BLIND SPEECH DEREVERBERATION USING CONVOLUTIVE NONNEGATIVE MATRIX FACTORIZATION WITH MIXED PENALIZATION

F. J. Ibarrola[†], L. E. Di Persia[†] and R. D. Spies[‡]

[†]*Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH-UNL/CONICET, Argentina. Ciudad Universitaria, CC 217, Ruta Nac. 168, km 472.4, (3000) Santa Fe, Argentina.*
fibarrola@sinc.unl.edu.ar, ldipersia@sinc.unl.edu.ar
[‡]*Instituto de Matemática Aplicada del Litoral, IMAL, CONICET-UNL, Centro Científico Tecnológico CONICET Santa Fe, Colectora Ruta Nac. 168, km 472, Paraje "El Pozo", (3000), Santa Fe, Argentina and Departamento de Matemática, Facultad de Ingeniería Química, Universidad Nacional del Litoral, Santa Fe, Argentina.*
rspies@santafe-conicet.gov.ar

Abstract: When a signal is recorded in an enclosed room, it typically gets affected by reverberation. This degradation represents a problem when dealing with audio signals, particularly in the field of speech applications, such as automatic speech recognition. Although there are some approaches to deal with this issue that are quite satisfactory under certain conditions, constructing a method that works well in a general context still poses a significant challenge. As an effort in this direction, we propose a method based on convolutive nonnegative matrix factorization that mixes two penalizers in order to impose certain characteristics over the time-frequency components of the restored signal and the reverberant components. An algorithm for finding such a solution is described and tested. The results show a significant improvement on the quality of the restored signals.

Keywords: *signal processing, dereverberation, regularization*
2000 AMS Subject Classification: 65F22 - 65T50

1 INTRODUCTION

When captured in enclosed rooms, audio recordings will most certainly be affected by reverberant components due to reflections of the sound waves in the walls, ceiling, floor or furniture. This can severely degrade the characteristics of the recorded signal, generating difficult problems for processing such a signal, particularly when required for certain speech applications. The goal of any dereverberation technique is to remove or attenuate the reverberant components to obtain a cleaner signal. The dereverberation problem is called "blind" when the available data consists only of the reverberant signal itself, and this is the problem we shall address on this work.

Depending on the problem, our observation might consist of a single or multi-channel signal. That is, we might have a signal recorded by one or more microphones. For the latter case, there are several proposed methods that work quite well ([1]). For the case of single-channel, although some methods perform reasonably well ([2], [3], [4]), there is still much room for improvement.

In this work we present a dereverberation method for single channel data based on the idea of penalizing different characteristics of the components of a convolutive nonnegative matrix factorization (NMF) representation model for the reverberation phenomenon.

1.1 A REVERBERATION MODEL

Let $s, x : \mathbb{R} \to \mathbb{R}$ with support in$[0, \infty)$ be the functions associated to the clean and reverberant signals, respectively. Then, the reverberation model can be written as

$$x(t) = (h * s)(t), \tag{1}$$

where $h : \mathbb{R} \to \mathbb{R}$ is the room impulse response (RIR) signal, and "$*$" denotes convolution.

When dealing with sound signals (particularly speech signals), it is often convenient to work with the associated spectrograms rather than the signals themselves. Thus, we make use of the short time Fourier

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
R. Spies, L. Di Persia & F. J. Ibarrola; "BLIND SPEECH DEREVERBERATION USING CONVOLUTIVE NONNEGATIVE MATRIX FACTORIZATION WITH MIXED PENALIZATION"
Proceedings of VI MACI, 2017.

transform (STFT) on the discretized version of (1) to obtain the corresponding complex time-frequency components, resulting in the model

$$\mathbf{x}_k[t] = \sum_{\tau=0}^{T_h-1} \mathbf{s}_k[t-\tau]\mathbf{h}_k[\tau], \tag{2}$$

where $t = 1 \dots T$, is a discretized time variable, $k = 1, \dots K$, denotes the frequency subband and T_h is a parameter of the model associated to the maximum expected duration of the reverberation phenomenon.

