BLASTER: An Off-Grid Method for Blind and Regularized Acoustic Echoes Retrieval – with supplementary material

Diego Di Carlo, Clément Elvira, Antoine Deleforge, Nancy Bertin, Rémi Gribonval

BLASTER: AN OFF-GRID METHOD FOR BLIND AND REGULARIZED ACOUSTIC ECHOES RETRIEVAL WITH SUPPLEMENTARY MATERIALS

Diego Di Carlo[†⋆], *Clément Elvira*[†⋆], *Antoine Deleforge*[‡], *Nancy Bertin*[†] *and Rémi Gribonval*[†§]

[†] Univ Rennes, Inria, CNRS, IRISA, France
[‡] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[§] Univ Lyon, Inria, CNRS, ENS de Lyon, UCB Lyon 1, LIP UMR 5668, F-69342, Lyon, France
[⋆] *authors contributed equally*

ABSTRACT

Acoustic echoes retrieval is a research topic that is gaining importance in many speech and audio signal processing applications such as speech enhancement, source separation, dereverberation and room geometry estimation. This work proposes a novel approach to blindly retrieve the *off-grid* timing of early acoustic echoes from a stereophonic recording of an unknown sound source such as speech. It builds on the recent framework of continuous dictionaries. In contrast with existing methods, the proposed approach does not rely on parameter tuning nor peak picking techniques by working directly in the parameter space of interest. The accuracy and robustness of the method are assessed on challenging simulated setups with varying noise and reverberation levels and are compared to two state-of-the-art methods.

***Index Terms*—** Blind Channel Identification, Super Resolution, Sparsity, Acoustic Impulse Response.

1. INTRODUCTION

In room acoustics and audio signal processing, the temporal structure of the room impulse response (RIR) plays a central role. It is the result of multiple (indirect) sound propagation paths due to specular and diffuse reflections on the room's surfaces, leading to reverberation [1]. In such conditions, the perceived sound quality is often considered degraded and it is common to observe a detrimental decrease of performance as reverberation increases for applications such as speech recognition [2] or music information retrieval [3].

On the other hand, RIRs contain very rich geometrical information about the acoustic scene. Recent *echo-aware* works have shown that the knowledge of the timing of early reflections may boost performance in many audio signal processing applications, from dereverberation [4, 5] to sound localization [6, 7] and separation [8, 9]. Moreover, it allows joint estimation of the receivers' positions [10], the reflective surfaces [11] and consequently the geometry of the room [12, 13].

Acoustic echo retrieval (AER) consists in estimating the properties of the early (strong) acoustic reflections only in multi-path environments [14], sometimes referred to as time delay estimation [15]. To achieve this, several methods rely on a known source signal [16, 17]. In contrast, when multiple receivers attend an unknown single source, AER can be seen as an instance of Single Input Multiple Output (Blind) Channel Estimation (SIMO-BCE) problem. A common approach for solving AER in the context of SIMO-BCE is to first blindly estimate a discrete version of the acoustic channels using the so-called cross-relation identity [18, 19]. The location of the echoes are then chosen among the strongest peaks with ad-hoc peak-picking techniques. However, in practice, the true timings of echoes rarely match the sampling grid, thus leading to pathological issues called basis-mismatch in the field of compressed sensing. To circumvent this issue, the authors of [14] proposed to leverage the framework of finite-rate-of-innovation sampling to make one step towards off-grid approaches. Despite promising results in the absence of noise and with synthetic data, the quality of the estimation highly relies on an initialization point.

Of particular interest in this paper is the recently proposed framework of continuous dictionaries (CD) [20]. By formulating an inverse problem as the recovery of a discrete measure over some parameter space, CD has allowed to overcome imaging device limitations in many applications such as super-resolution [20] or PALM/STORM imaging [21]. In this work, we formulate the problem of stereo AER within the framework of continuous dictionaries. The resulting optimization problem is convex and thus not prone to spurious minimizers. The proposed method is coined *Blind And Sparse Technique for Echo Retrieval* (BLASTER) and requires no parameter tuning. The method is compared to state-of-the art on-grid approaches under various noise and reverberation levels using simulated data. While comparable or slightly worse recovery rates are observed for the task of recovering 7 echoes or more, better results are obtained for fewer echoes and the off-grid nature of the approach yields generally smaller estimation errors.

2. BACKGROUND IN ACOUSTIC ECHO ESTIMATION

2.1. Signal and measurement model

Consider the common setup where a band-limited and square-integrable source signal s is emitted. Due to the geometry of the room, the latter signal is both reflected (several times) and attenuated before reaching a set of two microphones. The recorded signal at

The research presented in this paper is reproducible. Code and data are available at https://gitlab.inria.fr/panama-team/blaster microphone $i \in \{1,2\}$ reads

$$x_i = s * h_i^\star + n_i \tag{1}$$

where $*$ denotes the (continuous) convolution operator, n_i models some additive noise in the measurement process and $h_i^\star$ denotes the room impulse response (RIR). In the remainder of this paper, the superscript $\star$ refers to the ground truth. In AER, we are interested in RIRs that are streams of Diracs, *i.e.*,

$$h_i^\star(t) = \sum_{r=0}^{R_i-1} c_{i,r}\delta(t-\tau_{i,r}) \tag{2}$$

where R_i is the (unknown) number of echoes, $\{\tau_{i,r}\}_{r=0}^{R_i-1}$ models the echoes' delays, and $\{c_{i,r}\}_{r=0}^{R_i-1}$ are the corresponding non-negative attenuations. Note that $r=0$ defines the direct propagation path. In the noiseless case, that is when $n_i = 0$ for $i \in \{1,2\}$, we have the identity

$$x_1 * h_2^\star = x_2 * h_1^\star \tag{3}$$

by commutativity of the convolution operator. This result is dubbed cross-relation identity in the channel identification literature [18]. Hence, one can expect to recover the two filters by solving an optimization problem involving (3).

