Mixed penalization in convolutive nonnegative matrix factorization for blind speech dereverberation

Francisco J. Ibarrola * Leandro E. Di Persia * Ruben D. Spies †

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 signal processing, 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. 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 implementing the method is described and tested. Comparisons of the results against those obtained with state of the art methods are presented, showing significant improvement.

Keywords: signal processing, dereverberation, regularization.

1 Introduction

In recent years, many technological developments have attracted attention towards human-machine interaction. Since the most natural and easiest way of human communication is through speech, much research effort has been put into achieving the same natural interaction with machines. This effort has already generated many advances in a wide variety of fields such
as automatic speech recognition ([1]), automatic translation systems ([2]) and control of remote devices through voice ([3]), to name only a few. A significant amount of work has been recently devoted to produce robustness in speech recognition ([1]), resulting in several advances in the areas of speech enhancement ([1], [5]), multiple sources separation ([6], [7]), and particularly in dereverberation techniques ([8]), which constitute the topic of this work.

When recorded in enclosed rooms, audio signals 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 ([9]), generating difficult problems for its processing, particularly when required for certain speech applications ([10]). The goal of any dereverberation technique is to remove or to attenuate the reverberant components in order 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 deal with in 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, quite a few methods exist that work relatively well ([11], [12]).

For the single-channel case, we may distinguish between supervised and unsupervised approaches. The first kind refers to those that begin with a training stage that serves to learn some characteristics of the reverberation conditions, while the second kind alludes to those methods that can be implemented directly over the reverberant signal. Some supervised methods ([13], [14], [15]) appear to perform somewhat better than unsupervised ones, but they pose the disadvantage of needing learning data corresponding to the specific room conditions, microphone and source locations, and a previous process that might take a significant amount of time.

In the context of unsupervised blind dereverberation, although some recently proposed methods ([12], [16]) work reasonably well, there is still much room for improvement. Our work is based on a convolutive non-negative matrix factorization (NMF) reverberation model, as proposed by Kameoka et al ([16]), along with a Bayesian approach for building a generalized functional that mixes two types of penalizers over the elements of the representation model. Mixed penalization approaches have been recently used and successfully applied by several authors in many areas, mainly in signal and image processing applications ([17], [18], [19], [20], [21]). These techniques have shown to produce good results in terms of enhancing certain desirable characteristics on the solutions while precluding unwanted ones.
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. As it is customary, we shall assume that the reverberation process is well represented by a Linear Time-Invariant (LTI) system. Thus, the reverberation model can be written as

\[
x(t) = (h \ast s)(t),
\]

where \( h : \mathbb{R} \to \mathbb{R} \) is the room impulse response (RIR) signal, and \( \ast \) denotes convolution. This LTI hypothesis implies we are assuming the source and microphone positions to be static, and the energy of the signal to be low enough for the effect of the non-linear components to be relatively 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

\[
x_k(t) = \int_{-\infty}^{\infty} x(u)w(u-t)e^{-2\pi iu k}du,
\]

where \( w : \mathbb{R} \to \mathbb{R}_0^+ \) is a compactly supported, even function such that \( \|w\|_1 = 1 \). This function is called window.

In practice, we work with discretized versions of the signals involved (\( x[\cdot], h[\cdot], s[\cdot], \) and \( w[\cdot] \)). With this in mind, we shall define the discrete STFT as

\[
x_k[n] = \sum_{m=-\infty}^{\infty} x[m]w[m-n]e^{-2\pi i m k},
\]

where \( n = 1, \ldots, N \), is a discretized time variable that corresponds to window location, \( k = 1, \ldots, K \), denotes the frequency subband and \( N_h \) is a parameter of the model associated to the expected maximum duration of the reverberation phenomenon. The model is built as in [22], being the approximation due to the use of badn-to-band filters only. Later on, the values of \( n \) 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 $h_k[\tau] = |h_k[\tau]|e^{j\phi_k[\tau]}$. It is well known ([23]) 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 [16]) treating the $K \times N_k$ variables $\phi_k[\tau]$ as i.i.d. random variables with uniform distribution in $[-\pi, \pi)$. Denoting the complex conjugate by "$^\ast$" and the Kronecker delta by $\delta_{ij}$, the expected value of $|\tilde{x}_k[t]|^2$ is given by

$$E|\tilde{x}_k[n]|^2 = E \sum_{\tau,\tau'} s_k[n-\tau]s_k^*[n-\tau'] |h_k[\tau]| |h_k[\tau']|^2$$

$$= E \sum_{\tau,\tau'} s_k[n-\tau]s_k^*[n-\tau'] |h_k[\tau]| |h_k[\tau']| e^{j\phi_k[\tau']} e^{-j\phi_k[\tau']}$$

