COMPUTATIONALLY EFFICIENT MULTI-PITCH ESTIMATION USING SPARSITY

Shiwen Lei, Filip Elvander, Johan Swärd, Stefan I. Adalbjörnson, and Andreas Jakobsson

Centre for Mathematical Sciences, Lund University, Sweden.
email: {shiwen,filipelv,js,sia,aj}@maths.lth.se

ABSTRACT

In this work, we introduce a computationally efficient multi-pitch estimation algorithm making use of an approximative frequency domain reformulation of a recent block-sparse multi-pitch estimator. Different to most other pitch estimators, the proposed method does not require *a prior* knowledge of the number of sources present, nor the number of overtones of each such source. Evaluated on measured audio signals, the estimator is shown to offer excellent performance at a low computational cost.

1. INTRODUCTION

Recently, there has been a renewed interest in the estimation of the fundamental frequencies of harmonically related signals, such as those occurring, for instance, for voiced speech or tonal audio signal. The fundamental frequency, or pitch, is important in a wide range of applications, such as music information retrieval, source separation, enhancement, compression, and classification (see, e.g., [1–5] and the references therein), as well as in several biomedical, mechanical, and acoustic applications. Traditionally, estimation techniques have been mostly based on the signal correlation [6,7], or on different forms of filterbank or least squares techniques [3]. Typically, this form of techniques require prior knowledge of the model order of the signal, and often only allow for single source signals. Recently, several works have examined how these assumptions may be alleviated, exploiting the underlying sparsity of the harmonic signals to form reliable pitch estimates, without imposing explicit assumptions on the number of sources, or the number of harmonics of each present source [8, 9]. Due to the harmonic structure of pitch signals, these are not only sparse, but also exhibit a block sparse structure, such that a present pitch will activate an entire block of harmonically related components. This structure was exploited in [8], introducing the so-called PEBS (Pitch Estimation using Block Sparsity) estimator. This estimator has then been further refined; in [9], the PEBSI-Lite estimator presented a self-regularising technique for the two tuning parameters dictating the trade-off between sparsity and block-sparsity in the resulting solution. This allowed for a notably improved performance, without necessitating an otherwise typical cross-validation step to determine suitable tuning parameters. Regrettably, the resulting estimator is computationally cumbersome. In this work, we strive to address this problem by presenting a computationally efficient reformulation of the PEBS estimator. The resulting estimator required only a fraction of the computational requirements of PEBS, while achieving almost the same performance as the PEBSI-Lite estimator.

2. PROBLEM FORMULATION

Consider a complex-valued[1] signal consisting of K pitches, where the kth pitch is constituted by a set of L_k harmonically related sinusoids, defined by the component having the lowest frequency, ω_k, such that

$$y(t) = \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} \alpha_{k,\ell} e^{2i\pi f_k \ell t} + \epsilon(t) \tag{1}$$

for $t = 1, \ldots, N$, where $f_k \ell$ is the frequency of the ℓth harmonic in the kth pitch, with $a_{k,\ell}$ denoting the (k, ℓ):th complex-valued amplitude, and $\epsilon(t)$ an additive noise term detailing the non-tonal components in the signal. In most applications, both the number of sources, and the number of overtones of these sources, are typically unknown; generally, it is notoriously difficult to determine these model orders reliably. To avoid this difficulty, the signal model is expanded onto a predefined grid of $P \gg K$ candidate fundamentals, with each candidate having $L_{\text{max}} \geq \max_k L_k$ harmonics. As in [8, 9], L_{max} is selected to ensure that the corresponding highest frequency harmonic is limited by the Nyquist frequency, and varies depending on the considered candidate frequency. For notational simplicity, we will hereafter, without loss of generality, use the same L_{max} for all candidate frequencies. Assuming that the candidate fundamentals are chosen so numerous and so closely spaced that

This work was supported in part by the Swedish Research Council, Carl Trygger's foundation, and the Royal Physiographic Society in Lund.

[1] For notational simplicity and computational efficiency, we here use the discrete-time analytical signal formed from the measured (real-valued) signal (see, e.g., [3, 10]).

