# Switching divergences for spectral learning in blind speech dereverberation

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

September 19, 2018

**Abstract**

When recorded in an enclosed room, a sound signal will most certainly get affected by reverberation. This not only undermines audio quality, but also poses a problem for many human-machine interaction technologies that use speech as their input. In this work, a new blind, two-stage dereverberation approach based in a generalized $\beta$-divergence as a fidelity term over a non-negative representation is proposed. The first stage consists of learning the spectral structure of the signal solely from the observed spectrogram, while the second stage is devoted to model reverberation. Both steps are taken by minimizing a cost function in which the aim is put either in constructing a dictionary or a good representation by changing the divergence involved. In addition, an approach for finding an optimal fidelity parameter for dictionary learning is proposed. An algorithm for implementing the proposed method is described and tested against state-of-the-art methods. Results show improvements for both artificial reverberation and real recordings.

**Keywords**
signal processing, dereverberation, penalization



## 1 Introduction

Over the last years, with the technological advances and massive adoption of portable electronic devices with high computational capacity, the need for better human-machine interaction capabilities has emerged as a topic of interest. Since speech constitutes one of the most natural ways of human communication, trying to achieve a fluid interaction with machines by this mean has been the subject of much recent research. This need for improvement is inherent to a number of hot topics in the field of signal processing, including automatic translation systems ([1]), emotion and affective state recognition ([2]), digital personal assistants ([3]), to name just a few, that require the use of speech as inputs.

One of the main difficulties within this context comes from the fact that when recorded in enclosed rooms, audio signals are affected by reverberant components due to reflections of the sound waves in the walls, floor and ceiling. This can severely degrade the quality of the recorded signals (particularly when the microphones are far away from the sources, [4]), which in turn makes them unsuitable for direct use in certain speech applications ([5]). The goal of this work is to produce a dereverberation technique for removing or highly attenuating the reverberant components of a recorded signal in order to enhance its quality.

A speech dereverberation problem can be classified as "blind" whenever the available data consist only of the reverberant signal itself, or as "supervised" when information of the environment or the speakers is available. The problem can also be classified as single or multi-channel, depending on the number of microphones used for recording. In this work, we shall address the problem within a blind, single-channel setting, which is the most common in real-life problems, but also the most difficult, because of the scarce information.

Due to the characteristics of speech signals, most state-of-the-art methods deal with the dereverberation problem in a transformed domain, such as the one obtained by the Fan-Chirp Transform (see [6]) or the Short-Time Fourier Transform (STFT) ([7]). Some of these methods make use of non-negative matrix factorization (NMF) or its variants, such as convolutive NMF ([8]), along with Bayesian or penalization approaches. Although

*Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), UNL, CONICET, FICH, Ciudad Universitaria, CC 217, Ruta Nac. 168, km 472.4, (3000) Santa Fe, Argentina. (fibarrola@sinc.unl.edu.ar).

†Instituto de Matemática Aplicada del Litoral, IMAL, UNL, CONICET, 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.

such methods have shown to produce satisfactory results, they often neglect the relation between frequency components, for which some authors (e.g. [9]) have proposed an NMF model in which a *dictionary* is used for spectral modeling. The main problem with this kind of models within a blind setting has to do with the scarce available data. That is, the dictionary should be good for representing a clean signal, while learnt from a reverberant one.

This article begins by presenting a convolutive NMF reverberation representation that uses a dictionary for spectral modeling, and proposing a general form for a cost function with mixed penalization for characterizing the model. Different variants of that cost function are used for stating a two-stage method, where the first stage takes care of building a dictionary, while the second one is devoted to use such dictionary for getting an appropriate representation of the reverberation model. The main novelty of this work is that the process of learning the spectral structure (*i.e.* the first stage) is not aimed to obtain an optimal representation of the reverberant signal.

# 2 Reverberation Model

Let $s, x, h : \mathbb{R} \to \mathbb{R}$, supported in $[0, \infty)$, denote the functions associated to the clean and reverberant signals, and the room impulse response (RIR), respectively. As it is customary, we make the assumption that reverberation is well represented by a Linear Time-Invariant (LTI) system, which can be written as

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

where "$*$" denotes convolution. The use of this representation is underlaid by the hypotheses that the source and microphone positions are fixed, and the non-linear components are small enough to be neglected.

As we previously mentioned, when dealing with speech signals, it often results convenient to work with time-frequency representations rather than in the time domain. Thus, we shall make use of the Short Time Fourier Transform (STFT).

## 2.1 STFT-based reverberation model

The STFT of a function $x$ can be 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 prescribed even and compactly supported function such that $\|w\|_1 = 1$, called *window*.

Naturally, in practice we work with discretized versions of the signals, denoted as $x[\cdot]$, $h[\cdot]$, $s[\cdot]$, and $w[\cdot]$. The corresponding discrete STFT can be defined as

$$\mathbf{x}_k[n] \doteq \sum_{m=-\infty}^{\infty} x[m]w[m-n]e^{-2\pi imk},$$

where $n = 1, \ldots, N$, is a discrete time variable associated to the window locations, and $k = 1, \ldots, K$, denotes the frequency sub-band. Similarly, we denote by $\mathbf{s}_k[n]$ and $\mathbf{h}_k[n]$ the STFTs of $s$ and $h$, respectively. A discrete approximation of (1) in the STFT domain is given by

$$\mathbf{x}_k[n] \approx \tilde{\mathbf{x}}_k[n] \doteq \sum_{m=0}^{M-1} \mathbf{s}_k[n-m]\mathbf{h}_k[m], \quad n, k \in \mathbb{N}. \tag{2}$$

where $M$ is a given model parameter determined by the reverberation time. The model is built as in [10], where the approximation in (2) holds due to the use of band-to-band only filters. The window locations are chosen so that the support of the observed signal is contained in the union of the supports of the windows, and $K$ as to reach up to half the sampling frequency.

Since phase angles on the STFT components have been shown to be highly sensitive to mild variations on the associated signal ([11]), and within our blind setting we have no information about reverberation conditions, we proceed as in [12], by treating the phase angles $\phi_k[m]$ of $\mathbf{h}_k[m]$ as random variables. Let us assume them to be *i.i.d.* with uniform distribution in $[-\pi, \pi)$. Under this hypothesis, it can be shown ([7]) that the expected value of $|\tilde{\mathbf{x}}_k[t]|^2$ is given by

$$E|\tilde{\mathbf{x}}_k[n]|^2 = \sum_m |\mathbf{s}_k[n-m]|^2 \, |\mathbf{h}_k[m]|^2.$$

Note that the choice of $[-\pi, \pi)$ is arbitrary, since the equality holds for any $2\pi-$length interval. Finally, by defining $S_{k,n} \doteq |\mathbf{s}_k[n]|^2$, $H_{k,n} \doteq |\mathbf{h}_k[n]|^2$ and $X_{k,n} \doteq E|\tilde{\mathbf{x}}_k[n]|^2$, the convolutive NMF model reads

$$X_{k,n} = \sum_{m=0}^{M'} S_{k,n-m} H_{k,m}, \tag{3}$$

for $k = 1, \ldots, K,\ n = 1, \ldots, N$. Here, $M' \doteq \min\{M-1, n-1\}$, so we can treat $X$, $S$ and $H$ as nonnegative matrices with elements $X_{k,n}$, $S_{k,n}$ and $H_{k,n}$, respectively.