$$= \sum_{\tau,\tau'} s_k[n-\tau]s_k^*[n-\tau'] |h_k[\tau]| |h_k[\tau']| E e^{j(\phi_k[\tau]-\phi_k[\tau'])}$$

$$= \sum_{\tau,\tau'} s_k[n-\tau]s_k^*[n-\tau'] |h_k[\tau]| |h_k[\tau']| \delta_{\tau\tau'}$$

$$= \sum_{\tau} |s_k[n-\tau]|^2 |h_k[\tau]|^2.$$  

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 $S_k[n] = |s_k[n]|^2$, $H_k[n] = |h_k[n]|^2$ and $X_k[n] = E|\tilde{x}_k[n]|^2$. Then, our model reads

$$X_k[n] = \sum_{\tau} S_k[n-\tau]H_k[\tau],$$  

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

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

where $\epsilon_k[n]$ denotes the representation error. As shown in [16], this model is equivalent to a convolutive NMF ([21]) with diagonal basis. In the next section, we derive a cost function in order to find an appropriate convolutive representation that allows us to isolate the components $S_k[n]$.

2 A Bayesian approach

In the following, we will use a Bayesian approach to derive a cost function which we will then minimize in order to obtain our regularized solution. Let us begin by assuming, for every $k$, $\epsilon_k[n], S_k[n], H_k[n]$ are independent random variables, also independent with respect to $k$. Also, let us denote by $S, Y, X \in \mathbb{R}^{K \times N}$ and $H \in \mathbb{R}^{K \times N_k}$ the non-negative matrices whose $(k,n)$-th elements are $S_k[n], Y_k[n], X_k[n]$ and $H_k[n]$, respectively.
As it is customary ([16]), for the representation error, we assume \( \epsilon_k[n] \sim \mathcal{N}(0, \sigma^2) \), where \( \sigma > 0 \) is an unknown parameter, and the variables are non-correlated with respect to \( n \). Hence, it follows from ([1]) that the conditional distribution of \( Y \) given \( S \) and \( H \) (i.e. the likelihood distribution) is given by

\[
\pi_{\text{like}}(Y|S, H) = \prod_{k=1}^{K} \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(Y_k[n] - X_k[n])^2}{\sigma^2}\right).
\]

Let us now turn our attention to \( S \). Figure 1 depicts the log-spectrograms for a clean signal and its reverberant version. As it can be observed, while the spectrogram of the clean signal is somewhat sparse, the one corresponding to the reverberant signal presents a smoother or more diffuse structure. The presence of discontinuities in the spectrogram of the clean signal can be favored by assuming \( S \) follows a generalized Gaussian distribution ([25]). Namely,

\[
\pi_{\text{prior}}(S) = \prod_{k=1}^{K} \prod_{n=1}^{N} \frac{1}{2\Gamma(1 + 1/p)b_k} \exp\left(-\frac{|S_k[n]|^p}{b_k^p}\right),
\]

where \( p \in (0, 2) \) is a prescribed parameter and \( b_k > 0 \) is unknown.

In regards to \( H \), although no general conditions are expected on its individual components, we do expect its first order time differences to exhibit a certain degree of regularity (see Figures 2 and 3). In fact, if windows are set close enough relative to the duration of the reverberation phenomenon, then consecutive time components of \( H \) will capture overlapped information, which along with the exponential decay characteristic of the RIR ([26]) accounts for a somewhat smooth structure. Therefore, we define the time differences matrix \( V \in \mathbb{R}^{K \times (N_h-1)} \), with components \( V_k[n] = H_k[n] - H_k[n-1] \ \forall n = 1, \ldots, N_h - 1, \ k = 1, \ldots, K \). The regularity of

Figure 1: Spectrograms for a clean speech signal (left) and the corresponding reverberant speech signal (right).
these variations is contemplated by assuming $V$ follows a normal distribution:

$$
\pi_{\text{prior}}(V) = \prod_{k=1}^{K} \prod_{n=2}^{N_k} \frac{1}{\sqrt{2\pi\eta_k}} \exp \left( -\frac{V_k[n]^2}{\eta_k^2} \right).$$

Using Bayes’ theorem, the a posteriori joint distribution of $S$ and $H$ conditioned to $Y$ satisfies

$$
\pi_{\text{post}}(S, H | Y) \propto \pi_{\text{like}}(Y | S, H) \pi_{\text{prior}}(S) \pi_{\text{prior}}(H). \tag{5}
$$

