Penalized nonnegative representations for speech separation

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

In this work we address the problem of supervised audio source separation within a reverberant environment. We make use of a nonnegative representation in order to model the mixture along with reverberation. This kind of models often pose the problem that the number of variables to learn is large with respect to the data, which is to say there are many possible choices of the elements that result in the same approximation of the mixture. We use a probabilistic approach in order to derive a penalized cost function that aims to overcome this issue by inducing a certain structure over the representation elements. Preliminary results account for a considerable improvement in restoration quality with the introduction of penalizers.

Keywords: signal processing, bayesian model, reverberation
2000 AMS Subject Classification: 92C55 - 68T10

1 Introduction

One of the main problems arising in audio signal processing is that of source separation. That is, given a recording of sound coming from two or more sources, we want to isolate the signals produced by each one.

In this work we shall use a nonnegative matrix factorization (NMF) model to address the problem of speech separation. Some early approaches have made use of training data in order to build a dictionary that encompass the spectral characteristics of each speaker and separate the sources by isolating the activation components ([6]). This works reasonably well for purely additive mixtures, but in reality we are usually faced with reverberation and noisy environments. A more recent approach ([5]) made use of a mixture of NMF and convolutive NMF (CNMF) in order to address this issue. This kind of model entails learning many coefficients from limited data, and thus poses a practical problem: an observed power spectrogram can be accurately represented by a set of parameters that is not representative of the phenomenon we intend to model.

In this work we make use of a Bayesian approach to define proper penalization terms over a cost function. Then, we build an algorithm for minimizing such function which results in effectively learning a mixed NMF-CNMF representation by inducing certain structure over its elements.

2 A reverberant mixture model

Let us consider a setting with I speakers and R microphones. In order to model the reverberant mixture, we begin by defining the continuous, compactly supported functions $s_i, h_{r,i} : \mathbb{R} \to \mathbb{R},\ i =$

[*] fibarrola@sinc.unl.edu.ar

$1, \ldots, I,\ r = 1, \ldots R$, where s_i is the signal from the i-th source, and $h_{i,r}$ is the impulse response signal from the i-th source to the j-th microphone. Then, under the hypothesis that the phenomenon can be accurately represented by a linear time invariant (LTI) system, we can define

$$x_r(t) \doteq \sum_{i=1}^{I} (h_{r,i} * s_i)(t), \quad r = 1, \ldots, R, \tag{1}$$

where x_r is an approximation to the recording y_r, obtained from the r-th microphone.

Since speech signals present large oscillations, we switch to the time frequency domain by means of the Short Time Fourier Transform (STFT). That is, we define $Y_{k,n,r} \doteq |\hat{y}_{r;k}(n)|^2$, $X_{k,n,r} \doteq |\hat{x}_{r;k}(n)|^2$, $H_{k,n,r,i} \doteq |\hat{h}_{r,i;k}(n)|^2$ and $S_{k,n,i} \doteq |\hat{s}_{i;k}(n)|^2$, where $\hat{\cdot}_k(n)$ denotes the STFT at frequency k and time n. With these definitions, Equation 1 leads to

$$Y_{k,n,r} = X_{k,n,r} + \epsilon_{k,n,r} = \sum_{i=1}^{I} \sum_{m=1}^{M} H_{k,m,r,i} S_{k,n-m+1,i} + \epsilon_{k,n,r}, \tag{2}$$

where $\epsilon \in \mathbb{R}^{K \times N \times R}$ is a tensor modeling both the representation error and noise. Details on how this model can be built from (1) can be found in [2].

Finally, let us assume that each source signal can be well represented by using NMF. This means that $\exists W \in \mathbb{R}_{0,+}^{K \times J \times I}, U \in \mathbb{R}_{0,+}^{J \times N \times I}$ such that $S_{k,n,i} \approx \sum_j W_{k,j,i} U_{j,n,i}$. Making a small abuse of notation, model (2) now reads

$$Y_{k,n,r} = X_{k,n,r} + \epsilon_{k,n,r} = \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{j=1}^{J} H_{k,m,r,i} W_{k,j,i} U_{j,n-m+1,i} + \epsilon_{k,n,r}. \tag{3}$$

We now need a way to find representation elements in (3) that allow for a good separation.

3 Cost function

Given that we do not know the model components, we shall treat them as realizations of random variables, and from there, build a cost function whose minimizer will provide a good representation.