Since we intend to introduce a spectral modeling of the clean signal, we shall make use of an NMF approach over the clean spectrogram $S$.

### 2.2 NMF model

Let us assume that there exist $W \in \mathbb{R}_{0,+}^{K \times J}$, $\quad U \in \mathbb{R}_{0,+}^{J \times N}$, $(J < \min\{K, N\})$ that provide a "good" NMF representation for $S \in \mathbb{R}_{0,+}^{K \times N}$. That is,

$$S \cong WU.$$

The accuracy of this approximation can be defined in terms of the Euclidean distance or some divergence measure (details on this will be discussed later on). In order to keep the notation simple, we shall assume the latter approximation to hold exactly and replace $S$ in (3) by $WU$, which results in the model

$$X_{k,n} = \sum_{m=0}^{M'} \sum_{j=1}^{J} W_{k,j} U_{j,n-m} H_{k,m}. \tag{4}$$

Two remarks are in order: firstly, note that the approximation error in the assumption $S = WU$ will be taken into account by the representation error of $X$ with respect to the data, and hence the latter assumption poses no problem. Secondly, we note that the model (4) has a scale indeterminacy, in the sense that for any $\alpha > 0$, the matrices $\tilde{W} = \alpha W$, $\tilde{H} = \alpha H$, and $\tilde{U} = \alpha^{-2} U$ would give the same representation $X$. Hence, in order to avoid numerical issues, we add the constraints $\|W_j\|_1 = \|H_k^T\|_\infty = 1$, where $W_j,\ j = 1, \ldots, J$, are the columns of $W$ and $H_k,\ k = 1, \ldots, K$ are the rows of $H$. This means that the spectrogram $S$ is represented by a normalized dictionary and that reverberation preserves the signal's maximal energy.

In the next section, a fidelity term and penalizers for building an appropriate cost function $f$ will be defined. This cost function will then be minimized in order to obtain the desired matrices $\hat{W}$, $\hat{U}$ and $\hat{H}$, as follows:

*Algorithm overview*

1. Set the parameters of $f = f(Y, X)$ so as to prioritize spectral learning and minimize $f$ with respect to its arguments in order to find an appropriate dictionary $\hat{W}$.
2. Reset the parameters of $f$ in order to emphasize accuracy in the representation. Then minimize $f$ with respect to $U$ and $H$ subject to $W = \hat{W}$, to obtain $\hat{U}$ and $\hat{H}$.
3. Approximate the clean spectrogram $S$ using $\hat{W}$ and $\hat{U}$.

## 3 Cost function

### 3.1 Fidelity term

Given a reverberant (and possibly noisy) spectrogram $Y$, we intend to find matrices $W$, $U$ and $H$ that, while complying with certain desired characteristics, provide a representation $X$, as in (4), that accurately approximates $Y$.

Many ways of measuring the fidelity of that approximation have been proposed: the Euclidean distance ([12]), the Kullback-Leibler divergence ([9]), and the Itakura-Saito divergence ([13]) being the most commonly used. Assume we have a known clean spectrogram $S$ that we want to represent using an NMF factorization $WU$. Different choices of the fidelity measure will lead to dictionary atoms (column vectors of $W$) with different characteristics. As it can be seen in Fig. 1, a particular fidelity measure may emphasize the appearance of atoms that enable a good approximation in the higher energy zones while neglecting the low-energy ones, while another fidelity measure may result in the opposite.

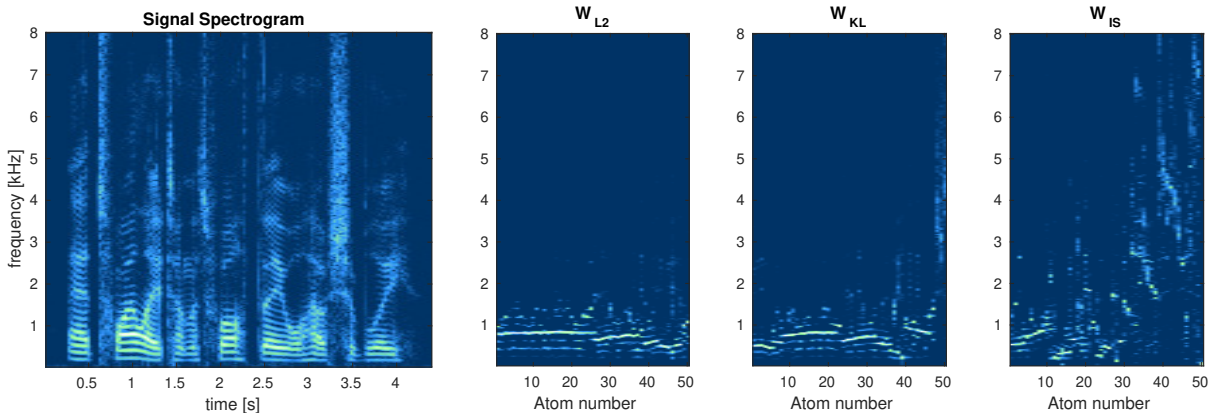


Figure 1: Left: The spectrogram of a clean signal, sampled at 16[kHz], using a 512 samples window with overlapping of 256. $W_{\mathrm{L2}}$: dictionary obtained using Frobenius norm. $W_{\mathrm{KL}}$: dictionary obtained using Kullback-Leibler divergence. $W_{\mathrm{IS}}$: dictionary obtained using Itakura-Saito divergence. All the dictionary atoms were ordered by correlation in order to help visualization.

In order to find an "optimal" dictionary $W$, we begin by recalling a generalized divergence, as introduced in [14]. For $X, Y \in \mathbb{R}_{0,+}^{K\times N}$ and $\beta \in \mathbb{R}_+ \backslash \{1\}$, the $\beta$-divergence of $X$ from $Y$ is defined as

$$D_\beta(Y||X) \doteq \sum_{k,n} \left( Y_{k,n} \frac{Y_{k,n}^{\beta-1} - X_{k,n}^{\beta-1}}{\beta(\beta-1)} + X_{k,n}^{\beta-1} \frac{X_{k,n} - Y_{k,n}}{\beta} \right).$$

This $\beta$-divergence generalizes all three aforementioned fidelity measures. In fact, it can be seen that $D_2(\cdot||\cdot)$ corresponds to (half) the squared Frobenius norm of $Y - X$, whereas $D_\beta(\cdot||\cdot)$ approaches the Kullback-Leibler divergence as $\beta \to 1$ and the Itakura-Saito divergence as $\beta \to 0$. An appropriate way of choosing the parameter $\beta$ will be discussed later on. We now proceed to introduce the penalization terms which shall embed the desired characteristics on the components that constitute the model.

