Multi-Pitch Estimation and Tracking Using Bayesian Inference in Block Sparsity

Karimian-Azari, Sam; Jakobsson, Andreas; Jensen, Jesper Rindom; Christensen, Mads Græsbøll

Published in:
2015 Proceedings of the 23rd European Signal Processing Conference (EUSIPCO 2015)

DOI (link to publication from Publisher):
10.1109/EUSIPCO.2015.7362336

Publication date:
2015

MULTI-PITCH ESTIMATION AND TRACKING USING BAYESIAN INFERENCE IN BLOCK SPARSITY

Sam Karimian-Azari[⋆]*, Andreas Jakobsson*[†]*, Jesper R. Jensen*[⋆]*, and Mads G. Christensen*[⋆]

[⋆] Audio Analysis Lab, AD:MT, Aalborg University, email: {ska, jrj, mgc}@create.aau.dk
[†] Dept. of Mathematical Statistics, Lund University, email: aj@maths.lth.se

ABSTRACT

In this paper, we consider the problem of multi-pitch estimation and tracking of an unknown number of harmonic audio sources. The regularized least-squares is a solution for simultaneous sparse source selection and parameter estimation. Exploiting block sparsity, the method allows for reliable tracking of the found sources, without posing detailed *a priori* assumptions of the number of harmonics for each source. The method incorporates a Bayesian prior and assigns data-dependent regularization coefficients to efficiently incorporate both earlier and future data blocks in the tracking of estimates. In comparison with fix regularization coefficients, the simulation results, using both real and synthetic audio signals, confirm the performance of the proposed method.

***Index Terms*—** Multi-pitch estimation, tracking, harmonic signal, regularized least-squares, sparsity

1. INTRODUCTION

Estimation of the fundamental frequency, or pitch, detailing a set of audio sources, is an important problem in a wide range of applications, such as source separation, music transcription, and enhancement [1–3]. In speech recognition, for example, reliable pitch estimates are required in a prosodic implementation. The topic has for this reason attracted much interest, in particular for single pitch estimation [4], but the more challenging problem of multi-pitch estimation has also been given notable attention [5–8]. Often, these methods make strong *a priori* assumptions on the number of measured sources, as well as on the model orders of these sources. To determine such model order information is well known to be challenging [6], although some efforts on joint pitch and model order estimator techniques have been presented for the single pitch case [9]. For joint multi-pitch and model order estimation of the given number of sources, the problem have been formulated for polyphonic music transcription [5].

The recent pitch estimation using block sparsity (PEBS) technique introduced in [8] avoids such assumptions by imposing a verity of sparsity constraints, such that from a large dictionary of feasible pitches, both the number of sources and the model order of each found source can be determined. In this work, we introduce an extension of the PEBS algorithm to allow the efficient tracking of audio sources. Given the natural behavior of audio signals, the pitch often changes smoothly over time. That makes pitch values in sequential data frames highly correlated, which is often exploited in pitch tracking [10–12]. To allow for such temporal smoothness, we introduce data-dependent regularization coefficients for the sparsity constraints in the PEBS method, such that the estimate for the currently processed data frame is affected by the local spectral neighborhood of both the past and future data frames. The approach builds on earlier work on the adaptive Lasso [13] and the Bayesian Lasso [14], as well as use a Gaussian smoothing kernel to regularize the corresponding components in the PEBS dictionary.

The remainder of this paper is organized as follows: In the next section, we present the signal model. In Section 3, we present the proposed multi-pitch estimation and tracking using Bayesian inference. Experimental results are presented in Section 4. Finally, we conclude on our work in Section 5.

