REVERBERATION SUPPRESSION BASED ON SPARSE LINEAR PREDICTION IN NOISY ENVIRONMENTS

Nicolás López[1,2], *Gaël Richard*[2], *Yves Grenier*[2], *Ivan Bourmeyster*[1]

[1] Arkamys - 31 rue Pouchet, 75017 Paris, France
[2] Institut Mines-Télécom; Télécom ParisTech; CNRS LTCI - 37/39 rue Dareau, 75014 Paris, France

ABSTRACT

We present a single channel method for late reverberation suppression. The proposed approach estimates late reverberation as a linear combination of previous time-frequency frames. We impose a sparsity constraint on the predictor in order to select the most relevant signal frames for the estimation. The dataset used for the evaluation is corrupted by background noise, thus we propose to jointly suppress background noise and late reverberation. This leads to an important improvement in the quality of the processed signals as well as an improvement of the automatic speech recognition scores. The method appears to be efficient mainly in far field conditions and in highly reverberant environments. In addition, it is suitable for real time processing.

Index Terms— Single Channel Speech Enhancement, Late Reverberation Estimation, Lasso, Sparse Linear Prediction, Noise Reduction

1. INTRODUCTION

In this paper we describe a method for single channel late reverberation suppression in noisy conditions. The proposed approach uses a speech enhancement framework to estimate the power spectral density (*psd*) of late reverberation in the time-frequency domain. This framework was originally proposed for dereverberation in [1] and extended later in [2].

The estimation is based on the principle of linear prediction: late reverberation is modeled as a linear combination of previously observed signal frames. In addition we assume that only a small set of those frames is relevant for the prediction, i.e., we introduce a sparse prior for the linear predictor. Under these assumptions, we can model the estimation of late reverberation as a regression problem that can be solved using the Lasso [3]. The estimated late reverberation is then used to design a time-frequency filter based on the Log Spectral Amplitude estimator [4]. The resulting estimator is studied in blind conditions: its settings are kept unchanged for every different acoustic environment. Thus, we do not need to estimate the acoustics parameters that are often required for dereverberation approaches, namely the reverberation time and the Direct to Reverberant Ratio (DRR).

In order to improve noise robustness, we couple our method with a background noise estimator to achieve joint noise and reverberation suppression. This joint approach was evaluated for both *Speech Enhancement* (SE) and *Automatic Speech Recognition* (*ASR*) tasks and the results were submitted to the REVERB Challenge[5]. We show that the method is able to reduce late reverberation in every condition. However, it is more efficient in far field conditions and in highly reverberant environments where the distortion levels are kept at very reasonable levels. The overall method is compatible with real-time processing.

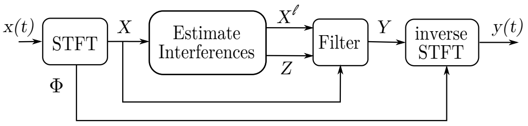

Fig. 1. Speech enhancement framework for single channel reverberation suppression.

2. PROPOSED METHOD

The proposed method is based on the speech enhancement framework depicted in Figure 1. We denote by X the magnitude of the Short Time Fourier Transform (STFT) of the noisy and reverberant signal $x(t)$. Late reverberation X^ℓ is estimated in the time-frequency domain as well as the background noise Z. Both interferences are used to design an enhancement filter. We use the noisy and reverberant phase Φ together with the enhanced spectral magnitude Y to produce the time domain output $y(t)$.

2.1. Late reverberation estimation

In the frequency domain, the reverberated signal at frequency k and time n is usually written as:

$$X_{k,n} = X^e_{k,n} + X^\ell_{k,n} \ , \tag{1}$$

where $X^e_{k,n}$ and $X^\ell_{k,n}$ represent respectively the early and late reverberation terms [2]. According to this model, late reverberation appears like an additive interference that can be suppressed by spectral subtraction techniques. As reverberation is produced by delayed and damped replicas of the direct sound, we propose to predict $X^\ell_{k,n}$ in each frequency channel as a linear combination of L signal frames preceding the current frame:

$$\hat{X}^\ell_{k,n} = \sum_{i=0}^{L-1} \alpha_i X_{k,n-i-\delta}. \tag{2}$$

Here we introduce a delay of δ frames intended to reduce the influence of the direct path on the prediction. As speech is known to be sparse in the time-frequency domain, we impose a sparsity constraint on the predictor $\boldsymbol{\alpha} = [\alpha_0 \ \ldots \ \alpha_{L-1}]^T$. Hence, we assume that only a small number of frames contribute significantly to the prediction. Under these assumptions, we state the problem of estimating late reverberation as an instance of the Lasso [3]:

$$\underset{\alpha}{\text{minimize}} \; ||X_{k,n} - \sum_{i=0}^{L-1} \alpha_i X_{k,n-i-\delta}.||^2 \quad \text{s.t.} \quad |\boldsymbol{\alpha}| \leq \lambda. \tag{3}$$

The proposed estimator predicts late reverberation as a redundancy term while early reflections are viewed as the residual of the prediction. The predictors are obtained by solving the Lasso with the Least Angle Regression (LARS) algorithm [6]. To this aim, we use the efficient *mexLasso* function from the SPAMS toolbox[1]. According to the model (3), the solver projects the current observation on the vector $V_{k,n} = [X_{k,n-\delta}, \ldots, X_{k,n-\delta-L+1}] \in \mathbb{R}^L$. The LARS algorithm selects at most as many active predictors as the smallest dimension of $V_{k,n}$. Thus, only one active prediction coefficient among L is selected in the configuration used in this paper, resulting in a sparse predictor. The main difference with the methods in [1, 2] is that the delay between the current observation and the active predictor is now allowed to change in the range $[\delta, \delta + L]$. The hyperparameter λ acts as an upper bound for the magnitude of the predictor and allows to control the maximal amount of energy associated to late reverberation. For the remaining of this paper, we will work in blind conditions, i.e., λ will be kept constant independently of the acoustic environment. Its value must be high enough to cover a wide range of reverberation times, so that the system adapts the magnitude of the predictor to the acoustic conditions under consideration.

We further reduce the complexity by subsampling the spectrogram along the frequency axis. For this, we define a set of 10 non overlapping octaves and we average the bins in each octave, to produce a 10-channel spectrogram. Finally, all the frequency bins within the same octave are associated to the same predictor to synthesize the late reverberation spectrogram as stated in (2).

2.2. Background noise estimation

When background noise is present, the performance of the estimator introduced above is degraded. To cope with this, we propose to independently estimate background noise and reverberation in order to build a joint enhancement filter.

Let Z be magnitude spectrum of the background noise. The corresponding noise *psd* is estimated as:

$$Z^2_{k,n} = \beta_Z \, Z^2_{k,n-1} + (1 - \beta_Z) X^2_{k,n}, \tag{4}$$

where β_Z is the smoothing constant for the recursive average related to the noise spectrum estimation. In order to improve the accuracy of this estimate, we use a simple Voice Activity Detector (VAD) based on hard thresholding. The noise *psd* is only updated when speech is absent. For all the experiments, we use $\beta_Z = 0.97$ which corresponds to a time constant of 371 ms for a sampling rate of 16 kHz.

2.3. Time-Frequency filtering

The time-frequency filter $G_{k,n}$ is based on the Ephraim and Malah's Log Spectral Amplitude (LSA) estimator [4] and defined as:

$$G_{k,n} = \frac{\xi_{k,n}}{1 + \xi_{k,n}} \exp \left\{ \frac{1}{2} \int_{\nu_{k,n}}^{\infty} \frac{e^{-t}}{t} \mathrm{d}t \right\}, \tag{5}$$

where $\xi_{k,n}$ and $\gamma_{k,n}$ denote respectively the *a priori* and the *a posteriori* Signal to Noise Ratios and $\nu_{k,n} = \frac{\xi_{k,n}}{1+\xi_{k,n}} \gamma_{k,n}$. This filter was originally proposed to suppress only background noise. In [7] the definitions of both $\gamma_{k,n}$ and $\xi_{k,n}$ are modified in order to handle the case of multiple interferences. The well-known *direction directed* approach is used to recursively estimate the *a priori* SNR related to late reverberation and background noise.