## 3.2 Penalizers

Clearly, there are many ways of building the matrices $W, U$ and $H$ leading to a representation with small divergence with respect to the observation. One way of narrowing down the possible choices is by introducing penalizing terms into our cost function for promoting certain desired features over its minimizers. In a quite general context, this leads to a cost function of the form

$$f(W, U, H) \doteq D_\beta(Y||X) + P_u(U) + P_h(H),$$

where $P_u : \mathbb{R}_{0,+}^{J\times N} \to \mathbb{R}_{0,+}$, and $P_h : \mathbb{R}_{0,+}^{K\times M} \to \mathbb{R}_{0,+}$ are penalizing functions, each one imposing a cost over the appearance of certain features on $U$ and $H$, respectively.

As it can be observed, while the spectrogram of the clean signal depicted in Fig. 2 presents a somewhat sparse structure, the one corresponding to the reverberant signal presents a smoother, more diffuse structure. As it is customary ([9]), we shall hinder the smoothness observed in the reverberant spectrogram from appearing in the restored spectrogram by defining a penalizer over the activation coefficients matrix $U$ of the form

$$P_u(U) \doteq \sum_{j,n} \lambda_n^{(u)} U_{j,n},$$

where $\lambda_n^{(u)} \geq 0$, $n = 1, \ldots, N$, are called penalization parameters for $P_u$. We let the penalizer depend on the time index $n$ as to allow for better compliance with the inherent silences of the recorded signals (more on this subject in Section 5.3.2).

In order to define a penalizer over $H$, we turn our attention to Fig. 3, that shows a simulated RIR in a room with a reverberation time of 450[ms]. The log-spectrogram exhibits a high-energy vertical band on the left, corresponding to the first echoes to reach the receiver, that slowly fades to the right, as deemed by a linear impulse response. The oblique straight lines of less energy correspond to an apparent frequency increase due to

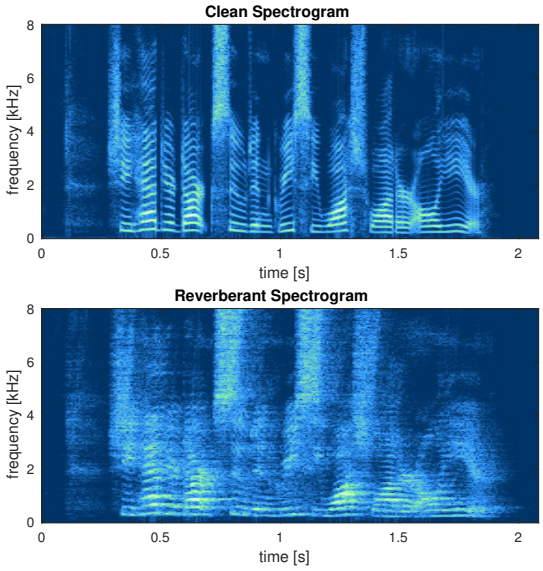


Figure 2: Top: spectrogram of a clean signal, sampled at 16[kHz], using a 512 samples window with overlapping of 256. Bottom: the spectrogram of a reverberant (600[ms]) version of the same signal.

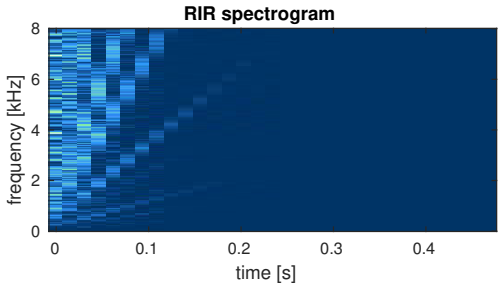


Figure 3: Log-spectrogram for an artificial 16 [kHz] RIR signal with reverberation time of 450 [ms]. The spectrogram was made using a Hanning window length of 512 and overlapping of 256.

the increasing rate at which echoes reach the microphone in rectangular rooms ([15]). From these characteristics, and the fact that the overlapping of windows results in consecutive time components of $H$ capturing common information, it is reasonable to expect the components of $H$ to exhibit a smooth decay over time ([16]). This structure can be promoted (see [7]) by introducing a penalizer of the form

$$P_h(H) \doteq \sum_k \lambda_k^{(h)} \|LH_k^T\|_2^2,$$

where $\lambda_k^{(h)} \geq 0$, $H_k \in \mathbb{R}_{0,+}^M$, $k = 1, \ldots, K$ are the rows of $H$, and $L \in \mathbb{R}^{(M-1)\times M}$ is a *finite difference matrix*, so that $[LH_k^T]_m = H_{k,m+1} - H_{k,m}$.

With all of the above, the cost function is defined as follows:

$$f(W, U, H) \doteq D_\beta(Y||X) + \sum_{j,n} \lambda_n^{(u)} U_{j,n} + \sum_k \lambda_k^{(h)} \|LH_k^T\|_2^2. \tag{5}$$

In the next section we state a two-stage optimization process in order to minimize $f$, first with respect to $W$, and then with respect to both $U$ and $H$. In-line with the core idea stated before, by appropriately tunning its parameters, the cost function (5) can be used for building a good dictionary in a first stage, and for seeking a good representation of the data in a second step.

# 4 Optimization

The optimization process that shall yield the restored spectrogram $\hat{S}$ is divided in two main steps: firstly, given the observed reverberant spectrogram $Y \in \mathbb{R}_{0,+}^{K\times N}$, a suitable dictionary $\hat{W} \in \mathbb{R}_{0,+}^{K\times J}$ that be able to provide a good representation of the target clean spectrogram $S$ is built. Once this is accomplished, the algorithm proceeds to find $\hat{U} \in \mathbb{R}_{0,+}^{J\times N}$ and $\hat{H} \in \mathbb{R}_{0,+}^{K\times M}$ minimizing $f$ given $\hat{W}$.

In order to minimize the cost function, we shall begin by introducing the concept of auxiliary function.

## 4.1 Auxiliary function

**Definition 4.1** *Let $\Omega \subset \mathbb{R}^P$ and $f : \Omega \to \mathbb{R}_0^+$. Then, $g : \Omega \times \Omega \to \mathbb{R}_0^+$ is called an* auxiliary function *for $f$ if $g(\omega,\omega) = f(\omega)$ and $g(\omega,\omega') \geq f(\omega)$, $\forall \omega,\omega' \in \Omega$.*

**Lemma 4.2** *If we let $f$ and $g$ be as in the definition above, $\omega^0 \in \Omega$ be arbitrary and*

$$\omega^t \doteq \arg\min_{\omega} g(\omega,\omega^{t-1}),\ t \in \mathbb{N}$$

*then it can be shown ([17]) that the sequence $\{f(\omega^t)\}_{t\geq 1}$ is non-increasing.*

The idea is to build an auxiliary function $g$ for $f$ with respect to each of its three arguments individually, and then use them iteratively for minimizing $f$.