Algorithm 1 The PEBS algorithm

```
Let $\mathbf{x}(0) = \mathbf{u}(0) = \mathbf{d}(0) = 0$ and form $\mathbf{A}^H\mathbf{y}$
Form $\mathbf{\Psi}_{\mathbf{A}} = (\mathbf{A}^H\mathbf{A} + \mathbf{I}_M)^{-1}$
for $l = 1, \cdots$ do
    $\mathbf{x}(l+1) = \mathbf{\Psi}_{\mathbf{A}}(\mathbf{A}^H\mathbf{y} - \mathbf{u}(l) + \mathbf{d}(l))$
    for $\mathcal{G} = 1, \cdots, P$ do
        $\mathbf{u}_{\mathcal{G}} = \overline{S}(S(\mathbf{x}_{\mathcal{G}}(l+1) - \mathbf{d}_{\mathcal{G}}(l), \lambda_1), \gamma_{\mathcal{G}})$
    end for
    $\mathbf{d}(l+1) = \mathbf{d}(l) - \mathbf{x}(l+1) + \mathbf{u}(l+1)$
end for
Form $\mathbf{x}$ using (10)
```

the approximation

$$y(t) \approx \sum_{p=1}^{P}\sum_{\ell=1}^{L_{\max}} a_{p,\ell}e^{2\pi i f_p \ell t} + \epsilon(t) \tag{2}$$

holds reasonably well, the number of non-zero amplitudes, $\alpha_{p,\ell}$, should be few, making the signal sparse. By further grouping all the harmonically related components resulting from each pitch candidate in a block structure, such that

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e} \tag{3}$$

where

$$\mathbf{y} = \left[\begin{array}{ccc} y(0) & \dots & y(N-1) \end{array}\right]^T \in \mathbb{C}^{N\times 1} \tag{4}$$

$$\mathbf{A} = \left[\begin{array}{ccc} \mathbf{A}_1 & \dots & \mathbf{A}_P \end{array}\right] \in \mathbb{C}^{N\times M} \tag{5}$$

$$\mathbf{A}_k = \left[\begin{array}{ccc} \mathbf{a}_{k,1} & \dots & \mathbf{a}_{k,L_{\max}} \end{array}\right] \in \mathbb{C}^{N\times L_{\max}} \tag{6}$$

$$\mathbf{a}_{k,\ell} = \left[\begin{array}{ccc} 1 & \dots & e^{2i\pi f_k \ell (N-1)} \end{array}\right]^T \in \mathbb{C}^{N\times 1} \tag{7}$$

$$\mathbf{x} = \left[\begin{array}{ccc} \mathbf{x}_1 & \cdots & \mathbf{x}_M \end{array}\right] \in \mathbb{C}^{M\times 1} \tag{8}$$

$$\mathbf{x}_p = \left[\begin{array}{ccc} a_{p,1} & \cdots & a_{p,L_{\max}} \end{array}\right] \in \mathbb{C}^{N\times 1} \tag{9}$$

where $\mathbf{e}$ formed reminiscent to $\mathbf{y}$, with $M = PL_{\max}$, and

$$\mathbf{x} = \underset{\mathcal{G}}{\cup}\mathbf{x}_{\mathcal{G}} \tag{10}$$

This allows the multi-pitch estimation problem to be formulated as [8]

$$\underset{\mathbf{x}}{\text{minimize}}\ \frac{1}{2}||\mathbf{y} - \mathbf{A}\mathbf{x}||_2^2 + \lambda_1||\mathbf{x}||_1 + \sum_{\mathcal{G}}\gamma_{\mathcal{G}}||\mathbf{x}_{\mathcal{G}}||_2 \tag{11}$$

where λ_1 and $\gamma_{\mathcal{G}}$ are user parameters regulating the relative sparsity and block-sparsity of the solution, respectively. Typically, these parameters are selected using cross-validation or using some simple heuristics. For completeness, the PEBS algorithm is summarised in Algorithm 1. Regrettably, the resulting optimisation is computationally cumbersome, even if implemented using the efficient scheme proposed in [8].

