MULTI-PITCH ESTIMATION VIA FAST GROUP SPARSE LEARNING

Ted Kronvall, Filip Elvander, Stefan Ingi Adalbjörnsson, and Andreas Jakobsson

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

ABSTRACT

In this work, we consider the problem of multi-pitch estimation using sparse heuristics and convex modeling. In general, this is a difficult non-linear optimization problem, as the frequencies belonging to one pitch often overlap the frequencies belonging to other pitches, thereby causing ambiguity between pitches with similar frequency content. The problem is further complicated by the fact that the number of pitches is typically not known. In this work, we propose a sparse modeling framework using a generalized chroma representation in order to remove redundancy and lower the dictionary's block-coherency. The found chroma estimates are then used to solve a small convex problem, whereby spectral smoothness is enforced, resulting in the corresponding pitch estimates. Compared with previously published sparse approaches, the resulting algorithm reduces the computational complexity of each iteration, as well as speeding up the overall convergence.

Index Terms— multi-pitch estimation, group lasso, data-adaptive dictionary, generalized chroma features

1. INTRODUCTION

Fundamental frequency estimation of sources consisting of harmonically related sinusoids is a problem frequently arising in areas such as audio processing, non-destructive testing, and biomedical modeling. For example, correctly determining the pitches present in a signal is a fundamental building block in many music information retrieval applications, such as automatic music transcription and genre classification [1]. However, pitch estimation for multi-pitch signals is a difficult problem. Non-parametric methods, such as autocorrelation-based methods (see, e.g., [2] and references therein), generally suffer from the drawback of being unable to distinguish between the fundamental pitch period and multiples of it. Parametric estimators, on the other hand, are more robust to such issues (see, e.g. [3]), but rely heavily on accurate *a priori* model order information of both the number of pitches present and the number of harmonic overtones for each pitch. Also, semi-parametric methods using sparse representations have been successfully implemented, such as in [4], where matching pursuit is used for single pitch estimation. For multiple pitches, the use of sparse reconstruction algorithms allow for estimators not requiring explicit knowledge of the number of sources or their harmonics (see, e.g., [5–7]). In these works, the necessary model order selections are instead formed via suitably chosen tuning parameters, indicating how appropriate a given pitch candidate is to be present in the signal. Typically, such parameters are set using some simple heuristics or via cross-validation, although some efforts have been made to formulate automatic tuning schemes as well [7]. One common difficulty of these methods is that of sub- and super-octave errors, i.e., the case when a 2^n multiple of the fundamental, f_0, is selected in place of the true fundamental, for some $n \in \mathbb{Z}$. In order to avoid these problems, different forms of group penalties have been proposed, as discussed further below, typically increasing the number of tuning parameters, thereby making the suitable selection of such parameters more difficult. In this work, we strive to improve upon our earlier efforts by reducing the required computational complexity of the resulting algorithm, while also introducing a self-regularizing scheme for the selection of the necessary tuning parameters. This is done by introducing a generalized chroma representation allowing suitable chroma candidates to be selected from a small data-adaptive dictionary, yielding a notable reduction in the problem size, as well as an improvement of the convergence speed due to an efficient re-parametrization. The found chroma candidates are then used to form the resulting pitch estimates via a further optimization step aimed at finding the appropriate octave of the found chroma candidates.

2. MULTI-PITCH ESTIMATION USING GENERALIZED CHROMAS

Consider N samples of a complex-valued[1] signal consisting of K sources, with each source being formed by harmonically related sinusoidal components, such that the signal may be well modeled as

$$x(t) = \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} a_{k,\ell} e^{i2\pi f_k \ell t} \qquad (1)$$

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 and computational simplicity, we here consider the discrete-time analytic signal of any real-valued measured signal.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

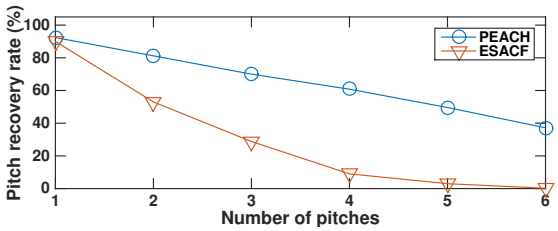

Fig. 1. Pitch recovery rate for the PEACH and ESACF algorithms, for a varying number of pitches. Here, ESACF has been given oracle order information of the number of sources.