However, in practice, only sampled versions of the two recorded signals are available. More precisely, we consider a measurement model where the incoming signal undergoes a (ideal) low-pass filter ϕ with frequency support $[-{}^{F_s}\!/_2, {}^{F_s}\!/_2]$ before being regularly sampled at the rate F_s. We denote $\mathbf{x}_1, \mathbf{x}_2 \in \mathbf{R}^{2N}$ the two vectors of $2N$ (consecutive) samples and $i \in \{1,2\}$ by

$$\mathbf{x}_i[n] = (\phi * x)\left(\frac{n}{F_s}\right) \qquad \forall n \in \{0,\dots,2N-1\}. \tag{4}$$

2.2. Existing works

Starting from the identity (3), the common SIMO BCE cross-relation framework aims to compute h_1, h_2 solving the following LASSO-type problem in the discrete-time domain:

$$\widehat{\mathbf{h}}_1, \widehat{\mathbf{h}}_2 = \operatorname*{arg\,min}_{\mathbf{h}_1,\mathbf{h}_2} \|\mathcal{T}(\mathbf{x}_1)\mathbf{h}_2 - \mathcal{T}(\mathbf{x}_2)\mathbf{h}_1\|_2^2 + \lambda\|\mathbf{h}\|_1 \quad \text{s.t.} \quad \mathbf{h}[0] = 1 \tag{5}$$

where $\mathbf{x}_i$ and $\mathbf{h}_i$ are the discrete, sampled version of x_i, h_i respectively and $\mathbf{h} = [\mathbf{h}_1^\intercal, \mathbf{h}_2^\intercal]$. $\mathcal{T}(\mathbf{x}_i)$ is the $(2N+L-1)\times L$ Toeplitz matrix[1] associated to convolution where $2N$ and L respectively denote microphone and filter signal length. The constraint $\mathbf{h}[0]=1$ is called an anchor constraint.

The accuracy of estimated RIRs has been subsequently improved using a priori knowledge of the filters: in particular, the authors of [22] have proposed to use sparsity penalty and non-negativity constraints to increase robustness to noise as well as Bayesian-learning methods to automatically infer the value of λ in [5]. Even if sparsity and non-negativity could be seen as a strong assumption, works in speech enhancement [6, 8] and room geometry [11, 13] estimation have proven the effectiveness of this approach. On a similar scheme, in [23], (5) is solved using an adaptive time-frequency-domain approach while [24] proposes to use the ℓ_p-norm instead of the ℓ_1-norm. A successful approach has been presented recently by Crocco *et al.* in [19], where the anchor constraint is replaced by an *iterative weighted* ℓ_1 equality constraint.

[1]The first row and column of $\mathcal{T}(\mathbf{x}_i)$ are respectively $[\mathbf{x}_i[2N-n], 0, \dots, 0]$ and $[\mathbf{x}_i[2N-n], \mathbf{x}_i[2N-n+1], \dots, \mathbf{x}_i[n], 0, \dots, 0]^\intercal$.

3. PROPOSED METHOD

3.1. Cross-relation in the Fourier domain

We first remark that the cross-relation identity (3) ensures that the relation $\phi * x_1 * h_2^\star = \phi * x_2 * h_1^\star$ holds, hence

$$\mathcal{F}(\phi * x_1)\cdot \mathcal{F}\,h_2^\star = \mathcal{F}(\phi * x_2)\cdot \mathcal{F}\,h_1^\star \tag{6}$$

where $\mathcal{F}$ denotes the Fourier transform (FT)

$$\forall\, f \in \mathbf{R}\,, \quad \mathcal{F}\,y(f) = \int_{-\infty}^{+\infty} y(t)\mathrm{e}^{-\mathrm{i}2\pi f t}\,\mathrm{d}t \tag{7}$$

for any signal or filter y (note that we use the same notation when referring to the Fourier transform of a function and a distribution).

