Sparse Modeling of Harmonic Signals

Elvander, Filip

2017

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Elvander, F. (2017). *Sparse Modeling of Harmonic Signals*. [Licentiate Thesis, Mathematical Statistics]. Lund University / Centre for Mathematical Sciences /LTH.

Total number of authors:
1

SPARSE MODELING OF HARMONIC SIGNALS

FILIP ELVANDER

LUND UNIVERSITY

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118
SE-221 00 Lund
Sweden

http://www.maths.lu.se/

Licentiate Theses in Mathematical Sciences 2017:1
ISSN 1404-028X

LUTFMS-2018-2017

Printed in Sweden by Media-Tryck, Lund 2017

Contents

Acknowledgements

I would like to thank my supervisor, Prof. Andreas Jakobsson, for all his help, support, super-human enthusiasm, and advice on matters small and large during these two years. I would also like to thank my co-authors, Dr. Stefan Ingi Adalbjörnsson, Johan Swärd, Ted Kronvall, Dr. Shiwen Lei, and Prof. Johan Karlsson, for giving substance to the expression ”the more, the merrier”. Especially, I would like to thank Stefan for giving me invaluable guidance concerning research, as well as supplying me with interesting audio books. Also, a special thanks to Mona Forsler, James Hakim, Joakim Lübeck, Maria Lövgren, and Lise-Lotte Mörner for invariably resolving all practical issues. To my parents, Christine and Johan, and my sister Linnéa, thank you for all support and for always being my sanity check. Lastly, to my girlfriend Anna-Lovisa, for being my center of gravity and partner in crime.

Lund, 2017 *Filip Elvander*

List of papers

This thesis is based on the following papers:

A Filip Elvander, Stefan Ingi Adalbjörnsson, Ted Kronvall, and Andreas Jakobsson, "An Adaptive Penalty Multi-Pitch Estimator with Self-Regularization". *Elsevier Signal Processing*, vol. 127, pp. 56-70, October 2016.

B Filip Elvander, Johan Swärd, and Andreas Jakobsson, "Online Estimation of Multiple Harmonic Signals". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 273-284, February 2017.

C Filip Elvander, Stefan Ingi Adalbjörnsson, Johan Karlsson, and Andreas Jakobsson, "Using Optimal Transport for Estimating Inharmonic Pitch Signals". *42nd IEEE International Conference on Acoustics, Speech, and Signal Processing*. New Orleans, USA, March 5-9, 2017.

Additional papers not included in the thesis:

1. Johan Swärd, Filip Elvander, and Andreas Jakobsson, "Designing Optimal Sampling Schemes", submitted to the *25th European Signal Processing Conference*, Kos island, Greece, August 28 - September 2, 2017.

2. Filip Elvander, Johan Swärd, and Andreas Jakobsson, "Time-Recursive Multi-Pitch Estimation Using Group Sparse Recursive Least Squares", *50th Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, USA, November 6-9, 2016.

3. Shiwen Lei, Filip Elvander, Johan Swärd, Stefan Ingi Adalbjörnsson, and Andreas Jakobsson, "Computationally Efficient Multi-Pitch Estimation Using Sparsity", *11th IMA International Conference on Mathematics in Signal Processing*, Birmingham, UK, November 12-14, 2016.

4. Filip Elvander, Stefan Ingi Adalbjörnsson, and Andreas Jakobsson, "Robust Non-Negative Least Squares Using Sparsity", *24th European Signal Processing Conference*, Budapest, Hungary, August 29 - September 2, 2016.

5. Ted Kronvall, Filip Elvander, Stefan Ingi Adalbjörnsson, and Andreas Jakobsson, "Multi-Pitch Estimation via Fast Group Sparse Learning", *24th European Signal Processing Conference*, Budapest, Hungary, August 29 - September 2, 2016.

6. Ted Kronvall, Filip Elvander, Stefan Ingi Adalbjörnsson, and Andreas Jakobsson, "An Adaptive Penalty Approach to Multi-Pitch Estimation", *23rd European Signal Processing Conference*, Nice, France, August 31 - September 4, 2015.

