AUDIO SOURCE SEPARATION BASED ON CONVOLUTIVE TRANSFER FUNCTION AND FREQUENCY-DOMAIN LASSO OPTIMIZATION

Xiaofei Li[1], *Laurent Girin*[1,2] *and Radu Horaud*[1]

[1]INRIA Grenoble Rhône-Alpes, [2]GIPSA-Lab & Univ. Grenoble Alpes

ABSTRACT

This paper addresses the problem of under-determined convolutive audio source separation in a semi-oracle configuration where the mixing filters are assumed to be known. We propose a separation procedure based on the convolutive transfer function (CTF), which is a more appropriate model for strongly reverberant signals than the widely-used multiplicative transfer function approximation. In the short-time Fourier transform domain, source signals are estimated by minimizing the mixture fitting cost using Lasso optimization, with a ℓ_1-norm regularization to exploit the spectral sparsity of source signals. Experiments show that the proposed method achieves satisfactory performance on highly reverberant speech mixtures, with a much lower computational cost compared to time-domain dual techniques.

Index Terms— Source separation, convolutive transfer function, ℓ_1-norm regularization

1. INTRODUCTION

In this paper we address the problem of multichannel audio source separation (MASS) from, possibly underdetermined, convolutive mixtures (the number of sensors is lower than the number of sources). This problem is often divided into two subproblems, which are both difficult: identification of mixing filters and estimations of source signals. This paper focuses on audio source estimation assuming that the mixing filters are either known or an estimation is available.

Most of convolutive source separation techniques are designed in the short time Fourier transform (STFT) domain. In this domain, the convolutive process is assumed to be well approximated at each TF bin by a product between the source STFT coefficient and the Fourier transform of the mixing filter. This assumption is called the multiplicative transfer function (MTF) approximation [1], or the narrowband approximation. Moreover, the sparsity of the audio signals in the time-frequency (TF) domain is a desirable effect. Based on these properties, the separation methods rely on binary masking of the mixture STFT bins, e.g. [2, 3], on the ℓ_1-norm minimization, e.g. [4], on probabilistic models for source signals, e.g. [5, 6], or on a combination of these methods.

The MTF approximation is theoretically valid only if the length of the mixing filter impulse response is smaller than the length of the STFT window. In practice, this is very rarely the case, even for moderately reverberant environments, since the STFT window is limited to assume local stationarity of audio signals. Hence the MTF can be a poor approximation, fundamentally endangering the separation performance; this becomes critical for strongly reverberant environments. Yet, the MTF is poorly questioned in the MASS literature, and only a few studies attempted to tackle its limitations. In [6], the use of a full-rank spatial covariance matrix for the source images, instead of the rank-1 matrix corresponding to the MTF model [5], is claimed to overcome to some extent the limitations of MTF. A more direct approach to the problem is proposed in [7] where the source signals are estimated in the time domain by minimizing a wide-band ℓ_2-norm mixture-fitting cost, in which the exact source-filter convolution is used, using a Lasso optimization technique. [8] improved this wide-band Lasso (W-Lasso) technique by a re-weighted scheme. The W-Lasso technique achieves quite good source separation performance in reverberant environments, at the price of a tremendous computation time.

To represent convolution more accurately in the STFT domain, especially for the long filter case, cross-band filters (CBFs) were introduced in [9] in the context of linear system identification, as an alternative to MTF. Using the CBFs, an output STFT coefficient is represented as a summation over frequency bins of multiple convolutions between the input STFT coefficients and the TF-domain filter impulse response, along the frame coordinate. A convolutive transfer function (CTF) approximation is further introduced in [10] to simplify the analysis. Here, at each frequency, the output STFT coefficient is modeled as a (unique) convolution of the input STFT coefficients and the CTF, along the frame axis. The CBFs were recently considered for solving MASS [11], in combination with a high-resolution non-negative matrix factorization model of the source signal. A variational EM algorithm was proposed to estimate the filters and infer the source signals. Unfortunately, this method was observed to perform well only in a semi-blind setup where both filters and source parameters are initialized from the individual source images.

This research has received funding from the ERC Advanced Grand VHIA (#340113) and from EU-FP7 STREP project EARS (#609465).

