SOURCE SEPARATION WITH SCATTERING NON-NEGATIVE MATRIX FACTORIZATION

Joan Bruna, Pablo Sprechmann and Yann LeCun

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

ABSTRACT

This paper presents a single-channel source separation method that extends the ideas 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 by 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, where signals are represented at several time resolutions. This new signal representation is invariant to smooth changes in the signal, consistent with its temporal dynamics. We evaluate the proposed approach in a speech separation task obtaining promising results.

Index Terms— source separation, scattering, non-negative matrix factorization.

1. INTRODUCTION

The problem of source separation has been widely studied in the speech processing community [1, 2]. It becomes particularly challenging when only one microphone is used, or in the presence 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 at an appropriate temporal resolution.

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. Nonnegative matrix factorization (NMF) [3], have been widely adopted in various audio processing tasks, including in particular source separation, see [4] for a recent review. There are many works that follow this line in speech separation [5, 6] and enhancement [7, 8].

In 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 temporally consistent at larger temporal scales. Standard NMF approaches treat different time-frames independently, ignoring the temporal dynamics of the signals. To capture temporal dependencies, works have consider convolutional extension of NMF [9], where the atoms correspond to patches comprised of several time-frequency frames. To avoid the increase in dimensionality, many works have proposed regularized extensions of NMF to promote learned structure in the codes. Examples of these approaches are, temporal smoothness of the activation coefficients [10], including co-occurrence statistics of the basis functions [11], and learned temporal dynamics [12, 13, 14].

NMF-based source separation methods can be thought as the concatenation of two operators. First, the signal is represented in a feature space given by a non-linear analysis operator, typically defined as 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. The separation is obtained by inverting these representations. Performing the separation in the non-linear 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, thus relieving the NMF from learning this irrelevant variability. This comes at the expense of inverting the unmixed estimates in the feature space, normally known as the phase recovery problem [15]. In the case of standard NMF, this is typically done via Wiener filtering.

In this work, rather than optimizing a coding scheme with improved temporal coherence, we concentrate in the extracton of discriminative and stable features. For that purpose, it is crucial to increase the temporal context of the representation, reducing uninformative variability while preserving distinctive speech characteristics. Increasing the temporal scale of STFT or MEL representations results in loss of important discriminative information [16]. In order to overcome the limitations of these shallow representations, scattering transforms [16, 17] cascade several stages of complex wavelet decompositions and complex modulus, yielding discriminative representations with the ability to capture temporal structures at larger scales, e.g. smooth changes in pitch and envelope. Scattering transforms achieve state-of-the-art results on auditory texture discrimination, and music genre recognition [16, 18]. Recently they have been considered in the setting of blind source separation in [19], but here we concentrate in the supervised (and semi-supervised) framework.

In source separation, it is desirable to have a representation with better spatio-temporal resolution than standard scattering features. In this work, we propose a scattering pyramid representation, consisting in a collection of scattering features at different temporal resolutions. With this representation, singals are modeled using a multilevel set of dictionaries each acting at a given temporal resolution. In this way, short-term temporal dynamics of the signal can be captured by the long-context model, capitalizing on the stability properties of scattering coefficients [17].

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 non-linear representation. Recent studies in textured sound synthesis from scattering moments have solved this problem successfully using gradient descent algorithms [20]. Sparse synthesis models with coherent dictionaries are known to be highly unstable representations [21]. Thus, training them to satisfy slowness and temporal consistency can be challenging. In contrast, analysis operators are stable by construc-

tion.

Recent works have started to use discriminative training in the context of source separation. Methods based on NMF [22, 23] and recurrent neural networks [24, 25]. These initial results show the benefits of this setting. The proposed method could also be framed into the discriminative setting.

Section 2 reviews non-negative matrix factorization, while Section 3 describes scattering representations for speech. Our source separation algorithm is described in Section 4 and numerical experiments on TIMIT and GRID datasets are reported in Section 5.

