On the use of convolutive nonnegative matrix factorization with mixed penalization for blind speech dereverberation

Francisco J. Ibarrola *[1], Ruben D. Spies[2], and Leandro E. Di Persia[1]

[1]Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL, CONICET, Ciudad Universitaria UNL, (3000) Santa Fe, Argentina.
[2]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.

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 for applications involving automatic speech and/or speaker recognition. There are some approaches to deal with this issue that are quite satisfactory when multi-channel recordings or learning data are available, but this is not the general case in most human-computer interaction applications, and constructing a method that works well in a general context still poses a significant challenge. In this article, 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. Comparisons of the results against state of the art methods are presented, showing significant improvement.

Keywords: signal processing, dereverberation, regularization.

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 ([1]), generating difficult problems for processing such a signal, particularly when required for certain speech applications ([2]). 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 ([3]). For the case of single-channel, although some methods perform reasonably well ([4], [5], [6]), there is still much room for improvement.

*fibarrola@sinc.unl.edu.ar

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.

March 20, 2019

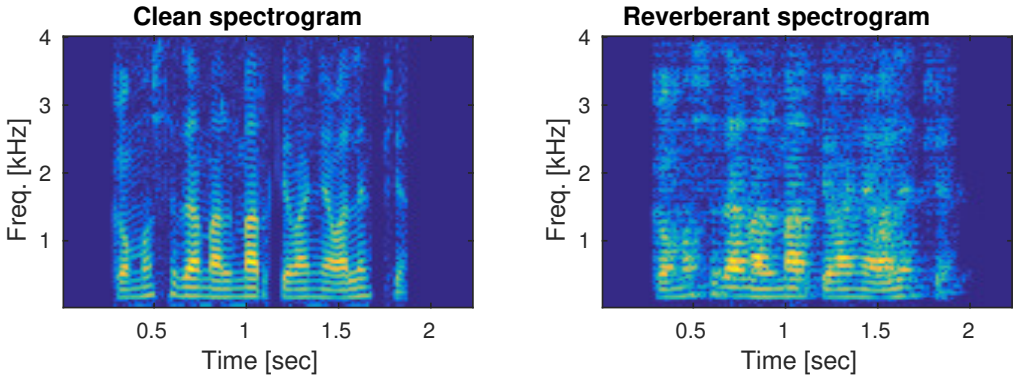

Figure 1: Spectrograms for a clean speech signal (left) and the corresponding reverberant speech signal (right).

Let $s, x : \mathbb{R} \to \mathbb{R}$, with support in $[0, \infty)$, be the functions associated with the clean and reverberant signals, respectively. Then, our 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. This model is valid under the hypothesis of a linear, time-invariant system. In practice, this implies we are assuming the source and microphone positions to be static, and the signal energy to be low enough for the effect of the non-linear components to be insignificant.

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 transform (STFT), defined as

$$\mathbf{x}_k[t] \doteq \int_{-\infty}^{\infty} x(u)w(u-t)e^{-2\pi iuk}du, \quad t, k \in \mathbb{R}$$

where $w : \mathbb{R} \to \mathbb{R}_0^+$ is a given *window* function. Denoting the STFTs of h and s by $\mathbf{s}_k[t]$ and $\mathbf{h}_k[t]$, respectively, a discretized approximation of the STFT model associated to (1) is given ([4]) by

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

where $t = 1, \ldots, T$, is a discretized time variable that corresponds to window locations, $k = 1, \ldots, K$, denotes the frequency subband and T_h is a parameter of the model associated to the expected maximum duration of the reverberation phenomenon. Later on, the values of t will be chosen in such a way that the union of the windows' supports contain the support of the observed signal, and the values of k in such a way that they cover the whole frequency spectrum, up to half the sampling frequency.

Now, let us write $\mathbf{h}_k[\tau] = |\mathbf{h}_k[\tau]|e^{j\phi_k[\tau]}$. It is well known that the phase angles $\phi_k[\tau]$ are highly sensitive with respect to mild variations on the reverberation conditions. To overcome the problems derived from this, we shall proceed (see [4]) to treat the $K \times T_h$ variables $\phi_k[\tau]$ as random variables

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
F. Ibarrola, L. Di Persia & R. Spies; "On the use of convolutive nonnegative matrix factorization with mixed penalization for blind speech dereverberation"
2017 XLIII Latin American Computer Conference (CLEI), pp. 1-4, 2017.