### 2.1 Mixed penalization

Our goal is to find $\hat{S}$ and $\hat{H}$ that are representative of the a posteriori distribution $(5)$. Although the immediate instinct might be to compute the expected value, there are quite a few other ways to proceed, with different degrees of reliability and complexity. In lights of the assumed distributions and the high dimensionality of the problem, the maximum a posteriori (MAP) estimator is a reasonable choice in this case. Note that maximizing $(4)$ is tantamount to minimizing $-\log \pi_{\text{post}}(S, H | Y)$. If we denote by $S_k, Y_k, X_k \in \mathbb{R}^N$, $H_k \in \mathbb{R}^{N_h}$ and $V_k \in \mathbb{R}^{N_h-1}$ the (transposed) rows of $S, Y, X, H$ and $V$, and define $L \in \mathbb{R}^{N_h-1 \times N_h}$ in such a way that $LH_k = V_k$, then

$$
-\log \pi_{\text{post}}(S, H | Y) = \sum_{k=1}^{K} \left[ \frac{1}{\sigma^2} ||Y_k - X_k||_2^2 + \frac{1}{b_k^p} \sum_n |S_k[n]|^p + \frac{1}{\eta_k^2} ||LH_k||_2^2 \right] + C,
$$

where $C$ is a constant which does not depend on $S$ nor $H$.

Finally, the latter equation leads to the cost function

$$
J(S, H) \doteq \sum_{k=1}^{K} \left( ||Y_k - X_k||_2^2 + \lambda_{s,k} ||S_k||_p^p + \lambda_{h,k} ||LH_k||_2^2 \right), \tag{6}
$$
Figure 3: Subband signals $H_{65}[n], \ n = 1, \ldots, N$, with window length 256 and different over- 
lappings. The signals show certain regularity, that increases with the window over- 
lapping.

which shall be minimized to find our regularized solution. In this context, $\lambda_{s,k}, \lambda_{h,k} \geq 0$ can be 
thought of as penalization parameters weighting both penalizers relative to the fidelity term, 
whereas the exponent $p \in (0, 2)$ is a tuning parameter. It is timely to point out that small 
values of $p$ 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 $||S_k||_\infty = ||Y_k||_\infty \ \forall k$, which means that the maximum values shall remain 
equal for every frequency.

2.2 Regularization parameters

As mentioned before, the parameters $\lambda_{h,k}, \lambda_{s,k}, \ k = 1, \ldots, K$, 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 $\lambda_{h,k} = \lambda_h \sum_{n=1}^{N} |Y_k[n]|^2$ and $\lambda_{s,k} = \lambda_s \forall k = 1, \ldots, K$ (note that the relation between $S_k$ and $Y_k$ is already contemplated in the constraint that intends to avoid scale indeterminacy). This means we only need to look for two parameters ($\lambda_h, \lambda_s$) and then multiply $\lambda_h$ by the energy of the signal associated to each row of $Y$.

Next, we present an algorithm for approximating matrices $H$ and $S$ minimizing $J$.

### 3 Updating rules

We shall build an iterative algorithm following the idea in [16], 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

\begin{align}
(i) \quad g(w, w) &= f(w) \quad \text{and} \quad (ii) \quad g(w, w') \geq f(w), \quad \forall w, w' \in \Omega.
\end{align}

Let $w^0 \in \Omega$ be arbitrary, and let

\begin{align}
w^j = \arg \min_w g(w, w^{j-1}).
\end{align}

With this definition, it can be shown ([27]) that the sequence $\{f(w^j)\}_j$ is non-increasing. We intend to use this property as a tool for alternatively updating the matrices $H$ and $S$. Let us begin by fixing $H = H'$, where $H'$ is an arbitrary $K \times N_h$ matrix. We will show that

\begin{align}
g_s(S, S') &= \sum_{k,n,\tau} S_k'[\tau] H_k'[n-\tau] \left( Y_k[n] - \frac{S_k[\tau]}{S_k'[\tau]} X_k'[n] \right)^2 + \sum_{k} \lambda_{h,k} ||LH_k'||_2^2
\end{align}

\begin{align}
+ \sum_{k,n} \lambda_{s,k} \left( \frac{p}{2} S_k'[n]^{p-2} S_k[n]^2 + |S_k'[n]|^p - \frac{p}{2} |S_k[n]|^p \right)
\end{align}

is an auxiliary function for $J$ (as defined in [3]) with respect to $S$. From this point on, we denote by $X_k'[n] = \sum_{\tau} S_k'[n - \tau] H_k'[\tau]$. The equality condition $(i)$ in (7) is rather straightforward. In
\[
g_s(S, S) = \sum_{k,n,\tau} S_k[\tau]H_k'[n-\tau] \left( Y_k[n] - \sum_{\nu} S_k[\nu]H_k'[n-\nu] \right)^2 + \sum_k \lambda_{h,k} ||LH_k'||_2^2 + \sum_{k,n} \lambda_{s,k} |S_k[n]|^p \\
= \sum_{k,n,\tau} \left( Y_k[n] - \sum_{\nu} S_k[\nu]H_k'[n-\nu] \right)^2 + \sum_k \lambda_{h,k} ||LH_k'||_2^2 + \sum_{k,n} \lambda_{s,k} |S_k[n]|^p \\
= J(S, H').
\]