Introduction

This thesis consists of three papers concerned with estimation of multi-pitch signals, i.e., signals consisting of one or more harmonic, or close-to-harmonic, structures characterized by a set of fundamental frequencies. Estimation in this context refers to the task of both finding the number of these structures constituting the signal, as well as their fundamental frequencies. The first paper considers stationary signal frames and exploits the assumed perfectly harmonic structure in order to extract the individual pitches. The second paper, while also relying on the harmonic assumption, presents a time-recursive pitch estimator, allowing for non-stationarities such as amplitude and frequency modulation. Lastly, the third paper presents a relaxation of the harmonic model, allowing the pitches to display various degrees of inharmonicity and thereby introduces a framework for robust multi-pitch estimation. The three papers have in common that they all rely on formulating multi-pitch estimation as a sparse reconstruction problem and utilize convex optimization techniques in order to produce the estimates.

1 Background

A *pitch* is defined as a set of sinusoids whose frequencies are integer multiples of a single frequency, referred to as the fundamental frequency or pitch frequency, i.e., the sinusoidal frequencies constituting a pitch with fundamental frequency f_k are contained in the set

$$\boldsymbol{\Omega}_k \subseteq \{ f_k \ell \mid \ell = 1, \ldots, L_k \} \tag{1}$$

where L_k is referred to as the harmonic order. This type of signal structure appears in a variety of applications, ranging from audio processing to radar applications with rotating or vibrating targets [1]. For example, the harmonic structure in (1) quite accurately captures the voiced part of human speech, e.g., sustained vowels, although the harmonic structure is in general not perfect due to the phenomenon of inharmonicity. This can be seen in Figure 1, showing the magnitude of the short-time Fourier transform (STFT) for a recording of a female voice, i.e., a

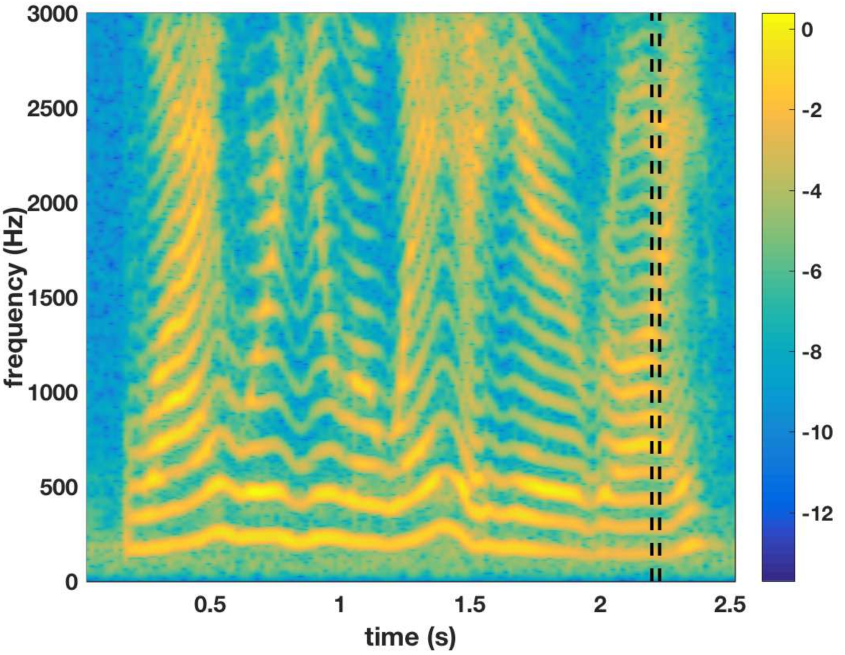

Figure 1: Magnitude of the STFT for a signal consisting of a female voice. The magnitude, in dB, is illustrated by the color of the plot.

single-pitch signal, sampled at 8 kHz, with the magnitude in dB illustrated by the color of the plot. As can be seen, at each time instant the power is distributed over evenly spaced frequencies, with the spacing corresponding to the fundamental frequency at that particular time instant. In Figure 2, the magnitude spectrum of a 30 ms excerpt of the full signal is shown, corresponding to the portion between the dashed lines in Figure 1. In the figure, one may note that the spectral peaks are separated by approximately 143 Hz, i.e., the fundamental frequency on this time interval is 143 Hz, although, there seems to be some inharmonicity present as the frequency of the tenth harmonic is slightly higher than the expected 1430 Hz. In this thesis, this type of harmonic signal is assumed to be well described by