i.i.d. with uniform distribution in $[-\pi, \pi)$. Denoting the complex conjugate as "$*$" and the Kronecker delta as δ_{ij}, the expected value of $|\tilde{\mathbf{x}}_k[t]|^2$ is given by

$$\begin{aligned} E|\tilde{\mathbf{x}}_k[t]|^2 &= E\Big(\sum_{\tau,\nu} \mathbf{s}_k[t-\tau]\mathbf{s}_k^*[t-\nu]\mathbf{h}_k[\tau]\mathbf{h}_k^*[\nu] \Big) \\ &= \sum_{\tau,\nu} \mathbf{s}_k[t-\tau]\mathbf{s}_k^*[t-\nu]|\mathbf{h}_k[\tau]||\mathbf{h}_k[\nu]|Ee^{j(\phi_k[\tau]-\phi_k[\nu])} \\ &= \sum_{\tau,\nu} \mathbf{s}_k[t-\tau]\mathbf{s}_k^*[t-\nu]\,|\mathbf{h}_k[\tau]|\,|\mathbf{h}_k[\nu]|\,\delta_{\tau\nu} \\ &= \sum_{\tau} |\mathbf{s}_k[t-\tau]|^2\,|\mathbf{h}_k[\tau]|^2. \end{aligned}$$

Note that the $[-\pi, \pi)$ interval choice for $\phi_k[\tau]$ is arbitrary, since this result holds for any $2\pi-$length interval. Finally, let us define $S_k[t] \doteq |\mathbf{s}_k[t]|^2$, $H_k[t] \doteq |\mathbf{h}_k[t]|^2$ and $X_k[t] \doteq E|\tilde{\mathbf{x}}_k[t]|^2$. 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

$$Y_k[t] = X_k[t] + \epsilon_k[t], \tag{4}$$

where $\epsilon_k[t]$ denotes the representation error. As shown in [4], this model is equivalent to a convolutive NMF ([11]) with diagonal basis. In the next section, we build a cost function in order to find an appropriate convolutive representation that allows us to isolate the components of $S_k[t]$.

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.

As it can be seen in Figure 1, for clean speech signals, the spectrogram is expected to have some sparse structure, which is not preserved under reverberant conditions. 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.

Based upon these ideas, we propose the following cost function:

$$J(H,S) \doteq \sum_{t,k} \left[(Y_k[t] - X_k[t])^2 + \lambda_{1,k}|H_k[t]|^{p_1} + \lambda_{2,k}|S_k[t]|^{p_2} \right],$$

where $\lambda_{1,k}, \lambda_{2,k} \geq 0$ 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. 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), we impose the (somewhat arbitrary) additional constraint $||S_k||_2 = ||Y_k||_2\ \forall k$, which means that the ℓ^2-norm (energy) shall remain equal for every frequency.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
F. Ibarrola, L. Di Persia & R. Spies; "On the use of convolutive nonnegative matrix factorization with mixed penalization for blind speech dereverberation"
2017 XLIII Latin American Computer Conference (CLEI), pp. 1-4, 2017.

2.1 Regularization parameters

As mentioned before, the parameters $\lambda_{1,k}, \lambda_{2,k},\ k = 1, \ldots, K$, weight the penalizers against the fidelity term. In this sense, the optimal weights of these regularization parameters might vary as a function of the frequency subband, and hence their proposed dependency on k. Since searching blindly for $2K$ parameters is non-viable in practice, we quantify this dependency by defining $\lambda_{1,k} \doteq \lambda_1 \sum_{t=1}^{T} |Y_k[t]|^2$ and $\lambda_{2,k} \doteq \lambda_2 \sum_{t=1}^{T} |Y_k[t]|^2$. This means we only need to look for two paramenters (λ_1, λ_2) and then multiply them by the energy of the signal associated to each row of Y.

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

3 Updating rules

We shall build an iterative algorithm following the idea in [4], 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 ([7]) that for any $w^0 \in \Omega$, the sequence $\{f(w^j)\}_{j=0}^{\infty}$ is non-increasing under the update rule

$$w^j = \arg\min_{w} g(w, w^{j-1}), \quad j = 1, \ldots, \infty. \tag{5}$$

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, if we let

$$X'_k[t] = \sum_{\tau} S'_k[\tau] H'_k[t - \tau],$$

it can be shown that the function g_s, defined as

$$\begin{aligned} g_s(S, S') \doteq & \sum_{k,t,\tau} \frac{S'_k[\tau] H'_k[t-\tau]}{X'_k[t]} \left(Y_k[t] - \frac{S_k[\tau]}{S'_k[\tau]} X'_k[t] \right)^2 \\ & + \sum_{k,t} \lambda_{1,k} |H'_k[t]|^{p_1} \\ & + \sum_{k,t} \lambda_{2,k} \Big(\frac{p_2}{2} S'_k[t]^{p_2 - 2} S_k[t]^2 + \Big(1 - \frac{p_2}{2}\Big) |S'_k[t]|^{p_2} \Big), \end{aligned}$$

is an auxiliary function for J with respect to S.