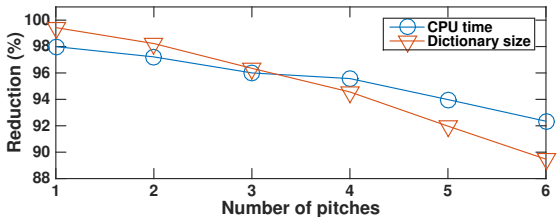

Fig. 2. Reduction in dictionary size and in the required computation time when using the data adaptive pruning scheme, for a varying number of pitches.

where f_k and $a_{k,l}$ denote the k:th normalized fundamental frequency and the complex valued amplitude for the ℓ:th harmonic in the k:th pitch, respectively, and with L_k denoting the number of harmonics of the k:th source. In this work, we strive to form estimates of the fundamental frequencies present in the signal, when the signal is measured in the presence of other signals and noise, here jointly denoted $e(t)$, such that the measured signal is $y(t) = x(t) + e(t)$, with $e(t)$ assumed to be reasonably well modeled as a white, circularly symmetric, Gaussian noise. Typically, both the number of sources, K, and the number of harmonics for each source, L_k are unknown, and may vary noticeably over the measured signal. It is also very common, for instance in audio signals, that some of the overtones for one source overlap with those of another source, complicating the problem further. Finally, the estimates should avoid making the above noted sub- and super octave errors, typically by assuming that the spectral envelope of the sources' harmonics are smooth, such that adjacent harmonics are of comparable magnitude [8].

In order to form the sought pitch estimates, we introduce a sparse reconstruction framework, forming an over-complete dictionary of $P \gg K$ pitch candidates, such that a small subset of these candidates well approximates (1). To allow for an unknown number of overtones, L_k, $\forall k$, we further let $L_{\max} \geq \max_k L_k$ be an upper bound for the number of harmonics in each pitch, allowing (1) to be well approximated as

$$x(t) \approx x_{\mathbf{\Psi}}(t) \triangleq \sum_{p=1}^{P} \sum_{\ell=1}^{L_{\max}} a_{p,\ell} e^{i2\pi f_p \ell t} \tag{2}$$

where $\mathbf{\Psi}$ denotes the set of candidate amplitudes, i.e.,

$$\mathbf{\Psi} = \left\{ \mathbf{\Psi}_{f_1}, \ldots, \mathbf{\Psi}_{f_P} \right\} \tag{3}$$

$$\mathbf{\Psi}_{f_k} = \left\{ a_{k,1}, \ldots, a_{k,L_{\max}} \right\} \tag{4}$$

This formulation allows the pitch estimates to be sought as the set of candidate pitches minimizing

$$\min_{\mathbf{\Psi}} \frac{1}{2} \sum_{t=1}^{N} |y(t) - x_{\mathbf{\Psi}}(t)|^2 + \lambda \sum_{p=1}^{P} \sqrt{\sum_{\ell=1}^{L_{\max}} |a_{p,\ell}|^2} \tag{5}$$

where λ sets the balance between residual fit and the solution assumptions being promoted.

By analyzing the Karush-Kuhn-Tucker condition for (5), one may note that due to the square root in the ℓ_2-norm, it is preferable to cluster components together rather than spreading them out to different groups [9]. Regrettably, the solution in (5) does not resolve the sub- and super octave problem, nor does it prevent the occurrence of elements with zero amplitudes within each active group. As proposed in [5], these issues may be incorporated in the minimization by including further penalties. However, such an extension will also require the need of setting the corresponding tuning parameters, causing a large increase in computational cost for the cross-validation step.