In this paper we propose to use the CTF for MASS quite differently than what was proposed in the past. Following the spirit of W-Lasso, the source signals are estimated, at each frequency, by minimizing a CTF-based ℓ_2-norm mixture fitting cost in the STFT domain. In addition, we add to the fitting cost a ℓ_1-norm regularizer such as to exploit the sparsity of TF-domain audio signals along the frames. By circumventing the MTF approximation, the proposed method achieves satisfactory source separation performance in reverberant environments, likewise W-Lasso, but since the STFT frame level is considered instead of the time-domain sample level, separation is obtained at a much lower computational burden, as shown by our experiments. Another potential advantage of the proposed method is that, compared to the time-domain mixing filters in W-Lasso, it may be easier to identify the CTF mixing filters from the mixture signals based on the TF-domain sparsity.

The rest of this paper is organized as follows. Section 2 presents the CTF model. The proposed source separation method is given in Section 3. Experiments are presented in Section 4. Section 5 concludes the paper.

2. CONVOLUTIVE TRANSFER FUNCTION

In a reverberant and noise-free environment, a source image $y(n)$ is, in time domain, given by

$$y(n) = a(n) \star s(n) \tag{1}$$

where $s(n)$ and $a(n)$ are the source signal and the impulse response of the propagating filter, respectively, and $\star$ denotes linear convolution. With the usual MTF approximation, (1) is approximated in the STFT domain as

$$y_{p,k} = a_k s_{p,k} \tag{2}$$

where $y_{p,k}$ and $s_{p,k}$ are the STFT of the corresponding signals, and a_k is the Fourier transform of the filter $a(n)$, $p \in [1, P]$ is the frame index, N is the frame length, and $k \in [0, N-1]$ is the frequency bin index. As discussed above, this approximation is only valid when the filter $a(n)$ is shorter than the STFT window, which is often questionable. In this paper we therefore use the CTF model, i.e. $y(n)$ is approximated in the STFT domain by:

$$y_{p,k} = \sum_{p'} a_{p',k} s_{p-p',k} = a_{p,k} \star s_{p,k}. \tag{3}$$

The filter CTF, i.e. the TF-domain impulse response $a_{p',k}$, is related to the time-domain impulse response $a(n)$ by:

$$a_{p',k} = (a(n) \star \zeta_k(n))|_{n=p'L}, \tag{4}$$

which represents the convolution with respect to the time index n evaluated at multiples of the frame step L, with

$$\zeta_k(n) = e^{j\frac{2\pi}{N}kn} \sum_{m=-\infty}^{+\infty} \overline{\omega}(m)\, \omega(n+m), \tag{5}$$

where $\overline{\omega}(n)$ and $\omega(n)$ denote the STFT analysis and synthesis windows, respectively. The CTF can be interpreted as: the time-domain convolution is transformed into a TF-domain convolution with a certain approximation error.

3. SEMI-BLIND SOURCE SEPARATION BASED ON CTF

3.1. Mixture model and source separation formulation

We consider a multi-channel underdetermined convolutive mixture with J sources and I sensors ($I < J$). Based on the CTF model (3), in the STFT domain the sensor signals $x^i_{p,k}, i \in [1, I]$ are given by:

$$x^i_{p,k} = \sum_{j=1}^{J} y^j_{p,k} + e^i_{p,k} = \sum_{j=1}^{J} a^{i,j}_{p,k} \star s^j_{p,k} + e^i_{p,k}, \tag{6}$$

where $a^{i,j}_{p,k}$ is the CTF from source j to sensor i, and $e^i_{p,k}$ denotes the noise signal. For frequency k, let $\mathbf{x} \in \mathbb{C}^{I\times P}$, $\mathbf{s} \in \mathbb{C}^{J\times P}$ and $\mathbf{e} \in \mathbb{C}^{I\times P}$ denote the matrices of sensor signals, source signals and noise signals, respectively, and let $\mathbf{A} \in \mathbb{C}^{I\times J\times P}$ denote the three-way CTF array. Since the proposed algorithm is frequency-wise, the frequency index k is omitted from now on. Then (6) can be written as:

$$\mathbf{x} = \mathbf{A} \star \mathbf{s} + \mathbf{e}. \tag{7}$$

For underdetermined mixtures, the source signals cannot be estimated by inverting the filters. Instead, we estimate the source signals by minimizing a ℓ_2-norm mixture fitting cost. Moreover, the TF-domain sparsity of the speech sources is enforced using an ℓ_1-norm regularization term. Overall, the source separation is carried out by solving the following convex optimization problem:

$$\min_{\mathbf{s}\in\mathbb{C}^{J\times P}} \parallel \mathbf{A} \star \mathbf{s} - \mathbf{x} \parallel_2^2 + \lambda \parallel \mathbf{s} \parallel_1, \tag{8}$$

where the free parameter λ is chosen as a trade-off between the fitting term and the regularization term. Note that both the ℓ_2 and ℓ_1 norms on matrices in (8) are defined here as vector norms. As mentioned in the introduction, this problem was considered in [7] in the time-domain, as a "wide-band" (convolutive) version of the Lasso problem [12] or of the basis pursuit denoising problem [13]. [7] involved a large data size, leading to a low optimization convergence rate and a huge computational cost. Applying the "convolutive" Lasso to MASS in the STFT domain, thanks to the CTF model proposed here, enables to considerably reduce the computational burden and to directly exploit the STFT sparsity of audio signals.

3.2. Optimization using FISTA

To solve the optimization problem (8), we adopt the fast iterative shrinkage-thresholding algorithm (FISTA) [14], as al-

ready done in [7] in the time-domain, with the notable difference that here the optimization process is carried on the complex domain. Let $\mathcal{F}(\mathbf{s}) = \| \mathbf{A} \star \mathbf{s} - \mathbf{x} \|_2^2$ denote the fitting cost function, which is L-Lipschitz differentiable. Its derivative is

$$\Delta\mathcal{F}(\mathbf{s}) = \tilde{\mathbf{A}} \star (\mathbf{A} \star \mathbf{s} - \mathbf{x}), \tag{9}$$

where the adjoint matrix $\tilde{\mathbf{A}}$ is obtained by conjugate transposing the source and channel indices, and then temporally reversing the filters.

Similar to [7], the Lipschitz constant L is computed using the power iteration algorithm, which is summarized in Algorithm 1. In this work, $\mathbf{v}$ is initialized as a matrix composed of I replications of the first channel sensor signal. The convergence criterion consists in testing if the difference in amplitude of the inner product between the values of $\mathbf{v}$, at the current and previous iterations, is larger than a threshold δ.

The ℓ_1-norm regularization term in (8) is a lower semi-continuous function, non-differentiable at 0. The proximity operator of the function $\gamma \| \cdot \|_1$ at point $\mathbf{z}$, aka the shrinkage operator, is defined as:

$$\text{Prox}_{\gamma\|\cdot\|_1}(\mathbf{z}) = \underset{\mathbf{y}}{\text{argmin}} \frac{1}{2} \| \mathbf{z} - \mathbf{y} \|_2^2 + \gamma \| \mathbf{y} \|_1 . \tag{10}$$

The closed-form entrywise solution is given by:

$$y_i = \frac{z_i}{|z_i|} \max(0, |z_i| - \gamma). \tag{11}$$

Based on the derivative operator (9), on the Lipschitz constant L, and on the shrinkage operator, FISTA is summarized in Algorithm 2. $\mathbf{s}_0$ is also initialized as the matrix composed of I replication of the first channel sensor signal. The convergence criterion is chosen as a threshold ϵ over the relative decrease of the objective cost in (8) at each iteration.

Algorithm 1 Power iteration.

Input: $\mathbf{A}$, $\tilde{\mathbf{A}}$
Initialization: $\mathbf{v} \in \mathbb{C}^{I \times P}$
repeat
  $\mathbf{w} = \tilde{\mathbf{A}} \star (\mathbf{A} \star \mathbf{v})$
  $\mathbf{v} = \mathbf{w} / \| \mathbf{w} \|_2$
until convergence
$L = \| \mathbf{w} \|_2$

4. EXPERIMENTAL EVALUATION

To test the efficiency of the proposed source separation method, experiments were conducted with simulated binaural signals, under various acoustic conditions. Binaural room impulse responses (BRIR) were generated with the ROOMSIM simulator [15] and with the head related impulse