$$x_k(t) = \sum_{\ell=1}^{L_k} a_{k,\ell} e^{i2\pi f_k \ell t} , \tag{2}$$

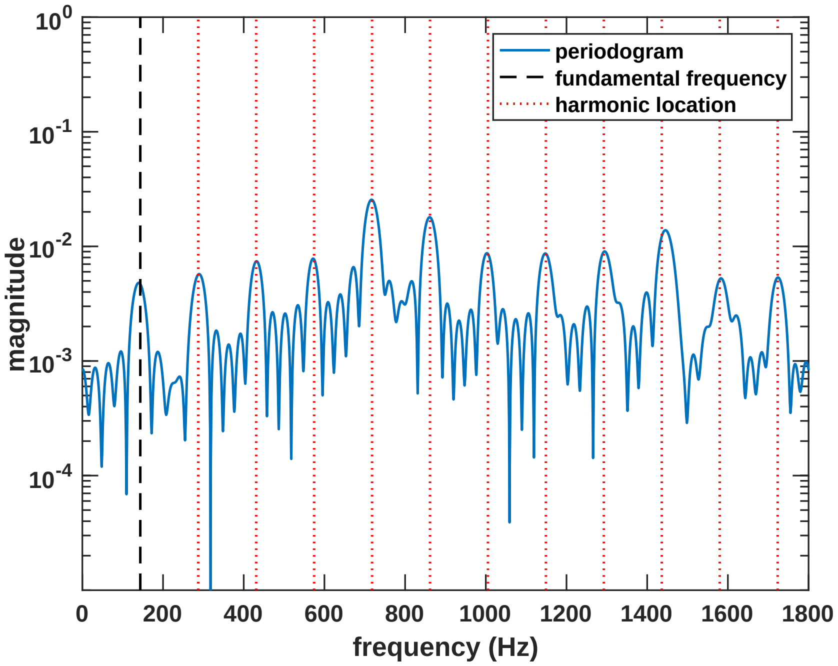

Figure 2: Magnitude spectrum of a signal consisting of a female voice. The location of the harmonics are indicated by the vertical lines.

i.e., as the superposition of L_k harmonically related complex sinusoids. Although, typically, measured signals are real-valued, for ease of notation and computational efficiency, we will herein instead consider the discrete-time, analytic representation [2]. From (2), it can be noted that the reciprocal of the fundamental frequency, f_k, corresponds to the period of the signal, i.e.,

$$x_k(t) = x_k\left(t + T_k\right) = x_k\left(t + 1/f_k\right) \ . \tag{3}$$

This is illustrated in Figure 3, showing the (real-valued) time-domain wave-form for the same signal as before. The zoomed-in portion of the figure corresponds to the same 30 ms as shown in Figure 2. As can be seen in the figure, the signal is approximately periodic, completing a bit more than four periods on the 30 ms interval. This observation can be utilized in single-pitch scenarios in order to estimate the fundamental frequency, f_k, based on the autocorrelation of the signal [3]. In contrast, this thesis is concerned with the more complex task of

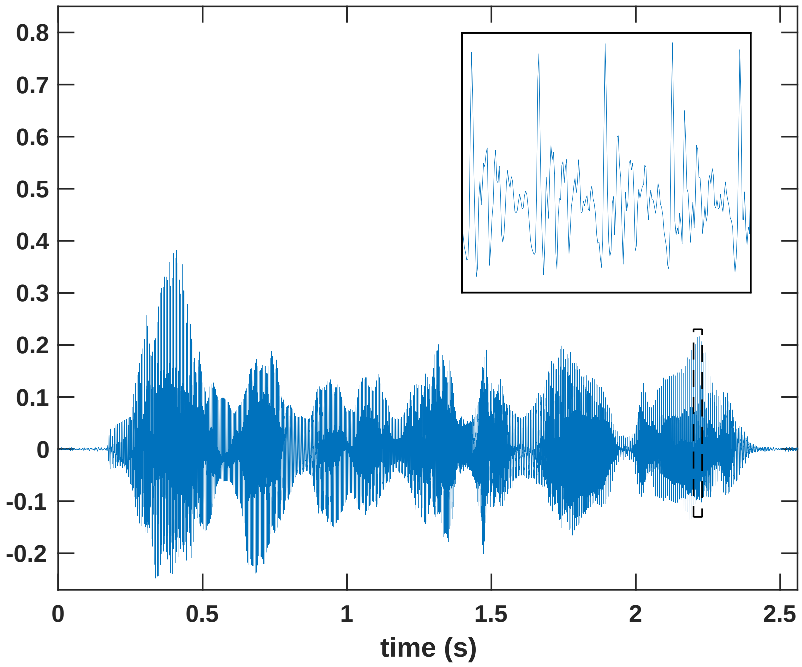

Figure 3: Single-pitch signal consisting of a female voice. The zoomed-in portion corresponds to the spectrum in Figure 2.