We will use this approach for the derivation of the enhancement filter. The smoothing constant for the estimation of the *a priori* SNR is set to $\beta_{snr} = 0.998$. We additionally define a lower bound G_{min} to the enhancement gain $G_{k,n}$ in order to avoid annoying musical noise.

3. REVERB CHALLENGE EVALUATION

In this section we evaluate the proposed single channel algorithm with the evaluation tools from the REVERB Challenge on both the *SimData* and the *RealData* datasets. We will focus on the *SE* task but we will also give the performance of the proposed system for the *ASR* task.

3.1. Settings

In the following *SE* experiments we want to evaluate the performance of the method for joint dereverberation and noise reduction. We will also evaluate each stage individually to give a better insight on the properties of our system. All the *SE* results are presented as follows:

- *Baseline*: unprocessed reverberant signals,
- *DRVNR*: joint dereverberation and noise reduction,
- *DRV*: only dereverberation,
- *NR*: only noise reduction.

In all cases, we use exactly the same settings for the estimation of the interferences and the design of the filter. The main parameters are summarized in Table 1. For the STFT filterbank, we use a 32 ms long Hamming window with 75% overlap.

Late reverberation estimation:		
$L = 10$	$\delta = 5$	$\lambda = 0.9$
Background noise estimation:		
$\beta_Z = 0.97$		
Filter design:		
$\beta_{snr} = 0.998$	G_{min} = -18 dB	

Table 1. Values of the parameters for the evaluation

3.2. Evaluation of the Speech Enhancement task

The *SE* task is twofold, it allows to assess the performance of a given method and also to study the ability of each objective measure to rate the quality of the resulting signals. We will analyze and comment the evaluation results obtained with each of the metrics.

3.2.1. Signal to Reverberant Modulation Ratio (SRMR)

The SRMR is non-intrusive measure based on the modulation envelope that was specially designed for the evaluation of reverberant signals. Higher values indicate reduced reverberation. The results are summarized in Table 2. Let us consider first the results obtained with the *SimData* dataset. We observe that the proposed *DRVNR* approach yields the higher average SRMR. For every considered room

[1] `http://spams-devel.gforge.inria.fr/`

SRMR	SimData							RealData		
	Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
	Near	Far	Near	Far	Near	Far		Near	Far	
Baseline	4.50	4.58	3.74	2.97	3.57	2.73	3.68	3.17	3.19	3.18
DRVNR	6.96	8.19	6.59	7.21	6.23	6.28	6.91	7.40	7.68	7.54
DRV	5.91	6.39	5.83	5.94	5.70	5.77	5.92	9.05	8.83	8.94
NR	4.70	5.09	4.28	4.01	4.28	3.76	4.35	4.62	4.76	4.69

Table 2. SRMR for SimData and RealData

there is an increase of the SRMR that confirms the ability of our method to blindly reduce reverberation under different acoustic conditions. Both the *DRV* and *NR* approaches manage to reduce reverberation but to a lower extent. It is not surprising that the *NR* approach increases the SRMR. Indeed, the modulation envelope of the signal is modified by the *NR* processing and leads to an increase of the SRMR. This shows that the SRMR is not able to evaluate only the effects of reverberation but it can also be influenced by other modulation artifacts introduced during the processing. Still, the largest increases on the SRMR appear when we include dereverberation in the processing which indicates that the measure is mostly sensitive to reverberation.
Regarding the *RealData* dataset, the *DRV* approach gives better results than *DRVNR*. However, as we will see below, it also introduces important distortions that are avoided with the *DRVNR* approach.

3.2.2. *Cepstral distance (CD) and Log Likelihood Ratio (LLR)*

Now we study two different distortion measures: the cepstral distance and the log likelihood ratio. For both measures lower values mean less distortion. The results are reported in Tables 3 and 4. We observe that when we individually apply *NR* or *DRV* the distortion levels are increased with respect to the *baseline*. This is due to the musical noise that appears when either noise or reverberation levels are high.