Firstly, given that no information is available on the error, we shall assume $\epsilon_{k,n,r}$ to be a realization of a zero-mean normal distribution. This corresponds to $Y_{k,n,r}$ having the following distribution, conditioned on $X_{k,n,r}$:

$$\pi_{like}(Y_{k,n,r}|X_{k,n,r}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y_{k,n,r} - X_{k,n,r})^2}{2\sigma^2}\right).$$

On the other hand, an NMF representation of a clean spectrogram is expected to exhibit a sparse use of the atoms (columns of W) to build a representation, unlike the smooth structure expected for a reverberant one. Sparsity can be favored by assuming the elements of U to be realizations of an exponential distribution. This corresponds to the following Probability Density Function (PDF):

$$\pi_{prior}(U_{j,n,i}) = \frac{\lambda_u}{2} \exp(-\frac{\lambda_u}{2} U_{j,n,i}).$$

Finally, we would expect the components of the impulse response tensor H to exhibit a smooth decay over time. This can be induced by assuming the time gradient of the associated random tensor has a normal distribution. If we let $\underline{H}_{k,r,i}$ be the row vector with components $H_{k,m,r,i}, m = 1, \ldots, M$, this leads to

$$\pi_{prior}(\underline{H}_{k,r,i}) = \frac{1}{\det(2\pi(L\Sigma^{-1}L^{T})^{-1})} \exp\left(-\frac{1}{2}\underline{H}_{k,r,i} L\Sigma^{-1}L^{T} \underline{H}_{k,r,i}^{T}\right),$$

where L is a finite difference matrix associated to the gradient, and $\Sigma \doteq \frac{1}{\lambda_h} I_{(M-1 \times M-1)}$, for some $\lambda_h > 0$.

In order to get a representation whose elements are representative of the aforementioned PDFs, we can find the *maximum-a-posteriori* (MAP) estimator, which amounts to minimizing

$$-\log \pi_{post}(X|Y) = -\log[\pi_{like}(Y|X)\pi_{prior}(U)\pi_{prior}(H)].$$

Under the assumption that the underlying random variables are uncorrelated, this is equivalent to minimizing

$$f(U,H) \doteq \sum_{k,n,r} (Y_{k,n,r} - X_{k,n,r})^2 + \sigma^2 \lambda_u \sum_{j,n,i} U_{k,n,i} + \sigma^2 \lambda_h \sum_{k,r,i} \|\underline{H}_{k,r,i} L\|^2. \tag{4}$$

Next, we introduce a procedure for minimizing this cost function.

4 Optimization

In order to minimize the cost function f, defined in (4), we resort to a minimization-majorization method. This essentially consists of building a new function in a larger space that is easier to minimize than f, and use it to iteratively approach a minimizer of f.

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 $g(\omega,\omega') \geq f(\omega), \ \ \forall \omega, \omega' \in \Omega$.

Then, given an arbitrary $\omega^{(0)} \in \Omega$ and $\omega^t \doteq \arg\min_\omega g(\omega, \omega^{t-1}), \ t \in \mathbb{N}$, it can be shown ([4]) that the sequence $\{f(\omega^t)\}_{t \geq 1}$ is non-increasing.

With this in mind, we can define an auxiliary function for f with respect to U as

$$g_u(U,U') \doteq \sum_{k,n,j,m,r} \frac{W_{k,j,i} U'_{j,m,i} H_{k,n-m+1,i,r}}{X'_{k,n,i}} \left(Y_{k,n,r} - X'_{k,n,r} \frac{U_{j,m,i}}{U'_{j,m,i}} \right)^2 + \sigma^2 \lambda_u \sum_{j,n,i} U_{j,n,i},$$

where $U'_{j,m,i}$ is arbitrary, and $X'_{k,n,r} \doteq \sum_{i,m,j} H_{k,m,r,i} W_{k,j,i} U'_{j,n-m+1,i}$. The proof that this is indeed an auxiliary function for f w.r.t. U is omitted due to space limitations, but the reader is referred to [3] for an analogous, detailed proof.

Now, given that g_u is quadratic w.r.t. U, it can be minimized simply by meeting its first order necessary condition, which readily leads to the updating rule

$$U^{(t)}_{j,m,i} = U^{(t-1)}_{j,m,i} \frac{\left[\sum\limits_{k,n,r} Y_{k,n,r} W_{k,j,i} H_{k,n-m+1,r,i} - \sigma^2 \lambda_u \right]_\epsilon}{\sum\limits_{k,n,r} X^{(t-1)}_{k,n,r} W_{k,j,i} H_{k,n-m+1,r,i}}, \tag{5}$$

where the operation $[\cdot]_\epsilon \doteq \max\{\cdot, \epsilon\}$ is meant to preclude the elements of U from dropping to zero (or below), given that within a multiplicative rule, if an element drops to zero it cannot regain positive values.

