SOUCE SEPARATION WITH SCATTERING NON-NEGATIVE MATRIX FACTORIZATION

Joan Bruna, Pablo Sprechmann, Yann LeCun

New York University
Courant Institute of Mathematical Sciences
{bruna, pablo, yann}@cs.nyu.edu

ABSTRACT

This paper presents a single channel source separation method that extendes the approach of Nonnegative Matrix Factorization (NMF). We interpret the approach of audio demixing via NMF as a cascade of a pooled analysis operator, given for example by the magnitude spectrogram, and a synthesis operators given bye the matrix decomposition. Instead of imposing the temporal consistency of the decomposition through sophisticated structured penalties in the synthesis stage, we propose to change the analysis operator for a deep scattering representation. This new signal representation is invariant to smooth changes in the signal, consistent with the temporal dynamics. The proposed approach is evaluate in a speech enhancement task producing promising results.

Index Terms— One, two, three, four, five

1. INTRODUCTION

The problem of isolating or enhancing a speech signal recorded in a noisy environment has been widely studied in the audio processing community [1, 2]. It becomes particularly challenging in the presences of non-stationary background noise, which is a very common situation in many applications encountered, e.g., in telephony. We approach this problem as a monaural source separation method by modeling the speech as one source, and the noise as the other. This is a natural approach when the characteristics of both the signal of interest and the noise vary throughout time [3, 4, 5, 6].

The decomposition of time-frequency representations, such as the power or magnitude spectrogram in terms of elementary atoms of a dictionary, has become a popular tool in audio processing. Non-negative matrix factorization (NMF) [7, 8], have been widely adopted in various audio processing tasks, including in particular source separation, see [9] for a recent review. There are many works that follow this line in speech separation [10, 11] and enhancement [4, 6, 12, 13].

In plain NMF, signals that can be well approximated with the learned dictionary are likely to resemble the training data on a frame by frame manner. They might, however, not be globally consistent. Standard NMF approaches treat different time-frames independently, ignoring the temporal dynamics of the signals. In other words, there is additional structure in speech at a time-scale larger than the frame-length that cannot be learned (or exploited) with NMF. In order to overcome this limitation, many works have proposed regularized extensions of NMF to promote learned (or designed) structure in the codes. Examples of these approaches are, temporal smoothness of the activation coefficients [14], including co-occurrence statistics of the basis functions [3], and learned temporal dynamics [5, 15, 16, 17].

We propose to think of the NMF-based unmixing process as the concatenation of two operators. First, the signal is represented in a feature space given by a poolled analysis operator given: the magnitude of a time-frequency representation such as the Short-Time Fourier Transform (STFT). Then a synthesis operator, given by the dictionary learning stage, is applied to produce an unmixing in the feature space. Finally, the separation is obtained by inverting these representations. Performing the separation in the pooled representation is key to the success of the algorithm. The magnitude STFT is in general sparse (simplifying the separation process) and invariant to variations in the phase (relieving the NMF from learning this irrelevant variability). This comes at the expense of inverting a pooled unmixing in the feature space, normally known as the phase recovery problem [**?**]. In the case of standard NMF, this is easily done via Wiener filtering technique discussed in Section 2.

In this work, rather than seeking for a coding scheme with temporal regularity, we seek to encode a representation of the audio signal at a larger scale in which natrual variability in speech is highly compressible. A sensibly designed scattering transform can compress changes that are temporally consistent with a speech singal, e.g. smooth changes in pitch and envelope. A dictionary learnt to represent the signal in this deep representation would implicitly be learning the short term temporal dynamics of the signal.

Scattering transforms have recently been introduced [**?**] to represent audio signals and images, achieving state-of-the-art results for texture discrimination, and music genre recognition [**?**]. A scattering transform iterates on complex wavelet transforms and modulus operators which compute their envelop. It has close relations with psychophysical and physiological models [**?**].

Our claim is that an important part of the consistency that is imposed via structured NMF, can be eliminated with a better signal representation. In this new setting one can learn the temporal dynamics with a very simple NMF encoding. However, the problem that becomes more difficult is that of inverting the representation. Recent studies in textured sound synthesis from scattering moments have solved this problem successfully using gradient descent algorithms [18].

