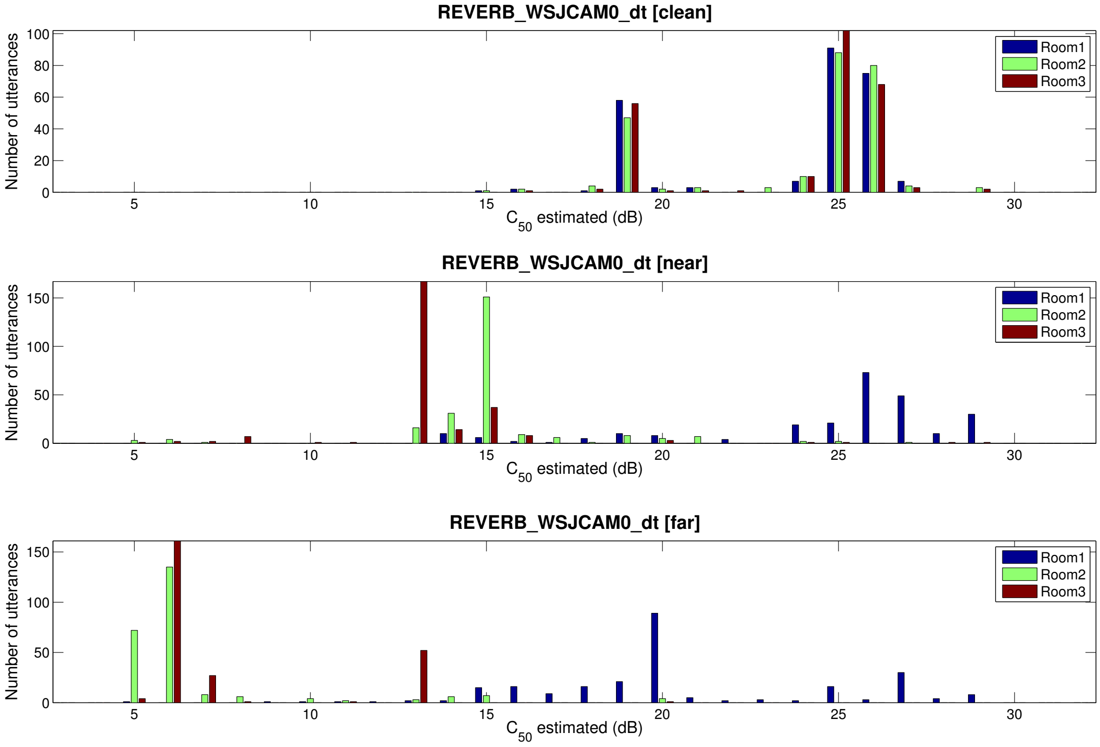

Fig. 2. Estimated C_{50} distribution of the simulated data subset of development test set. First plot represents the C_{50} distribution for clean data; second chart shows the C_{50} distribution for near distance recordings; and the third graph is the C_{50} distribution for far distance recordings. Blue bars represent the small room; green bars represent medium room; and red bars represent large room.

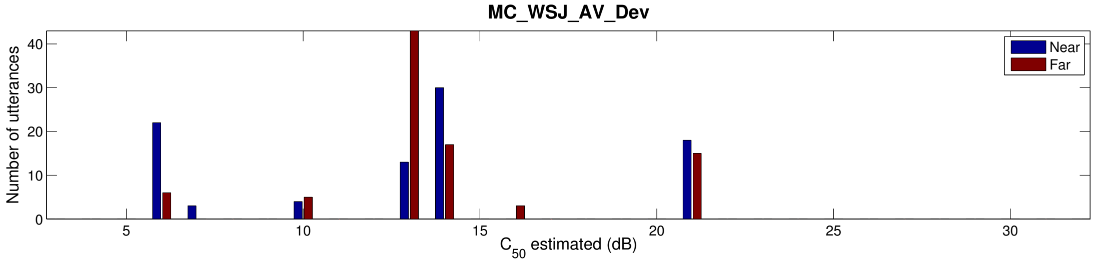

Fig. 3. Estimated C_{50} values of the real data subset of development test set. Blue bars represent near distances between speaker and microphone; and red bars represent far distances.

model for each group. The most reverberant model is trained with the RIRs that have C_{50} lower than 10 dB. The second acoustic model is trained with RIRs that have C_{50} between 10 dB and 20 dB. Finally the third model, which represents the least reverberant conditions, is trained with those RIRs with a C_{50} higher than 20 dB. These acoustic models are selected in the recognition stage by applying exactly the same training thresholds. The first chart in Figure 4 represents this configuration.