2.1. A generalized chroma representation

As an alternative way of forming the pitch estimates, we here propose the use of a generalized chroma representation. We group pitches together in chromas, such that each chroma collects the pitches that have a largely overlapping frequency content, i.e., the pitches for which the ambiguity problem is the worst. Then, in a post-processing step further detailed below, the most suitable pitches are selected from the found chromas. Let $[f_{\min}, f_{\max})$ denote the range of considered fundamental frequencies, chosen somewhere in the interval $[f_s/N, f_s/2]$, where $f_s/2$ corresponds to the Nyquist frequency. Each candidate fundamental frequencies may then be expressed as

$$f_p = f_{\min} 2^{p/Q}, \quad p = 0, \ldots, P-1 \tag{6}$$

where Q denotes the chroma resolution, stating the number of grid points in a doubling interval $[f, 2f)$, and where $P = \lfloor Q \log_2(f_{\max}/f_{\min}) \rfloor$ is the number of candidate pitches, with $\lfloor \cdot \rfloor$ denoting the truncation operator. We denote the lowest frequency in each generalized chroma group the chroma frequency, all of which are contained within $[f_{\min}, 2f_{\min})$. Thus, $p = c/Q + m$, with $c = p \bmod Q$ and $m = \lfloor p/Q \rfloor$, where $x \bmod y$ denotes the remainder of x after division with y, yielding the chroma indices $c = 0, \ldots, Q-1$, with as many octaves, m, as will fit for $f_p \in [f_c, f_{\max})$, allowing the chroma

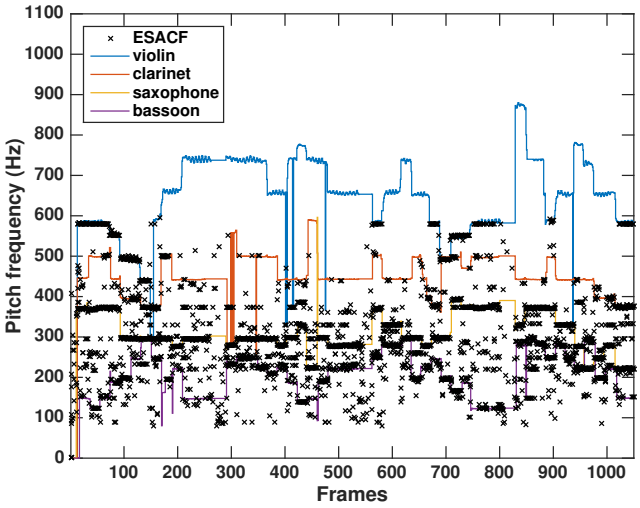

Fig. 3. Pitch tracks produced by ESACF when applied to a 30 second excerpt of J. S. Bach's *Für deinen Thron*, performed by a violin, a clarinet, a saxophone, and a bassoon.

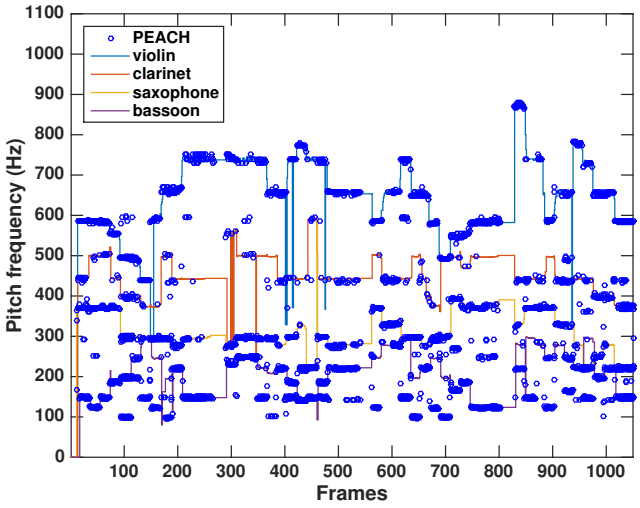

Fig. 4. Pitch tracks produced by PEACH when applied to a 30 second excerpt of J. S. Bach's *Für deinen Thron*, performed by a violin, a clarinet, a saxophone, and a bassoon.