Synthesis models with coherent dictionaries are known to be highly unstable representations [19]. Thus, training them to satisfy slowness and temporal consistency can be challenging. For example, in [16] the authors explain that the learning the temporal dynamics in NMF via a Kalman type of model, can become very difficult when the coding is sparse due to the instability and jitter in the codes. In contrast, analysis operators are sable by construction.

2. NMF SPEECH ENHANCEMENT

We consider the setting in which we are given a noisy time-domain signal $x(t)$ that is the sum of a speech signal $s(t)$ and a non-stationary noise $n(t)$,

$$x(t) = s(t) + n(t),$$

and we aim at finding a clean estimate $\hat{s}(t)$ of $s(t)$. NMF-based denoising techniques typically operate on a (non-negative) time-frequency representation of $s(t)$ (such as the spectrogram or the power spectrum) that we denote as $\mathbf{V} \in \mathbb{R}^{m\times n}$, comprising m frequency bins and n temporal frames. NMF attempts to find the non-negative activations $\mathbf{H}_{\rm s} \in \mathbb{R}^{q\times n}$ and $\mathbf{H}_{\rm n} \in \mathbb{R}^{r\times n}$ best representing the speech and the noise components, respectively, in two fixed dictionaries $\mathbf{W}_{\rm s} \in \mathbb{R}^{n\times q}$ and $\mathbf{W}_{\rm n} \in \mathbb{R}^{n\times r}$. This task is achieved through the solution of

$$\min_{\mathbf{H}_{\rm s},\mathbf{H}_{\rm n}\geq \mathbf{0}} D(\mathbf{V}|\mathbf{W}_{\rm s}\mathbf{H}_{\rm s} + \mathbf{W}_{\rm n}\mathbf{H}_{\rm n}) + \lambda\,\psi(\mathbf{H}_{\rm s},\mathbf{H}_{\rm n}). \qquad (1)$$

The first term in the optimization objective measures the dissimilarity between the input data and the estimated channels. Frequent choices of D are the squared Euclidean distance, the Kullback-Leibler divergence, and the Itakura-Saito divergence, for which there exist standard optimization algorithms [20]. In this work we concentrate on a (weighted) Euclidean distance, but any other option could be used instead. The second (optional) term in the minimization objective is included to promote some desired structure of the activations. This is done using a designed regularization function ψ and its relative importance is controlled by the parameters λ.

Once the optimal activations are solved for, the spectral envelopes of the speech and the noise are estimated as $\mathbf{W}_{\rm s}\mathbf{H}_{\rm s}$ and $\mathbf{W}_{\rm n}\mathbf{H}_{\rm n}$, respectively. Since these estimated speech spectrum envelope contains no phase information, they are used to build soft masks to filter the signals mixture signal [12].

In the semi-supervised setting, it is assumed that the model for one of the observed signals (speech or noise) is not available beforehand and is estimated along in the inference. Following [], we adopt the semi-supervised setting where $\mathbf{W}_{\rm n}$ and $\mathbf{H}_{\rm n}$ are unknown and with no particular structure, and learned from the data.

3. SCATTERING TRANSFORM

The rythmic and modulation structure characteristic of acoustic pulse trains can be efficiently extracted with the scattering transform [**?**, **?**], computed by iterating wavelet transforms and complex modulus nonlinearitites. This section reviews its definition and main properties.

3.1. Wavelet Filter Bank

A wavelet $\psi(t)$ is a band-pass filter with good frequency and spatial localization. We consider a complex wavelet with a quadrature phase, whose Fourier transform satisfies $\widehat{\psi}(\omega) \approx 0$ for $\omega < 0$. We assume that the center frequency of $\widehat{\psi}$ is 1 and that its bandwidth is of the order of Q^{-1}. Wavelet filters centered at the frequencies $\lambda = 2^{j/Q}$ are computed by dilating ψ:

$$\psi_\lambda(t) = \lambda\,\psi(\lambda\,t) \;\text{ and hence }\; \widehat{\psi}_\lambda(\omega) = \widehat{\psi}(\lambda^{-1}\omega)\,. \qquad (2)$$