2. NMF SPEECH SOURCE SEPARATION

We consider the setting in which we are given a temporal signal $x(t)$ that is the sum of two speech signals $x_i(t)$, $i = 1, 2$:

$$x(t) = x_1(t) + x_2(t) \,, \tag{1}$$

and we aim at finding estimates $\widehat{x_i}(t)$. NMF-based source separation techniques typically operate on a non-negative time-frequency representation of $x(t)$, such as the spectrogram or the power spectrum, that we denote as $\Phi(x) \in \mathbb{R}^{m\times n}$, comprising m frequency bins and n temporal frames. NMF attempts to find the non-negative activations $Z_i \in \mathbb{R}^{q\times n}$, $i = 1, 2$ best representing the different speech components in two dictionaries $D_i \in \mathbb{R}^{m\times q}$. This task is achieved through the solution of

$$\min_{Z_i\geq 0} D(\Phi(x)|\sum_{i=1,2} D_i Z_i) + \lambda \sum_{i=1,2} \mathcal{R}(Z_i) \,. \tag{2}$$

The first term in the optimization objective measures the dissimilarity between the input data and the estimated channels. Frequent choices of μ are the squared Euclidean distance, the Kullback-Leibler divergence, and the Itakura-Saito divergence, for which there exist standard optimization algorithms [26]. In this work we concentrate on a reweighted Euclidean distance, but any other option could be used instead. The second term in the minimization objective is included to promote some desired structure of the activations. Once the optimal activations are solved for, the spectral envelopes of the speech and the noise are estimated as $D_i Z_i$. Since these estimated speech spectrum envelope contain no phase information, they are used to build soft masks to filter the mixture signal [27].

3. SCATTERING TRANSFORM

Discriminative features having longer temporal context can be constructed with the scattering transform [16, 17]. While these features have shown excellent performance in various classification tasks, in the context of source separation we require a representation that not only captures long-range temporal structures, but also preserves as much discriminability as possible. For this reason, we construct a multi-level representation consisting of a pyramid of scattering coefficients with different temporal resolutions at each level. This section reviews its definition and main properties when applied to speech signals.

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 $\mathcal{F}\psi(\omega) \approx 0$ for $\omega < 0$. We assume that the center frequency of $\mathcal{F}\psi$ 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: $\psi_\lambda(t) = \lambda\,\psi(\lambda\, t)$, and hence $\mathcal{F}\psi_\lambda(\omega) = \widehat{\psi}(\lambda^{-1}\omega)$. 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:

$$\forall\omega \geq 0 \,,\; 1-\epsilon \leq |\mathcal{F}\phi(\omega)|^2 + \frac{1}{2}\sum_{\lambda\in\Lambda} |\mathcal{F}\psi_\lambda(\omega)|^2 \leq 1 \,, \tag{3}$$

for some $\epsilon < 1$, where $\phi(t)$ is the lowpass filter carrying the low frequency information at scales larger than 2^J. The resulting filter bank has a constant number Q of bands per octave. The wavelet transform of a signal $x(t)$ is

$$Wx = \{x * \phi(t) \,,\; x * \psi_\lambda(t)\}_{\lambda\in\Lambda} \,.$$

Thanks to (3), one can verify that

$$\|x\|^2(1-\epsilon) \leq \|x*\phi\|^2 + \sum_{\lambda\in\Lambda} \|x*\psi_\lambda\|^2 \leq \|x\|^2 \,. \tag{4}$$