The joint *DRVNR* method gives good performance compared to the unprocessed signals. We observe lower distortion levels for rooms 2 and 3 while they are increased in room 1, the less reverberant enclosure. This shows that the settings we chose for the proposed method are best suited for big rooms. In the smallest room, late reverberation is over estimated and leads to excessive spectral subtraction. This introduces distortion artifacts that affect the quality of the processed audio. This effect can be limited by setting a lower value for the hyperparameter λ.

To conclude this part, we suggest that CD and LLR are equally informative and thus we can use just one of them to evaluate the distortion.

CD	Room 1		Room 2		Room 3		Ave.
	Near	Far	Near	Far	Near	Far	
Baseline	1.99	2.67	4.63	5.21	4.38	4.96	3.97
DRVNR	2.67	3.03	4.32	4.87	4.14	4.63	3.94
DRV	3.88	4.21	4.65	5.22	4.61	5.07	4.61
NR	4.45	4.82	4.41	5.35	4.86	5.64	4.92

Table 3. SimData: Cepstral distance [dB]

3.2.3. *Frequency Weighted Signal to Noise Ratio (FWSNR)*

The results of the evaluation with the frequency weighted signal to noise ratio are presented in Table 5. In average, the three approaches under evaluation improve the FWSNR. A more detailed view shows that we actually improved the FWSNR in Rooms 2 and 3 but we failed in Room 1. In addition, the larger improvements were obtained in far field conditions. This is coherent with the previous observations. The proposed method has a satisfactory performance in large rooms. In addition, it is more efficient if we place ourselves in far field conditions but it is still able to enhance the signals in near field conditions as long as the reverberation time of the room is high.

LLR	Room 1		Room 2		Room 3		Ave.
	Near	Far	Near	Far	Near	Far	
Baseline	0.35	0.38	0.49	0.75	0.65	0.84	0.58
DRVNR	0.42	0.45	0.51	0.72	0.67	0.81	0.60
DRV	0.79	0.83	0.81	1.02	0.94	1.06	0.91
NR	0.78	0.86	1.01	1.23	1.09	1.28	1.04

Table 4. SimData: Log Likelihood Ratio

Regarding the measure itself, it behaves similarly to the SRMR in Rooms 2 and 3. However in Room 1 we did not observe the improvement that was assessed by the SRMR. We suggest that the FWSNR measures a trade-off between the reduction of reverberation and the distortions introduced.

FWSNR	Room 1		Room 2		Room 3		Ave.
	Near	Far	Near	Far	Near	Far	
Baseline	8.12	6.68	3.35	1.04	2.27	0.24	3.62
DRVNR	6.47	6.29	4.05	2.91	3.51	2.42	4.27
DRV	4.95	4.63	5.16	3.90	4.62	3.54	4.47
NR	6.18	5.50	5.69	1.06	3.56	0.93	3.82

Table 5. SimData: Frequency Weighted segmental Signal to Noise Ratio [dB]

3.2.4. *Wall clock time*

We ran a Matlab implementation of the proposed method and we observed a real time factor of 9.41 % for the *SimData* dataset and of 9.29% for *RealData* dataset. For this experiment we used a computer running under Windows 7 Professional (SP 1, 64 bits). The CPU was Intel Core i7 (1core) M640 CPU at 2,80GHz and 4,0 Gb of RAM memory were available. The reference beamforming system made available for the REVERB Challenge showed real time factors of 2.07% and 2.10% in each of the previous datasets (this in only given for normalization purposes). Our blind approach for joint noise and reverberation suppression is able to run in real-time.

3.3. Evaluation of the Automatic Speech Recognition task

For the *ASR* task we only evaluate the performance of joint noise and reverberation suppression method (*DRVNR*). *ASR* systems are highly

WER	SimData							RealData		
	Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
	Near	Far	Near	Far	Near	Far		Near	Far	
Baseline	12.93	17.72	24.03	72.54	30.46	79.72	39.53	83.16	84.48	83.81
AM_{clean}	17.54	22.42	24.04	45.60	30.78	56.92	32.87	74.58	71.71	73.14
AM_{mmc}	19.13	21.42	21.00	29.89	24.45	35.24	25.35	52.06	51.08	51.57