To prove condition (ii) in (7) we begin by defining
\[
P_{k,n} = \sum_{\tau} \frac{S_k'[\tau]H_k'[n-\tau]}{X_k'[n]} \left( Y_k[n] - \frac{S_k[\tau]}{S_k'[\tau]} X_k'[n] \right)^2,
\]
\[
R_{k,n} = (Y_k[n] - \sum_{\tau} S_k[\tau]H_k'[n-\tau])^2,
\]
and \( Q : \mathbb{R}^+ \to \mathbb{R} \) such that \( Q(x) = \frac{x^p}{2}x^p - \frac{x^p}{2}x^p \). With these definitions, we can write
\[
g_s(S, S') = \sum_k \left( \sum_n (P_{k,n} + \lambda_{s,k} Q(S_k'[n])) + \lambda_{h,k} ||LH_k'||_2^2 \right),
\]
and
\[
J(S, H') = \sum_k \left( \sum_n (R_{k,n} + \lambda_{s,k} |S_k[n]|^p) + \lambda_{h,k} ||LH_k'||_2^2 \right).
\]
Hence, to prove that \( g_s(S, S') \geq J(S, H') \forall S, S' \) it is sufficient to show that \( P_{k,n} \geq R_{k,n} \) and
To prove that \( Q(S'_k[n]) \geq |S_k[n]|^p \) \( \forall n = 1, \ldots, N, k = 1, \ldots, K \). In fact,

\[
P_{k,n} - R_{k,n} = \sum_{\tau} \frac{S'_k[\tau]H'_k[n - \tau]}{X'_k[n]} \left( Y_k[n] - \frac{S_k[\tau]}{S'_k[\tau]} X'_k[n] \right)^2 \]

\[
- (Y_k[n] - \sum_{\tau} S_k[\tau] H'_k[n - \tau])^2
\]

\[
= \sum_{\tau} \frac{H'_k[n - \tau] S_k[\tau]^2 X'_k[n]}{S'_k[\tau]} - \left( \sum_{\tau} S_k[\tau] H'_k[n - \tau] \right)^2
\]

\[
= \sum_{\tau, \nu} \frac{H'_k[n - \tau] S_k[\tau]^2 H'_k[n - \nu] S'_k[\nu]}{S'_k[\tau]} - \sum_{\tau, \nu} S_k[\tau] H'_k[n - \tau] S_k[\nu] H'_k[n - \nu]
\]

\[
= \sum_{\tau < \nu} \frac{H'_k[n - \tau] H'_k[n - \nu]}{S'_k[\tau] S'_k[\nu]} \left( \frac{S_k[\tau]^2 S'_k[\nu]}{S'_k[\tau]} - 2 S_k[\tau] S_k[\nu] + S_k[\nu]^2 S'_k[\tau] \right) + \sum_{\tau < \nu} \frac{H'_k[n - \tau] H'_k[n - \nu]}{S'_k[\nu] S'_k[\tau]} (S_k[\tau] S'_k[\nu] - S_k[\nu] S'_k[\tau])^2 \geq 0.
\]

To prove that \( Q(S'_k[n]) \geq |S_k[n]|^p \), we begin by noting that \( Q \in C^\infty(\mathbb{R}^+) \). Then, the first order necessary condition for \( Q \) yields

\[
0 = \frac{\partial Q}{\partial x} = \frac{p(p - 2)}{2} x^{p - 3} S_k[n]^2 + p x^{p - 1} - \frac{p^2}{2} x^{p - 1} - \frac{p(p - 2)}{2} x^{p - 1} (x^{-2} S_k[n]^2 - 1),
\]

meaning the only point at which the derivative of \( Q \) equals zero is at \( x = S_k[n] \). Furthermore, \( \frac{\partial^2 Q}{\partial x^2} Q(S_k[n]) = S_k[n]^{p - 2} (2p - p^2) > 0 \) \( \forall p \in (0, 2) \), meaning that \( Q(S_k[n]) = |S_k[n]|^p \) is the global minimum of \( Q \). This yields

\[
g_s(S, S') = \sum_k \left( \sum_n \left( P_{k,n} + \lambda_{s,k} Q(S'_k[n]) \right) + \lambda_{h,k} ||LH'_k||_2^2 \right)
\]

\[
\geq \sum_k \left( \sum_n \left( R_{k,n} + \lambda_{s,k} |S_k[n]|^p \right) + \lambda_{h,k} ||LH'_k||_2^2 \right) = J(S, H').
\]

In an analogous way, it can be shown that if we let \( S = S' \) be fixed, where \( S' \) is an arbitrary
$K \times N$ matrix, then

$$g_h(H, H') = \sum_{k,n,\tau} S_k'[n - \tau] H_k'[\tau] \left( \frac{Y_k[n] - H_k'[\tau]}{H_k'[\tau]} X_k'[n] \right)^2 + \sum_{k} \lambda_{s,k} ||S_k'||_p + \sum_{k} \lambda_{h,k} ||LH||_2^2$$

is an auxiliary function for $J(S', H)$ with respect to $H$.