Algorithm 2 The F-PEBS algorithm

```
for  all pitch candidates do
    Compute $\mathbf{A}_{\mathbf{y},\mathcal{G}}$ using the FFT
    Calculate $S(\mathbf{A}^*_{\mathbf{y},\mathcal{G}}, \lambda_1)$ using (19)
    Calculate $\mathbf{x}_{\mathcal{G}} = \overline{S}(S(\mathbf{A}_{\mathbf{y},\mathcal{G}}, \lambda_1), \gamma_{\mathcal{G}})$ using (22)
end for
Form $\mathbf{x}$ using (10)
```

3. THE FREQUENCY-DOMAIN PEBS ESTIMATOR

In order to speed up the minimisation of (11), we proceed to introduce the frequency-domain PEBS estimator, approximating (11) with the optimisation problem

$$\underset{\mathbf{x}}{\text{minimize}}\ \frac{1}{2}||\mathbf{A}^H\mathbf{y} - \mathbf{x}||_2^2 + \lambda_1||\mathbf{x}||_1 + \sum_{\mathcal{G}}\gamma_{\mathcal{G}}||\mathbf{x}_{\mathcal{G}}||_2 \tag{12}$$

which may be viewed as a frequency-domain representation of the measurement being matched to a line model. Given the similarities in the formulation, it may be expected that the minimums of (11) and (12) should be close to each other. Let $\mathbf{A}_{\mathbf{y},\mathcal{G}}$ be the sub-vector of $\mathbf{A}^H\mathbf{y}$ corresponding to the sub-vector $\mathbf{x}_{\mathcal{G}}$. Using (10), the minimisation in (12) may then be expressed as the sum of P sub-problems as

$$\underset{\mathbf{x}_{\mathcal{G}}}{\text{minimize}}\sum_{\mathcal{G}}\ \frac{1}{2}||\mathbf{A}_{\mathbf{y},\mathcal{G}} - \mathbf{x}_{\mathcal{G}}||_2^2 + \lambda_1||\mathbf{x}_{\mathcal{G}}||_1 + \gamma_{\mathcal{G}}||\mathbf{x}_{\mathcal{G}}||_2 \tag{13}$$

As these problems are convex, each minimisation may be formed using the sub-gradient equations with respect to the variable $\mathbf{x}_{\mathcal{G}}$, such that

$$-\left(\mathbf{A}_{\mathbf{y},\mathcal{G}} - \mathbf{x}_{\mathcal{G}}\right)^* + \lambda_1\mathbf{p} + \gamma_{\mathcal{G}}\mathbf{v} = 0 \tag{14}$$

where $(\cdot)^*$ denotes the Hermitian, with $\mathbf{p}$ and $\mathbf{v}$ being the sub-gradients of $||\mathbf{x}_{\mathcal{G}}||_1$ and $||\mathbf{x}_{\mathcal{G}}||_2$, respectively, formed as

$$\mathbf{p}^{(i)} = \begin{cases} \dfrac{\mathbf{x}_{\mathcal{G}}^{(i)*}}{||\mathbf{x}_{\mathcal{G}}^{(i)}||_2}, & \mathbf{x}_{\mathcal{G}} \neq 0 \\ \{\mathbf{p}_{\mathcal{G}}^{(i)} : ||\mathbf{p}_{\mathcal{G}}^{(i)}||_2 \leq 1\}, & \mathbf{x}_{\mathcal{G}} = 0 \end{cases} \tag{15}$$

$$\mathbf{v} = \begin{cases} \dfrac{\mathbf{x}_{\mathcal{G}}^*}{||\mathbf{x}_{\mathcal{G}}||_2}, & \mathbf{x}_{\mathcal{G}} \neq 0 \\ \{\mathbf{v}_{\mathcal{G}} : ||\mathbf{v}_{\mathcal{G}}||_2 \leq 1\}, & \mathbf{x}_{\mathcal{G}} = 0 \end{cases} \tag{16}$$

where $\mathbf{p}^{(i)}$ and $\mathbf{x}_{\mathcal{G}}^{(i)}$ are the i:th elements of the vector $\mathbf{p}$ and the i:th elements of vector $\mathbf{x}_{\mathcal{G}}$, respectively. Substituting (15) and (16) into (14), one obtains

$$-\left(\mathbf{A}_{\mathbf{y},\mathcal{G}} - \mathbf{x}_{\mathcal{G}}\right)^* + \lambda_1\mathbf{p} + \gamma_{\mathcal{G}}\frac{\mathbf{x}_{\mathcal{G}}^*}{||\mathbf{x}_{\mathcal{G}}||_2} = 0 \tag{17}$$