We will proceed in a similar fashion than in [18]. Firstly, let us notice that $\forall Y \in \mathbb{R}_{0,+}^{K\times N}$, $D_\beta(Y||\cdot) \in \mathcal{C}^\infty(\mathbb{R}_+^{K\times N})$, and

$$\frac{\partial^2 D_\beta(Y||X)}{\partial X_{k,n}^2} = (\beta-1)X_{k,n}^{\beta-2} + (2-\beta)X_{k,n}^{\beta-3}Y_{k,n}. \tag{6}$$

By defining

$$\check{D}_\beta(Y||X) \doteq \sum_{k,n}\left(\frac{\chi_{\beta>1}(\beta)}{\beta}X_{k,n}^\beta - \frac{\chi_{\beta\leq 2}(\beta)}{\beta-1}Y_{k,n}X_{k,n}^{\beta-1} + \frac{1}{\beta(\beta-1)}Y_{k,n}^\beta\right),$$

and

$$\hat{D}_\beta(Y||X) \doteq \sum_{k,n}\left(\frac{\chi_{\beta<1}(\beta)}{\beta}X_{k,n}^\beta - \frac{\chi_{\beta>2}(\beta)}{\beta-1}Y_{k,n}X_{k,n}^{\beta-1}\right),$$

we have $D_\beta = \check{D}_\beta + \hat{D}_\beta$, where $\check{D}_\beta$ is convex and $\hat{D}_\beta$ is concave (both w.r.t. $X$). In the following, we will make use of this decomposition in order to build auxiliary functions for updating each one of the components of $X$.

## 4.2 Building $\hat{W}$

As mentioned before, the parameters required for building a proper dictionary $\hat{W}$ are not necessarily the same as those leading to an optimal representation. Thus, we begin by fixing $H_{k,n} = 1$ if $n = 1$ and $H_{k,n} = 0, \forall n = 2,\ldots,M,\ \ k = 1\ldots,K$. This means that we are precluding $H$ from modeling reverberation, and henceforth it does not make sense to promote temporal sparsity over $U$, and so we set $\lambda_n^{(u)} = 0$, $\forall n = 1,\ldots,N$, only for the first stage.

Now, provided we have found adequate parameters (what we address in Section 5.3.2), the problem of finding an appropriate dictionary reduces to minimizing (5) with respect to $W$ and $U$ subject to $H$ and $\lambda_n^{(u)}$ be set as above. To do so, we begin by finding an auxiliary function for (5) w.r.t. $W$. Let $W' \in \mathbb{R}_+^{K\times J}$, and let us denote

$X'_{k,n} = \sum_{j,m} W'_{k,j} U_{j,n-m} H_{k,m}$. Then,

$$\begin{aligned}
\check{D}_\beta(Y_{k,n}||X_{k,n}) &= \check{D}_\beta\left(Y_{k,n}\middle|\middle|\sum_{j,m} W_{k,j}U_{j,n-m}H_{k,m}\right) \\
&= \check{D}_\beta\left(Y_{k,n}\middle|\middle|\frac{\sum_{j,m} W_{k,j}U_{j,n-m}H_{k,m}X'_{k,n}\frac{W'_{k,j}}{W'_{k,j}}}{X'_{k,n}}\right) \\
&= \check{D}_\beta\left(Y_{k,n}\middle|\middle|\frac{\sum_{j,m} W'_{k,j}U_{j,n-m}H_{k,m}X'_{k,n}\frac{W_{k,j}}{W'_{k,j}}}{\sum_{j,m} W'_{k,j}U_{j,n-m}H_{k,m}}\right) \\
&\leq \sum_{j,m} \frac{W'_{k,j}U_{j,n-m}H_{k,m}}{X'_{k,n}} \check{D}_\beta\left(Y_{k,n}\middle|\middle|X'_{k,n}\frac{W_{k,j}}{W'_{k,j}}\right), \qquad (7)
\end{aligned}$$

where the last step is due to Jensen's inequality.

In regard to $\hat{D}_\beta$, since it is concave w.r.t. $X$, it follows that

$$\hat{D}_\beta(Y_{k,n}||X_{k,n}) \leq \hat{D}_\beta(Y_{k,n}||X'_{k,n}) + \frac{\partial \hat{D}_\beta(Y_{k,n}||X'_{k,n})}{\partial X_{k,n}} \sum_{j,m} (W_{k,j} - W'_{k,j}) U_{j,n-m} H_{k,m}. \qquad (8)$$

Given $U$ and $H$ fixed, let us define $g_w : \mathbb{R}_+^{K\times J} \times \mathbb{R}_+^{K\times J} \to \mathbb{R}$ by

$$\begin{aligned}
g_w(W, W') \doteq & \sum_{k,n,j,m} \frac{W'_{k,j}U_{j,n-m}H_{k,m}}{X'_{k,n}} \check{D}_\beta\left(Y_{k,n}\middle|\middle|X'_{k,n}\frac{W_{k,j}}{W'_{k,j}}\right) \\
& + \sum_{k,n,j,m} \frac{\partial \hat{D}_\beta(Y_{k,n}||X'_{k,n})}{\partial X_{k,n}} (W_{k,j} - W'_{k,j}) U_{j,n-m} H_{k,m} \\
& + \sum_{k,n} \hat{D}_\beta(Y_{k,n}||X'_{k,n}).
\end{aligned}$$

Then, it follows from (7) and (8) that $g_w$ is an auxiliary function for $f$ w.r.t. $H$. Note that the equality condition in Definition 4.1 also holds.

Since $g_w(W, W')$ is convex with respect to $W$, it can be minimized by equating its gradient to zero, what leads to

$$0 = \left(\frac{W_{k,j}}{W'_{k,j}}\right)^{\alpha_1} \sum_{n,m} X'^{\beta-1}_{k,n} U_{j,m} H_{k,n-m} - \left(\frac{W_{k,j}}{W'_{k,j}}\right)^{\alpha_2} \sum_{n,m} X'^{\beta-2}_{k,n} Y_{k,n} U_{j,m} H_{k,n-m},$$

where $\alpha_1 = (\beta-1)\chi_{\beta>1}(\beta)$, and $\alpha_2 = (\beta-2)\chi_{\beta\leq 2}(\beta)$. This automatically leads to the updating equation

$$W^{(t)}_{k,j} = W^{(t-1)}_{k,j} \frac{\left[\left(\sum_{m,n} \left(X^{(t-1)}_{k,n}\right)^{\beta-2} Y_{k,n} U_{j,m} H_{k,n-m}\right)^{\eta}\right]_\epsilon}{\left(\sum_{m,n} \left(X^{(t-1)}_{k,n}\right)^{\beta-1} U_{j,m} H_{k,n-m}\right)^{\eta}}, \qquad (9)$$

where $\eta \doteq \frac{1}{\alpha_1 - \alpha_2}$. Here, the supra index $t$ denotes the iteration number and $[\cdot]_\epsilon$ denotes the operation $\max\{\cdot, \epsilon\}$ , with $\epsilon$ being a small constant ($\sim 10^{-10}$). This is used to avoid the elements of $W$ from dropping to 0 (or below), as once an element is null, it cannot regain positive values by a multiplicative updating procedure (see [19]). For simplicity of notation, we have avoided the use of superscripts in all the variables that do not depend directly on $W$.