2. SIGNAL MODEL

Consider a sum of M harmonic audio sources, each with a fundamental frequency ω_m, and containing L_m harmonics, for $m = 1, 2, \ldots, M$, and Let

$$\mathbf{y}_n = \begin{bmatrix} y(n) & y(n+1) & \ldots & y(n+N-1) \end{bmatrix}^T \quad (1)$$

denote the data frame processed at time n, with N being the length of the frame. To simplify the notation and to reduce the resulting computational complexity, we here model the discrete-time analytical signal of the measured signals, as obtained using the method detailed in [15] (see also [6]). Thus, $\mathbf{y}_n$ may be well modeled as

$$\mathbf{y}_n \triangleq \sum_{m=1}^{M} \mathbf{Z}_m \mathbf{b}_m + \mathbf{v} = \mathbf{Z}\mathbf{b} + \mathbf{v} \quad (2)$$

This work was funded in part by the Villum Foundation, Carl Trygger's foundation, the Swedish research council, and the Danish Council for Independent Research, grant ID: DFF 1337-00084.

where

$$\begin{aligned}\mathbf{Z} &= \begin{bmatrix}\mathbf{Z}_1 & \mathbf{Z}_2 & \dots & \mathbf{Z}_M\end{bmatrix}\\ \mathbf{Z}_m &= \begin{bmatrix}\mathbf{z}_m & \mathbf{z}_m^2 & \dots & \mathbf{z}_m^{L_m}\end{bmatrix}\\ \mathbf{z}_m^l &= \begin{bmatrix}1 & e^{jl\omega_m} & \dots & e^{jl\omega_m(N-1)}\end{bmatrix}^T\\ \mathbf{b} &= \begin{bmatrix}\mathbf{b}_1^T & \mathbf{b}_2^T & \dots & \mathbf{b}_M^T\end{bmatrix}^T\\ \mathbf{b}_m &= \begin{bmatrix}b_{m,1} & b_{m,2} & \dots & b_{m,L_m}\end{bmatrix}^T\end{aligned}$$

and $(\cdot)^T$ denotes the transpose. The matrix $\mathbf{Z}$ contains the $L_{\text{tot}} = \sum_{m=1}^{M} L_m$ complex-valued sinusoids, with the corresponding complex amplitudes $\mathbf{b}$, and is formed out of sub-basis matrices, $\mathbf{Z}_m$, detailing the tones presented in each of the M sources. The additive noise, $\mathbf{v}$, is here formed similar to $\mathbf{y}_n$ in (1), and is assumed to be a circularly symmetric Gaussian distributed white noise, i.e., $E\{\mathbf{v}(n)\mathbf{v}^H(n)\} = \sigma_v^2\mathbf{I}_N$, where $E\{\cdot\}$ denotes the expectation.

3. MULTI-PITCH ESTIMATION AND TRACKING

Consider the problem of spectral amplitude estimation of multiple sinusoids from the observed signal $\mathbf{y}_n$. For the given (known) basis matrix $\mathbf{Z}$, with $N \gg L_{\text{tot}}$, and where the complex basis vectors $\mathbf{z}_m^l$ are assumed to be independent, one may form an estimate of the unknown pitch frequencies using the ordinary least-squares (LS) method, minimizing the sum of squared residuals such that $\hat{\mathbf{b}} = (\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{y}_n$. However, such a solution requires knowledge of both the number of sources and the number of harmonics for each source. To avoid these assumptions, we define a (large) dictionary matrix over the considered range of frequencies, $\omega_r \in [\omega_{\min}, \omega_{\max}]$, and harmonics, such that the allowed number of harmonics for the dictionary elements $r = 1, 2, \cdots, S$ are limited to $L_r = \lfloor \pi/\omega_r \rfloor$, where $\lfloor\cdot\rfloor$ denotes the truncation operation to the nearest lower integer. Accordingly,

$$\mathbf{y}_n \triangleq \mathbf{W}\mathbf{a} + \mathbf{v} \tag{3}$$

where the $N \times S$ dictionary matrix is formed as

$$\mathbf{W} = \begin{bmatrix}\mathbf{Z}_1 & \mathbf{Z}_2 & \dots & \mathbf{Z}_S\end{bmatrix} \tag{4}$$

where $S \gg M$. The spectral amplitudes of the $L_{\text{ext}} = \sum_{r=1}^{S} L_r$ sinusoids of the dictionary, i.e.,

$$\mathbf{a} = \begin{bmatrix}\mathbf{a}_1^T & \mathbf{a}_2^T & \dots & \mathbf{a}_S^T\end{bmatrix}^T \tag{5}$$

are exceedingly sparse, containing only L_{tot} non-zero values. Then, for the problem of multi-pitch estimation, we form an estimate of the pitch frequencies by maximizing the likelihood of the spectral amplitude estimates, $\hat{\mathbf{a}}$, of the corresponding frequencies, such that

$$\hat{\mathbf{\Omega}} = \arg\max_{\mathbf{\Omega}} P\big(\{\|\hat{\mathbf{a}}_r\|_2\}_{r=1}^{S}\big|\mathbf{\Omega}\big) \tag{6}$$

where $\mathbf{\Omega} = \begin{bmatrix}\omega_1 & \omega_2 & \dots & \omega_{\tilde{M}}\end{bmatrix}^T$, for a given $\tilde{M}$, which may differ from the true number of sources, M.