In an analogous way, fixing $S = S'$, an auxiliary function for J with respect to H is given by g_h, defined as

$$\begin{aligned} g_h(H, H') \doteq & \sum_{k,t,\tau} \frac{S'_k[t-\tau] H'_k[\tau]}{X'_k[t]} \left(Y_k[t] - \frac{H_k[\tau]}{H'_k[\tau]} X'_k[t] \right)^2 \\ & + \sum_{k,t} \lambda_{1,k} \left(\frac{p_1}{2} H'_k[t]^{p_1 - 2} H_k[t]^2 + \frac{2 - p_1}{2} |H'_k[t]|^{p_1} \right) \\ & + \sum_{k,t} \lambda_{2,k} |S'_k[t]|^{p_2}. \end{aligned}$$

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 (5). This leads to the following updating rules:

$$S_k[\tau] = S'_k[\tau] \frac{\sum_t H'_k[t-\tau] Y_k[t]}{\sum_t H'_k[t-\tau] X'_k[t] + \frac{\lambda_{2,k}}{2} p_2 |S'_k[\tau]|^{p_2 - 1}},$$

$$H_k[\tau] = H'_k[\tau] \frac{\sum_t S'_k[t-\tau] Y_k[t]}{\sum_t S'_k[t-\tau] X'_k[t] + \frac{\lambda_{1,k}}{2} p_1 |H'_k[\tau]|^{p_1 - 1}}.$$

In order to avoid the aforementioned scale indeterminacy, every updating step is to be followed by scaling S_k so that its ℓ^2 norm coincides with that of the observation Y_k. In principle, the algorithm is run until $\|S - S'\|_F^2$ decreases below an established threshold value, although it is worth noting that other stopping criteria might also be suitable.

4 Experimental results

For the experiments, we took 110 speech signals from the TIMIT database[1], recorded at 16 KHz, and we artificially made them reverberant using the software Room Impulse Response Generator by E.A.P. Habets[2], based on the model in [9]. Each signal was degraded under different reverberation conditions: three different room sizes, each with three different microphone positions and four different reverberation times.

In order to avoid preprocessing, the choice of the regularization parameters was made *a priori* by means of empirical rules, based upon signals from a different database. This is supported by the fact that the parameters were observed to be rather robust with respect to variations of the reverberation conditions, and hence they were chosen simply as $\lambda_1 = 1$ and $\lambda_2 = 10^{-4}$. The rest of the model parameters were chosen as specified in Table 1.

p_1	p_2	T_h	window size	win. overlapping	max. iter.
1.8	1	15	256 samples	128 samples	20

Table 1: Model parameter values

As previously discussed, the choice of $p_1 = 1.8$ is meant to promote smoothness over H, while the choice of $p_2 = 1$ aims to induce sparsity over S.

In order to evaluate the performance of our model, we made comparisons against two state of the art methods under the same conditions: the one proposed by Kameoka *et al* in [4], and the one proposed by Wisdom et. al. in [10] (with a window length of 2048), choosing all the parameters as suggested by the authors.

To measure performance, following [8], we made use of the frequency weighted segmental signal-to-noise ratio (fwsSNR) and the cepstral distance. The results for each performance measure are stated in Tables 2 and 3 and depicted in Figure 2, the different reverberation times: 300[ms], 450[ms], 600[ms] and 750[ms]. Notice that for the case of fwsSNR, higher values correspond to better performance, while for the cepstral distance, small values indicate higher quality.

Rev. time	Rev. Signal	Kameoka's	Wisdom's	Mixed pen.
300 [ms]	8.102(1.96)	7.950(1.73)	8.262(1.53)	**9.148**(1.71)
450 [ms]	4.815(1.42)	5.127(1.36)	5.771(1.28)	**6.458**(1.45)
600 [ms]	3.082(1.20)	3.358(1.19)	4.140(1.17)	**4.547**(1.31)
750 [ms]	1.998(1.11)	2.184(1.10)	3.013(1.12)	**3.239**(1.22)

Table 2: Mean and (standard deviation) of fwsSNR for each method and reverberation time.