In a similar fashion, it can be shown that an auxiliary function for $f$ with respect to $U$ is given by

$$\begin{aligned} g_u(U,U') \doteq & \sum_{k,n,j,m} \frac{W_{k,j}U'_{j,m}H_{k,n-m}}{X'_{k,n}} \check{D}_\beta\left(Y_{k,n}\middle|\middle|X'_{k,n}\frac{U_{j,m}}{U'_{j,m}}\right) \\ & + \sum_{k,n,j,m} \frac{\partial \hat{D}_\beta(Y_{k,n}||X'_{k,n})}{\partial X_{k,n}} W_{k,j}(U_{j,m}-U'_{j,m})H_{k,n-m} \\ & + \sum_{k,n} \hat{D}_\beta(Y_{k,n}||X'_{k,n}) + \sum_{j,n} \lambda_n^{(u)} U_{j,n}. \end{aligned}$$

Here again, since $g_u(U,\cdot)$ is convex, it can be minimized by equating its gradient to zero, which is tantamount to solving

$$U_{j,m} = U'_{j,m}\left(\frac{\sum_{k,n} X'^{\beta-2}_{k,n} Y_{k,n} W_{k,j} H_{k,n-m} - \lambda_m^{(u)}\left(\frac{U'_{j,m}}{U_{j,m}}\right)^{\alpha_2}}{\sum_{k,n} X'^{\beta-1}_{k,n} W_{k,j} H_{k,n-m}}\right)^{\eta}.$$

Let us notice that this is an implicit equation with respect to $U_{j,m}$ for $\beta < 2$ (and $\lambda_j^{(u)} \neq 0$), but since $g_u$ is an auxiliary function for $f$ w.r.t. $U$, Lemma 4.2 guarantees that $U^{(t)}$ approaches a limit $\hat{U}$ as $t$ tends to infinity, and so the quotient $U^{(t)}_{j,m}/U^{(t-1)}_{j,m}$ should approach 1. Henceforth, the approximation $U^{(t)}_{j,m}/U^{(t-1)}_{j,m} \approx 1$ yields the following multiplicative updating rule:

$$U^{(t)}_{j,m} = U^{(t-1)}_{j,m} \frac{\left[\left(\sum_{k,n}\left(X^{(t-1)}_{k,n}\right)^{\beta-2} Y_{k,n} W_{k,j} H_{k,n-j} - \lambda_m^{(u)}\right)^{\eta}\right]_\epsilon}{\left(\sum_{k,n}\left(X^{(t-1)}_{k,n}\right)^{\beta-1} W_{k,j} H_{k,n-j}\right)^{\eta}}. \tag{10}$$

The dictionary $\hat{W} = \arg\min_W f(W,U,H)$ can thus be obtained by alternatively updating $W$ and $U$ using (9) and (10), respectively, until convergence.

Once $\hat{W}$ is obtained, we proceed to find $\hat{U}$ and $\hat{H}$ that be able to effectively model reverberation.

### 4.3 Building $\hat{U}$ and $\hat{H}$

Unlike in the first step, now we do want to impose a sparse structure over $U$, and so $\lambda_n^{(u)}$ should no longer be null for every $n = 1,\ldots,N$. Furthermore, it should be pointed out that the value of $\beta$ in this stage is not necessarily the same as in the previous one (and in fact they will be chosen differently in practice).

The updating rule for $U$ is exactly the same as stated in (10). In regard to $H$, we define the auxiliary function

$$\begin{aligned} g_h(H,H') \doteq & \sum_{k,n,j,m} \frac{W_{k,j}U_{j,n-m}H'_{k,m}}{X'_{k,n}} \check{D}_\beta\left(Y_{k,n}\middle|\middle|X'_{k,n}\frac{H_{k,m}}{H'_{k,m}}\right) \\ & + \sum_{k,n,j,m} \frac{\partial \hat{D}_\beta(Y_{k,n}||X'_{k,n})}{\partial X_{k,n}} (H_{k,m}-H'_{k,m}) W_{k,j}U_{j,n-m} \\ & + \sum_{k,n} \hat{D}_\beta(Y_{k,n}||X'_{k,n}) + \sum_{k} \lambda_k^{(h)} \|LH_k^T\|^2. \end{aligned}$$

By equating its gradient (with respect to $H_{k,m}$) to zero, we obtain, for every $k = 1,\ldots,K, m = 1,\ldots,M$,

$$\begin{aligned} 0 = & \sum_{j,n} W_{k,j}U_{j,n-m}\left(X'_{k,n}\right)^{\alpha_1}\left(\frac{H_{k,m}}{H'_{k,m}}\right)^{\alpha_1} - \sum_{j,n} W_{k,j}U_{j,n-m}Y_{k,n}\left(X'_{k,n}\right)^{\alpha_2}\left(\frac{H_{k,m}}{H'_{k,m}}\right)^{\alpha_2} \\ & - 2\lambda_k^{(h)}[L^T LH_k^T]_m. \end{aligned}$$

It has been observed that using a multiplicative updating rule analogous to those used for $W^{(t)}$ and $U^{(t)}$ usually results in undesired oscillations in the elements of $H^{(t)}$. This is most likely due to the alternating signs in the

rows of $L^T L$. In order to overcome this potential drawback, for every $k = 1, \ldots, K$, we define the diagonal matrix $A^{(k)} \in \mathbb{R}^{M \times M}_{0,+}$ with $A^{(k)}_{m,m} = \sum_{j,n} W_{k,j} U_{j,n-m} \left( X^{(t-1)}_{k,n} \right)^{\alpha_1} / H^{(t-1)}_{k,m}$ and define the vector $b^{(k)} \in \mathbb{R}^M_{0,+}$ as $b^{(k)} = \sum_{j,n} W_{k,j} U_{j,n-m} Y_{k,n} \left( X^{(t-1)}_{k,n} \right)^{\alpha_2}$. Then, under the same approximation used for arriving at (10), we can update $H$ by solving for $H^{(t)}_k$, $k = 1, \ldots, K$, the linear system

$$\left( A^{(k)} + 2\lambda^{(h)}_k L^T L \right) H^{(t)}_k = b^{(k)}. \tag{11}$$

It can be shown that the matrix $A^{(k)} + 2\lambda^{(h)}_k L^T L$ is strictly positive definite (unless $A^{(k)}$ is null), and hence the linear system (11) has a unique solution, whose elements are non-negative.

### 4.4 Additional considerations

Our approximate solution could be defined simply as $\hat{S} = \hat{W}\hat{U}$, but although this clearly leaves out reverberation (which is captured by $\hat{H}$), this low-rank approximation still entails some error. In order to avoid this, we estimate the clean spectrogram by multiplying the data elements $Y_{k,n}$ by a time-varying gain function $G_{k,n} \doteq \frac{\sum_j \hat{W}_{k,j} \hat{U}_{j,n}}{\sum_{j,m} \hat{W}_{k,j} \hat{U}_{j,n-m}, \hat{H}_{k,m}}$, as suggested in [9].

All steps necessary for our dereverberation method are summarized in Algorithm 1.[1]

Next, we proceed to show some experimental results.

## 5 Experimental results

In this section we present a series of experiments, firstly for parameter search and then for validating our method. All signals used in the experiments were taken from the TIMIT database ([20]), sampled at 16[kHz]. For the artificial RIR signals we made use of the software Room Impulse Response Generator[2].