Having defined auxiliary functions, we will use the updating rule derived from (8) to build an algorithm for iteratively approaching matrices $S$ and $H$ minimizing $J$. Notice this requires minimizing $g_s$ and $g_h$ with respect to the updating variables, but since $g_s$ is quadratic with respect to $S$ and $g_h$ is quadratic with respect to $H$, we can simply use the first order necessary conditions in both cases. From this point on, in the context of the iterative updating process, $S'$ and $H'$ will refer not to arbitrary nonnegative matrices, but to those estimations of $S$ and $H$ obtained in the immediately previous step.

### 3.1 Updating rule for $S$

Firstly, we shall derive an updating rule for $S_k[\tau]$. That is, we wish to minimize $g_s$ w.r.t. $S$. The first order necessary condition on $g_s$ yields

$$0 = \frac{\partial g_s(S, S')}{\partial S_k[\tau]}$$

$$= -2 \sum_n H_k'[n - \tau] Y_k[n] + \frac{S_k[\tau]}{S_k'[\tau]} \sum_n H_k'[n - \tau] X_k'[n] + \frac{\lambda_{s,k}}{2} p S_k'[\tau]^{p-1} S_k[\tau]$$

$$= -S_k'[\tau] \sum_n H_k'[n - \tau] Y_k[n] + \left( \sum_n H_k'[n - \tau] X_k'[n] + \frac{\lambda_{s,k}}{2} p S_k'[\tau]^{p-1} \right) S_k[\tau],$$

which easily leads to the multiplicative updating rule

$$S_k[\tau] = S_k'[\tau] \frac{\sum_n H_k'[n - \tau] Y_k[n]}{\sum_n H_k'[n - \tau] X_k'[n] + \frac{\lambda_{s,k}}{2} p |S_k'[\tau]|^{p-1}}.$$

In order to avoid the aforementioned scale indeterminacy, every updating step is to be followed by scaling $S_k$ so that its $\ell^\infty$ norm coincides with that of the observation $Y_k$. 

11
3.2 Updating rule for H

In order to find an updating rule for \( H \), we shall write \( g_h \) as a function of the transposed rows \( H_k \). We begin by noting

\[
g_h(H, H') = \sum_{k,n,\tau} S_k'[n - \tau] H_k'[\tau] X_k'[n] \left( Y_k[n] - \frac{H_k[\tau]}{H_k'[\tau]} X_k'[n] \right)^2 \\
+ \sum_k \lambda_s,k||S_k'||_p^p + \sum_k \lambda_{h,k}||LH_k||_2^2 \\
= \sum_{k,n,\tau} S_k'[n - \tau] H_k'[\tau] Y_k^2[n] - 2 \sum_{k,n,\tau} S_k'[n - \tau] Y_k[n] H_k[\tau] \\
+ \sum_{k,n,\tau} \frac{S_k'[n - \tau] X_k'[n] H_k^2[\tau]}{H_k'[\tau]} \\
+ \sum_k \lambda_s,k||S_k'||_p^p + \sum_k \lambda_{h,k}||LH_k||_2^2.
\]

Next, we define the diagonal matrices \( A^k, B^k \in \mathbb{R}^{N_h \times N_h} \), whose diagonal elements are \( A^k_{\tau,\tau} = \sum_n S_k'[n - \tau] X_k'[n] \) and \( B^k_{\tau,\tau} = H_k'[\tau] \), and the vector \( \zeta^k \in \mathbb{R}^{N_h} \) with components \( \zeta^k_\tau = \sum_n S_k'[n - \tau] Y_k[n] \). With these definitions, we can write

\[
g_h(H, H') = \sum_{k,n,\tau} S_k'[n - \tau] H_k'[\tau] X_k'[n] \left( Y_k[n] - \frac{H_k[\tau]}{H_k'[\tau]} X_k'[n] \right)^2 \\
- 2 \sum_k H_k^T \zeta^k + \sum_k H_k^T A^k \left( B^k \right)^{-1} H_k \\
+ \sum_k \lambda_s,k||S_k'||_p^p + \sum_k \lambda_{h,k} H_k^T L^T LH_k.
\]