An analogous procedure leads to an iterative updating rule for H. By defining, for every k, r and i, the diagonal matrix $A^{(k)} \in \mathbb{R}^{M \times M}_{0,+}$ with $A^{(k,r,i)}_{m,m} = \sum_{j,n} W_{k,j} U_{j,n-m} X^{(t-1)}_{k,n}$ and the vector $b^{(k,r,i)} \in \mathbb{R}^M_{0,+}$ as $b^{(k,r,i)}_m = \sum_{j,n} W_{k,j,i} U_{j,n-m+1,i} Y_{k,n,r} H^{(t-1)}_{k,m,r,i}$. Then, H can be updated by solving for $\underline{H}^{(t)}_{k,r,i}$, the linear system

$$\left(A^{(k)} + \sigma^2 \lambda_u L^T L \right) \left(\underline{H}^{(t)}_{k,r,i} \right)^T = b^{(k)}. \tag{6}$$

It can be shown that the matrix $A^{(k)} + \sigma^2 \lambda_u L^T L$ is strictly positive definite (unless $A^{(k)}$ is null), and hence the linear system (6) has a unique solution, whose elements are non-negative.

All of the above shows that the cost function f can be minimized by iteratively and sequentially updating U and H according to (5) and (6).

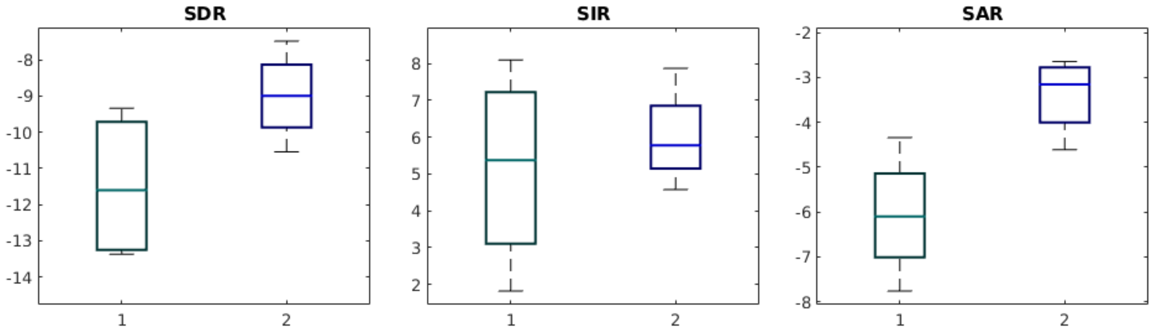

Figure 1: Separation measures results. Those obtained without penalization are on the left, and those using penalization on the right on every plot.

5 Experiments

In order to analyze the improvement (if any) accomplished by the introduction of the penalizers into the model, we tested our method against the one described in [5]. To do so, we randomly chose two male and two female speakers from the TIMIT database ([7]). We then chose a signal from each and built an artificial mixture by convolution with impulse responses generated with the software Room Impulse Response Generator[1], based on the model in [1]. The reverberation time was set to 450[ms].

The dictionaries for each speaker were built from seven signals, different from those used for the mixture. We do not delve into details on this matter due to space limitations, but the reader is again referred to [3] for details.

In order to evaluate the results, we used three standard separation measures: the Signal-to-Distortion Ratio (SDR), the Signal-to-Interference Ratio (SIR) and the Signal-to-Amplitude Ratio (SAR). Figure 1 depicts the obtained results, which clearly suggest an improvement when using our penalization approach.

6 Discussion

A penalization model based on a Bayesian approach over a mixed NMF model was introduced and tested. Although the results are preliminary, they clearly suggest a quality increment over the standard NMF-CNMF approach.

There is certainly much room for improvement. For one thing, exploring the use of probability density functions that take the correlation between the variables into account. Also, the model parameters can be set to depend on the speaker or frequency band and a way to optimally choose them is yet to be found. Finally, it is worth mentioning that the model can be easily extended to use a generalized β-divergence as a fidelity measure.

Acknowledgements

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.

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

References

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

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

[3] ——, *Switching divergences for spectral learning in blind speech dereverberation*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, (in press, 2019).

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

[5] N. Murata, H. Kameoka, K. Kinoshita, S. Araki, T. Nakatani, S. Koyama, and H. Saruwatari, *Reverberation-robust underdetermined source separation with non-negative tensor double deconvolution*, in Signal Processing Conference (EUSIPCO), 2016 24th European, IEEE, 2016, pp. 1648–1652.

[6] M. N. Schmidt and R. K. Olsson, *Single-channel speech separation using sparse non-negative matrix factorization*, in Ninth International Conference on Spoken Language Processing, 2006.

[7] V. Zue, S. Seneff, and J. Glass, *Speech database development at MIT: TIMIT and beyond*, Speech Communication, 9 (1990), pp. 351–356.