In order to measure the quality of the restored signals, we used the well known frequency weighted segmental signal-to-noise ratio (fwsSNR) and the cepstral distance ([21]). Additionally, we have computed the values of the speech-to-reverberation modulation energy ratio (SRMR, [22]). However, since the SRMR is non intrusive, its values must be used carefully for comparison purposes, keeping in mind that the resemblance of a restoration with the corresponding clean signal is not taken into account.

### 5.1 Parameter estimation

We begin by addressing the main parameter estimation problem for Stage 1 of Algorithm 1. Namely, finding an optimal value of $\beta$ for building a dictionary whose atoms (columns) be able to provide a good representation of a clean spectrogram. In order to evaluate whether a given parameter $\beta_1$ is good for dictionary building, we take a reverberant spectrogram $Y$, build a dictionary $W^{(\beta_1)}$ by minimizing $D_{\beta_1}(Y||WU)$, and then proceed to check how well can $W^{(\beta_1)}$ represent the corresponding clean spectrogram $S$. To do this, given $\beta^*$, we minimize $D_{\beta^*}(S||W^{(\beta_1)}U)$ with respect to $U$. It is important to point out that in this second step, $\beta^*$ is not necessarily the same as $\beta_1$, and hence the two steps above are performed for every pair $(\beta_1, \beta^*)$ in order to find the optimal one.

To do this, we have taken five random clean signals and made them reverberant by means of a discrete convolution with an artificial RIR. For each reverberant spectrogram $Y$ and each admissible pair $(\beta_1, \beta^*)$, we have taken the following steps:

1. Build a dictionary $W^{(\beta_1)} = \arg\min_{W,U} D_{\beta_1}(Y||WU)$.
2. Use $W^{(\beta_1)}$ to find a representation $\hat{S} = W^{(\beta_1)}\hat{U}$ for the associated clean spectrogram $S$, where $\hat{U} = \arg\min_U D_{\beta^*}(S||W^{(\beta_1)}U)$.
3. Test the accuracy of the representation $\hat{S}$ by computing the cepstral distance with respect to $S$.

[1] To try online: http://sinc.unl.edu.ar/web-demo/beta-dereverberation/

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

**Algorithm 1** Variable $\beta$-divergence dereverberation

**Preliminaries**

Given a speech signal $y$, build $Y_{k,n} = |\mathrm{STFT}(y)_{k,n}|^2$.

**Stage 1**

Set $\beta = \beta_1$ and $\lambda_n^{(u)} = 0,\ \forall n$.
Let $H_{k,n} = 1$ if $n = 1$ and $H_{k,n} = 0, \forall n \geq 2, \forall k$.
Initialize $W^{(0)}$ and $U^{(0)}$ randomly.
Let $t = 0$,
**while** $\|W^{(t)} - W^{(t-1)}\|_F^2 > \delta$
  $t \leftarrow t + 1$
  Update $W^{(t)}$ as stated in (9).
  Update $U^{(t)}$ as stated in (10).
**end while**
Let $\hat{W} = W^{(t)}$

**Stage 2**

Set $\beta = \beta_2$ and reset $\lambda_n^{(u)}\ \forall n$.
Let $H_{k,n}^{(0)} = \exp(1 - n),\ \forall n, k$.
Initialize $U^{(0)}$ as the last approximation in Stage 1.
Let $t = 0$,
**while** $\|S^{(t)} - S^{(t-1)}\|_F^2 > \delta$
  $t \leftarrow t + 1$
  Update $U^{(t)}$ as stated in (10).
  Update $H^{(t)}$ as stated in (11).
**end while**
Let $\hat{U} = U^{(t)}$
Let $\hat{H} = H^{(t)}$

**Reconstruction**

Let $G_{k,n} \doteq \sum_j \hat{W}_{k,j}\hat{U}_{j,n} / \left(\sum_{j,m} \hat{W}_{k,j}\hat{U}_{j,n-m}, \hat{H}_{k,m}\right)$.
Let $\hat{S}_{k,n} = G_{k,n}Y_{k,n}$.
Define $Z \in \mathbb{C}^{K \times N}$ by $Z_{k,n} = \sqrt{\hat{S}_{k,n}}\arg(Y_{k,n})$.
Define the restored signal in the time domain as
$\hat{s} \doteq \mathrm{ISTFT}(Z)$.

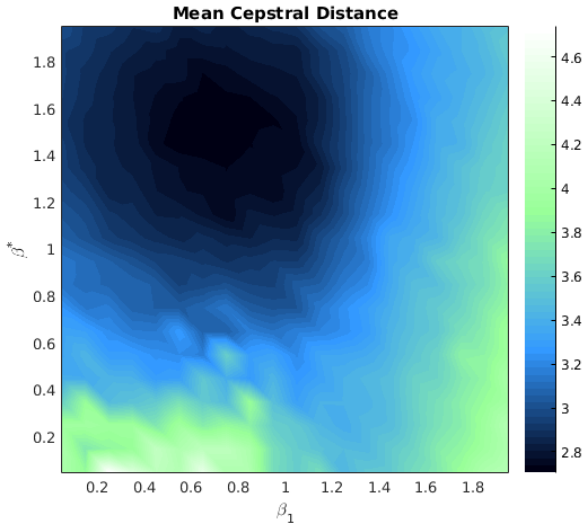


Figure 4: Mean cepstral distance values obtained from a representation of a clean signal using a $\beta^*$ divergence, with a dictionary built from a reverberant version using $\beta_1$ . Smaller values correspond to better results.

Fig. 4 depicts the resulting mean cepstral distance (over five trials over each of the five signals) as a function of the parameters $\beta_1$ and $\beta^*$. The minimizer is reached at (0.75,1.45), showing that $\beta_1 = 0.75$ is the best parameter choice for Stage 1 of Algorithm 1. Note that this does not necessarily mean that $\beta_2 = 1.45$ is the best choice for the second stage of Algorithm 1, since here we are minimizing $D_\beta(S||\hat{S})$ whereas the second step of the dereverberation method requires minimizing Equation (5).

It should be pointed out that functional (5) is a generalization of a Bayesian approach (similar to the one in [7]) if $U$ and $\nabla_t H$ are treated as random variables with exponential and normal *a-priori* distributions, respectively. In fact, by choosing $\beta = 2$, the minimizer of (5) corresponds to a *maximum-a-posteriori* (MAP) estimator, given proper choices of the penalization parameters. Therefore, we have chosen $\beta = 2$ for Stage 2 of Algorithm 1, which in fact was observed to lead to better results than $\beta = 1.45$.

A few relevant conclusions can be derived by observing Fig 4. First, that the values of $(\beta_1, \beta^*)$ leading to the smallest cepstral distances are away from the diagonal, thus corroborating our original conjecture that using different parameter values for the learning and representation steps could lead to improved results. Furthermore, note that better results are obtained for values of $(\beta_1, \beta*)$ in the top left area. This most probably reflects the fact that small values of $\beta_1$ lead to dictionaries which take all the frequency range into account, whereas high values of $\beta^*$ promote fidelity on the high-energy zones of the represented spectrogram.