In regard to the fwsSNR peformance measure, the values in Table 2 give account of a strong improvement of our proposed method with respect to the others. As for the cepstral distance, although our method outperforms the other two, an improvement with respect to the reverberant signal is

[1] https://catalog.ldc.upenn.edu/ldc93s1

[2] https://github.com/ehabets/RIR-Generator

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
F. Ibarrola, L. Di Persia & R. Spies; "On the use of convolutive nonnegative matrix factorization with mixed penalization for blind speech dereverberation"
2017 XLIII Latin American Computer Conference (CLEI), pp. 1-4, 2017.

Table 3: Mean and (standard deviation) of cepstral distance for each method and reverberation time.

Rev. time	Rev. Signal	Kameoka's	Wisdom's	Mixed pen.
300 [ms]	3.440(0.44)	4.057(0.45)	3.908(0.48)	**3.566**(0.44)
450 [ms]	4.264(0.44)	4.636(0.42)	4.511(0.41)	**4.124**(0.41)
600 [ms]	4.716(0.46)	5.006(0.42)	4.860(0.40)	**4.519**(0.41)
750 [ms]	5.011(0.48)	5.264(0.43)	5.089(0.40)	**4.807**(0.42)

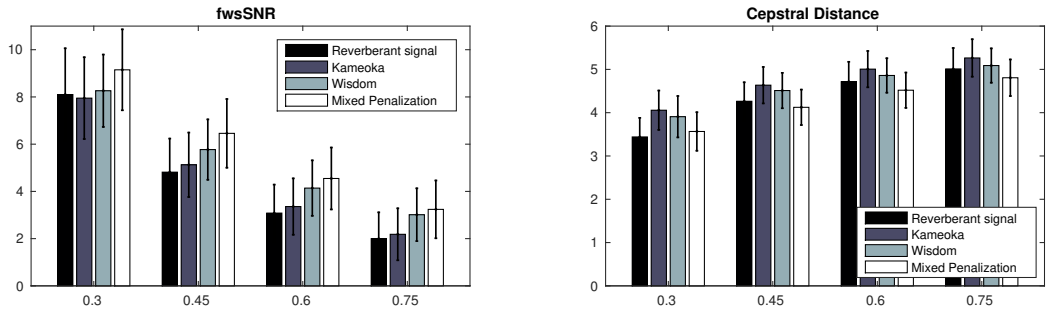

Figure 2: Mean and standard deviations of performance measures for different reverberation times.

observed only for reverberation times of 450[ms] or greater. A t-test with significance level $\alpha = 0.05$ was done using all the obtained results, showing statistical significance of the improvement on the performance of our method with respect to the reverberant signal and the other methods.

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 two state of the art methods. The results show that our mixed penalization method 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, different ways to optimize the model parameters can be sought, etcetera.

Acknowledgments

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 N°50120110100519 "Procesamiento de Señales Biomédicas." and CAI+D-UNL 2016, PIC 50420150100036LI "Problemas Inversos y Aplicaciones a Procesamiento de Señales e Imágenes".

References

[1] I. Tashev, *Sound Capture and Processing: Practical Approaches*, John Wiley & Sons, New Jersey, 2009.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
F. Ibarrola, L. Di Persia & R. Spies; "On the use of convolutive nonnegative matrix factorization with mixed penalization for blind speech dereverberation"
2017 XLIII Latin American Computer Conference (CLEI), pp. 1-4, 2017.

[2] X. Huang, A.Acero and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, New Jersey, 2001.

[3] 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).

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

[5] 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.

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

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

[8] Y. Hu and P. C. Loizou, *Evaluation of objective quality measures for speech enhancement*, IEEE Trans. Audio, Speech, Lang. Process. (2008), 16, pp. 229-238.

[9] J.B. Allen and D.A. Berkley, *Image method for efficiently simulating small-room acoustics*,Journal Acoustic Society of America. (1979), 65, pp. 943-950.

[10] S. Wisdom, T. Powers, L. Atlas and J. Pitton, *Enhancement of reverberant and noisy speech by extending its coherence*, Proceedings of Reverb Challenge (2014).

[11] P. Smaragdis, *Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs* , Fifth International Conference on Independent Component Analysis, LNCS 3195 (2004), pp 494-499.