While the FT of $h_i^\star$ can be expressed in closed-form (see (10) below), the FT of $\phi * x_i$ is not available due to the measurement process. To circumvent this issue, we use the approximation

$$\mathcal{F}(\phi * x_i)(\tfrac{k}{2N}F_s) \simeq X_i[k] \tag{8}$$

for all integers $k \in \{0,\dots,N\}$, where

$$\mathbf{X}_i[k] = \sum_{n=0}^{2N-1} \mathbf{x}_i[n]\mathrm{e}^{-\mathrm{i}2\pi\frac{kn}{2N}} \tag{9}$$

is the discrete Fourier transform of the real vector $\mathbf{x}_i$ for positive frequencies only. The FT of $h_1^\star, h_2^\star$ (see (2)) can be expressed in closed-form. Denoting Δ_τ the following parametric vector of complex exponential

$$\Delta_\tau \triangleq \left(\mathrm{e}^{-\mathrm{i}2\pi\frac{k}{2N}F_s\tau}\right)_{0\le k\le N} \in \mathbf{C}^{N+1}, \tag{10}$$

equation (6) evaluated at $f = \frac{k}{2N}F_s$ where $k \in \{0,\dots,N\}$ reads

$$\sum_{r=0}^{R_2-1} c_{1,r}\mathbf{X}_1 \odot \Delta_{\tau_{2,r}} = \sum_{r=0}^{R_1-1} c_{2,r}\mathbf{X}_2 \odot \Delta_{\tau_{1,r}} \tag{11}$$

where $\odot$ denotes the component-wise Hadamard product.

3.2. Echo localization with continuous dictionaries

By interpreting the FT of a Dirac as a parametric atom, we propose to cast the problem of RIR estimation into the framework of continuous dictionaries. To that aim, let us define the so-called *parameter set*

$$\Theta \triangleq [0,T]\times\{1,2\} \tag{12}$$

where T is the length (in time) of the filter. Then, the two desired filters $h_1^\star, h_2^\star$ given by (2) can be uniquely[2] represented by the following discrete measure over Θ

$$\mu^\star = \sum_{i=1}^{2}\sum_{r=0}^{R_i-1} c_{i,r}\delta_{(\tau_{i,r},i)}. \tag{13}$$

[2]Uniqueness is ensured as soon as we impose $c_{i,r} > 0\ \forall i, r$.

The rationale behind (12) and (13) is as follows. A couple of filters is now represented by a single stream of Diracs, where we have considered an augmented variable i indicating to which filter the spike belongs. For instance, a Dirac at $(\tau, 1)$ indicates that the first filter contains a Dirac at τ.

The set $\mathcal{M}_+(\Theta)$ of all unsigned and discrete Radon measures over Θ (*i.e.*, the set of all couples of filters) is equipped with the total-variation norm (TV-norm) $\|\mu\|_{\text{TV}}$. See [25] for a rigorous construction of measures set and the TV-norm. We now define the *linear* observation operator $\mathcal{A}\colon \mathcal{M}_+(\Theta) \to \mathbf{C}^{N+1}$, which is such that

$$\mathcal{A}\delta_{(\tau,i)} = \begin{cases} -\mathbf{X}_1 \odot \Delta_\tau & \text{if } i = 1 \\ +\mathbf{X}_2 \odot \Delta_\tau & \text{if } i = 2. \end{cases} \tag{14}$$

$\forall(\tau, i) \in \Theta$ where the two complex vectors $\mathbf{X}_1, \mathbf{X}_2$ have been defined in (9) and $\mathcal{F}_N\delta_\tau$ in (10). Then, by linearity of the observation operator $\mathcal{A}$, the relation (11) can be rewritten as

$$\mathcal{A}\mu^\star = \mathbf{0}_{N+1}. \tag{15}$$

Before continuing our exposition, we note that the anchor constraint can be written in a more convenient way. Indeed, the constraint $\mu(\{(0,1)\}) = 1$ ensures the existence of a Dirac at 0 in the filter 1. Then, the targeted filter reads

$$\mu^\star = \delta_{(0,1)} + \widetilde{\mu}^\star \tag{16}$$

where $\widetilde{\mu}^\star$ is a (finite) discrete measure verifying $\widetilde{\mu}^\star(\{(0,1)\}) = 0$. Denoting $\mathbf{y} \triangleq -\mathcal{A}\delta_{(0,1)} \in \mathbf{C}^{N+1}$, the relation (15) becomes

$$\mathcal{A}\widetilde{\mu}^\star = \mathbf{y}. \tag{17}$$

For the sake of clarity, we use these conventions hereafter and omit the tilde. Now, following [20, 26], one can expect to recover the desired filter $\mu^\star$ by solving

$$\widehat{\mu} = \underset{\mathcal{M}_+(\Theta)}{\arg\min}\ \|\mu\|_{\text{TV}} \quad \text{s.t.} \quad \begin{cases} \mathcal{A}\mu = \mathbf{y} \\ \mu(\{(0,1)\}) = 0. \end{cases} \tag{18-$\mathcal{P}^0\text{TV}$}$$

Note that (18-$\mathcal{P}^0\text{TV}$) has to be interpreted as a natural extension of the well-known *basis pursuit* problem to the continuous setting. Indeed, for *any* finite discrete measure $\mu = \sum_{r=0}^{R-1} c_r\delta_{(\tau_r,i_r)}$, the TV-norm of μ returns to the ℓ_1-norm of the coefficients, *i.e.*, $\|\mu\|_{TV} = \sum_{r=0}^{R-1}|c_r|$.