Algorithm 2 FISTA

Input: $\mathbf{x}$, $\mathbf{A}$, $\tilde{\mathbf{A}}$, L
Initialization: $k = 1$, $\mathbf{s}_0 \in \mathbb{C}^{I \times P}$, $\mathbf{z}_0 = \mathbf{s}_0$, $t_0 = 1$
repeat
  1. $\Delta\mathcal{F}(\mathbf{s}_{k-1}) = \tilde{\mathbf{A}} \star (\mathbf{A} \star \mathbf{s}_{k-1} - \mathbf{x})$
  2. $\mathbf{s}_k = \text{Prox}_{(\lambda/L)\|\cdot\|_1}(\mathbf{z}_{k-1} - \Delta\mathcal{F}(\mathbf{s}_{k-1})/L)$
  3. $t_k = (1 + \sqrt{(1 + 4t_{k-1}^2)})/2$
  4. $\mathbf{z}_k = \mathbf{s}_k + ((t_{k-1} - 1)/t_k)(\mathbf{s}_k - \mathbf{s}_{k-1})$
  5. $k = k + 1$
until convergence

response (HRIR) of a KEMAR dummy head [16]. The size of the room was 8 m $\times$ 5 m $\times$ 3 m. The KEMAR dummy head was located at (4 m, 1 m, 1.5 m). Speech signals from the TIMIT dataset [17] and then sampled at 16 kHZ were used as sources convolved with the simulated BRIRs to generate sensor signals. The speech sources were located at 1 m away from the dummy head with azimuth directions varying from -90° to 90°, spaced by 5°, and at an elevation of 0°. Three reverberation times were tested, namely $T_{60} = 0.22$ s, 0.5 s and 0.79 s. Moreover, the anechoic case was also tested. A set of underdetermined mixtures with 3, 4 and 5 sources were processed. For each experiment, 50 mixtures were generated. The STFT window was a Hamming window of 512 samples (32 ms), with 50% overlap. The free parameter λ was set at a fixed value of 10^{-3} through the whole set of experiments, as it was shown to be suitable for all the tested conditions. The power iteration convergence threshold δ was set to 0.999. The FISTA convergence threshold was set to 10^{-6}. The CTF coefficients were computed from the known filter impulse responses using (4).

For comparison, we tested four state-of-the-art source separation methods, all using in the same semi-blind configuration as the proposed method (i.e. with known mixing filters): the degenerate unmixing estimation technique (DUET) [2], which is based on time-frequency masking, assuming only a single source is active in each TF bin, the ℓ_1-norm minimization method (ℓ_1-MIN) [4], which assumes that at most I sources are active in each TF bin, the full-rank spatial covariance matrix (FR-SCM) method of [6], and the wide-band Lasso (W-Lasso) method of [7] with ℓ_1-norm regularization term on source STFT coefficients, and trade-off factor set to 10^{-5}. DUET and ℓ_1-MIN methods are based on the MTF approximation, i.e. the instantaneous mixing matrix in each frequency bin is employed. Since the BRIRs are longer than the STFT window, they have to be truncated to generate the mixing matrix with the Fourier transform. However we obtained better results using the Fourier transform of the HRIRs. For FR-SCM, the SCMs were individually estimated using each separate image source signal, following the line of the semi-oracle experiments in [6]. Then an EM was applied with the SCMs being kept fixed to the semi-oracle

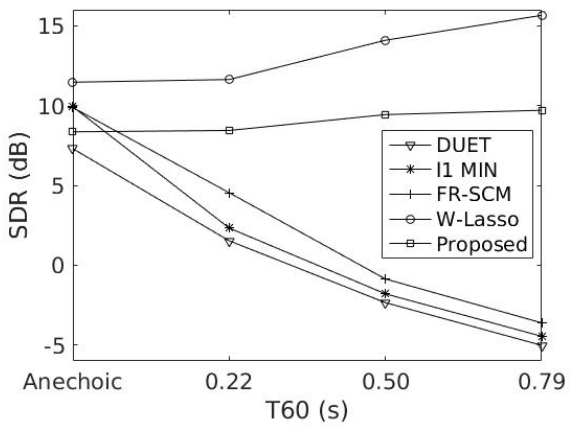

Fig. 1: Source separation performance (SDR) for 3-source mixtures as a function of reverberation time.

Table 1: Computation times, in seconds, for a 3-source mixture and for various reverberation times. All the algorithms were implemented in Matlab.

	T_{60} (s)			
Methods	Anechoic	0.22	0.50	0.79
DUET [2]	0.01	0.02	0.02	0.02
ℓ_1 MIN [4]	25.1	23.5	24.9	24.6
FR-SCM [6]	345	361	373	396
W-Lasso [7]	2694	2810	2975	3232
Proposed	18.1	22.2	28.1	32.9

values. Note that we did not compare our method with [11], since this method involves a specific rank-1 source model that is poorly appropriate for speech signals.