multi-pitch estimation, i.e., estimating a set of fundamental frequencies, $\{f_k\}_{k=1}^{K}$, from noise corrupted signal measurements $y(t)$ (for an overview on multi-pitch estimation, see, e.g., [4]). Specifically, a stationary frame of the signal is assumed to be well described by

$$y(t) = x(t) + e(t) \tag{4}$$

where

$$x(t) = \sum_{k=1}^{K} x_k(t) = \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} a_{k,\ell} e^{i2\pi f_k \ell t} , \tag{5}$$

and where $e(t)$ denotes an additive noise. The estimators presented in the three papers in this thesis will not assume any knowledge of the number of pitches, K, nor the number of harmonics, L_k, of each pitch and will instead consider these to be unknowns to be estimated alongside the fundamental frequencies, $\{f_k\}_{k=1}^{K}$.

2 Maximum likelihood estimation for pitches

Assume that one measures N noise corrupted samples, $y(t_n)$, from the model (4), collected in the vector

$$\mathbf{y} = \left[\begin{array}{ccc} y(t_1) & \dots & y(t_N) \end{array} \right]^T . \tag{6}$$

Also, assume for now that the number of pitches K, as well as the number of harmonics L_k for each pitch, are known so that only the fundamental frequencies, $\{f_k\}_{k=1}^K$, and the complex amplitudes of the harmonics, $a_{k,\ell}$, are the unknown parameters to be estimated. Define the parameter vector $\boldsymbol{\vartheta}$ as

$$\boldsymbol{\vartheta} = \left[\begin{array}{cc} \mathbf{f}^T & \mathbf{a}^T \end{array} \right]^T \tag{7}$$

where

$$\begin{aligned} \mathbf{f} &= \left[\begin{array}{cccc} f_1 & f_2 & \cdots & f_K \end{array} \right]^T && (8) \\ \mathbf{a} &= \left[\begin{array}{cccc} \mathbf{a}_1^T & \mathbf{a}_2^T & \cdots & \mathbf{a}_K^T \end{array} \right]^T && (9) \\ \mathbf{a}_k &= \left[\begin{array}{cccc} a_{k,1} & a_{k,2} & \dots & a_{k,L_k} \end{array} \right]^T . && (10) \end{aligned}$$

Then, one may write $\mathbf{y}$ as

$$\mathbf{y} = \mathbf{Z}(\mathbf{f})\mathbf{a} + \mathbf{e} \tag{11}$$

where

$$\begin{aligned} \mathbf{Z}(\mathbf{f}) &= \left[\begin{array}{cccc} \mathbf{Z}_1(f_1) & \mathbf{Z}_2(f_2) & \cdots & \mathbf{Z}_K(f_K) \end{array} \right] && (12) \\ \mathbf{Z}_k(f_k) &= \left[\begin{array}{cccc} \mathbf{z}(f_k) & \mathbf{z}(2f_k) & \cdots & \mathbf{z}(L_k f_k) \end{array} \right] && (13) \\ \mathbf{z}(f_k) &= \left[\begin{array}{cccc} \mathrm{e}^{i2\pi f_k t_1} & \mathrm{e}^{i2\pi f_k t_2} & \cdots & \mathrm{e}^{i2\pi f_k t_N} \end{array} \right]^T && (14) \\ \mathbf{e} &= \left[\begin{array}{cccc} e(t_1) & e(t_2) & \cdots & e(t_N) \end{array} \right]^T . && (15) \end{aligned}$$

A common assumption regarding the additive noise component, $e(t)$, is that it may be well modeled as being a white, circularly symmetric Gaussian noise with variance σ^2. Under this assumption, the probability density function (PDF) of the sample $\mathbf{y}$ is

$$p\left(\mathbf{y}; \boldsymbol{\vartheta}, \sigma^2\right) = \frac{1}{\left(\pi\sigma^2\right)^N} \exp\left\{ \frac{1}{\sigma^2} \left\| \mathbf{y} - \mathbf{Z}(\mathbf{f})\mathbf{a} \right\|_2^2 \right\} . \tag{16}$$