Finally, (17) can be exploited to take into account noise during the measurement process (*i.e.*, $n_i \neq 0$ in (1)), as well as approximation errors (see (8)-(11)). In that case, the first equality constraint in (18-$\mathcal{P}^0\text{TV}$) is relaxed, leading to the so-called Beurling-LASSO (BLASSO) problem

$$\begin{aligned} \widehat{\mu} = \underset{\mu\in\mathcal{M}_+(\Theta)}{\arg\min}\ & \tfrac{1}{2}\|\mathbf{y} - \mathcal{A}\mu\|_2^2 + \lambda\|\mu\|_{\text{TV}} \\ \text{s.t.}\quad & \mu(\{(0,1)\}) = 0. \end{aligned} \tag{19-$\mathcal{P}^\lambda_{\text{TV}}$}$$

We emphasize that although continuous Radon measures may potentially be admissible, the minimizers of (19-$\mathcal{P}^\lambda_{\text{TV}}$) are *guaranteed* to be streams of Dirac*s* [27, Theorem 4.2]. In addition, although problem (19-$\mathcal{P}^\lambda_{\text{TV}}$) seems to depend on some regularization parameter λ, we describe in Section 4 a procedure to automatically tune it to recover a desired number of spikes.

Finally, note that problem (19-$\mathcal{P}^\lambda_{\text{TV}}$) is convex with linear constraints. In this work, we particularize the sliding Frank-Wolfe algorithm proposed in [21] to solve (19-$\mathcal{P}^\lambda_{\text{TV}}$). Detailed descriptions of the steps of the algorithm are given in Appendix A.

4. EXPERIMENTS

The proposed method (BLASTER) is compared against the non-negative ℓ_1-norm method (BSN) of [22] and the iterative ℓ_1-norm approach (IL1C) described in [19]. The problem is formulated as estimating the time location of the first $R = 7$ strongest components of the RIRs for 2 microphones listening to a single sound source in a shoebox room. It corresponds to the challenging task of estimating first-order early reflections. The robustness of the methods is tested against different level of noise (SNR) and reverberation time (RT_{60}).

We propose to compute a *path of solutions* to automatically estimate the regularization parameter λ in (19-$\mathcal{P}^\lambda_{\text{TV}}$). More precisely, let $\lambda_{\max}$ be the smallest value of λ such that the null measure is the solution to (19-$\mathcal{P}^\lambda_{\text{TV}}$). It can be shown that $\lambda_{\max}$ is upper bounded by $\max_{\theta\in\Theta}|\mathbf{y}^\intercal\mathcal{A}\delta_\theta|$. Starting from $\ell = 1$ and the empty filter, we consider a sequential implementation where the solution of (19-$\mathcal{P}^\lambda_{\text{TV}}$) is computed for $\lambda_\ell = 10^{-0.05\ell}\lambda_{\max}$ until the desired number of spikes is found in each channel when incrementing ℓ. For each λ_ℓ, we search for a solution of (19-$\mathcal{P}^\lambda_{\text{TV}}$) with the solution obtained for $\lambda_{\ell-1}$ as a warm start.

The quality of the AER estimation is assessed in terms of precision[3] in percentage as in the literature of onset detection [28] and the root-mean-square-error (RMSE) in samples. Both metrics evaluate only the *matched* peaks, where a *match* is defined as being within a small window τ_{max} of a reference delay. These two metrics are similar to the ones used in [29].

For this purpose we created three synthetic datasets of 1000 observations each: $\mathcal{D}^{\text{(valid)}}$ is used for tuning the hyperparameter λ and the peak-picking parameters for IL1C and BSN using RT_{60} and SNR randomly drawn from $\mathcal{U}[0,1]$ (sec) and $\mathcal{U}[0,20]$ (dB) respectively; $\mathcal{D}^{\text{SNR}}$ features SNR value uniformly sampled in $[0, 6, 14, 20, \infty]$ while the RT_{60} is kept fixed to 400 ms; akin the $\mathcal{D}^{\text{RT}_{60}}$ is built sampling RT_{60} value uniformly in $[200, 400, 600, 800, 1000]$ ms keeping SNR fix to 20 dB. Moreover, while for $\mathcal{D}^{\text{(valid)}}$ broadband signals (white noise) are used as the source, for $\mathcal{D}^{\text{SNR}}$ and $\mathcal{D}^{\text{RT}_{60}}$ speech utterances from the TIMIT dataset are also included. The signal duration is kept fixed to 1 s with sampling frequency $F_s = 16$ kHz.

For a given RT_{60} value and room with random dimensions, a unique absorption coefficient is assigned to all surfaces based on the Sabine's formula. Then, the two microphones and the source are randomly positioned inside the room. The parameters of such audio scene are then passed as input to the `pyroomacoustic` simulator [30], which returns the corresponding RIRs as well as the *off-grid* echo delays and attenuation coefficients computed with the Image Method [31]. Note that when generating the data, no samples have been pruned to match any minimal separation condition.

To generate the microphone signals, an oversampled version of the source signal is convolved with ideal RIRs at high frequency ($F_s = 1024$ kHz) made up of on-grid Dirac*s*. The results are later resampled to meet the original F_s and Gaussian white noise is added to meet the given SNR value.

Quantitative results are reported in Fig. 1, Fig. 2 and Tab. 1. Here, for both RMSE and Precision and for both broadband and speech signal, the metrics are displayed against the dataset parameters. We observe that BSN performs worst in all tested conditions,

[3] Since only K time locations are considered in both the ground truth and the estimation, precision and recall are equal.