Next configuration (MS5) includes a new idea in the training: overlap training data to build models. In all cases the overlapping used was approximately 50% of the size of the neighbouring models. This configuration keeps the same previous models (MS3) and adds two additional models in the transitions. These two models are trained with data already included in the original models and located in the transition area between two neighbour acoustic models in terms of C_{50} which provides a smoother transition between acoustic models. The most representative model to the reverberation level estimated from the utterance is selected in the recognition phase. The bottom plot of Fig. 4 represents this idea. This chart shows that HMM number 1, 3 and 5 are still trained as HMM number 1, 2 and 3 of MS3. The difference is in the thresholds used to select these models in the recognition

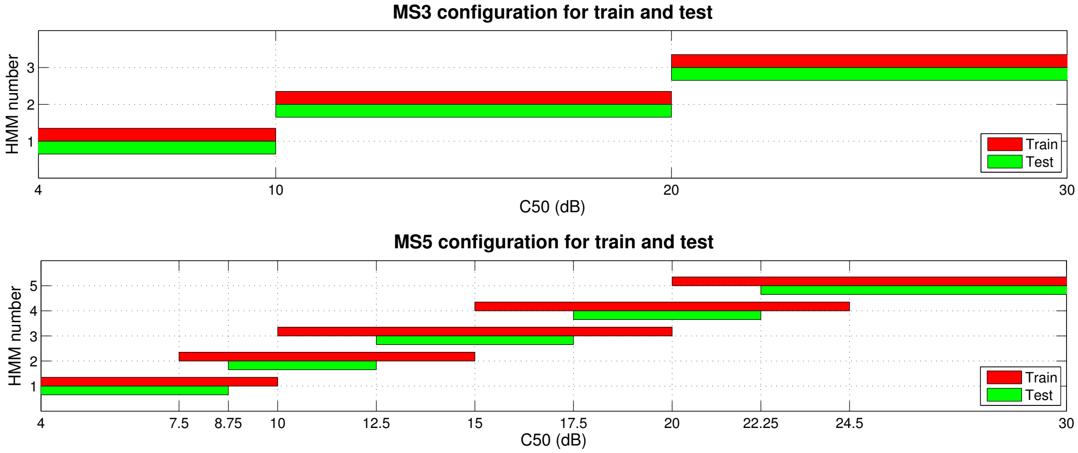

Fig. 4. Comparison of MS3 and MS5 configurations for training the acoustic (red bars) models and recognizing testing data (green bars) according to C_{50}. The difference is in the overlapping of the training data for MS5 configuration.

stage (green bars) and the incorporation of overlapped models (HMM number 2 and 4).

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 width of the training data of each model is decreased in terms of C_{50} which creates acoustic models more specific for each reverberant environment. Figure 5 shows the settings used for MS11.

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

This method combines two different approaches described before: C_{50}HLDA and model selection. Figure 6 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 which is then included in the feature vector before applying the HLDA transformation and also used to select the most suitable acoustic model.

Three different numbers of acoustic models are tested: 3 (MS3+ C_{50}HLDA), 5 (MS5+C_{50}HLDA) and 11 (MS11+ C_{50}HLDA) following the configuration presented in Figure 4 and Figure 5 respectively.

5. RESULTS & DISCUSSION

In this section we present the results of the methods described in the previous section and we compare the performance of each in terms of word error rate (WER). Table 1 presents the average of WER achieved with the non-reverberant recordings (Clean), simulated reverberant recordings (Sim.) and real reverberant recordings (Real), whereas Table 2 shows with more detail these results for each subset of the evaluation test set including the average of all subsets in the last column. Moreover, Figure 7 summarizes these results displaying the average WER for development test set and evaluation test set.

	Clean	**Sim.**	**Real**
	Avg.	*Avg.*	*Avg.*
Clean-cond.	10.94	51.86	88.51
Multi-cond.	30.16	29.52	56.95
Clean&Multi cond.	**18.26**	29.22	56.95
C_{50}HLDA	26.41	28.02	56.12
MS3	28.00	27.93	59.59
MS3+C_{50}HLDA	24.41	25.70	57.00
MS5	23.22	26.81	57.88
MS5+C_{50}HLDA	20.93	25.22	55.97
MS8	23.14	26.17	56.40
MS11	22.07	26.40	56.80
MS11+C_{50}HLDA	20.55	**24.52**	**54.21**
MS14	22.85	26.31	57.48
MS18	23.95	26.51	58.06

Table 1. WER (%) averages obtained in evaluation dataset. First two rows correspond to the baseline methods and the remainder are the methods proposed in this work.