This yields that the maximum-likelihood estimator (MLE) of $\boldsymbol{\vartheta}$ is the solution to the non-linear least squares problem

$$\underset{\boldsymbol{\vartheta}}{\text{minimize}}\ \frac{1}{N}\left\|\mathbf{y}-\mathbf{Z}(\mathbf{f})\mathbf{a}\right\|_2^2 \ . \tag{17}$$

Let the minimizer of (17) with respect to $\mathbf{f}$ be denoted $\mathbf{f}^\star$. Then, the minimizing amplitude vector is given by

$$\mathbf{a}^\star = \mathbf{Z}^\dagger(\mathbf{f}^\star)\mathbf{y} \tag{18}$$

where $\mathbf{Z}^\dagger(\mathbf{f}^\star)$ denotes the Moore-Penrose pseudo-inverse of $\mathbf{Z}(\mathbf{f}^\star)$, which in the case $\mathbf{Z}(\mathbf{f}^\star)$ is a full rank matrix has the closed-form expression

$$\mathbf{Z}^\dagger(\mathbf{f}^\star) = \left(\mathbf{Z}^H(\mathbf{f}^\star)\mathbf{Z}(\mathbf{f}^\star)\right)^{-1}\mathbf{Z}^H(\mathbf{f}^\star) \ . \tag{19}$$

Substituting this into (17), one may define the cost function

$$J(\mathbf{f};\mathbf{y}) = \frac{1}{N}\left\|\left(\mathbf{I}-\mathbf{Z}(\mathbf{f})\mathbf{Z}^\dagger(\mathbf{f})\right)\mathbf{y}\right\|_2^2 \tag{20}$$

which only depends on the vector of fundamental frequencies, $\mathbf{f}$. We thus have the MLEs of $\mathbf{f}$ and $\mathbf{a}$ according to

$$\hat{\mathbf{f}}_{\text{MLE}} = \arg\min_{\mathbf{f}} J(\mathbf{f};\mathbf{y}) \tag{21}$$

$$\hat{\mathbf{a}}_{\text{MLE}} = \mathbf{Z}^\dagger\left(\hat{\mathbf{f}}_{\text{MLE}}\right)\mathbf{y} \ . \tag{22}$$

However, finding the minimizer of $J(\mathbf{f};\mathbf{y})$ is a non-trivial task as $J(\mathbf{f};\mathbf{y})$ is non-convex in $\mathbf{f}$. This is illustrated in Figure 4, showing the values of $J(\mathbf{f};\mathbf{y})$ when evaluated on $f \in (0, 0.018)$ for a single-pitch signal with fundamental frequency $f_k = 0.01$ (in normalized frequency) and $L_k = 6$ harmonics. As can be seen, the cost function has a clear global minimum close to the true fundamental frequency. However, there are also numerous local minima, preventing the use of straightforward optimization schemes such as gradient descent. Also, without accurate prior knowledge of the locations of the fundamental frequencies, using a grid-search to find f_k is only computationally feasible in the single-pitch case as each evaluation of $J(\mathbf{f};\mathbf{y})$ requires $\mathcal{O}(N^3)$ operations for the calculation of the pseudo-inverse. This makes multi-pitch estimation for $K > 1$ pitches a computationally daunting task, as this would require a K-dimensional grid search.