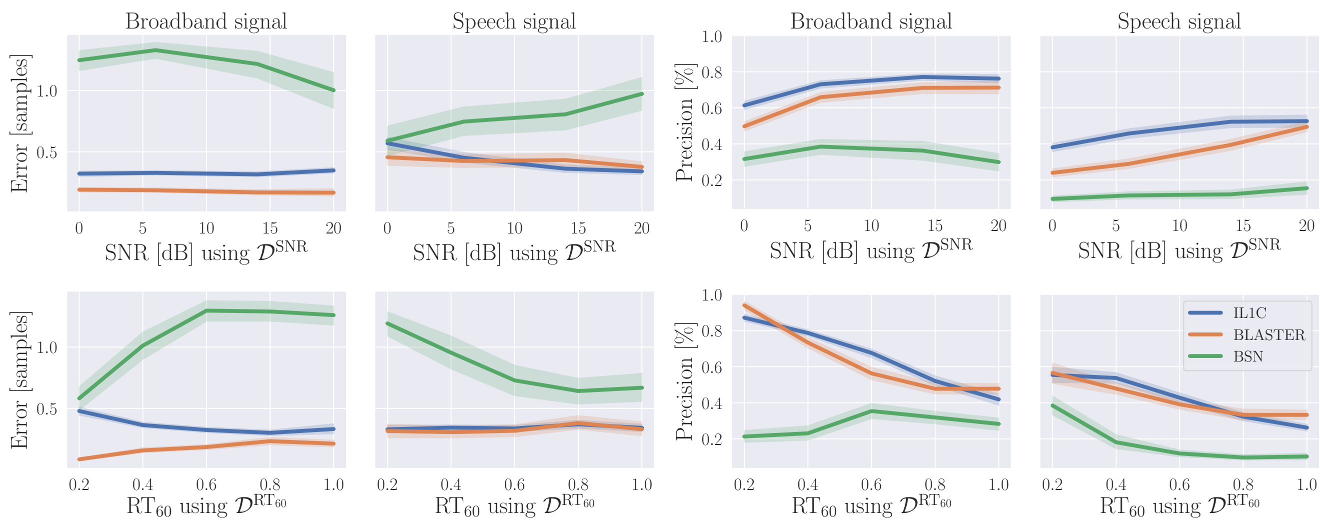

Fig. 1. Line plot with error bands for error (left) and precision (right) versus SNR level (top) and RT_{60} level (bottom) using broadband and speech signals for the task of recovering $R = 7$ echoes. A threshold of $\tau_{\text{max}} = 2$ samples is used to compute the precision.

	Precision [%]									
	R = 2 echoes					R = 7 echoes				
$\tau_{\max}$	0.5	1	2	3	10	0.5	1	2	3	10
BSN	8	9	27	46	62	5	8	38	54	73
IL1C	51	55	55	56	58	42	53	55	56	58
BLASTER	**68**	**73**	**74**	**75**	**75**	46	53	56	57	61

Table 1. Precision for different threshold $\tau_{\max}$ in samples for the recovery of $R = 2$ and 7 echoes, RT_{60} = 200 ms and SNR = 20 dB.

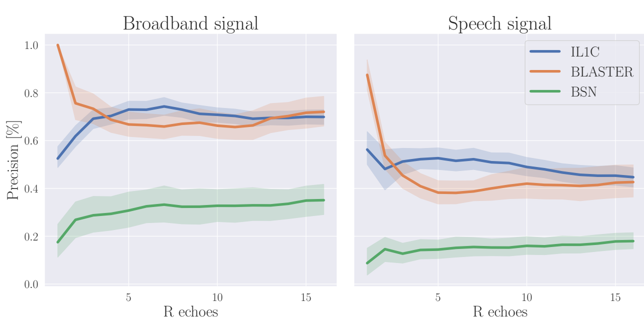

Fig. 2. Line plots with error bands of precision versus number of echoes R to be retrieved for broadband (left) and speech (right) signals with RT_{60} = 400 ms and SNR = 20 dB.

possibly due to its strong reliance on the peak picking step. For $R = 7$ or higher, BLASTER yields similar or slightly worse performance than IL1C for the considered noise and reverberation levels, with decreasing performance for both as these levels increase. Using speech rather than broadband signals also yields worse results for all methods. However, the echo timing RMSE is significantly smaller using BLASTER due to its off-grid advantage. We also note that BLASTER significantly outperforms IL1C on the task of recovering $R = 2$ echoes. As showed in Tab. 1, in mild conditions, up to 68% of echoes can be retrieved by BLASTER with errors lower than half a sample in that case. This is promising since the practical advantage of knowing the timing of two echoes per channel has been demonstrated in [7, 9].

5. CONCLUSIONS

A novel blind, off-grid, multichannel echo retrieval method has been proposed based on the framework of continuous dictionaries. Comparisons with state-of-the-art approaches on various noise and reverberation conditions show that this method performs best when the number of echoes to retrieve is small. While some robustness to noise, reverberation, and non-broadband signals is observed, our experiments reveal that room for improvement exists for this challenging and emerging topic. Future works will include an extension to more than two channels and experiments on real-world data.