in case $\mathbf{x}_{\mathcal{G}} \neq 0$. Extending the condition in [11] from the real to the complex-valued case, this implies that $||\mathbf{x}_{\mathcal{G}}||_2 = 0$ if

$$|\mathbf{A}_{\mathbf{y},\mathcal{G}}| \leq \lambda_1 \tag{18}$$

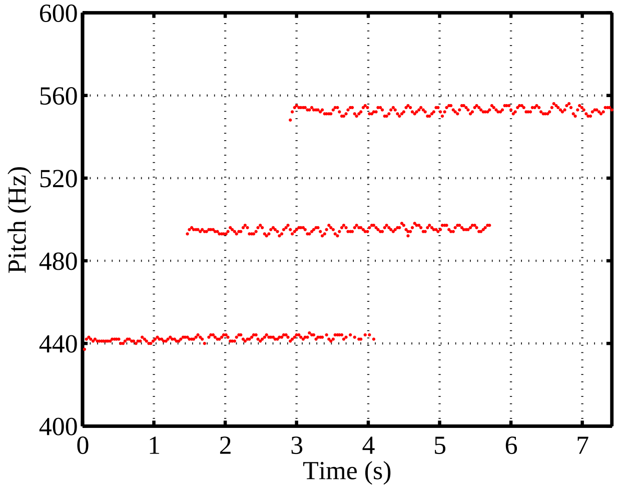

Fig. 1. The F-PEBS estimates for the trumpet signal.

Ensuring that (18) holds implies that

$$\begin{aligned}\mathbf{A}^*_{\mathbf{y},\mathcal{G}} - \lambda_1 \mathbf{p} &= \frac{\mathbf{A}^*_{\mathbf{y},\mathcal{G}}}{||\mathbf{A}_{\mathbf{y},\mathcal{G}}||_2} \max\left\{|\mathbf{A}_{\mathbf{y},\mathcal{G}}| - \lambda_1, 0\right\} \\ &\triangleq S(\mathbf{A}^*_{\mathbf{y},\mathcal{G}}, \lambda_1) \end{aligned} \tag{19}$$

When $\mathbf{x}_{\mathcal{G}} \neq 0$, (17) implies that

$$\mathbf{x}^*_{\mathcal{G}} = \frac{S(\mathbf{A}^*_{\mathbf{y},\mathcal{G}}, \lambda_1)}{1 + \gamma_{\mathcal{G}}/||\mathbf{x}_{\mathcal{G}}||_2} \tag{20}$$

Thus,

$$\begin{aligned} ||\mathbf{x}_{\mathcal{G}}||_2 &= \max\left\{ \max\{|\mathbf{A}_{\mathbf{y},\mathcal{G}}| - \lambda_1, 0\} - \gamma_{\mathcal{G}}, 0\right\} \\ &= \max\left\{||S(\mathbf{A}^*_{\mathbf{y},\mathcal{G}}, \lambda_1)||_2 - \gamma_{\mathcal{G}}, 0\right\} \end{aligned} \tag{21}$$

suggesting that the solution to (12) may be formed as

$$\begin{aligned} \mathbf{x}_{\mathcal{G}} &= \frac{\max\{||S(\mathbf{A}_{\mathbf{y},\mathcal{G}}, \lambda_1)||_2 - \gamma_{\mathcal{G}}, 0\} S(\mathbf{A}_{\mathbf{y},\mathcal{G}}, \lambda_1)}{\max\{||S(\mathbf{A}_{\mathbf{y},\mathcal{G}}, \lambda_1)||_2 - \gamma_{\mathcal{G}}, 0\} - \gamma_{\mathcal{G}}} \\ &\triangleq \overline{S}(S(\mathbf{A}_{\mathbf{y},\mathcal{G}}, \lambda_1), \gamma_{\mathcal{G}}) \end{aligned} \tag{22}$$