frequency to be expressed as $f_c = f_{\min} 2^{c/Q}$. This reformulation allows (2) to be expressed as

$$x_{\mathbf{\Psi}}(t) \triangleq \sum_{c=0}^{Q-1} \sum_{\ell \in \tilde{\mathcal{I}}_c} a_{c,\ell} e^{i2\pi f_c \ell t} \tag{7}$$

where the set of harmonics for all octaves in the chroma is denoted

$$\tilde{\mathcal{I}}_c = \left\{1 \cdot 2^m, 2 \cdot 2^m, \ldots, L_{\max} \cdot 2^m\right\}_{m=0,\ldots,M_c-1} \tag{8}$$

with M_c being the number of octaves considered in the c:th chroma group. By then estimating $\mathbf{\Psi}$ using (5), with $x_{\mathbf{\Psi}}(t)$ formed as in (7), the frequency content of the signal is clustered into a few chroma groups, from which, as we show below, the corresponding pitches can be readily found. It is worth noting that by using (7), we collect all the highly coherent pitches, i.e., the octaves, within the same group, and thereby reduce the block-coherence of the dictionary. As we initially only strive to find the active chromas in the signal, we may thus restrict our attention to only the subset $\mathcal{I}_c \subset \tilde{\mathcal{I}}_c$, containing only the unique harmonics for the c:th chroma, thereby reducing the number of parameters per chroma by almost a factor 2, as well as increasing the convergence speed and estimation performance due to the resulting reduced dictionary coherence (see also, e.g., [10]).

2.2. Chroma to pitch mapping

Using the estimated chromas, we proceed to mapping these to their suitable octaves, thereby obtaining the desired pitch estimates. This is done by ordering the amplitudes corresponding to the frequencies of the found chroma harmonics, merging any cluster of amplitudes with frequencies that are too closely spaced into a single component (with frequency equal to the mean of each such cluster); this is done to remedy the power leakage into false chromas that can occur due to the limited frequency resolution and off-grid effects. As the expected resolution of the lasso is of the order of $f_s/5N$ [11], all components with frequencies within this limit are thus combined to a single component, with an amplitude equal to the sum of the merged components. The resulting set of amplitudes, $\{d_i\}$, with corresponding frequencies $\{f_i\}$, will thereby avoid the spectral leakage that may be expected by the overlapping, or closely overlapping, harmonics of the different sources. Introducing $b_{p,\ell} \triangleq |a_{p,\ell}|$, the octaves are then found as the solution to the convex optimization problem

$$\min_{\{b_{p,\ell}\},\forall p,\ell} \quad \frac{1}{2}\sum_{i=1}^{J}\left||d_i| - \sum_{\{p,\ell\}\in\mathcal{J}_i} b_{p,\ell}\right|^2 + \kappa \sum_{p=1}^{P}\sum_{\ell=1}^{L_{\max}+1}\left|b_{p,\ell} - b_{p,\ell-1}\right| \tag{9}$$

where $b_{p,0} = b_{p,L_{\max}+1} \triangleq 0, \forall p$, and $\mathcal{J}_i$ is the set of pairs $\{p,\ell\}$ which fulfill

$$|f_{\min} 2^{p/Q}\ell - f_i| \leq f_s/5N \tag{10}$$

The minimization (9) is thus formed such that the found chromas explain the spectral peaks as well as possible, while still promoting solutions where the harmonics are spectrally smooth. Here, we set $\kappa = 0.1 \cdot 2\max_i(|a_i|)$, which corresponds to only accepting a pitch with spectrally smooth magnitudes if its largest magnitude is at least 10% of the largest amplitude cluster.