Table 6. Word Error Rate for SimData and RealData

dependent on the acoustic models used for training so we consider two different acoustic models in this evaluation. First, we use the acoustic model trained on clean data and made available for the REVERB Challenge, we will refer to this case as AM_{clean}. For the second acoustic model, we processed the reverberant data from the *Multi Condition* dataset with our enhancement approach. We trained a new acoustic model with this data in order to learn the characteristics of our approach. We will refer to this model as AM_{mmc}. Finally, we denote as *Baseline* the scores obtained with the unprocessed data and the clean acoustic model. All the *ASR* results presented here are obtained after applying CMLLR to the recognized data in order to improve the matching between the acoustic features and the training data. We obtained the best *ASR* results with this configuration.

We first analyze the behavior of the *Baseline* system. The *ASR* system has a fair performance in the smallest room but it rapidly degrades as the room becomes bigger. The system is particularly bad in far field conditions.

Now, we consider the AM_{clean} case. In room 1 from the *SimData* dataset, the distortions introduced by our approach (see above) provoke a degradation in the recognition accuracy. In rooms 2 and 3 there is a slight degradation in near field conditions but we observe a significant improvement in far field conditions. In the *RealData* dataset there was an improvement in near field conditions but the scores in far field conditions are still better. We conclude that our method is mostly suitable to enhance highly reverberated signals with low Direct to Reverberation Ratio. Finally, we obtain our best recognition results in the AM_{mmc} case. The overall behavior of the method is the same as in the AM_{clean}. However, the WER significantly dropped with this modified acoustic model. This shows that it is mandatory to train the acoustic model with reverberant data that has been processed by our approach.

3.4. Discussion

Here we summarize the main conclusions from the evaluation.

Regarding the proposed method, we showed that it is able to reduce reverberation in every condition but it introduces annoying distortions in near field conditions and in small rooms. This is in part explained because we are working in blind conditions and thus late reverberation is overestimated in small enclosures. Reducing the hyperparameter λ allows to limit those distortions. The *ASR* experiments shows that joint noise and reverberation suppression can significantly improve the WER. It is recommended to train the acoustic model with reverberant data that has been processed by our enhancement method.

Regarding the measures considered for this evaluation, the SRMR is well suited to evaluate the level of reverberation but care must be taken when using other processings that modify the modulation envelope of the signals. We suggest that the CD and the LLR are both well adapted to evaluate distortions. The FWSNR behaves similarly to the SRMR but it is also able to evaluate the trade-off between speech enhancement and distortions. Finally, there is a correspondence between the observed distortion measures (CD, LLR) and the WER. We believe that these measures could be used to predict the performance of the *ASR* system.

4. CONCLUSION

We presented a novel method for single channel late reverberation suppression. Late reverberation is estimated as a sparse linear combination of previous frames before being suppressed by a state-of-the-art time-frequency filter. While this method showed an improvement in the SRMR of the processed signals, it also introduced distortions to the speech signal. To cope with this we proposed to jointly suppress background noise and late reverberation. According to this framework, both interferences are estimated individually and used afterwards to design a joint filter. The overall performances were neatly improved for both Speech Enhancement and Automatic Speech Recognition tasks, specially for data recorded in far field conditions and in large rooms. The resulting system was able to process noisy and reverberant data in real-time while keeping reasonable distortion levels.

5. REFERENCES

[1] K. Lebart, J. M. Boucher, and P. N. Denbigh, “A New Method Based on Spectral Subtraction for Speech Dereverberation,” *Acta Acustica United With Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[2] E. Habets, “Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement,” Ph.D. dissertation, Technische Universiteit Eindhoven, 2007.

[3] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[4] Y. Ephraim and D. Malah, “Speech Enhancement Using a Minimum Mean Square Error Log Spectral Amplitude Estimator,” *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, 1985.

[5] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, A. S. Volker Leutnant, 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. Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, U.S.A., 2013.

[6] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least Angle Regression,” *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[7] E. Habets, I. Cohen, and S. Gannot, “MMSE Log-Spectral Amplitude Estimator for Multiple Interferences,” in *Proc. International Workshop on Acoustic Echo and Noise Control*, Paris, France, 2006.