Under the assumption of circularly symmetric Gaussian noise, the spectral amplitude estimates may be formed using the maximum likelihood (ML) method, such that

$$\hat{\mathbf{a}}_{\text{ML}} = \arg\max_{\mathbf{a}} \log P(\mathbf{y}_n\big|\mathbf{a}, \sigma_v) \tag{7}$$

where

$$P(\mathbf{y}_n|\mathbf{a}, \sigma_v) = \frac{1}{(2\pi\sigma_v^2)^{N/2}} \exp\Big(-\frac{1}{2\sigma_v^2}\|\mathbf{y}_n - \mathbf{W}\mathbf{a}\|_2^2\Big)$$

is the likelihood function, with $\|\cdot\|_2$ denoting the ℓ_2-norm. Given that the additive noise is assumed to be white, the resulting ML estimate coincides with the standard LS estimate, and may thus be efficiently formed accordingly. However, in order to avoid over-fitting, one often instead forms the regularized LS estimate (see, e.g., [16]). The least absolute shrinkage and selection operator (Lasso) [17] is a well known regularized LS estimator that shrinks the sum of absolute values of the amplitudes toward zero. Imposing a Laplace distribution on the amplitudes, the likelihood for those may be expressed as [14]

$$P(a_{r,l_r}|\tau_{r,l_r}, \sigma_v) = \frac{\tau_{r,l_r}}{2\sigma_v} \exp\Big(-\frac{\tau_{r,l_r}}{\sigma_v}|a_{r,l_r}|\Big). \tag{8}$$

Interpreting the Lasso as a Bayesian posteriori estimator, we express the probability of the spectral amplitudes, given the observations, and using the parameter vector $\mathbf{\Psi} = \big\{\bigcup_{r=1}^{S}\bigcup_{l_r=1}^{L_r}\tau_{r,l_r}\big\}$, as

$$\begin{aligned}P(\mathbf{a}|\mathbf{y}_n, \mathbf{\Psi}, \sigma_v) &\propto \prod_{r=1}^{S}\prod_{l_r=1}^{L_r} P(\mathbf{y}_n\big|a_{r,l_r}, \tau_{r,l_r}, \sigma_v) P(a_{r,l_r}\big|\tau_{r,l_r}, \sigma_v)\\ &\propto \exp\Big(-\frac{1}{2\sigma_v^2}\|\mathbf{y}_n - \mathbf{W}\mathbf{a}\|_2^2\Big) \prod_{r=1}^{S}\prod_{l_r=1}^{L_r} \exp\Big(-\frac{\tau_{r,l_r}}{\sigma_v}|a_{r,l_r}|\Big).\end{aligned}$$

As noted in [8], one may further include the group sparsity constraint to restrict the number of variable solutions (see also [18, 19]). Therefore, we extend on this notation by expressing the probability of the grouped variables, using the parameter vector $\mathbf{\Psi}_r = \big\{\bigcup_{l_r=1}^{L_r}\tau_{r,l_r}\big\}$, as

$$P(\mathbf{a}|\mathbf{y}_n, \mathbf{\Psi}, \sigma_v) \propto \exp\Big(-\frac{1}{2\sigma_v^2}\|\mathbf{y}_n - \mathbf{W}\mathbf{a}\|_2^2\Big) \prod_{r=1}^{S} P(\mathbf{a}_r\big|\mathbf{\Psi}_r, \sigma_v)$$

where $P(\mathbf{a}_r\big|\mathbf{\Psi}_r, \sigma_v) \propto \exp\Big(-\frac{\|\mathbf{\Psi}_r\|_2}{\sigma_v}\|\mathbf{a}_r\|_2\Big)$. Herein, we take into consideration the spectral neighborhood as it evolves over time, such that