2.3. Implementation

As the proposed minimization step in (5) is convex, it may be solved using one of the many freely available interior point methods, such as, e.g., SeDuMi [12] and SDPT3 [13]. Alternatively, computationally more efficient methods may be derived, reminiscent to those based on the ADMM introduced in, e.g., [5,6]. Such methods will require somewhere between $\mathcal{O}\left((PL_{\max})^2\right)$ and $\mathcal{O}\left((PL_{\max})^3\right)$ operations, depending on the problem [14]. Even though a single such optimization is quickly solved, the repeated evaluation required for select an appropriate λ using cross-validation quickly becomes computationally cumbersome. In this section, we therefore proceed to introduce an efficient implementation scheme that first reduce the size of the optimization problem, thereby speeding up the chroma estimation, and then efficiently update the found solution in order to form the cross-validation using the earlier found solution.

As the estimation problem in (5) is posed, it is at least K/P-sparse, and so we propose to reduce the number of components in the dictionary by pruning frequencies which likely have zero amplitude. In order to do so, we proceed to determine the total number of dominant frequency components in the signal, not distinguishing between sources or imposing any pitch structure. This may be done is various way, see, e.g., [15, 16]. Here, we use the well-known BIC rule for complex-valued sinusoids [15], selecting the total number of sinusoids, $\hat{J}$, as

$$\hat{J} = \arg\min_j \text{BIC}(j) \triangleq 2N \log \sigma_j^2 + (5j+1)\log N \quad (11)$$

where σ_j^2 is the residual variance when assuming that the signal consist of j sinusoids. To form the residual variance for each assumed order, we here use the MUSIC algorithm and solve for the unknown amplitudes using least squares (see, e.g., [17]). We then prune all harmonics in $\mathcal{I}_c$ corresponding to a frequency which lie further away than $f_j 2^{\pm\delta/Q}$ from the j:th frequency, for any of the $\hat{J}$ frequencies found. For $\delta = 2$, this corresponds to a large reduction in the number of components in the dictionary, typically by factor 10^1 to 10^2, which is illustrated in Figure 2.

We proceed to determine a suitable regularization parameter, λ, required to form (5), using an R-fold cross-validation over a set of S potential candidates $\lambda \in (0, \lambda_{\max}]$, where the validation is initiated using the largest value first, and then evaluated over the decreasing parameter values sequentially. Here,

$$\lambda_{\max} = \max_c \sqrt{\sum_{\ell \in \mathcal{I}_c} \left| \sum_{t=1}^{N} y(t) e^{-i2\pi f_c \ell t} \right|^2} \quad (12)$$

corresponds to the level of regularization where the entire solution becomes zero. As the main cost of the ADMM solver is to factorize and invert the dictionary matrix, this is performed only once prior to the cross-validation, which is then warm-started for each new λ using the latest solution, as a small change in λ will not change the solution much. Thus, the lasso estimation for the entire path of λ values may be done at approximately the same cost as for a single λ [18]. The resulting Pitch Estimation using Adaptive Chroma Heuristics (PEACH) algorithm is summarized in Algorithm 1.

Algorithm 1 The proposed PEACH algorithm

```
Create a generalized chroma dictionary and remove non-unique elements within each chroma
Find the number of frequencies in the signal using BIC, and estimate their locations
Remove all elements from the dictionary not being close to any found frequencies
for all R folds of cross-validation scheme do
    Factorize the dictionary, invert, and store
    Initiate estimation using a zero chroma solution
    for all candidate λ in the regularization path do
        Load inverted dictionary and latest chroma solution
        Do sparse chroma estimation at current λ and fold r
    end for
end for
Map active chroma to pitches
return pitch frequency estimates
```

3. NUMERICAL VALIDATION

In this section, we evaluate the efficiency of the proposed method using both synthetic and real data. Initially, we evaluate the proposed method on a 30 ms simulated signal constituted by 1 to 6 pitches, sampled at 44.1 kHz. In each simulation, every pitch frequency was drawn uniformly on the interval $[50, 1200]$ Hz, with each pitch containing (a uniform distribution of) 7 to 10 harmonics. Figure 1 shows the pitch recovery rate (PRR) for the PEACH algorithm and the ESACF estimator [2], clearly showing a preferred performance from the proposed algorithm. Here, each octave has been divided into $Q = 96$ chromas, we use R = 10 fold cross-validation for λ, and the PRR has been defined as the fraction of the simulations in which the correct number of pitches was found and where the estimated fundamental frequencies differed with less than 2 grid points, i.e., $\frac{1/4}{12}$ of an octave, from the ground truth. These results have been obtained using 500 Monte-Carlo simulations for each number of pitches, with both methods assuming a maximum harmonic order $L_{\max} = 15$ for each source. Here, ESACF has been allowed oracle information of the number of sources present, whereas PEACH has determined this as part of the estimation procedure.