Now, let us write $\mathbf{h}_k[\tau] = |\mathbf{h}_k[\tau]|e^{j\phi_k[\tau]}$. To overcome the problems derived from the well known sensitivity of the phase angle $\phi_k[\tau]$ with respect to variations of the reverberation conditions, we shall proceed as in [2], treating $\phi_k[\tau]$ as a random variable with distribution $\mathcal{U}[-\pi, \pi)$. Denoting the complex conjugate as * and the Kronecker delta as δ_{ij}, the latter assumption yields

$$\begin{aligned} E|\mathbf{x}_k[t]|^2 &= E\sum_{\tau,\tau'} \mathbf{s}_k[t-\tau]\mathbf{s}_k^*[t-\tau']\mathbf{h}_k[\tau]\mathbf{h}_k^*[\tau'] \\ &= \sum_{\tau,\tau'} \mathbf{s}_k[t-\tau]\mathbf{s}_k^*[t-\tau']\,|\mathbf{h}_k[\tau]|\,|\mathbf{h}_k[\tau']|\,Ee^{j(\phi_k[\tau]-\phi_k[\tau'])} \\ &= \sum_{\tau,\tau'} \mathbf{s}_k[t-\tau]\mathbf{s}_k^*[t-\tau']\,|\mathbf{h}_k[\tau]|\,|\mathbf{h}_k[\tau']|\,\delta_{\tau\tau'} \\ &= \sum_{\tau} |\mathbf{s}_k[t-\tau]|^2\,|\mathbf{h}_k[\tau]|^2. \end{aligned}$$

Finally, let us define $S_k[t] \doteq |\mathbf{s}_k[t]|^2$, $H_k[t] \doteq |\mathbf{h}_k[t]|^2$ and $X_k[t] \doteq E|\mathbf{x}_k[t]|^2$. Then, our model reads

$$X_k[t] = \sum_{\tau} S_k[t-\tau]H_k[\tau], \tag{3}$$

and the square magnitude of the observed spectrogram components can be written as

$$Y_k[t] = X_k[t] + \epsilon_k[t],$$

where $\epsilon_k[t]$ denotes the representation error. As shown in [2], this model is equivalent to a convolutive NMF with diagonal basis.

2 Mixed Penalization

As a way of measuring the representation error, we will use the square of the Frobenius norm $||Y-X||_F^2$, where Y and X are the matrices whose (k,t) components are $Y_k[t]$ and $X_k[t]$, respectively.

Since we are dealing with a blind dereverberation problem, we have no information on the structure of the matrix H (with elements $H_k[t]$). Hence, we must impose some conditions on the representation (3) in order to ensure that S and H will provide a satisfactory representation for our dereverberation problem.

For clean speech signals, the spectrogram is expected to have some sparse structure, which is not preserved under reverberant conditions (see Figure 1). This sparsity can be regained by introducing a penalization term over the matrix S. In a similar fashion, certain regularity conditions over the matrix H can be imposed to improve its correspondence with a room impulse response (RIR) signal.

Following these ideas, we propose the following cost function:

$$f(H,S) \doteq \sum_{t,k}(Y_k[t] - X_k[t])^2 + \lambda_1 \sum_{t,k} |H_k[t]|^{p_1} + \lambda_2 \sum_{t,k} |S_k[t]|^{p_2},$$

where $\lambda_1, \lambda_2 \geq 0$ are penalization parameters that quantify the weights of both penalizers relative to the fidelity term, whereas the exponents $p_1, p_2 \in (0,2)$ are tunning parameters. Note that small values of these parameters will promote sparsity, whereas values close to 2 will promote smoothness. Since there is a clear scale indeterminacy in the representation (3), the additional constraint $\sum_{\tau=1}^{T_h} H_k[\tau] = 1\ \forall k$ shall be imposed.

Next, we present an algorithm for approximating the matrices H and S that minimize f.

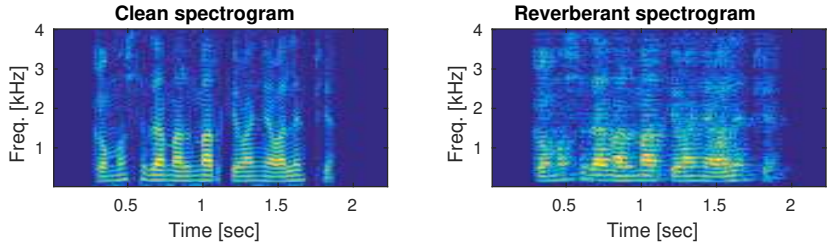

Figure 1: Spectrograms for a clean speech signal (left) and the corresponding reverberant speech signal (right).