3.2. Joint Time-Frequency Pyramid Scattering

Scattering coefficients provide a nonlinear representation computed by iterating over wavelet transforms and complex modulus nonlinearities. We start by removing the complex phase of wavelet coefficients in W^1x with a complex modulus nonlinearity. We arrange these first layer coefficients as nodes in the first level of the tree,

$$|W^1|x = \{x_i^1\}_{i=1\ldots 1+|\Lambda|} = \{x{*}\phi_1(\Delta_1 n) \,,\; |x{*}\psi_{1,\lambda_1}(\Delta_1 n)|\}_{\lambda\in\Lambda} \,.$$

where Δ_1 is the critical sampling rate of the highest frequency wavelet sub-band (the reciprocal of the largest bandwidth present in the filter bank) and ψ_1 has bandwidth Q_1^{-1}. These first layer coefficients give localized information both in time and frequency, with a trade-off dictated by the Q factor, Q_1, that adjusts the frequency resolution of these wavelets. For speech a typical choice is around $Q_1 = 32$.

In order to increase the robustness of the representation, we transform each of the down sampled signals from this first layer with a new wavelet filter bank and take the complex modulus of the oscillatory component. In order to sample each channel using the same temporal resolution, this time we apply the lowpass anti-aliasing filter to the demodulated channels:

$$|W^2|x = \{x_i^1{*}\phi_2(\Delta_2 n) \,,\; |x_i^1{*}\psi_{2,\lambda_2}|{*}\phi_2(\Delta_2 n)|\}_{i=1\ldots|W^1|} \,, \tag{5}$$

where Δ_2 is this time the critical sampling rate of the averaging filter ϕ_2. The multiscale variations of each envelope specify the amplitude modulations of $x(t)$ [16] and thus have the capacity to detect rhythmic structures appearing at different frequency bands. The Q-factor Q_2 of the second family of wavelets ψ_{2,λ_2} controls the time-frequency resolution of the transform. Since the envelopes $|x*\psi_{\lambda_1}|$ have bandwidth $\sim 2^{-j}Q_1^{-1}$, one typically chooses dyadic $Q_2 = 1$ second order wavelets. Scattering coefficients have a negligible amplitude for $\lambda_2 > \lambda_1$ because $|x*\psi_{\lambda_1}|$ is a regular envelop whose frequency support is below λ_2 [17]. Scattering coefficients are thus computed only for $\lambda_2 < \lambda_1$.

Scattering transforms have been extended along the frequency variables to capture the joint time-frequency variability of spectral envelopes and therefore provide representations locally stable to pitch variations [16]. We denote $\gamma = \log_2 \lambda_1$, and consider the scalogram as a two-dimensional function of γ and t:

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

In this work, we consider a second layer scattering with a separable wavelet transform $F * \overline{\psi}_{\gamma_2,\lambda_2}(\gamma,t)$, with

$$\overline{\psi}_{\gamma_2,\lambda_2}(\gamma,t) = \widetilde{\psi}_{\gamma_2}(\gamma)\psi_{2,\lambda_2}(t)\ ,\ \overline{\psi}_{0,\lambda_2}(\gamma,t) = \widetilde{\phi}(\gamma)\psi_{2,\lambda_2}(t)$$

$$\overline{\psi}_{\gamma_2,0}(\gamma,t) = \widetilde{\psi}_{\gamma_2}(\gamma)\phi_2(t)\ ,\ \overline{\phi}(\gamma,t) = \widetilde{\phi}(\gamma)\phi_2(t)$$

By replacing ψ and ϕ in (5) by $\overline{\psi}$ and $\overline{\phi}$ respectively, we obtain the joint scattering pyramid transform. In this implementation, we choose temporal wavelets $\psi(t)$ to be dyadic complex Morlet wavelets, and $\widetilde{\psi}$ to be dyadic real Haar wavelets to preserve good frequency localization, with no frequency downsampling.