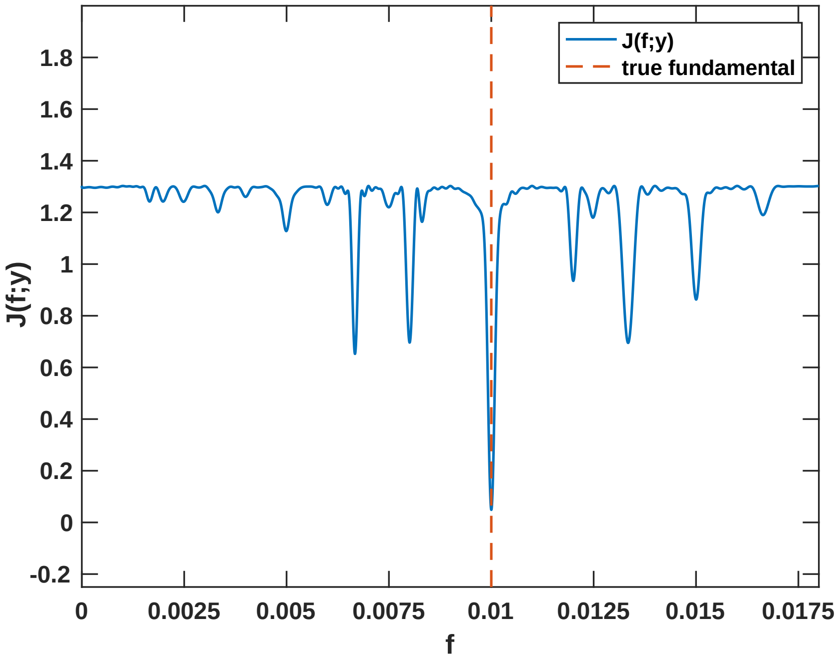

Figure 4: The cost function $J(\mathbf{f};\mathbf{y})$ evaluated for frequencies $f \in (0, 0.018)$ for a single-pitch signal with fundamental frequency $f_k = 0.01$ and $L_k = 6$ harmonics.

2.1 Approximating the MLE

In order to overcome the computational problems of finding the MLE by exact minimization of the cost function $J(\mathbf{f};\mathbf{y})$, one may instead proceed to use an approximation. First, assuming that the sampling is uniform, one may note that, asymptotically

$$\frac{1}{N}\mathbf{z}^H(f_k)\mathbf{z}(f_\ell) = \begin{cases} 1 & \text{if } k = \ell \\ 0 & \text{if } k \neq \ell \end{cases} \tag{23}$$

as $N \to \infty$, i.e.,

$$\lim_{N\to\infty} \frac{1}{N}\mathbf{Z}^H(\mathbf{f})\mathbf{Z}(\mathbf{f}) = \mathbf{I} \Rightarrow \lim_{N\to\infty} \mathbf{Z}^\dagger(\mathbf{f}) = \frac{1}{N}\mathbf{Z}^H(\mathbf{f}) \tag{24}$$

where $\mathbf{I}$ is the identity matrix, assuming that no harmonics of the different pitches overlap, i.e., $f_{k,\ell} \neq f_{p,n}$ unless $k = p$ and $\ell = n$. Using this approximation, one may reformulate the cost function, by expanding the expression for $J(\mathbf{f};\mathbf{y})$, obtaining

$$\tilde{J}(\mathbf{f};\mathbf{y}) = \frac{1}{N}\left\|\left(\mathbf{I} - \frac{1}{N}\mathbf{Z}(\mathbf{f})\mathbf{Z}^H(\mathbf{f})\right)\mathbf{y}\right\|_2^2 \tag{25}$$

$$= \frac{1}{N}\left(\|\mathbf{y}\|_2^2 - \frac{1}{N}\left\|\mathbf{Z}^H(\mathbf{f})\mathbf{y}\right\|_2^2\right) \tag{26}$$

$$= \frac{1}{N}\left(\|\mathbf{y}\|_2^2 - \frac{1}{N}\sum_{k=1}^{K}\left\|\mathbf{Z}_k^H(f_k)\mathbf{y}\right\|_2^2\right) . \tag{27}$$

As this cost function is separable in the K pitches, each individual fundamental frequency can be found as

$$\hat{f}_k = \arg\max_{f} \ \left\|\mathbf{Z}_k^H(f)\mathbf{y}\right\|_2^2 \tag{28}$$

which due to the structure of $\mathbf{Z}_k(f)$ can be implemented efficiently using the Fast Fourier Transform (FFT) [5] together with a grid search. However, when estimating the fundamental frequencies of mixtures of $K > 1$ pitches, one has to be careful not to pick the same maximum of (28) more than once, as it for some signals is possible that

$$\arg\max_{f} \ \left\|\mathbf{Z}_k^H(f)\mathbf{y}\right\|_2^2 = \arg\max_{f} \ \left\|\mathbf{Z}_\ell^H(f)\mathbf{y}\right\|_2^2 \ , \ k \neq \ell \, . \tag{29}$$

As an illustration, consider a signal consisting of two pitches with fundamental frequencies $f_1 = 0.01$ and $f_2 = 0.0095$, and $L_1 = 6$ and $L_2 = 4$ harmonics, respectively. Also, let the second pitch have higher power than the first, i.e.,

$$\sum_{\ell=1}^{6} |a_{1,\ell}|^2 < \sum_{\ell=1}^{4} |a_{2,\ell}|^2 \ . \tag{30}$$