Now, the first order necessary condition for \( g_h \) with respect to \( H_k \) is given by

\[
0 = \frac{\partial g_h(H, H')}{\partial H_k} = -2\zeta^k + 2A^k \left( B^k \right)^{-1} H_k + 2\lambda_{h,k} L^T LH_k,
\]

which readily leads to an updating rule consisting of solving the linear system

\[
(A^k + \lambda_{h,k} L^T L) H_k = B^k \zeta^k.
\]

Let us notice that under the assumption that the diagonal elements of \( A^k \) and \( B^k \) are strictly positive, and since \( L^T L \) is positive-semidefinite, \( (B^k)^{-1} A^k + \lambda_{h,k} L^T L \) is positive-definite, and hence the linear system has a unique solution. The assumption of \( A^k_{\tau,\tau} > 0 \) is adequate, since these elements correspond to the discrete convolution of \( S_k' \) and \( X_k' \). Although the validity of the hypothesis over \( B^k_{\tau,\tau} \) is not so clear, in practice, the matrix in system (11) has turned out to be non-singular. Nonetheless, \( H_k \) can be computed as the best approximate solution in the
least-squares sense. Then, solving this $N_h \times N_h$ linear system entails no challenge, since $N_h$ is usually chosen relatively small, depending on the window step and the reverberation time.

All the steps for the dereverberation process are stated in Algorithm 1. Note that in the initialization we define the clean spectrogram $S$ equal to the observation, which is natural since in a way they both correspond to the same signal, and $H_k$ as a vector with exponential time decay, which is an expected characteristic of a RIR. Finally, we set the stopping criterion over the decay of the norm of two consecutive approximations of $S$. This has shown to work quite well, although other stopping criteria might be considered.

Results to illustrate the performance of the algorithm are presented in the next section.

4 Experimental results

For the experiments, we took 110 speech signals from the TIMIT database (28), recorded at 16 KHz, and artificially made them reverberant by convolution with impulse responses generated with the software Room Impulse Response Generator1, based on the model in 29. Each signal was degraded under different reverberation conditions: three different room sizes, each with three different microphone positions and four different reverberation times2.

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 $\lambda_h = 1$ and $\lambda_s = 10^{-4}$. The rest of the model parameters were chosen as specified in Table 1.

| $p$ | $N_h$ | win. window size | win. overlap. | $\delta$ | max. iter. |
|-----|-------|------------------|---------------|---------|------------|
| 1   | 15    | Hann 512 samples | 256 samples   | $\|Y\|_F \times 10^{-3}$ | 20         |

Table 1: Model parameter values

Let us point out that the choice of $N_h$ was done as to allow $H$ to capture early reverberation while precluding overlapped representations. In the first place, it is desirable for $H$ to represent the RIR along the full Early Decay Time (EDT), the time period in which the reverberation phenomenon alters the clean signal the most, so its effect can be effectively nullified. On the other hand, if we were to choose $N_h$ too large, it might lead certain similarities in the observation $Y$ within a fixed frequency range to be represented as echoes from high energy components of

1https://github.com/ehabets/RIR-Generator
2A web demo for our algorithm can be found in http://fich.unl.edu.ar/sinc/web-demo/blindder/
Algorithm 1 Mixed penalization dereverberation

1: Initializing
2: \( S \leftarrow Y \)
3: \( H_k[n] \leftarrow \exp(-n) \quad \forall k = 1 \ldots K, n = 1 \ldots N \)
4: MAIN LOOP
5: for \( i = 1 \ldots \text{maxiter} \)
6: \( S' = S \).
7: \( X_k[n] \leftarrow \sum_{\tau} S_k[n - \tau] H_k[\tau] \quad \forall k = 1 \ldots K, n = 1 \ldots N \)
8: for \( k = 1 \ldots K \)
9: for \( \tau = 1 \ldots N \)
10: \( S_k[\tau] \leftarrow S_k[\tau] + \frac{\sum_n H_k[n - \tau] Y_k[n]}{\sum_n H_k[n - \tau] X_k[n] + \frac{\lambda_{h,k}}{2} p \| S_k[\tau] \|^{p-1}} \).
11: end for
12: \( S_k \leftarrow S_k \| Y_k \|_{\infty} \).
13: end for
14: for \( k = 1 \ldots K \)
15: Build the diagonal matrices \( A^k, B^k \in \mathbb{R}^{N_h \times N_h} \) :
16: \( A_{\tau,\tau}^k = \sum_n S_k[n - \tau] X_k[n] \),
17: \( B_{\tau,\tau}^k = H_k[\tau] \).
18: Build the vector \( \zeta^k \) :
19: \( \zeta^k_{\tau} = \sum_n S_k[n - \tau] Y_k[n] \)
20: Solve for \( H_k \) :
21: \( (A^k + \lambda_{h,k} B^k L^T L) H_k = B^k \zeta^k \).
22: end for
23: if \( \| S - S' \|_F \leq \delta \)
24: return
25: end if
26: end for
S. It is worth mentioning, however, that the performance of our dereverberation method has shown no high sensitivity with respect to the choice of $N_h$.