References

[1] D. Wang and G. J. Brown, "Reverberation," in *Computational Auditory Scene Analysis*, pp. 209–250. IEEE, 2011.

[2] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, Tomohiro N., and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[3] M. Barthet and M. Sandler, "On the effect of reverberation on musical instrument automatic recognition," in *Audio Engineering Society Convention 2010*. may 2010, vol. 1, pp. 658–665, Audio Engineering Society.

[4] M. Wu and D. L. Wang, "A two-stage algorithm for one-

microphone reverberant speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 774–784, 2006.

[5] Y. Lin, J. Chen, Y. Kim, and D. Lee, "Blind channel identification for speech dereverberation using l 1 -norm sparse learning," in *Proc. of Conf. on Advances in Neural Information Processing Systems*, 2008, pp. 1–2.

[6] F. Ribeiro, D. Ba, C. Zhang, and D. Florêncio, "Turning enemies into friends: Using reflections to improve sound source localization," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, July 2010, pp. 731–736.

[7] D. D. Carlo, A. Deleforge, and N. Bertin, "Mirage: 2d source localization using microphone pair augmentation with echoes," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proces.*, May 2019, pp. 775–779.

[8] I. Dokmanić, R. Scheibler, and M. Vetterli, "Raking the Cocktail Party," *IEEE Journal on Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 825–836, 2015.

[9] R. Scheibler, D. Di Carlos, A. Deleforge, and I. Dokmanic, "Separake: Source separation with a little help from echoes," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proces.*, April 2018, pp. 6897–6901.

[10] D. Salvati, C. Drioli, and G. Foresti, "Sound Source and Microphone Localization from Acoustic Impulse Responses," *IEEE Sig. Process. Letters*, vol. 23, no. 10, pp. 1459–1463, 2016.

[11] F. Antonacci, J. Filos, M. R.P. Thomas, E. A.P. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 10, pp. 2683–2695, 2012.

[12] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proc. of the National Academy of Sciences*, vol. 110, no. 30, pp. 12186–12191, 2013.

[13] M. Crocco, A. Trucco, and A. Del Bue, "Uncalibrated 3D room geometry estimation from sound impulse responses," *Journal of the Franklin Institute*, vol. 354, no. 18, pp. 8678–8709, 2017.

[14] H. P. Tukuljac, A. Deleforge, and R. Gribonval, "MULAN: A Blind and Off-Grid Method for Multichannel Echo Retrieval," in *Proc. of the Conf. on Neur. Inf. Process. Sys.*, 2018.

[15] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 1, pp. 026503, Dec 2006.

[16] Y. Park, W. Seong, and Y. Choo, "Compressive time delay estimation off the grid," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. EL585–EL591, 2017.

[17] J. Rindom Jensen, U. Saqib, and S. Gannot, "An EM method for multichannel TOA and DOA estimation of acoustic echoes," in *IEEE Workshop on Applications of Signal Processing To Audio and Acoustics (WASPAA)*, 2019.

[18] G. Xu, H. Liu, L. Tong, and T. Kailath, "A Least-Squares Approach to Blind Channel Identification," *IEEE Transactions on Signal Processing*, vol. 43, no. 12, pp. 2982–2993, 1995.

[19] M. Crocco and A. Del Bue, "Estimation of TDOA for room reflections by iterative weighted l1 constraint," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proces.*, 2016, pp. 3201–3205.

[20] E. J. Cands and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Communications on Pure and Applied Mathematics*, vol. 67, no. 6, pp. 906–956, 2014.

[21] Q. Denoyelle, V. Duval, G. Peyré, and E. Soubies, "The Sliding Frank-Wolfe Algorithm and its Application to Super-Resolution Microscopy," *Inverse Problems*, 2019.

[22] Y. Lin, J. Chen, Y. Kim, and D. Lee, "Blind sparse-nonnegative (BSN) channel identification for acoustic time-difference-of-arrival estimation," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 106–109, 2007.

[23] K. Kowalczyk, E. Habets, W. Kellermann, and P. Naylor, "Blind system identification using sparse learning for TDOA estimation of room reflections," *IEEE Signal Processing Letters*, vol. 20, no. 7, pp. 653–656, 2013.

[24] A. Aïssa-El-Bey and K. Abed-Meraim, "Blind SIMO channel identification using a sparsity criterion," *IEEE Workshop on Signal Processing Advances in Wireless Communications, SPAWC*, pp. 271–275, 2008.

[25] W. Rudin, *Real and Complex Analysis, 3rd Ed.*, McGraw-Hill, Inc., New York, NY, USA, 1987.

[26] Y. de Castro and F. Gamboa, "Exact reconstruction using beurling minimal extrapolation," *Journal of Mathematical Analysis and Applications*, vol. 395, no. 1, pp. 336 – 354, 2012.

[27] K. Bredies and M. Carioni, "Sparsity of solutions for variational inverse problems with finite-dimensional data," 2018.

[28] S. Böck, F. Krebs, and M. Schedl, "Evaluating the online capabilities of onset detection methods," *Proc. of International Society for Music Information Retrieval Conference, (ISMIR 2012)*, pp. 49–54, 2012.