where the sub-vector $\mathbf{A}_{\mathbf{y},\mathcal{G}}$ may be efficiently formed using the Fast Fourier transform (FFT). The resulting estimator, here termed the Frequency-domain PEBS (F-PEBS), is summarised in Algorithm 2. Comparing the PEBS and F-PEBS algorithms, one may note that the calculation of $\mathbf{\Psi_A} = (\mathbf{A}^H\mathbf{A} + \mathbf{I}_M)^{-1}$ in the PEBS algorithm requires $\mathcal{O}(M^3)$ operations, followed by the computation of $\mathbf{A}^H\mathbf{y}$, requiring $\mathcal{O}(NM)$ operations per iteration (typically, $Q =$ 15-30 iterations are sufficient for convergence). The computation of F-PEBS initially requires the forming of $\mathbf{A}_{\mathbf{y},\mathcal{G}}$ for each pitch candidate, requiring $\mathcal{O}(M \log M)$ operations per iteration; this is the most computational complicated part in the PEBS. Overall, the PEBS and F-PEBS algorithms thus require $\mathcal{O}(MNQ)$ and $\mathcal{O}(M \log_2 M)$ operations, respectively. The computational costs for the two algorithms are summarised in Table 3. The complexity of PEBSI-Lite is less straight-forward, as it will depend on the structure of the observed signal, but may be expected to be at least as cumbersome as that of PEBS.

Complexity	PEBS	F-PEBS
$(\mathbf{A}^H\mathbf{A} + \mathbf{I}_M)^{-1}$	$\mathcal{O}(M^3)$	None
$\mathbf{A}^H\mathbf{y}$	$\mathcal{O}(M \log_2(M))$	$\mathcal{O}(M \log_2(M))$
$\mathbf{\Psi_A y_A}$	$\mathcal{O}(MN)$	None
$\overline{S}(\cdot)/S(\cdot)$	$\mathcal{O}(M)$	$\mathcal{O}(M)$
Totally	$\mathcal{O}(MNQ)$	$\mathcal{O}(M \log_2(M))$

Table 1. Complexity: PEBS VS. F-PEBS

4. NUMERICAL EXAMPLES

We proceed to examine the performance of the proposed estimator using real audio signals, comparing with the PEBS [8], the PEBSI-Lite [9], and the ESACF [7] estimators. To allow for its best possible performance, ESACF was given oracle knowledge of the number of present pitches, whereas PEBS, PEBSI-Lite, and F-PEBS was not provided any information of the number of sources present. The algorithms have been evaluated on two different data sets, namely that of a sum of three trumpet signals, and that of a set of Bach pieces. Each data was sampled at 44.1 kHz and divided into non-overlapping frames of length 30 ms.

In the first example, we consider a signal containing three trumpets playing the tones A4, B4, and C4,with corresponding fundamental frequencies 440 Hz, 493.88 Hz, and 554.37 Hz (these frequencies were obtained using the approximate nonlinear square method applied to the each trumpet signal individual), lasting 4.3405 s, 4.7161 s, and 4.6177 s, respectively. The three trumpets signals have been added together with the time delay between each two adjacent signals being randomly selected in $\mathcal{U}([0.5, 1.5])$ s. For PEBS, $\lambda_1 = 0.02||\mathbf{A}^H\mathbf{y}||_2^2/2/N$ and $\gamma_{\mathcal{G}} = 0.002$, while for F-PEBS, $\lambda_1 = 0.1||\mathbf{A}^H\mathbf{y}||_2^2/2/N$ and $\gamma_{\mathcal{G}} = 1.5$. Figure 1 shows the resulting pitch estimates for the F-PEBS estimator, illustrating how it is well able to follow the vibrato in the trumpet signal; in this case, the PEBSI-Lite, PEBS and F-PEBS estimators have about the same resolution, whereas ESACF estimates, suffers a little bit degradations (the corresponding figures for the ESCAF and PEBSI-Lite estimators may be found in [9]).

We proceed to evaluated the performance of the estimators on the Bach10 dataset [12]. This dataset consists of ten string quartets composed by Johann Sebastian Bach. The parts are performed by a violin, a clarinet, a saxophone, and a bassoon, with each piece being approximately 30 second-

Performance	F-PEBS	PEBS	PEBSI-Lite	ESACF
Accuracy	0.84	0.86	0.96	0.82
Precision	0.99	0.99	0.98	0.99
Recall	0.84	0.89	0.98	0.82
Time (s)	3.8	878	24060	0.965
Accuracy	0.41	0.39	0.45	0.27
Precision	0.59	0.56	0.63	0.47
Recall	0.58	0.51	0.61	0.37
Time (s)	7.8	3876	25983	2.3