The signal-to-distortion ratio (SDR) [18] in decibels (dB), averaged over 50 mixtures for each condition, is used as the separation performance metric (only 5 mixtures are tested for W-Lasso due to its high computation time). Fig. 1 plots the SDR obtained for 3-source mixtures and for the 3 reverberation times. It can be seen in these plots that all five methods achieve high SDR in the anechoic case. DUET performs the worst, due to its limiting assumption that only a single source is active at each TF bin. ℓ_1-MIN and FR-SCM perform better by assuming that more sources can co-exist in a TF bin. W-Lasso achieves the highest SDR, thanks to the exact (time-domain) convolution model. The proposed method achieves a lower SDR than W-Lasso due to the CTF approximation error. As the reverberation time increases, the SDRs of DUET, of ℓ_1-MIN and to a least extent of FR-SCM, dramatically decrease. For DUET and for ℓ_1-MIN, the MTF approximation is no longer suitable when the filter impulse response is (much) longer than the STFT window. FR-SCM mitigates the problem by using a full-rank spatial covariance matrix, which models the reverberations, although to a limited extend. In contrast to these three methods, both W-Lasso and the proposed method achieve remarkably stable performances: the SDR actually increases with T_{60}, which is a bit surprising at first sight. This can be explained as follows: (i) the mixture models in W-Lasso and in the proposed method fit the actual mixture better than the MTF, and (ii) given a good fit of the mixture model, the longer the filter, the more information availalble to discriminate and separate different sources. Again, due to the CTF approximation error, the proposed method performs worse than W-Lasso by 2 to 5 dB, depending on T_{60}.

Table 2: Source separation performance (SDR in dB) for various number of sources ($T_{60} = 0.5$ s).

	Number of sources		
Methods	3	4	5
DUET [2]	-2.35	-4.54	-5.64
ℓ_1 MIN [4]	-1.79	-3.56	-4.79
FR-SCM [6]	-0.86	-2.50	-4.75
W-Lasso [7]	13.87	7.71	5.58
Proposed	9.43	5.94	4.46

Table 1 shows the average computation time needed to process one mixture (averaged over the 50 test mixtures; the average mixture duration is about 4 s) for each method and each tested T_{60}. We can see in this table that the W-Lasso method is much more time-consuming than the others. DUET is the fastest. The computation time of FR-SCM is about 12% of the computation time of W-Lasso. The computation time of the proposed method and of ℓ_1-MIN is comparable, and is less than 1% of the computation time of W-Lasso.

Finally, Table 2 displays the SDR for various number of sources, for $T_{60} = 0.5$ s. As expected, the SDR of all five methods degrades when the source number increases. W-Lasso has the fastest degradation with increasing number of sources, whereas the proposed method seems more robust, so that for 5 sources, W-Lasso achieves only about 1.1 dB SDR improvement over the proposed method.

5. CONCLUSION

In this paper, we have proposed a semi-blind source separation method based on the CTF model and STFT-domain Lasso optimization. Overall, the proposed method and the time-domain W-Lasso perform prominently better than DUET, ℓ_1-MIN and FR-SCM. The proposed method drastically improves the computation efficiency over W-Lasso with an acceptable decrease in separation performance. It thus seems to be an excellent trade-off between the fast but inaccurate MTF model, and the exact but highly greedy time-domain W-Lasso, especially for filters that are much longer than the STFT window. Recently, it was proposed a CTF-based method to extract the direct path of a single source [19]. Based on the STFT-domain sparsity, the extension to multiple sources is under development, which altogether will enable the development of a CTF-based blind source separation method.

6. REFERENCES

[1] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, 2007.

[2] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on signal processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[3] M. I. Mandel, R. J. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[4] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 81–81, 2007.

[5] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[6] N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[7] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1818–1829, 2010.

[8] S. Arberet, P. Vandergheynst, J-P. Carrillo, R. E.Thiran, and Y. Wiaux, "Sparse reverberant audio source separation via reweighted analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1391–1402, 2013.

[9] 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.

[10] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

[11] R. Badeau and M. D. Plumbley, "Multichannel high-resolution NMF for modeling convolutive mixtures of non-stationary signals in the time-frequency domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 11, pp. 1670–1680, 2014.

[12] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[13] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[14] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[15] D. R. Campbell, "The ROOMSIM user guide (v3.3)," 2004.

[16] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR dummy-head microphone," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995.

[17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," Tech. Rep. 107, National Institute of Standards and Technology, Gaithersburgh, MD, 1988.

[18] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 11, pp. 2171–2186, 2016.