$$\begin{aligned}\hat{\mathbf{a}} &= \arg\max_{\mathbf{a}} \log P(\mathbf{a}\big|\mathbf{y}_n, \mathbf{\Psi}, \sigma_v)\\ &= \arg\min_{\mathbf{a}} \frac{1}{2}\|\mathbf{y}_n - \mathbf{W}\mathbf{a}\|_2^2 + J\end{aligned} \tag{9}$$

where J denotes the imposed constraints, formed as

$$J = \|\boldsymbol{\psi}_{\mathrm{L}} \odot \mathbf{a}\|_1 + \sum_{r=1}^{S} \|\boldsymbol{\psi}_{\mathrm{GL},r}\|_2 \|\mathbf{a}_r\|_2 \tag{10}$$

and with $\odot$ denoting the element-wise matrix product. To allow for the required sparsity constraints [8], the penalty term J involves both the ℓ_1-norm penalty for the ordinary Lasso and the ℓ_2-norm penalty for the group-Lasso. The real-valued and non-negative regularization coefficients

$$\boldsymbol{\psi}_{\mathrm{L}} = \begin{bmatrix} \boldsymbol{\psi}_{\mathrm{GL},1}^T & \boldsymbol{\psi}_{\mathrm{GL},2}^T & \dots & \boldsymbol{\psi}_{\mathrm{GL},S}^T \end{bmatrix}^T \tag{11}$$

$$\boldsymbol{\psi}_{\mathrm{GL},r} = \begin{bmatrix} \psi_{\mathrm{GL},r,1} & \psi_{\mathrm{GL},r,2} & \dots & \psi_{\mathrm{GL},r,L_r} \end{bmatrix}^T \tag{12}$$

are assigned to the individual and grouped sinusoids, respectively, to make a trade-off between the residual and penalties. In [8], the PEBS estimator was formulated using common regularization coefficients for the two norms, such that

$$J = \lambda_{\mathrm{L}} \|\mathbf{a}\|_1 + \sum_{r=1}^{S} \lambda_{\mathrm{GL},r} \|\mathbf{a}_r\|_2 \tag{13}$$

where $\lambda_{\mathrm{L}} = \tau\sigma_v$ and $\lambda_{\mathrm{GL},r} = \tau\sigma_v\sqrt{L_r}$ with the common shrinkage coefficient τ.

As shown in [8], the resulting minimization may suffer from spurious estimates for weak signals and/or onsets, occasionally resulting in an overestimation of the model order. To reduce the occurrence of such spurious estimates, and to allow for a smooth spectral evaluation over frames, we in the following expand on the penalties in (13) to instead allow for more flexible penalty terms. In order to do so, we introduce adaptive weighting of the penalty terms in PEBS, using the notation of an adaptive Lasso, as introduced in [13], such that

$$\|\boldsymbol{\psi}_{\mathrm{GL},r}\|_2 = \frac{\hat{\sigma}_v}{(\|\tilde{\mathbf{a}}_r\|_2)^k} \tag{14}$$

$$\psi_{\mathrm{GL},r,l_r} = \frac{\hat{\sigma}_v}{(|\tilde{a}_{r,l_r}|)^k} \tag{15}$$

where $k > 0$ is a user defined parameter, and with the noise variance being estimated as $\hat{\sigma}_v \cong \|\mathbf{y}_n - \mathbf{W}\tilde{\mathbf{a}}\|_2$, and $\tilde{\mathbf{a}} = E\{\mathbf{a}\big|\boldsymbol{\Psi}, \sigma_v\}$, with $E\{\cdot\}$ denoting the expectation. The resulting adaptive penalty thereby offers a more flexible trade-off between the mean-squared error (MSE) and the bias. The introduced penalty is reminiscent of the iterative re-weighting adaptive Lasso [13], wherein the bias is similarly reduced by applying less shrinkage to the important predictors.