The baseline methods considered to compare the performance consist of decoding the data using the two acoustic models provided in the REVERB challenge: the acoustic model trained with clean data (Clean-cond.) and the acoustic model trained with reverberant data (Multi-cond.). The performance of these baselines are shown in the first two rows of Table 1 and Table 2. Clean-cond. models provide a better performance in non-reverberant environments whereas using Multi-cond. models a significant decrease of WER is achieved for reverberant environments.

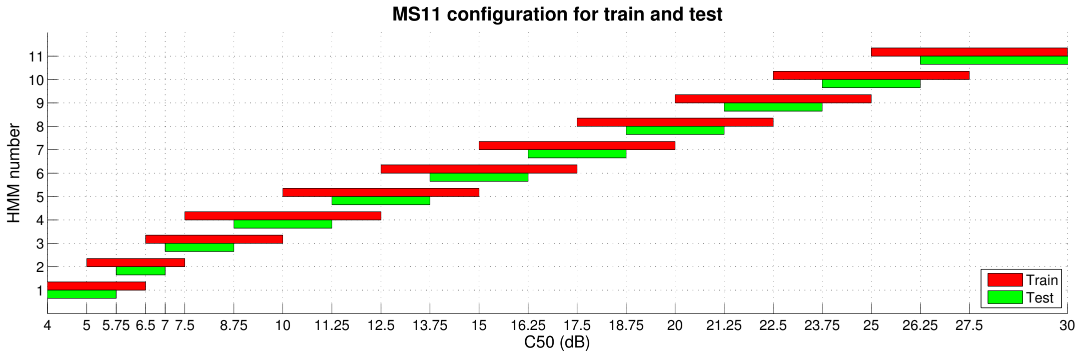

Fig. 5. MS11 configurations for training the acoustic models (red bars) overlapping of the training data and recognizing testing data (green bars) according to C_{50}.

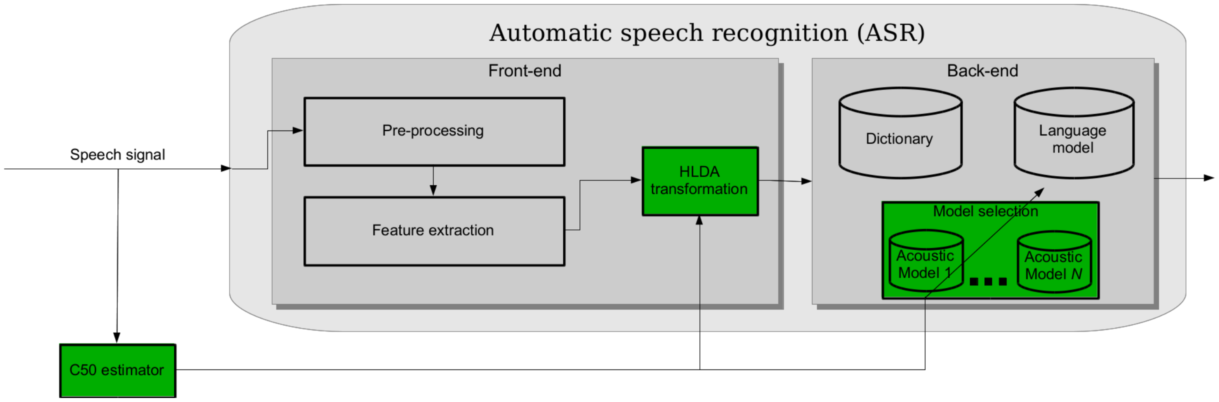

Fig. 6. Diagram of the reverberant speech recognition highlighting in green the proposed modifications.