We denote by Λ the index set of $\lambda = 2^{j/Q}$ over the signal frequency support, with $j \leq J$, and we impose that these filters fully cover the positive frequencies

$$\forall\omega \geq 0\,,\; 1-\epsilon \leq |\widehat{\phi}(\omega)|^2 + \frac{1}{2}\sum_{\lambda\in\Lambda} |\widehat{\psi}_\lambda(\omega)|^2 \leq 1\,, \qquad (3)$$

for some $\epsilon < 1$, where $\phi(t)$ is the lowpass filter carrying the low frequency information at scales larger than 2^J. The wavelet transform of a signal $I(t)$ is

$$WI = \{I \ast \phi(t)\,,\; I \ast \psi_\lambda(t)\}_{\lambda\in\Lambda}\,.$$

Thanks to (3), one can verify that

$$\|I\|^2(1-\epsilon) \leq \|I\ast\phi\|^2 + \sum_{\lambda\in\Lambda} \||I\ast\psi_\lambda|^2\|^2 \leq \|I\|^2\,. \qquad (4)$$

3.2. Joint Time-Frequency Scattering

Scattering coefficients provide a nonlinear representation computed by iterating over wavelet transforms and a modulus. First order scattering coefficients are local averages of wavelet coefficient amplitudes:

$$\forall\lambda\in\Lambda\;\;,\;\; SI(\lambda;t) = |\,I\ast\psi_\lambda|\ast\phi(t)\,.$$

The Q-factor Q_1 adjusts the frequency resolution of these wavelets. Due to the temporal average, first order scattering coefficients provide no information on the time-variation of the scalogram $|I * \psi_{\lambda_1}(t)|$ at temporal scales smaller than 2^J. It averages all modulations and transient events, and thus lose perceptually important information.

Second order scattering coefficients recover information on audio-modulations and transients by computing the wavelet coefficients of each envelope $|I * \psi_{\lambda_1}|$, and their local averagest:

$$\forall \lambda_2 \;\;,\;\; SI(\lambda_1, \lambda_2; t) = ||\, I * \psi_{\lambda_1}| * \psi_{\lambda_2}| * \phi(t) \;.$$

These multiscale variations of each envelope $|I * \psi_{\lambda_1}|$ specify the amplitude modulations of $I(t)$ [**?**] and thus have the capacity to detect rythmic structures appearing at different frequency bands. The Q-factor Q_2 of the second family of wavelets ψ_{λ_2} controls the time-frequency resolution of the transform. Smaller Q_2 results in wavelets with good temporal resolution, and thus allows us to accurately measure the sharp transitions of amplitude modulations. On the other hand, large Q_2 factors are useful to detect regular and precise rythmic structures present in the envelopes $|I * \psi_{\lambda_1}|$. Scattering coefficients have a negligible amplitude for $\lambda_2 > \lambda_1$ because $|I * \psi_{\lambda_1}|$ is then a regular envelop whose frequency support is below λ_2. Scattering coefficients are thus computed only for $\lambda_2 < \lambda_1$.

Applying more wavelet transform envelopes defines scattering moments at any order $m \geq 1$:

$$SI(\lambda_1, ..., \lambda_m; t) = |\; |I * \psi_{\lambda_1}| * ...| * \psi_{\lambda_m}| * \phi(t) \;. \qquad (5)$$

By iterating on the inequality (4), one can verify [**?**] that the Euclidean norm of scattering coefficients

$$\|SI\|^2 = \sum_{m=1}^{\infty} \sum_{(\lambda_1,...,\lambda_m)\in\Lambda_m} \|SI(\lambda_1, ..., \lambda_m; \cdot)\|^2 \qquad (6)$$

satisfies

$$\|SI\|^2 \leq \|I\|^2 \;.$$

For most audio signals, the energy of the scattering vector $\|SI\|^2$ is concentrated over first and second layers. In practice, we thus only compute $SI(\lambda_1)$ and $SI(\lambda_1, \lambda_2)$ for $1 \leq \lambda_1 = 2^{j_1/Q_1} \leq N$ and $1 \leq \lambda_2 = 2^{j_2/Q_2} < \lambda_1$.

Scattering transforms have been extended along the frequency variables to capture frequency variability and provide transposition invariant representations [**?**]. Transpositions refer to translations along a log frequency variable. We denote $\gamma = \log_2 \lambda_1$, and define wavelets $\bar\psi_{\bar\lambda}(\gamma)$ having an octave bandwidth of $Q = 1$. The corresponding wavelet transform is thus computed with convolutions along the log-frequency variable γ.