As the frequency content of most audio signals are piecewise smooth [20, 21], it is reasonable to model the dominant components in each frame as being close to those in the earlier and the following frames. Thus, the neighboring frames can be expected have nearly the same expectation of the absolute values, i.e., $E\big\{|\mathbf{a}(n+t)|\big|\boldsymbol{\Psi}, \sigma_v\big\} \simeq E\big\{|\mathbf{a}(n)|\big|\boldsymbol{\Psi}, \sigma_v\big\}$. In practice, one may apply time averaging over $2T+1$ initial

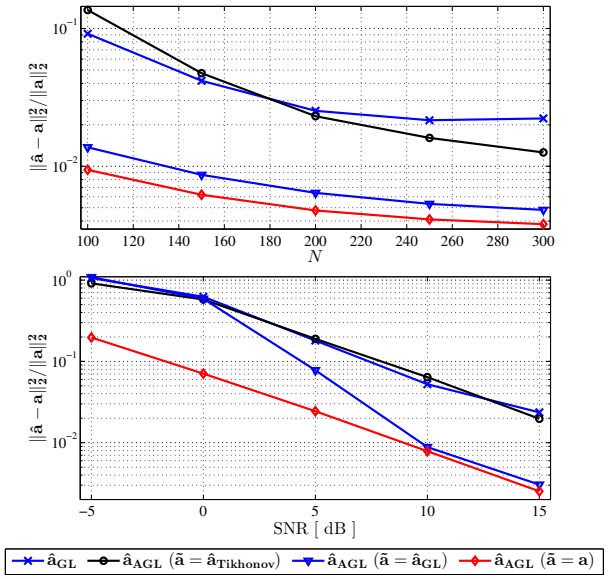

Fig. 1. Normalized MSE of the spectral amplitude estimates versus the sample length N, at SNR $= 10$ dB (top), and versus SNR, using $N = 150$ (bottom).

estimates of $\mathbf{a}(n)$ to find an estimate of the expectation at the time instance n, such that

$$E\{\mathbf{a}(n)\big|\boldsymbol{\Psi}, \sigma_v\} \approx \frac{1}{2T+1} \sum_{t=-T}^{T} \hat{\mathbf{a}}(n+t) \odot \mathbf{h}(t) \tag{16}$$

where $\mathbf{h}(t)$ is a phase shift vector depending on the specific frequencies of the dictionary with unit absolute values, and where $\hat{\mathbf{a}}(n)$ denotes the estimated amplitude vector at time n, as obtained from the initialization or the earlier processed frames. For fast varying spectral content, as well as for poor initial or earlier spectral estimates, we include a spectral smoothing, formed using kernel regression. Here, we make use of the Nadaraya-Watson method introduced in [22], which use a monotonic decay over spectral neighborhood of the considered centroid, such that

$$\bar{a}_{r,l_r} = \frac{\sum_{g=1}^{S}\sum_{l_g=1}^{L_g} K_{\boldsymbol{\Sigma}}(\mathbf{x}_g - \mathbf{x}_r)\ \tilde{a}_{g,l_g}}{\sum_{g=1}^{S}\sum_{l_g=1}^{L_g} K_{\boldsymbol{\Sigma}}(\mathbf{x}_g - \mathbf{x}_r)} \tag{17}$$

where the kernel function is defined as

$$K_{\boldsymbol{\Sigma}}(\mathbf{x}_g - \mathbf{x}_r) = \exp\Big(-\frac{1}{2}(\mathbf{x}_g - \mathbf{x}_r)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_g - \mathbf{x}_r)\Big)$$

with $\boldsymbol{\Sigma}$ denoting the diagonal covariance matrix, giving more weight to the amplitudes $\tilde{a}_{g,l_g}$ at the data point $\mathbf{x}_g = [\omega_g, l_g\omega_g]^T$ that has a smaller Euclidean distance to $\mathbf{x}_r = [\omega_r, l_r\omega_r]^T$.