3 UPDATING RULES

We shall build an iterative algorithm following the idea in [2], which is based on the auxiliary function technique.

Let $\Omega \subset \mathbb{R}$ and $f : \Omega \to \mathbb{R}_0^+$. Then, $g : \Omega \times \Omega \to \mathbb{R}_0^+$ is called an *auxiliary function* for f if $\forall w, w' \in \Omega$, $g(w, w') \geq f(w)$ and $g(w, w) = f(w)$. With this definition, it can be shown ([5]) that the sequence $\{f(w^j)\}_j$ is non-increasing under the update rule

$$w^j = \arg\min_w g(w, w^{j-1}). \tag{4}$$

We will use this approach to alternatively update the matrices H and S. Let us begin by fixing $H = H'$, where H' is an arbitrary $K \times T_h$ matrix. Then, it can be shown that an auxiliary function for f with respect to S is given by

$$\begin{aligned} g_s(S, S') \doteq & \sum_{k,t,\tau} \frac{S'_k[\tau]H'_k[t-\tau]}{X'_k[t]} \left(Y_k[t] - \frac{S_k[\tau]}{S'_k[\tau]} X'_k[t] \right)^2 + \lambda_1 \sum_{k,t} |H'_k[t]|^{p_1} \\ & + l + \lambda_2 \sum_{k,t} \left(\frac{p_2}{2} S'_k[t]^{p_2-2} S_k[t]^2 + |S'_k[t]|^{p_2} - \frac{p_2}{2} |S'_k[t]|^{p_2} \right), \end{aligned}$$

where $X'_k[t] = \sum_\tau S'_k[\tau]H'_k[t-\tau]$. In an analogous way, fixing $S = S'$, an auxiliary function for f with respect to H is given by

$$\begin{aligned} g_h(H, H') \doteq & \sum_{k,t,\tau} \frac{S'_k[t-\tau]H'_k[\tau]}{X'_k[t]} \left(Y_k[t] - \frac{H_k[\tau]}{H'_k[\tau]} X'_k[t] \right)^2 + \lambda_2 \sum_{k,t} |S'_k[t]|^{p_2} \\ & + \lambda_1 \sum_{k,t} \left(\frac{p_1}{2} H'_k[t]^{p_1-2} H_k[t]^2 + |H'_k[t]|^{p_1} - \frac{p_1}{2} |H'_k[t]|^{p_1} \right). \end{aligned}$$

Now, since g_s is quadratic with respect to S and g_h is quadratic with respect to H, we can use the first order necessary conditions to find the minimizers complying with the update rule (4). This leads to the following updating rules:

$$S_k[\tau] = S'_k[\tau] \frac{\sum_t H'_k[t-\tau]Y_k[t]}{\sum_t H'_k[t-\tau]X'_k[t] + \frac{\lambda_2 p_2}{2} |S'_k[\tau]|^{p_2-1}}, \qquad H_k[\tau] = H'_k[\tau] \frac{\sum_t S'_k[t-\tau]Y_k[t]}{\sum_t S'_k[t-\tau]X'_k[t] + \frac{\lambda_1 p_1}{2} |H'_k[\tau]|^{p_1-1}}.$$

Every updating step must be followed by a normalization of the rows of H to avoid the aforementioned scale indeterminacy issue. In principle, the algorithm is run until $\|S - S'\|_F^2$ reachs an established threshold value, but it is worth noting that other stopping criteria might also be suitable.

4 EXPERIMENTAL RESULTS

We begin by showing an example of the performance of our method. Starting from a clean speech signal sampled at 8kHz, we have artificially constructed a reverberant version by discrete convolution with a RIR signal from a simulated enclosed room with 400ms of reverberation time. The spectrogram was then computed using STFT with 256 window length and overlapping of 128 samples. Figure 2 shows the clean speech spectrogram, together with its reverberant version and a restoration using our method with parameters $p_1 = 1.8$ and $p_2 = 1.2$, meaning we impose some sparsity to S and a mild smoothness to H.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
R. Spies, L. Di Persia & F. J. Ibarrola; "BLIND SPEECH DEREVERBERATION USING CONVOLUTIVE NONNEGATIVE MATRIX FACTORIZATION WITH MIXED PENALIZATION"
Proceedings of VI MACI, 2017.