We can reapply the same operator as many times k as desired until reaching a temporal context $T = \Delta_k$, but in this implementation we demonstrate the method with $k \leq 2$. If the wavelet filters are chosen such that they define a non-expansive mapping [17], it results that every layer defines a metric which is increasingly contracting:

$$\||W^k|x - |W^k|x'\| \leq \||W^{k-1}|x - |W^{k-1}|x'\| \leq \|x - x'\|\ .$$

Every layer thus produces new feature maps at a lower temporal resolution. In the end we obtain a tree of different representations,

$$\Phi_j(x) = |W^j|x, \quad \text{with} \quad j = 1,\dots,k.$$

4. SOURCE SEPARATION ALGORITHM

We show in this section how the inverse problem of source separation can be solved via a sparse NMF in the pyramid scattering domain, followed by phase recovery. We consider the supervised monoaural source separation problem (1), in which the components x_i, $i = 1,2$ come from sources for which we have training data $X_i = \{x_{ij}\}_{j\leq K}$, and one is asked to produce estimates $\widehat{x}_i$.

Let us consider for simplicity the features $\Phi_j(x_i)$, $j = 1,2$, $i = 1,2$, $x_i \in X_i$, obtained by localizing the scattering features of two different resolutions at their corresponding sampling rates. Therefore, Φ_1 is equivalent to a CQT and carries more localized information than Φ_2. On the other hand, Φ_2 is stable at representing longer temporal contexts. We train independent models for each source at each resolution by solving,

$$\min_{D_j^i \geq 0} \sum_{x\in X_i} \min_{z\geq 0} \frac{1}{2}\|\Phi_j(x) - D_i^j z\|^2 + \lambda_j\|z\|_1\ .$$

where D_i^j represents the non-negative dictionary of source i at resolution j. This model exploits the linearization properties of scattering coefficients since it searches low-dimensional linear approximations.

At test time, given and input x, x_1 and x_2 are estimated by minimizing

$$\min_{x_i', z_i} \sum_{j=1,2}\sum_{i=1,2} \frac{1}{2}\|\Phi_j(x_i') - D_i^j z_i^j\|_2^2 + \lambda\|z_i^j\|_1\ \ s.t.\ x = x_1' + x_2'\ . \tag{6}$$

Problem (6) is minimized with an alternating gradient descent between x_i' and z_i^j. Fixing the z_i^j variables and minimizing with respect to x_i' requires locally inverting the scattering operators Φ_j, which amounts to solve an overcomplete phase recovery problem and can be solved with gradient descent, as shown in [20]. Fixed x_i', solving for z_i^j are four independent ℓ_1 non-negative sparse coding problems, which can be solved efficiently. In this work, we use the SPAMS package [28]. Note that unlike standard NMF problems, the optimization in 6 is carried out directly on the temporal signal.

When the analysis operators Φ_j are able to produce sparse representations of the sources, then at each level we have

$$\sum_{i=1,2} \|\Phi_j(x_i') - D_i^j z_i^j\|_2^2 \approx \|\Phi_j(x) - D_1^j z_1 - D_2^j z_2\|_2^2\ , \tag{7}$$

which can be used in practice to produce a greedy initialization for (6) as follows. We obtain an estimator $\widehat{\Phi_1(x_i)} = D_i^1 z_i^{1*}$, where the z_i^{1*} are defined as

$$z_i^{1*} = \arg\min_{z_i\geq 0} \frac{1}{2}\|\widehat{\Phi(x)} - \sum_{i=1,2} D_i^1 z_i\|_2^2 + \lambda\|z_i\|_1\ .$$

We can produce an estimate $\widehat{x}_i$ from $\widehat{\Phi_1(x_i)}$ by using the complex phases of $W_1 x$. This can be thought as running standard NMF with CQT features.

In the proposed framework, the decomposition needs to be coherent at both levels of temporal resolution. The first level representation is well located temporally and allows for a high resolution rerepentation of the signals, while the deeper representations can be thought as a regularizer imposing temporal coherence. In the deep representation, intra-class variability given by small pitch and timber variations is linearized up to temporal scales 2^J without loosing as much discriminative information as the spectrogram [16, 17].