4. EXPERIMENTAL RESULTS

To investigate the performance of the extended PEBS method, we conducted simulations using both synthetic and real audio

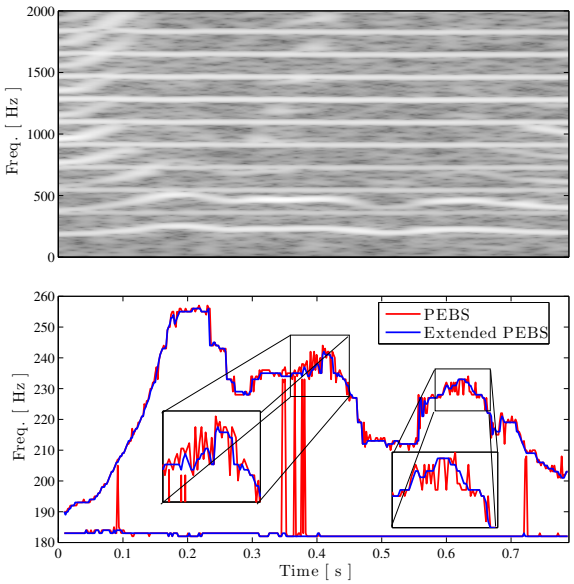

Fig. 2. Spectrogram of the examined speech and trumpet signals (top), and the resulting multi-pitch estimates (bottom).

signals. Since the PEBS method preferably outperforms most stat-of-the-art methods, such as Capon, ANLS, and ORTH [8], we here only compare the found results with the PEBS method. In these simulations, we solved the convex minimization in (9) using the Matlab CVX package [23].

In the first experiment, we estimate the spectral amplitudes of a single-source synthetic signal for varying number of samples and signal-to-noise ratio (SNR). The synthetic signal was generated using the signal model in (2). The fundamental frequency of these signals were uniformly drawn on $\omega_1 \in [160, 290] \times (2\pi/f_s)$, with a uniformly distributed number of harmonics $L_1 \in \mathcal{U}\{5, \lfloor \pi/\omega_1 \rfloor\}$, unit amplitudes, and sampling frequency $f_s = 8.0$ kHz. The used dictionary contained $S = 130$ candidate pitches. The expectation in (14) and (15) was approximated using (16) with $k = 0.5$. Fig. 1 shows the resulting normalized MSE as obtained from 100 Monte-Carlo simulations. As comparison, the figure shows the amplitude estimates of the PEBS method with the adaptive penalties, $\hat{\mathbf{a}}_{\text{AGL}}$, using different initiation estimates: the PEBS amplitude estimates with common penalties ($\tilde{\mathbf{a}} = \hat{\mathbf{a}}_{\text{GL}}$), the Tikhonov[1] amplitude estimates ($\tilde{\mathbf{a}} = \hat{\mathbf{a}}_{\text{Tikh}}$), and the actual amplitudes ($\tilde{\mathbf{a}} = \mathbf{a}$). Here, the user parameters have been set as $\delta = 0.1$, $\lambda_{\text{L}} = 0.12$, and $\lambda_{\text{GL},r} = 0.12\sqrt{L_r}$. As is clear from the figure, the extended PEBS method offers an improved performance as compared to the regular PEBS algorithm, over all considered data lengths (except for the initial estimates using the Tikhonov estimator) and SNRs.

We proceed to examine a real audio signal consisting of a mixture of a female voice and a trumpet signal, corrupted

[1]The Tikhonov estimator is formed as a regularized LS estimate such that $\hat{\mathbf{a}}_{\text{Tikh}} = (\mathbf{W}^H\mathbf{W} + \delta\mathbf{I})^{-1}\mathbf{W}^H\mathbf{y}_n$, where $\delta \geq 0$ is the regularization coefficient, and $\mathbf{I} \in \mathbb{R}^{L_{\text{ext}}}$ is an identity matrix.

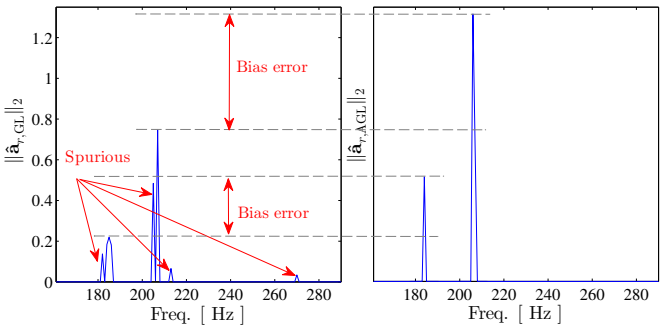

Fig. 3. The ℓ_2-norm of spectral amplitude estimates using the common PEBS method (left), and the extended PEBS method (right).