[29] M. Crocco and A. Del Bue, "Room impulse response estimation by iterative weighted L1-norm," in *Proc. Europ. Sig. Proces. Conf.*, 2015, pp. 1895–1899.

[30] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proces.*, Apr 2018.

[31] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[32] K. Bredies and Hanna K. Pikkarainen, "Inverse problems in spaces of measures," *ESAIM: Control, Optimisation and Calculus of Variations*, vol. 19, no. 1, pp. 190–218, 2013.

[33] N. Rao, P. Shah, and S. Wright, "Forwardbackward greedy algorithms for atomic norm regularization," *IEEE Transactions on Signal Processing*, vol. 63, no. 21, pp. 5798–5811, Nov 2015.

A. SLIDING FRANK-WOLFE ALGORITHM

Among all the methods that address the resolution of (19-$\mathcal{P}_{\text{TV}}^{\lambda}$), a significant number of them are based on variations of the well-known Frank-Wolfe iterative algorithm, see, *e.g.*, [21, 32, 33]. In this paper, we particularize the *sliding Frank-Wolfe* (SFW) algorithm proposed in [21]. Starting from an initial guess (*e.g.*, the null measure), SFW repeats the four following steps until convergence:

1. add a parameter (position of echo) to the support of the solution,
2. update all the coefficients solving a (finite dimensional) Lasso,
3. update jointly the position of the echoes and the coefficients,
4. eventually remove parameters (echoes) associated to coefficients equal to zero.

Finally, SFW stops as soon as an iterate satisfies the first order optimality condition associated to the convex problem (19-$\mathcal{P}_{\text{TV}}^{\lambda}$). More particularly, denoting $\mu^{(t)}$ the estimated filters at iteration t, SFW stops as soon as $\mu^{(t)}$ satisfies [32, Proposition 3.6]

$$\sup_{\theta\in\Theta} \lambda^{-1}\left|\left\langle \mathcal{A}\delta_\theta, \mathbf{y} - \mathcal{A}\mu^{(t)}\right\rangle\right| \leq 1. \tag{20}$$

The complete SFW method for echo estimation is described by Algorithm 1. We now provide additional details about the implementation of each step.

Non-negative Blasso. To take into account the non-negative constraint on the coefficients, the authors of [21] have proposed to slightly modify the SFW algorithm by *i)* removing the absolute value in (20) and *ii)* adding the non-negativity constraints at step 2 and 3 (see lines 14 and 15 of Algorithm 1). The reader is referred to [21, remark 8 in Section 4.1] for more details.

***Real part in* (20).** We have shown earlier that SFW stops as soon as an iterate $\mu^{(t)}$ satisfies (20) at some iteration t. Since the estimated coefficients $\left\{c_r^{(t)}\right\}_{r=1}^{R}$ are (non-negative) scalars, (20) can be rewritten as

$$\sup_{\theta\in\Theta} \lambda^{-1}\,\mathrm{Re}(\langle \mathcal{A}\delta_\theta, \mathbf{y} - \mathcal{A}\mu^\star\rangle) \leq 1. \tag{21}$$

In particular, using the real part in the implementation allows to remove the imaginary part that may appear due to the imprecision.

Precision of the stopping criterion. Unfortunately, condition (20) cannot be met due to the machine precision, *i.e.*, the solution of (19-$\mathcal{P}_{\text{TV}}^{\lambda}$) is computed up to some prescribed accuracy. In this paper, we say that the algorithm stops as soon as

$$\sup_{\theta\in\Theta} \lambda^{-1}\,\mathrm{Re}(\langle \mathcal{A}\delta_\theta, \mathbf{y} - \mathcal{A}\mu^\star\rangle) \leq 1 + \varepsilon \tag{22}$$

where ε is a positive scalar set to $\varepsilon = 10^{-3}$.

Finding new parameters (Line 7). The new parameter is found by solving

$$\arg\max_{\theta\in\Theta} \mathrm{Re}(\langle \mathcal{A}\delta_\theta, \mathbf{y} - \mathcal{A}\widehat{\mu}\rangle). \tag{23}$$

Algorithm 1: Sliding Frank-Wolfe algorithm for solving (19-$\mathcal{P}_{\text{TV}}^{\lambda}$).

Input: Observation operator $\mathcal{A}$, positive scalar λ, precision ε

Output: Channels represented as a measure $\widehat{\mu}$

// Initialization

1 $\mathbf{y} \leftarrow -\mathcal{A}\delta_{(0,1)}$ **// observation vector**

2 $\mu^{(0)} = 0_{\mathcal{M}}$ **// estimated filters**

3 $\mathcal{E}^{(0)} = \emptyset$ **// estimated echoes**

4 $x_{\max} = (2\lambda)^{-1}\|\mathbf{y}\|_2^2$;

// Starting algorithm

5 **repeat**

6 $t \leftarrow t + 1$ **// Iteration index**

// 1. Add new element to the support

7 Find $\theta^{\text{new}} \in \arg\max_{\theta\in\Theta} \mathrm{Re}\Big(\Big\langle \mathcal{A}\delta_\theta, \mathbf{y} - \mathcal{A}\mu^{(t-1)}\Big\rangle\Big)$;