## 5.2 Illustration

Before beginning with the actual experiments we show how the method works by plotting the result obtained for just one signal. The signal corresponds to a female speaker pronouncing the sentence "She had your dark suit in greasy wash water all year", from the TIMIT database, recorded in an office room (Room 1, in Table 4) in real-life conditions, as specified in Section 5.3.2. All representation elements are depicted in Fig. 5. It can be seen that at the end of Stage 1, a dictionary $W^{(1)}$ is built while reverberation is captured in the coefficient matrix $U^{(1)}$. In the second stage, reverberation is mostly represented by $H^{(2)}$, thus allowing the coefficients in $U^{(2)}$ to provide a good representation $S^{(2)}$ of the clean spectrogram $S$.

## 5.3 Validation

We have chosen two different settings for the validation experiments. The first one using simulations in order have a large number of trials available, and the second one using real recordings to guarantee the method is applicable in real-life conditions.

The model parameters used for all the experiments are detailed in Table 1.

In order to evaluate the performance of our method, comparisons against two state-of-the-art methods applicable under the same conditions were made. The first one was proposed in [7], and it has shown to perform

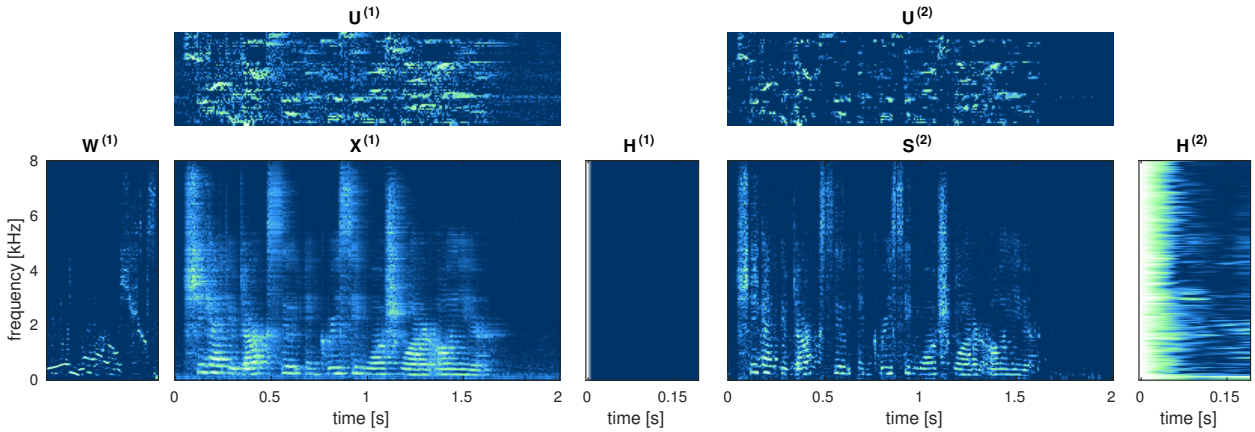


Figure 5: Representation elements obtained with the proposed method. $W^{(1)}$, $U^{(1)}$, $H^{(1)}$, and $S^{(1)}$=$W^{(1)}U^{(1)}$ are the matrices at the end of Stage 1, and $U^{(2)}$, $H^{(2)}$, and $S^{(2)} = W^{(1)}U^{(2)}$ are the matrices at the end of the dereverberation process. All the elements are in log scale, in amplitude.

Table 1: Model parameters

| win. size | win. overl. | $J$ | $M$ | $\beta_1$ | $\beta_2$ |
|---|---|---|---|---|---|
| 512 | 256 | 64 | 20 | 0.75 | 2 |

| $\lambda_n^{(u)}$ | $\lambda_k^{(h)}$ | $\delta$ |
|---|---|---|
| $mean(Y) \times 10^{-3}$ | $0.3\|Y_k\|^2$ | $\|Y\| \times 10^{-3}$ |

Table 2: Simulated room settings

| | Length | Width | Height |
|---|---|---|---|
| Room 1 dimensions | 5.00 [m] | 4.00 [m] | 6.00 [m] |
| Room 2 dimensions | 4.00 [m] | 4.00 [m] | 3.00 [m] |
| Room 3 dimensions | 10.0 [m] | 4.00 [m] | 5.00 [m] |
| Source position | 2.00 [m] | 3.50 [m] | 2.00 [m] |
| Microphone 1 position | 2.00 [m] | 1.50 [m] | 1.00 [m] |
| Microphone 2 position | 2.00 [m] | 2.00 [m] | 1.00 [m] |
| Microphone 3 position | 2.00 [m] | 2.00 [m] | 2.00 [m] |

quite well. The other one was proposed by Wisdom *et al* in [6], and showed an excellent performance in the Reverb Challenge ([23]).

### 5.3.1 Simulated experiments

For the simulations, 110 speech signals from the TIMIT database were taken, and made reverberant by convolution with artificial impulse responses. The artificial RIRs were generated varying the microphone positions and room dimensions, as specified in Table 2. The reverberation time was set at either 450[ms], 600[ms] or 750[ms], resulting in 27 different reverberation conditions, and hence a total of 2970 reverberant signals for testing.

Table 3 and Fig. 6 show the results obtained with each performance measure and each one of the methods. Note that our proposed method (labeled "Beta") outperforms ($p < 0.01$) the other two in terms of fwsSNR and cepstral distance, but not the Bayesian ([7]) in terms of SRMR. However, taking into account that SRMR quantifies the extent to which a signal "seems" reverberant, but not how much such a restoration resembles the corresponding clean signal, it should only be considered as a complement to the other two measures.

### 5.3.2 Experiments using recordings

In order to test whether our method works in real-life situations, we made recordings in two of our own office rooms, during standard office hours and with air conditioners and computers left on. The offices' dimensions

Table 3: Mean and standard deviation (between parenthesis) of performance measures for each method, using simulations. Best results are shown in boldface.

| Measure | fwsSNR | Cepstral Dist. | SRMR |
|---|---|---|---|
| Reverberant | 5.377 (1.70) | 5.308 (0.61) | 2.470 (1.01) |
| Wisdom | 5.593 (1.67) | 5.279 (0.60) | 2.898 (1.14) |
| Bayesian | 7.604 (1.60) | 4.614 (0.52) | **4.423** (1.48) |
| Beta | **8.153** (1.51) | **4.573** (0.48) | 3.751 (1.21) |

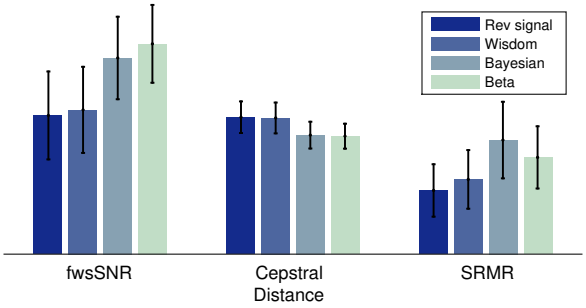


Figure 6: Mean and standard deviation of performance measures for each method, using simulations.