	Clean			**Sim.**						**Real**		
	Room1	*Room2*	*Room3*	*Room1*		*Room2*		*Room3*		*Room1*		
				near	*far*	*near*	*far*	*near*	*far*	*near*	*far*	Avg.
Clean-cond.	10.50	11.51	10.81	15.29	25.29	43.90	85.80	51.95	88.90	88.71	88.31	47.36
Multi-cond.	30.29	30.07	30.11	20.60	21.15	23.70	38.72	28.08	44.86	58.45	55.44	34.67
Clean&Multi cond.	**17.67**	**18.25**	**18.87**	18.69	21.11	23.78	38.72	28.14	44.86	58.45	55.44	31.27
C_{50}HLDA	26.33	26.82	26.09	18.57	19.48	21.21	37.74	27.85	43.29	57.84	54.39	32.69
MS3	28.11	27.22	28.66	17.76	21.09	22.19	36.39	29.07	41.07	61.45	57.73	33.70
MS3+C_{50}HLDA	24.40	24.12	24.72	16.50	19.45	20.45	33.51	26.89	37.38	58.67	55.33	31.03
MS5	22.77	22.96	23.94	16.44	19.01	20.78	36.95	26.97	40.73	59.57	56.18	31.48
MS5+C_{50}HLDA	20.72	20.66	21.41	16.59	17.30	19.92	33.56	25.39	38.56	57.30	54.63	29.64
MS8	22.77	22.19	24.45	16.35	18.49	20.98	34.62	26.87	39.70	57.59	55.20	30.83
MS11	22.48	21.40	22.33	16.64	18.42	20.97	35.99	26.58	39.82	58.80	54.79	30.74
MS11+C_{50}HLDA	20.69	20.73	20.22	**15.54**	**17.10**	**19.63**	**33.00**	**25.39**	**36.43**	**55.57**	**52.84**	**28.83**
MS14	23.09	22.48	22.98	17.35	18.35	21.14	35.39	25.76	39.87	58.70	56.25	31.03
MS18	23.38	23.83	24.64	16.93	18.30	21.37	35.63	26.86	39.96	59.47	56.65	31.54

Table 2. WER (%) obtained in evaluation dataset. First two rows correspond to the baseline methods and the remainder are the methods proposed in this work.

The method C_{50}FV provides a similar performance compared with the baselines. This outcome is due to the fact that we are using diagonal covariance matrix to build the acoustic model. Therefore this feature only provides information

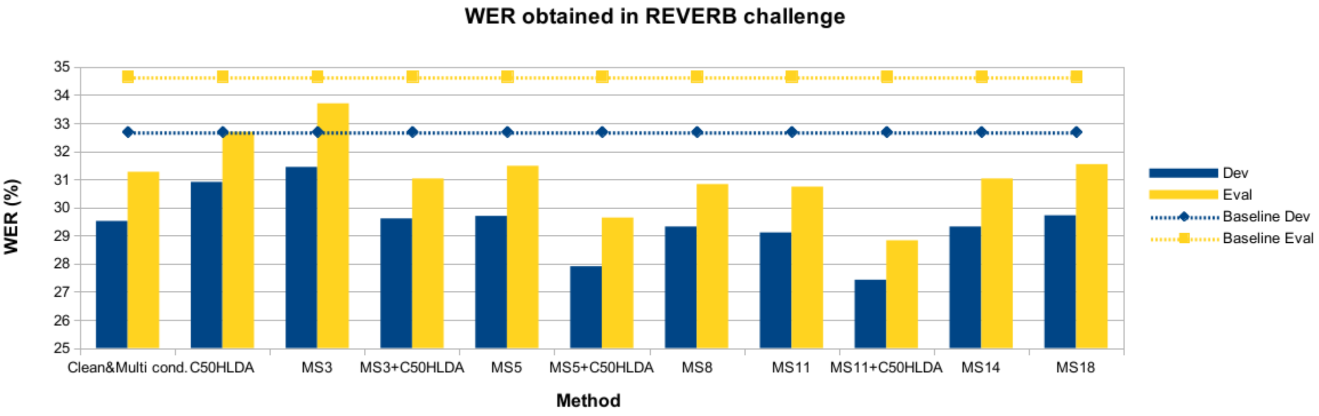

Fig. 7. Comparison of the ASR performance of several methods (bars) against the baselines (dotted lines) for development test set (blue) and evaluation test set (yellow).

regarding the probability of the acoustic unit to be seen in this reverberant environment not taking into account possible dependences with the MFCC. C_{50}PCA adds C_{50} estimate in the feature vector but the performance achieved is significantly lower due to the computation of the transformation matrix followed by PCA. These results are excluded in Table 1 and Table 2 because of the poor performance. On the other hand, the last method described in section 4.1 (C_{50}HLDA) 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.