The scalogram is now considered as a function of γ and t:

$$F(\gamma, t) = |I * \psi_{2^\gamma}(t)| \;.$$

Second order scattering can then be generalized by computing them as first order coefficients of $F(\gamma, t)$ computed with a separable wavelet transform:

$$SI(\lambda_1; t; \lambda_2, \bar\lambda_2) = |F * \bar\Psi_{\lambda_2, \bar\lambda_2}| * \Phi((\log_2 \lambda_1, t) \;,$$

where $\Psi(\lambda, \bar\lambda)(\gamma, t) = \psi_\lambda(t)\overline{\psi_{\bar\lambda}}(\gamma)$ and $\Phi(\gamma, t)$ is a two-dimensional blurring kernel.

4. ALGORITHM

This section explains how to solve the ill-posed inverse problem of source separation via a sparse coding in the scattering domain followed by phase recovery.

Write down the energy minization problem.

Explain the greedy algorithm

5. EXPERIMENTAL RESULTS

Data sets. We evaluated the separation performance of the proposed methods on a subset of the GRID dataset [21]. Three randomly chosen sets of distinct clips each were used for training (500 clips), validation (10 clips), and testing (50 clips). The clips were resampled to 8 KHz. For the noise signals we used the AURORA corpus [22], which contains six categories of noise recorded from different real environments (street, restaurant, car, exhibition, train, and airport). As before, three sets of distinct clips each were used for training (15 clips), validation (3 clips), and testing (15 clips).

Evaluation measures. As the evaluation criteria, we used the *source-to-distortion ratio* (SDR), *source-to-interference ratio* (SIR), and *source-to-artifact ratio* (SAR) from the BSS-EVAL metrics [23]. We also computed the standard *signal-to-noise ratio* (SNR). When dealing with several frames, we computed a global score (GSDR, GSIR, GSAR and GSNR) by averaging the metrics over all test clips from the same speaker and noise weighted by the clip duration.

Training setting. The same training settings were used in all experiments. We used dictionaries of size 60 and 10 atoms for representing the speech and noise, respectively. These values were obtained using cross-validation. We used $\lambda_{\rm s} = 0.1$ and $\lambda_{\rm n} = 0$ (which means that no sparsity was promoted in the representation of the noise) and $\mu = 0.001$. As the example, we used $\beta = 1$ and $\beta = 0$, and $\alpha = 0$ in the high level cost (**??**). For the SGD algorithm we used $\eta = 0.1$ and minibatch of size 50. These were obtained by trying several values of during a small number of iterations, keeping those producing the lowest error on a small validation set. All training signals where mixed at 5 dB.

Results. Figure **??** shows the evolution of the high level cost (**??**) and the SDR on the validation set with the SGD iterations. The algorithm converges to a dictionary that achieves about 2 dB better SDR on the validation set. Tables 1 and

Table 1: Average performance (in dB) for NMF and proposed supervised NMF methods measured in terms of SDR, SIR, SAR and SNR. Speech and noise were mixed at $5dB$ of SNR. The standard deviation of each result is shown between brackets.

	SDR	SIR	SAR	SNR
NMF $\beta = 1$	7.5 [1.5]	13.7 [0.9]	8.9 [1.7]	8.2 [1.3]
TS-NMF $\beta = 1$	9.5 [1.4]	15.2 [0.7]	11.0 [1.7]	10.0 [1.2]
TS-NMF $\beta = 0$	8.6 [1.3]	14.1 [1.2]	10.3 [1.5]	9.1 [1.1]

Table 2: See description of Table 1. In this case, speech and noise were mixed at $0dB$ of SNR.

	SDR	SIR	SAR	SNR
NMF $\beta = 1$	4.5 [1.1]	9.3 [0.9]	6.8 [1.2]	5.8 [0.8]
TS-NMF $\beta = 1$	6.3 [1.0]	11.9 [0.7]	8.0 [1.1]	7.2 [0.8]
TS-NMF $\beta = 0$	5.2 [1.2]	12.0 [1.7]	6.6 [1.2]	6.3 [0.9]

2 show some initial results for the proposed approach. We compare the performance of standard supervised sparse-NMF (referred simply as NMF) against the performance of the same sparse-NMF model trained on a task-specific manner (referred as TS-NMF) on denoising two with different SNR levels. Observe that the task-specific supervision leads to improvements in performance, maintaining (at $5dB$ SNR) the improvements observed on the validation set. Interestingly, the method also works when using $\beta = 0$ (Itakura-Saito), even if the developments in Section **??** are technically not valid in this case, since the divergence is not convex. In future work we plan to analyze what happens when a non-speaker specific dictionaries are trained. We expect to observe similar improvements, if the training data is diverse enough.