Table 4: Office rooms settings

| | Length | Width | Height |
|---|---|---|---|
| Room 1 dimensions | 4.15 [m] | 3.00 [m] | 3.00 [m] |
| Source 1 position | 3.60 [m] | 1.50 [m] | 1.50 [m] |
| Microphone 1 position | 1.10 [m] | 1.50 [m] | 1.50 [m] |
| Room 2 dimensions | 5.85 [m] | 4.55 [m] | 3.00 [m] |
| Source 2 position | 1.10 [m] | 1.50 [m] | 1.50 [m] |
| Microphone 2 position | 1.10 [m] | 4.00 [m] | 1.50 [m] |

Table 5: Mean and standard deviation (between parenthesis) of performance measures for each method. Best results are shown in boldface.

| Measure | fwsSNR | Cepstral Dist. | SRMR |
|---|---|---|---|
| Reverberant | 3.613 (1.52) | 4.994 (0.56) | 2.756 (0.75) |
| Wisdom | 4.917 (1.37) | 4.577 (0.43) | 3.222 (0.77) |
| Bayesian | 6.254 (1.33) | 4.769 (0.60) | **4.809** (1.10) |
| Beta | **6.678** (1.18) | **4.524** (0.53) | 4.036 (0.84) |

are shown in Table 4, along with the speaker and microphone positions. The reverberation times of the rooms turned out to be of 460[ms] in Room 1 and of 440[ms] in Room 2, as measured using sine sweeps ([24]). Four speakers (two male and two female) were randomly selected from the TIMIT database, and 10 speech signals from each were recorded in each room, with a sampling frequency of 16[kHz].

As it is customary, the clean speech sources had their low-frequency components filtered out. Hence, we pre-processed our reverberant recordings using a 5000 tap FIR high-pass filter with cut-off frequency of 30[Hz] to mitigate the low frequency noise. For the comparisons to be fair, all the methods were tested after this pre-processing was made.

In order to better cope with the noise, the penalization parameters for $U$ were reset to $\lambda_n^{(u)} = \frac{mean(Y)}{\|U_n^1\|_1} \times 10^{-1}$, where $U_n^1$ is the $n$-th column of $U$ as estimated at the end of Stage 1 of Algorithm 1. This prevents the model from attempting to represent ambient noise during speech silences.

Results are depicted in Table 5 and illustrated in Figure 7. Once again, we see that our proposed method outperforms the others in terms of the fwsSNR, but loses to the Bayesian in terms of SRMR. As for the cepstral distance, the improvement between our proposed method and Wisdom's is the only one not reaching statistical significance ($p > 0.01$).

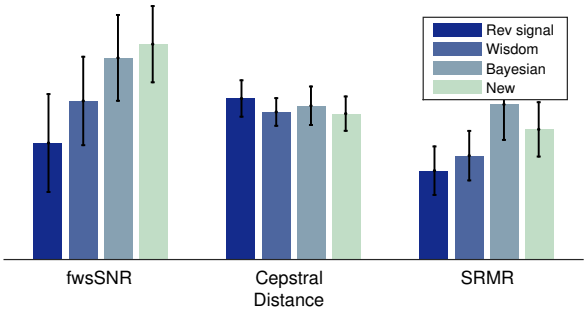


Figure 7: Mean and standard deviation of performance measures for each method, using recordings.

# 6 Conclusions

In this work, a new blind, single channel dereverberation method in the time-frequency domain that makes use of variable $\beta$-divergence as a cost function was presented and tested. The method comprises two stages: one for learning the spectral structure into a dictionary, and a second one for using such a dictionary to build an accurate representation by means of a convolutive NMF model. The corresponding algorithm for implementing the method was introduced and tested. Additionally, a method for finding an optimal learning divergence was introduced.

Results show that the proposed method improves restoration quality with respect to state-of-the-art methods, as measured by the fwsSNR and cepstral distance. Improvement in regard to SRMR is only partial, but being this a non-intrusive measure, that is not too much of a drawback.

There is certainly much room for improvement. For instance, exploring the use of penalization terms at the learning stage and other ways of enhancing the quality of the dictionary, as well as generating atoms for specifically modeling (and then removing) noise and incorporating specific initialization methods. All this is subject of future study.

Finally, although our method is constructed for a blind setting, it is worth noting that it can be easily adapted to be supervised by modifying the learning stage, provided speaker information is available.

# Acknowledgments

This research was funded by ANPCyT under projects PICT 2014-2627 and PICT 2015-0977, by UNL under projects CAI+D 50420150100036LI, CAI+D 50020150100059LI, CAI+D 50020150100055LI and CAI+D 50020150100082LI.

# References

[1] S. Yun, Y. J. Lee, and S. H. Kim, "Multilingual speech-to-speech translation system for mobile consumer devices," *IEEE Transactions on Consumer Electronics*, vol. 60, no. 3, pp. 508–516, 2014.

[2] L. D. Vignolo, S. R. M. Prasanna, S. Dandapat, H. L. Rufiner, and D. H. Milone, "Feature optimisation for stress recognition in speech," *Pattern Recognition Letters*, vol. 84, pp. 1–7, 2016.

[3] R. Sarikaya, P. A. Crook, A. Marin, M. Jeong, J.-P. Robichaud, A. Celikyilmaz, Y.-B. Kim, A. Rochette, O. Z. Khan, X. Liu *et al.*, "An overview of end-to-end language understanding and dialog management for personal digital assistants," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 391–397.

[4] I. J. Tashev, *Sound capture and processing: practical approaches*. John Wiley & Sons, 2009.

[5] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR Upper Saddle River, 2001, vol. 95.

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

[7] F. Ibarrola, L. Di Persia, and R. Spies, "A bayesian approach to convolutive nonnegative matrix factorization for blind speech dereverberation," *Signal Processing*, vol. 151, pp. 89–98, 2018.

[8] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," *Proceedings of the 5th Conference on Independent Component Analysis and Blind Signal Separation*, pp. 494–499, 2004.

[9] N. Mohammadiha, P. Smaragdis, and S. Doclo, "Joint acoustic and spectral modeling for speech dereverberation using non-negative representations," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4410–4414.

[10] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[11] B. Yegnanarayana, P. S. Murthy, C. Avendaño, and 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.

[12] H. Kameoka, T. Nakatani, and 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.

[13] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[14] R. Kompass, "A generalized divergence measure for nonnegative matrix factorization," *Neural computation*, vol. 19, no. 3, pp. 780–791, 2007.

[15] E. De Sena, N. Antonello, M. Moonen, and T. Van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 4, pp. 774–786, 2015.

[16] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien Jr, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[17] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2001, pp. 556–562.

[18] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the $\beta$-divergence," *Neural computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[19] S. Choi, A. Cichocki, H.-M. Park, and S.-Y. Lee, "Blind source separation and independent component analysis: A review," *Neural Information Processing-Letters and Reviews*, vol. 6, no. 1, pp. 1–57, 2005.

[20] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[21] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[22] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[23] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[24] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Audio Engineering Society Convention 122*. Audio Engineering Society, 2007.