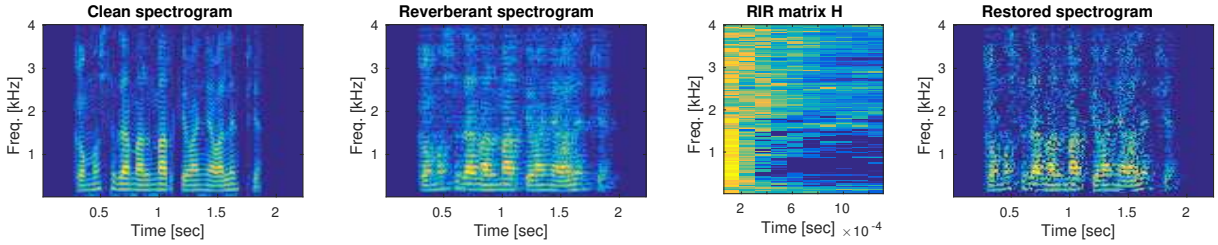

Figure 2: Spectrograms for a clean speech signal, its reverberant version, the RIR matrix and the obtained restoration.

To measure the performance of the method, we used the *frequency weighted segmental SNR* (fwsSNR) for its relevance for speech applications such as automatic speech recognition ([6]). The fwsSNR values are 16.20 for the reverberant signal and 17.41 for the restored example, indicating a significant improvement.

Next, we compare our method with the one proposed by Kameoka *et. al.* ([2]), which essentially consists of single penalization based on a Bayesian approach. To do so, both methods were run on artificially constructed reverberant signals with six different RIRs (three different microphone/source positions and two reverberation times) from a database of 20 clean speech signals. The parameters of the model were set as: $T_h = 10$, $p_1 = 1.8$ and $p_2 = 1.2$. For the sake of comparison, λ_2 and the maximum number of iterations (set as 20) were chosen as in [2] and λ_1 was chosen as $\lambda_2 \times 10^3$. The results of the experiment are summarized in Table 1, where improvements on the mean fwsSNR values (over the 20 signals) can be seen.

	RIR1	RIR2	RIR3	RIR4	RIR5	RIR6
Reverberant signal	16.93	14.19	17.86	15.19	18.00	15.76
Kameoka's restoration	17.38	14.58	18.38	15.66	18.49	16.25
Mixed pen. restoration	17.60	14.76	18.53	15.88	18.52	16.48

Table 1: Experimental results measures: mean fwsSNR values for speech dereverberation.

5 Conclusions

In this work we presented a model for signal dereverberation based on convolutive NMF with mixed penalization. An iterative updating algorithm was introduced and its performance was tested and compared with a state of the art method. The results show that our mixed penalization improves the quality of the restorations.

Although these preliminary results are promising, there is still room for improvement. For instance, other types of penalizing terms can be used, and different ways to optimize the model parameters can be sought.

Acknowledgements

This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET through PIP 2014-2016 N 11220130100216-CO, the Air Force Office of Scientific Research, AFOSR/SOARD, through Grant FA9550-14-1-0130 and by Universidad Nacional del Litoral, UNL, through CAID-UNL 2011 N50120110100519 "Procesamiento de Señales Biomédicas."

References

[1] M. Delcroix, T. Yoshioka, A. Ogawam Y. Kubo, M. Fujimoto, I. Nobutaka, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, A. Namakura, *Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB Challenge*, Proceedings of Reverb challenge 02.3 (2014).

[2] H. Kameoka, T. Nakatani, T Yoshioka, *Robust speech dereverberation based on non-negativity and sparse nature o speech spectrograms*, ICASSP (2009), pp. 45-48.

[3] S. Xizhong and M Guang, *Complex cepstrum based single channel speech dereverberation*, Proceedings of 4th International Conference on Computer Science & Education (2009), pp. 7-11.

[4] M. Moshirynia, F. Razzazi, A Haghbin, *A speech dereverberation method using adaptive sparse dictionary learning*, REVERB Workshop (2014), pp. 1-7.

[5] D. D. Lee, H. S. Seung, *Algorithms for non-negative matrix factorization*, NIPS (2000), pp. 556-562.

[6] Y. Hu and P. C. Loizou, *Evaluation of objective quality measures for speech enhancement*, IEEE Trans. Audio, Speech, Lang. Process. (2008), 16, pp. 229238.