6. CONCLUSION

In this work we presented an algorithm for the task-supervised training of NMF models. Unlike standard supervised NMF, the proposed approach matches the optimization objective used at the train and testing stages. In this way, the dictionaries can be trained in a task-specific manner. We cast this problem as bilevel optimization that can be efficiently solved via stochastic gradient descent. The proposed approach allows non-Euclidean data terms such as β-divergences. A limited case study of sparse NMF with task specific supervision demonstrates promising results. Including temporal dynamics into this model is the subject of ongoing research.

7. REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, vol. 30, CRC, 2007.

[2] E. Hänsler and G. Schmidt, *Speech and Audio Processing in Adverse Environments*, Springer, 2008.

[3] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, “Speech denoising using nonnegative matrix factorization with priors,” in *ICASSP*, 2008, pp. 4029–4032.

[4] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, “Real-time speech separation by semi-supervised nonnegative matrix factorization,” in *LVA/ICA*, 2012, pp. 322–329.

[5] G. J. Mysore and P. Smaragdis, “A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics,” in *ICASSP*, 2011, pp. 17–20.

[6] Z. Duan, G. J. Mysore, and P. Smaragdis, “Online plca for real-time semi-supervised source separation,” in *LVA/ICA*, 2012, pp. 34–41.

[7] D.D. Lee and H.S. Seung, “Learning parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[8] P. Smaragdis, B. Raj, and M. Shashanka, “A probabilistic latent variable model for acoustic modeling,” *Advances in models for acoustic processing, NIPS*, vol. 148, 2006.

[9] Paris Smaragdis, Cedric Fevotte, G Mysore, Nasser Mohammadiha, and Matthew Hoffman, “Static and dynamic source separation using nonnegative factorizations: A unified view,” *Signal Processing Magazine, IEEE*, vol. 31, no. 3, pp. 66–75, 2014.

[10] M. N. Schmidt and R. K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization,” in *INTERSPEECH*, Sep 2006.

[11] M. V. S. Shashanka, B. Raj, and P. Smaragdis, “Sparse Overcomplete Decomposition for Single Channel Speaker Separation,” in *ICASSP*, 2007.

[12] M. N. Schmidt, J. Larsen, and F.-T. Hsiao, “Wind noise reduction using non-negative sparse coding,” in *MLSP*, Aug 2007, pp. 431–436.

[13] Nasser Mohammadiha, Paris Smaragdis, and Arne Leijon, “Supervised and unsupervised speech enhancement using nonnegative matrix factorization,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 2140–2151, 2013.

[14] C. Févotte, “Majorization-minimization algorithm for smooth itakura-saito nonnegative matrix factorization,” in *ICASSP*. IEEE, 2011, pp. 1980–1983.

[15] J. Han, G. J. Mysore, and B. Pardo, "Audio imputation using the non-negative hidden markov model," in *LVA/ICA*, 2012, pp. 347–355.

[16] C. Févotte, J. Le Roux, and J. R. Hershey, "Non-negative dynamical system with application to speech and audio," in *ICASSP*, 2013.

[17] Nasser Mohammadiha and Arne Leijon, "Nonnegative hmm for babble noise derived from speech hmm: Application to speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 5, pp. 998–1011, 2013.

[18] Joan Bruna and Stéphane Mallat, "Audio texture synthesis with scattering moments," *arXiv preprint arXiv:1311.0407*, 2013.

[19] Rodolphe Jenatton, Rémi Gribonval, and Francis Bach, "Local stability and robustness of sparse dictionary learning in the presence of noise," *arXiv preprint arXiv:1210.0685*, 2012.

[20] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[21] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. of the Acoustical Society of America*, vol. 120, pp. 2421, 2006.

[22] D. Pearce and H.-G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *INTERSPEECH*, 2000, pp. 29–32.

[23] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 14, no. 4, pp. 1462–1469, 2006.