In order to evaluate the performance of our model, we made comparisons against two state of the art methods that work under the same conditions. The one proposed by Kameoka et al in [16], choosing all the parameters as suggested, and the one proposed by Wisdom et al in [12], with a window length of 2048.

To measure performance, following [30], we made use of the frequency weighted segmental signal-to-noise ratio (fwsSNR) and cepstral distance. Furthermore, we also measured the speech-to-reverberation modulation energy ratio (SRMR, [31]), which has the advantage of being non-intrusive (it does not use the clean signal as an input). The results for each performance measure are stated in Tables 2-4 and depicted in Figures 4-6, classified in function of the reverberation times: 300[ms], 450[ms], 600[ms] and 750[ms]. Notice that for the cases of fwsSNR and SRMR, higher values correspond to better performance, while for the cepstral distance, small values indicate higher quality.

![fwsSNR](image.png)

Figure 4: Mean and standard deviations of performance fwsSNR for different reverberation times.

In regard to the fwsSNR performance measure, the values in Table 2 (Figure 4) give account of significant improvements of our proposed method with respect to the other two. This improvement becomes more evident as the reverberation time increases. As for the cepstral distance, although the results in Table 3 (Figure 5) account for a better performance of our proposed method, the quality with respect to the reverberant signal is improved only for reverberation times of 450[ms] or more. Finally, the SRMR also shows an improvement with respect
Table 2: Mean and (standard deviation) of fwsSNR for each method and reverberation time (best results in boldface).

| Rev. time [ms] | Rev. Signal | Kameoka der. | Wisdom der. | Mixed penalization |
|----------------|-------------|--------------|--------------|--------------------|
| 300            | 8.102 (1.96)| 7.950 (1.73) | 8.262 (1.53) | **8.658 (1.59)**   |
| 450            | 4.815 (1.42)| 5.127 (1.36) | 5.771 (1.28) | **6.539 (1.56)**   |
| 600            | 3.082 (1.20)| 3.358 (1.19) | 4.140 (1.17) | **4.732 (1.43)**   |
| 750            | 1.998 (1.11)| 2.184 (1.10) | 3.013 (1.12) | **3.440 (1.31)**   |

Figure 5: Mean and standard deviations of cepstral distance for different reverberation times.

to the other methods for reverberation times of 450[ms] or greater (see Table 4, Figure 6).

5 Conclusions

In this work, a new blind dereverberation method for speech signals based on regularization over a convolutive NMF representation of the signal spectrograms was introduced and tested. Results show a significant improvement over the state of the art methods, specially for high reverberation times. There is certainly much room for improvement, e.g. finding ways of optimally choosing the regularization parameters, exploring the use of other penalizers, etc.
Table 3: Mean and (standard deviation) of cepstral distance for each method and reverberation time (best results in boldface).

| Rev. time [ms] | Rev. Signal | Kameoka der. | Wisdom der. | Mixed penalization |
|----------------|-------------|--------------|--------------|--------------------|
| 300            | 3.440 (0.44)| 4.057 (0.45) | 3.908 (0.48) | 3.521 (0.35)       |
| 450            | 4.264 (0.44)| 4.636 (0.42) | 4.511 (0.41) | 3.985 (0.39)       |
| 600            | 4.716 (0.46)| 5.006 (0.42) | 4.860 (0.40) | 4.370 (0.40)       |
| 750            | 5.011 (0.48)| 5.264 (0.43) | 5.089 (0.40) | 4.657 (0.41)       |

Figure 6: Mean and standard deviations of SRMR for different reverberation times.

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, 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”.

17
Table 4: Mean and (standard deviation) of SRMR for each method and reverberation time (best results in boldface).

References

[1] M. Kim, H.-M. Park, Efficient online target speech extraction using DOA-constrained independent component analysis of stereo data for robust speech recognition, Signal Processing 117 (2015) 126–137.

[2] S. Yun, Y. J. Lee, S. H. Kim, Multilingual speech-to-speech translation system for mobile consumer devices, IEEE Transactions on Consumer Electronics 60 (3) (2014) 508–516.

[3] R. Neßelrath, M. M. Moniri, M. Feld, Combining speech, gaze, and micro-gestures for the multimodal control of in-car functions, in: Intelligent Environments (IE), 2016 12th International Conference on, IEEE, 2016, pp. 190–193.

[4] L. Di Persia, D. Milone, H. L. Rufiner, M. Yanagida, Perceptual evaluation of blind source separation for robust speech recognition, Signal Processing 88 (10) (2008) 2578–2583.