Proceeding, we evaluate the discussed algorithms on real audio, using a recording of J. S. Bach's *Für deinen Thron*, performed by a violin, a clarinet, a saxophone, and a bassoon. The recording was taken from the Bach10 dataset [19].

	PEACH	ESACF
Accuracy	0.492	0.353
Precision	0.722	0.562
Recall	0.607	0.486

Table 1. Performance measures for the PEACH and ESACF algorithms, when evaluated on J. S. Bach's *Für deinen Thron*.

The signal was sampled at 44.1 kHz, then decimated to 22.05 kHz, and divided into frames of length 30 ms. As before, we compare the performance of PEACH to that of ESACF, with both algorithms being given the maximum harmonic order $L_{\max} = 10$, and ESACF oracle information of the number of pitches in *each* processed frame. Figures 3 and 4 show pitch tracks obtained using PEACH and ESACF, respectively, together with the ground truth estimates of each instrument's pitch frequency, obtained by applying YIN [20] to each single-source channel. As can be seen, PEACH produces considerably more consistent pitch estimates than ESACF, which has a tendency to erroneously pick sub-octaves instead of present pitches. Thereby, ESACF mostly misses the high-pitched violin part of the music piece, whereas PEACH tracks the violin quite accurately. From Figures 3 and 4, it can be seen that both algorithms have trouble with estimating the pitch frequency of the clarinet. This is caused by the clarinet's harmonic structure; most of the power is concentrated to the first harmonic, making it susceptible to being picked up by one of the other pitches. Performance measures, as defined in [21], are presented in Table 1, confirming the superior performance of the proposed PEACH algorithm.

4. REFERENCES

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

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

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

[4] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 101–111, jan. 2003.

[5] 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.

[6] T. Kronvall, F. Elvander, S. I. Adalbjörnsson, and A. Jakobsson, "An Adaptive Penalty Approach to Multi-Pitch Estimation," in *23rd European Signal Processing Conference*, Nice, France, Aug. 31 - Sept. 4 2015.

[7] F. Elvander, T. Kronvall, S. I. Adalbjörnsson, and A. Jakobsson, "An Adaptive Penalty Multi-Pitch Estimatior with Self-Regularization," *to appear in Elsevier Signal Processing*.

[8] 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.

[9] X. Lv, G. Bi, and C. Wan, "The Group Lasso for Stable Recovery of Block-Sparse Signal Representations," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1371–1382, 2011.

[10] Y. V. Eldar, P. Kuppinger, and H. Bolcskei, "Block-Sparse Signals: Uncertainty Relations and Efficient Recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.

[11] J. Karlsson and L. Ning, "On Robustness of l1-Regularization Methods for Spectral Estimation," in *IEEE 53rd Annual Conference on Decision and Control*, Dec 2014, pp. 1767–1773.

[12] J. F. Sturm, "Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11-12, pp. 625–653, August 1999.

[13] R. H. Tutuncu, K. C. Toh, and M. J. Todd, "Solving semidefinite-quadratic-linear programs using SDPT3," *Mathematical Programming Ser. B*, vol. 95, pp. 189–217, 2003.

[14] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[15] P. Stoica and Y. Selén, "Model-order Selection — A Review of Information Criterion Rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, July 2004.

[16] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Sinusoidal Order Estimation using Angles between Subspaces," *EURASIP Journal on Advances in Signal Processing*, 2009, Article ID 948756, 11 pages.

[17] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Prentice Hall, Upper Saddle River, N.J., 2005.

[18] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

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

[20] 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.

[21] 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.