5. EXPERIMENTAL RESULTS

Evaluation settings. We evaluated the proposed method in two settings: speaker-specific and multi-speaker. In the first setting we trained a speaker-specific model for each speaker in the mixture and tested it using sentences (from the same speakers) outside the training set. In the second setting, we trained a generic model on a mixed group of male and female speakers, none of which were included in the test set. All signals where mixed at 0 dB and clips resampled to 16 KHz.

Data sets. We used a subset of speakers (3 males and 3 females) of the GRID dataset [29] for evaluating the speaker-specific setting. For each speaker, 500 randomly-chosen clips were used for training (around 25 minutes) and 200 clips were used for testing. For the multi-speaker case we used a subset of the TIMIT dataset. We adopted the standard test-train division, using all the training recordings for bulding the models and a subset of 12 different speakers (6 males and 6 females) for testing. For each speaker we randomly chose two clips and compared all female-male combinations (144 mixtures).

Evaluation measures. We used the *source-to-distortion ratio* (SDR), *source-to-interference ratio* (SIR), and *source-to-artifact ratio* (SAR) from the BSS-EVAL metrics [30].

Training setting. We evaluated the proposed scattering NMF model with pyramids of one and two layers, reffered as *scatt-NMF$_1$* and *scatt-NMF$_2$* respectively. As a baseline we used standard NMF with frame lengths of 1024 samples and 50% overlap. The dictionaries in standard NMF were chosen with 200 and 400 atoms for the speaker-specific and multi-speaker settings respectively. These values were obtained using cross-validation on a few clips separated from the training as a validation set. In all cases, we applied *scatt-NMF* using a scattering transforms with resolution $Q_1 = 32$ and $Q_2 = 1$. The resulting representation had 175 coefficients for

	Speaker-Specific			Multi-Speaker		
	SDR	SIR	SAR	SDR	SIR	SAR
NMF	5.6 [1.8]	13.4 [2.8]	6.9 [1.3]	6.1 [2.9]	14.1 [3.8]	7.4 [2.1]
scatt-NMF$_1$	8.6 [1.7]	**16.8** [3.3]	9.6 [1.4]	6.2 [2.8]	13.5 [3.5]	7.8 [2.2]
scatt-NMF$_2$	**8.9** [1.5]	**16.8** [2.4]	**9.9** [1.3]	**6.9** [2.7]	**16.0** [3.5]	**7.9** [2.2]

Table 1: Separation with speakers-specific and multi-speaker settings. Average SDR, SIR and SAR (in dB) for NMF and proposed and *scatt-NMF$_2$*. Standard deviation of each result shown between brackets.

	SDR	SIR	SAR
NMF-KL	5.4	7.3	7.8
RNN [25]	6.0	8.1	**8.1**
RNN joint disc. training [25]	**7.4**	**11.8**	7.5
scatt-NMF$_2$	6.7	11.1	6.9

Table 2: Comparison RNN based separation, with and without joint discriminative training with soft masks [25].

the first level and around 2000 for the second layer. For the single speaker case we trained dictionaries with 200 atoms for *scatt-NMF$_1$* and 800 atoms for *scatt-NMF$_2$*. While for the multi-speaker case we used 400 atoms for *scatt-NMF$_1$* and 1000 atoms for *scatt-NMF$_2$*. In all cases, the features were frame-wise normalized and we used $\lambda = 0.1$.

Results. Table 1 shows the results obtained for the speaker-specific and multi-speaker settings. [1] In all cases we observe that the one layer scattering transform outperforms the STFT in terms of SDR. These results go in line with the ones obtained in [24], showing that the choice of stable features has a strong effect in the performance. Furthermore, there is a tangible gain in including a deeper representation; *scatt-NMF$_2$* performs always better than *scatt-NMF$_1$*. As expected, the results obtained with the speaker-specific setting are better than those of the more challanging problem of multi-speaker setting.