[5] C. E. Martnez, J. Goddard, L. E. Di Persia, D. H. Milone, H. L. Rufiner, Denoising sound signals in a bioinspired non-negative spectro-temporal domain, Digital Signal Processing 38 (2015) 22–31.

[6] L. Di Persia, D. Milone, M. Yanagida, Indeterminacy free frequency-domain blind separation of reverberant audio sources., IEEE Trans. Audio, Speech and Lang. Proc. 17 (2) (2009) 299–311.

[7] L. E. Di Persia, D. H. Milone, Using multiple frequency bins for stabilization of FD-ICA algorithms, Signal Processing 119 (2016) 162–168.

[8] A. Tsilfidis, J. Mourjopoulos, Signal-dependent constraints for perceptually motivated suppression of late reverberation, Signal Processing 90 (3) (2010) 959–965.

[9] I. J. Tashev, Sound Capture and Processing, John Wiley & Sons, Ltd, 2009.
[10] X. Huang, A. Acero, H.-W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, 1st Edition, Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001.

[11] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, et al., Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge, in: REVERB Workshop, 2014.

[12] S. Wisdom, T. Powers, L. Atlas, J. Pitton, Enhancement of reverberant and noisy speech by extending its coherence, in: Proceedings of REVERB Challenge Workshop, Florence, Italy, 2014, pp. 1–8.

[13] M. Moshirynia, F. Razzazi, A. Haghbin, A speech dereverberation method using adaptive sparse dictionary learning, in: Proceedings of REVERB Challenge Workshop, p1, Vol. 2, 2014.

[14] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E.-S. Chng, H. Li, The ntu-adsc systems for reverberation challenge 2014, in: Proc. REVERB challenge workshop, 2014.

[15] K. Nathwani, R. M. Hegde, Joint source separation and dereverberation using constrained spectral divergence optimization, Signal Processing 106 (2015) 266–281.

[16] H. Kameoka, T. Nakatani, T. Yoshioka, Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms, in: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, pp. 45–48.

[17] F. J. Ibarrola, G. L. Mazzieri, R. D. Spies, K. G. Temperini, Anisotropic bv?l2 regularization of linear inverse ill-posed problems, Journal of Mathematical Analysis and Applications 450 (1) (2017) 427 – 443.

[18] D. Lazzaro, L. B. Montefusco, S. Papi, Blind cluster structured sparse signal recovery: A nonconvex approach, Signal Processing 109 (2015) 212–225.

[19] F. J. Ibarrola, R. D. Spies, A two-step mixed inpainting method with curvature-based anisotropy and spatial adaptivity, Inverse Problems and Imaging 11 (2) (2017) 247–262.

[20] V. Peterson, H. L. Rufiner, R. D. Spies, Generalized sparse discriminant analysis for event-related potential classification, Biomedical Signal Processing and Control 35 (2017) 70–78.
[21] G. L. Mazzieri, R. D. Spies, K. G. Temperini, Mixed spatially varying l2-bv regularization of inverse ill-posed problems, Journal of Inverse and Ill-posed Problems 23 (6) (2015) 571–585.

[22] Y. Avargel, I. Cohen, System identification in the short-time fourier transform domain with crossband filtering, IEEE Transactions on Audio, Speech, and Language Processing 15 (4) (2007) 1305–1319.

[23] B. Yegnarayana, P. S. Murthy, C. Avendaño, H. Hermansky, Enhancement of reverberant speech using lp residual, in: Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, Vol. 1, IEEE, 1998, pp. 405–408.

[24] P. Smaragdis, Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs, Independent Component Analysis and Blind Signal Separation (2004) 494–499.

[25] C. Bouman, K. Sauer, A generalized gaussian image model for edge-preserving map estimation, IEEE Transactions on Image Processing 2 (3) (1993) 296–310.

[26] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien Jr, C. R. Lansing, A. S. Feng, Blind estimation of reverberation time, The Journal of the Acoustical Society of America 114 (5) (2003) 2877–2892.

[27] D. D. Lee, H. S. Seung, Algorithms for non-negative matrix factorization, in: Advances in neural information processing systems, 2001, pp. 556–562.

[28] V. Zue, S. Seneff, J. Glass, Speech database development at mit: Timit and beyond, Speech Communication 9 (4) (1990) 351–356.

[29] J. B. Allen, D. A. Berkley, Image method for efficiently simulating small-room acoustics, The Journal of the Acoustical Society of America 65 (4) (1979) 943–950.

[30] Y. Hu, P. C. Loizou, Evaluation of objective quality measures for speech enhancement, IEEE Transactions on audio, speech, and language processing 16 (1) (2008) 229–238.

[31] T. H. Falk, C. Zheng, W.-Y. Chan, A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech, IEEE Transactions on Audio, Speech, and Language Processing 18 (7) (2010) 1766–1774.