by an additive white noise, with SNR $= 10$ dB, using $N = 150$ samples per frame. We apply $2T+1 = 3$ initial estimates in (16), using the regular PEBS estimates, and with $\mathbf{\Sigma} = \text{diag}\{6.25, 0.01\} \times (2\pi/f_s)^2$ in the kernel smoother, where $\text{diag}\{\cdot\}$ denotes a diagonal matrix formed from a vector argument. Fig. 2 shows the spectrogram of the examined signal, together with the resulting pitch estimates of the two audio sources. As can be seen from the figure, the extended PEBS method estimates and tracks the audio sources smoothly, whereas the PEBS method suffer from some overshoots. For instance, at time 0.09 sec, the PEBS estimate finds the pitch of one of the sources close to the other, clearly mistakenly the spectral sidelobes of the first source for the pitch of the other signal source (see also Fig. 3). As can be seen from the Fig. 3, the spectral amplitude estimates using the common PEBS method have some spurious non-zeros, and bias in comparison with the extended PEBS method.

5. CONCLUSION

In this work, we have presented a method for multi-pitch estimation and tracking of audio signals such as voiced speech and harmonic musical instruments, without assuming detailed prior knowledge about the signal sources. We have applied a general dictionary consisting of a set of groups for feasible fundamental frequencies and harmonics. Using ℓ_1-norm penalties is a well known solution for such the sparse signal formulation for both the individual and grouped sinusoids. We have shown that the regularization coefficients of the penalty terms should not be identical for all components of the dictionary, and assigned data-dependent regularization coefficients incorporated with an expectation on individual and grouped sinusoids. Experimental results have confirmed that the data-dependent regularization coefficients have a lower bias in comparison with the fixed ones. To track the pitch values smoothly over time, we have also applied a low-pass filter on the expected values to assign monotonic regularization coefficients regarding the spectral and temporal neighborhoods.

6. REFERENCES

[1] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.*, pp. 177–180, IEEE, 2003.

[2] B. Kostek, "Musical instrument classification and duet analysis employing music information retrieval techniques," *Proc. IEEE*, vol. 92, no. 4, pp. 712–729, 2004.

[3] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 731–740, Oct. 2001.

[4] L. Rabiner, M. Cheng, A. E. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 5, pp. 399–418, 1976.

[5] M. Davy, S. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *J. Acoust. Soc. Am.*, vol. 119, pp. 2498–2517, Apr. 2006.

[6] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Process.*, vol. 5, no. 1, pp. 1–160, 2009.

[7] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 982–994, 2007.

[8] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Multi-pitch estimation exploiting block sparsity," *Signal Processing*, vol. 109, pp. 236–247, 2015.

[9] M. G. Christensen, J. L. Højvang, A. Jakobsson, and S. H. Jensen, "Joint fundamental frequency and order estimation using optimal filtering," *EURASIP J. on Applied Signal Processing*, vol. 2011, pp. 1–18, Jun. 2011.

[10] D. Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[11] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, pp. 1917–1930, Apr. 2002.

[12] H. Ney, "A dynamic programming algorithm for nonlinear smoothing," *Signal Processing*, vol. 5, pp. 163–173, March 1983.

[13] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[14] T. Park and G. Casella, "The bayesian lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.

[15] S. L. Marple, Jr., "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Signal Process.*, vol. 47, pp. 2600–2603, Sep. 1999.

[16] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

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

[18] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[19] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *arXiv preprint arXiv:1001.0736*, 2010.

[20] S. Karimian-Azari, N. Mohammadiha, J. R. Jensen, and M. G. Christensen, "Pitch estimation and tracking with harmonic emphasis on the acoustic spectrum," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, 2015.

[21] C. Dubois and M. Davy, "Joint detection and tracking of time-varying harmonic components: a flexible bayesian approach," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 4, pp. 1283–1295, 2007.

[22] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, *The elements of statistical learning*, vol. 2. Springer, 2009.

[23] M. Grant and S. Boyd, "Matlab software for disciplined convex programming," *Online accessiable: http://cvxr.com/cvx*, 2013.