8 $\eta^{(t)} \leftarrow \lambda^{-1}\,\mathrm{Re}\Big(\Big\langle \mathcal{A}\delta_{\theta^{\text{new}}}, \mathbf{y} - \mathcal{A}\mu^{(t-1)}\Big\rangle\Big)$;

9 **if** $\eta^{(t)} \leq 1 + \varepsilon$ **then**

10 Stop and return $\widehat{\mu} = \mu^{(t-1)}$ is a solution ;

11 **end**

12 $\mathcal{E}^{(t-½)} \leftarrow \mathcal{E}^{(t-½)} \cup \{\theta^{\text{new}}\}$;

13 $R^{(t)} \leftarrow \mathtt{card}(\mathcal{E}^{(t-½)})$ **// Number of detected echoes**

// 2. Lasso update of the coefficients

14 $\mathbf{c}^{(t-½)} \leftarrow \arg\min\limits_{\mathbf{c}\in\mathbf{R}_+^{R^{(t)}}} \frac{1}{2}\Big\|\mathbf{y} - \sum\limits_{\theta\in\mathcal{E}^{(t-½)}} c_\theta \mathcal{A}\delta_\theta\Big\|_2^2 + \lambda\|\mathbf{c}\|_1$ approximated using a proximal gradient algorithm ;

// 3. Joint update for a given number of spikes

15 $\mathcal{E}^{(t)}, \mathbf{c}^{(t)} \leftarrow \arg\min\limits_{\boldsymbol{\theta}\in\Theta^{R^{(t)}}, \mathbf{c}\in[0,x_{\max}]^{R^{(t)}}} \frac{1}{2}\Big\|\mathbf{y} - \sum\limits_{r=1}^{R^{(t)}} \mathbf{c}_r\mathcal{A}\delta_{\theta_r}\Big\|_2^2 + \lambda\|\mathbf{c}\|_1$ approximated using a non-convex solver initialized with $(\mathcal{E}^{(t-½)}, \mathbf{c}^{(t-½)})$;

// 4. Eventually remove zero amplitude Dirac masses

16 $\mathcal{E}^{(t)} \leftarrow \Big\{\theta_r^{(t)} \in \mathcal{E}^{(t)} \mid \mathbf{c}_r^{(t)} \neq 0\Big\}$;

17 $\mathbf{c}^{(t)} \leftarrow \Big\{\mathbf{c}_r^{(t)} \mid \mathbf{c}_r^{(t)} \neq 0\Big\}$;

18 $\mu^{(t)} \leftarrow \sum\limits_{r=1}^{\mathtt{card}(\mathcal{E}^{(t)})} \mathbf{c}_r^{(t)}\delta_{\theta_r^{(t)}}$;

19 **until** *until convergence*;

To solve this optimization problem, we first find a maximizer on a thin grid made of 20000 points. We then proceed to a local refinement using the `scipy` optimization library[4].

Nonnegative Lasso (Line 14). The nonnegative Lasso is solved using a custom implementation of a proximal gradient algorithm. In particular, the procedure stops as soon as a stopping criterion in terms of duality gap is reached (10^{-6}).

Joint update (Line 15). In order to ease the numerical resolution, we show that given a positive integer R, the solution of

$$\underset{\boldsymbol{\theta}\in\Theta^R,\mathbf{c}\in\mathbf{R}^R}{\arg\min}\ \tfrac{1}{2}\Big\|\mathbf{y}-\sum_{r=1}^{R}\mathbf{c}_r\mathcal{A}\delta_{\theta_r}\Big\|_2^2+\lambda\|\mathbf{c}\|_1 \tag{24}$$

is equivalent to the solution of

$$\underset{\boldsymbol{\theta}\in\Theta^R,\mathbf{c}\in[0,x_{\max}]^R}{\arg\min}\ \tfrac{1}{2}\Big\|\mathbf{y}-\sum_{r=1}^{R}\mathbf{c}_r\mathcal{A}\delta_{\theta_r}\Big\|_2^2+\lambda\|\mathbf{c}\|_1 \tag{25}$$

where

$$x_{\max}=\frac{1}{2\lambda}\|\mathbf{y}\|_2^2. \tag{26}$$

Indeed, let us denote $\boldsymbol{\theta}^\star$, $\mathbf{c}^\star$ the minimizers of (24). For any $\boldsymbol{\theta}\in\Theta^R$, the couple $\boldsymbol{\theta},\mathbf{0}_R$ is admissible for (24) so we have by definition

$$\tfrac{1}{2}\Big\|\mathbf{y}-\sum_{r=1}^{R}\mathbf{c}_r^\star\mathcal{A}\delta_{\theta_r^\star}\Big\|_2^2+\lambda\|\mathbf{c}^\star\|_1\leq\tfrac{1}{2}\big\|\mathbf{y}\big\|_2^2. \tag{27}$$

Hence

$$0\leq c_r^\star\leq\|\mathbf{c}^\star\|_1\leq\tfrac{1}{2\lambda}\big\|\mathbf{y}\big\|_2^2\triangleq x_{\max}. \tag{28}$$

Finally, the joint update of the coefficients and parameters is performed using the Sequential Least SQuares Programming (SLSQP) implemented in the `scipy` optimization library, see footnote 4.

[4] https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html.