Table 2. Performance measures for the F-PEBS, PEBS, PEBSI-Lite, and the ESACF estimators. The top part of the table corresponds to the results from the trumpet signals, whereas the lower part corresponds to Bach pieces. In the lower part, the accuracy, precision and recall are given for the entire Bach10 pieces, whereas the running times are given only for the first 15 s of one of the pieces, namely *Ach, lieben Christen*.

s long. The data set contains several sequences where the overtones of the instruments overlap. Estimates of the ground truth fundamental frequencies in each frame were obtained by applying YIN [6] to each individual channel. Obvious errors in the YIN estimates were then corrected manually. Figure 2 shows the resulting F-PEBS estimates (the corresponding figures for the PEBSI-Lite and ESACF estimators may be found in [9]). Comparing these results with those of the PEBS-Lite and ESCAF estimators (given as Figures 17 and 18 in [9]), one may note that F-PEBS performs much better than the ESCAF estimators, although worse than the PEBS-Lite estimator. This is also confirmed in Table 2, which summarise the accuracy, precision, and recall (see, e.g., [13]), as well as the relative computational complexity (measured as the relative execution time as compared to the ESACF estimate), for the discussed estimators.

5. CONCLUSIONS

In this work, we have derived a frequency-domain multi-pitch estimator, exploiting the block sparsity of the signal of interest. The resulting estimator, F-PEBS, is found to yield performance similar to the recent PEBSI-Lite estimator, although at a fraction of the computational complexity.

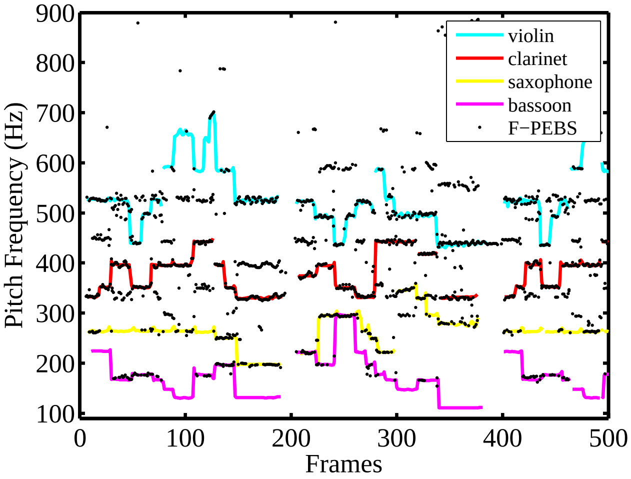

Fig. 2. Pitch tracks produced by F-PEBS when applied to the first 15 s of J.S. Bach's *Ach, lieben Christen*, performed by a violin, a clarinet, a saxophone, and a basson. The ground truth has been formed using the YIN estimator [6] applied to the single source signals.

6. REFERENCES

[1] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Signal Processing*, vol. 88, no. 4, pp. 972–983, April 2008.

[2] A. Klapuri and M. Davy, *Signal Processing Methods for Music Transcription*, Springer, 2006.

[3] M. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, Morgan & Claypool, San Rafael, Calif., 2009.

[4] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 11, no. 6, pp. 804–816, 2003.

[5] M. Müller, D. P. W. Ellis, A. Klapuri, and G. Richard, "Signal Processing for Music Analysis," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1088–1110, 2011.

[6] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[7] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 708–716, 2000.

[8] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Multi-Pitch Estimation Exploiting Block Sparsity," *Elsevier Signal Processing*, vol. 109, pp. 236–247, April 2015.

[9] F. Elvander, T. Kronvall, S. I. Adalbjörnsson, and A. Jakobsson, "An Adaptive Penalty Multi-Pitch Estimator with Self-Regularization," *Elsevier Signal Processing*, vol. 127, pp. 56–70, October 2016.

[10] S. L. Marple, "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2600–2603, September 1999.

[11] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A Sparse-Group Lasso," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.

[12] Z. Duan and B. Pardo, "Bach10 dataset," http://music.cs.northwestern.edu/data/Bach10.html, Accessed December 2015.

[13] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of Multiple-F0 Estimation and Tracking Systems," in *International Society for Music Information Retrieval Conference*, Kobe, Japan, October 2009.