We also compared the proposed approach with the speaker-specific setting discussed in [25]. In this work the authors investigate several alternatives of using Recurrent Neural Networks (RNN) for speech separation. Several optimization settings are evaluated on two given speakers of the TIMIT dataset, some of which aim at learning short-term temporal dynamics. This is a very challenging setting due to the very small available training data (less than 10 seconds per speaker). The evaluations of *scatt-NMF$_2$* were performed using the setting provided in [25] (with the corresponding training, developement and testing data) while their results are taken from the paper. *scatt-NMF$_2$* outperforms the benchmark KL-NMF in SDR and SIR, and is competitive with the best performing networks reported in [25], with and without joint discriminative training, see Table 2.

In summary, these results confirm that inverse problems such as speech source separation can benefit from the properties of stable and highly discriminative non-linear representations, such as scattering operators. Sparse inference is able to extract more relevant information thanks to the stability to time-frequency deformations, while the phase recovery can still be efficiently performed with gradient descent.

[1] Audio samples are available at `www.cims.nyu.edu/~bruna/scatt_source_separation`.

6. DISCUSSION

NMF-based audio source separation techniques can be thought as applying a synthesis operator on a feature space given by a pooled analysis operator. Leveraging recent developemnts in signal processing, we propose to substitute the first stage with a deep scattering transform. The obtained features are designed to capture the joint time-frequency variability of speech signals and efficiently represent a longer temporal context. Experimental evaluation shows that using deeper representations leads to a tangible improvement in performance in challenging source separation settings. A natural extension of this work is to investigate the use of learned representations instead, or on top of, the designed ones. We expect further gains by applying discriminative dictionary learning [23]. Future work includes testing more thoroughly the potential of the proposed model in combination with convolutional neural networks, which have been very successful in other signal and image processing problems.

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] 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.

[4] P. Smaragdis, C. Fevotte, G Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *Signal Processing Magazine, IEEE*, vol. 31, no. 3, pp. 66–75, 2014.

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

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

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

[8] N. Mohammadiha, P. Smaragdis, and A. 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.

[9] Paris Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 1–12, 2007.

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

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

[12] 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.

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

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

[15] R. W. Gerchberg and W. Owen Saxton, "A practical algorithm for the determination of the phase from image and diffraction plane pictures," *Optik*, vol. 35, pp. 237–246, 1972.

[16] J. Andén and S. Mallat, "Deep scattering spectrum," *arXiv preprint arXiv:1304.6763*, 2013.

[17] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1872–1886, 2013.

[18] J. Bruna, *Scattering Representations for Recognition*, Ph.D. thesis, Palaiseau, Ecole polytechnique, 2013.

[19] G. Wolf, S. Mallat, and S. Shamma, "Audio source separation with time-frequency velocities," *International Workshop on Machine Learning for Signal Processing*, 2014.

[20] J. Bruna and S. Mallat, "Audio texture synthesis with scattering moments," *arXiv preprint arXiv:1311.0407*, 2013.

[21] R. Jenatton, R. Gribonval, and F. Bach, "Local stability and robustness of sparse dictionary learning in the presence of noise," *arXiv preprint arXiv:1210.0685*, 2012.

[22] Felix Weninger, Jonathan Le Roux, John R Hershey, and Shinji Watanabe, "Discriminative nmf and its application to single-channel source separation," *Proc. of ISCA Interspeech*, 2014.

[23] P. Sprechmann, A. M. Bronstein, and G. Sapiro, "Supervised non-euclidean sparse NMF via bilevel optimization with applications to speech enhancement," in *HSCMA*. IEEE, 2014, pp. 11–15.

[24] Felix Weninger, Jonathan Le Roux, John R Hershey, and Björn Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. IEEE GlobalSIP 2014 Symposium on Machine Learning Applications in Speech Processing*, 2014.

[25] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *ICASSP*, 2014, pp. 1562–1566.

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

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

[28] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *ICML*, 2009, pp. 689–696.

[29] 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.

[30] 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.