Table 1 and Table 2 also display the performance obtained with the methods described in section 4.2 based on model selection. It shows that using C_{50} to select between the acoustic models provided by REVERB challenge (i.e., Clean&Multi cond.) a lower WER than using only one of them is achieved. Further improvement can be achieved by training more reverberant models. MS3 configuration employs three reverberant models (upper plot in Figure 4) and the performance in reverberant conditions has been improved in most of the situations but on average the error rate has been increased with respect to Clean&Multi cond. mainly due to the poor performance in clean environments. The performance of this configuration is improved with more than 2% of WER by only overlapping the training data to build the acoustic models (MS5). Increasing the number of models trained using the overlapping of the reverberant data technique (i.e., MS8, MS11, MS14 and MS18) results in a further reduction of WER. These results show that the best performance is obtained with MS11, while after this point an increase in the number of models produces an increase in WER. This could be due to an insufficient accuracy of the C_{50} estimator.

Finally, the system presented in Figure 6 is tested by training 3 reverberant models (MS3+C_{50}HLDA), 5 (MS5+ C_{50}HLDA) and 11 (MS11+C_{50}HLDA). The last two configurations are trained using the overlapping of the training data. A significant improvement is obtained by combining both methods; the WER decreases by 2% with respect to the error achieved using only model selection. As is clearly shown in Figure 7, the best performance is obtained with MS11+C_{50}HLDA which approximately outperforms the best baseline method (Multi-cond.) by 6% in both test sets.

Table 1 and Table 2 highlight in bold the lowest WER obtained in each data set. MS11+C_{50}HLDA presents the best performance in reverberant conditions but Clean&Multi cond. shows the best performance in clean condition. This is mainly because all the data used to train MS11+C_{50}HLDA is reverberant data while Clean&Multi cond uses reverberant and clean data to train the acoustic models. Therefore MS11+C_{50}HLDA could be further improved including a clean acoustic model to recognize non reverberant data.

6. CONCLUSIONS

In this paper we have shown various approaches for single-channel reverberant speech recognition using the C_{50} measure. One approach investigated was to include the C_{50} as an additional feature in the ASR system. This approach helped to improve the ASR performance of the best baseline by a relative word error rate reduction (WERR) of 5.71%. Another approach was to use the C_{50} information to perform acoustic model selection, which in turn gave a WERR of 11.33%. The best performance was achieved by combining both approaches, leading to a WERR of 16.84% (6% absolute). These results clearly indicate that C_{50} can be successfully used for reverberant speech recognition tasks.

It was also shown that overlapping the training data in the creation of reverberant acoustic models (according to the C_{50} value) can significantly improve ASR performance.

7. REFERENCES

[1] 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.

[2] 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.

[3] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[4] R. Haeb-Umbach and A. Krueger, *Reverberant Speech Recognition*, pp. 251–281, John Wiley & Sons, 2012.

[5] W. Li, L. Wang, F. Zhou, and Q. Liao, "Joint sparse representation based cepstral-domain dereverberation for distant-talking speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7117–7120.

[6] T. Yoshioka and T. Nakatani, "Noise model transfer using affine transformation with application to large vocabulary reverberant speech recognition," in *Proc. Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7058–7062.

[7] Y. Tachioka, S. Watanabe, and J.R. Hershey, "Effectiveness of discriminative training and feature transformation for reverberated and noisy speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6935–6939.

[8] A. Sehr, R. Maas, and W. Kellermann, "Model-based dereverberation in the logmelspec domain for robust distant-talking speech recognition," in *Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 4298–4301.

[9] L. Couvreur and C. Couvreur, "Blind model selection for automatic speech recognition in reverberant environments," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 36, no. 2-3, pp. 189–203, 2004.

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

[11] K. Kondo, Y. Takahashi, T. Komatsu, T. Nishino, and K. Takeda, "Computationally efficient single channel dereverberation based on complementary wiener filter," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7452–7456.

[12] T. Ishii, H. Komiyama, T. Shinozaki, Y. Horiuchi, and S. Kuroiwa, "Reverberant speech recognition based on denoising autoencoder," in *Proc. INTERSPEECH*, 2013, pp. 3512–3516.

[13] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S.-J. Hahm, and A. Nakamura, "Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds," *Computer Speech & Language*, vol. 27, no. 3, pp. 851–873, 2013.

[14] Michael L. Seltzer and Richard M. Stern, "Subband likelihoodmaximizing beamforming for speech recognition in reverberant environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 2109–2121, 2006.

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

[16] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 26, no. 4, pp. 283 – 297, 1998.

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

[18] L. Olshen, Breiman J. H., Friedman R. A., and Charles J. Stone, "Classification and regression trees," *CRC Press*, 1984.

[19] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAMO: a british english speech corpus for large vocabulary continuous speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and SignalProcessing (ICASSP)*, 1995, vol. 1, pp. 81–84.