The exact cost function $J(\mathbf{f};\mathbf{y})$ for this case is shown in Figure 5, whereas the approximate cost function $\tilde{J}(\mathbf{f};\mathbf{y})$ is shown in Figure 6. As can be seen, $J(\mathbf{f};\mathbf{y})$ attains its global minimum close to $\mathbf{f} = \begin{bmatrix} f_1 & f_2 \end{bmatrix}^T$, whereas $\tilde{J}(\mathbf{f};\mathbf{y})$ attains its global minimum close to $\mathbf{f} = \begin{bmatrix} f_2 & f_2 \end{bmatrix}^T$, due to f_2 maximizing both $\left\|\mathbf{Z}_1^H(f)\mathbf{y}\right\|_2^2$

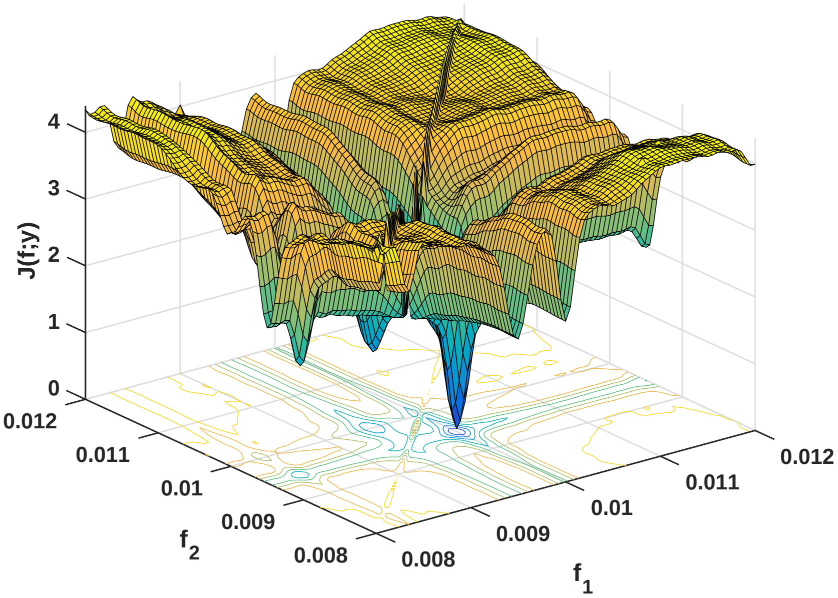

Figure 5: The cost function $J(\mathbf{f};\mathbf{y})$ evaluated for frequencies $\mathbf{f} \in (0.008, 0.012)^2$ for a two-pitch signal with fundamental frequencies $f_1 = 0.01$, $f_2 = 0.0095$, having $L_1 = 6$ and $L_2 = 4$ harmonics, respectively.

and $||\mathbf{Z}_2^H(f)\mathbf{y}||_2^2$. The ground truth fundamental frequencies f_1 and f_2 are instead located close to a local minimum of $\tilde{J}(\mathbf{f};\mathbf{y})$. Thus, estimating fundamental frequencies by minimizing $\tilde{J}(\mathbf{f};\mathbf{y})$ requires the use of a scheme for avoiding picking the same pitch more than once, which is a non-trivial task for the case of noisy signals and closely spaced pitches. Also, it should be noted that the success of such a pitch picking scheme depends on the order in which the pitches are selected; in our two-pitch example, starting with selecting an estimate of f_1 will produce the estimate $\hat{\mathbf{f}} \approx \begin{bmatrix} f_2 & f_1 \end{bmatrix}^T$, whereas starting with f_2 will produce $\hat{\mathbf{f}} \approx \begin{bmatrix} f_1 & f_2 \end{bmatrix}^T$. A refinement of this estimate by minimizing the exact cost function $J(\mathbf{f};\mathbf{y})$ in a neighborhood of $\hat{\mathbf{f}}$ would then only be successful in the second case, as the first case, where the pitches have been swapped, will incur a bias in the estimates due to the mis-match of the harmonic orders. It should also be noted that $\tilde{J}(\mathbf{f};\mathbf{y})$ might be a poor approximation of $J(\mathbf{f};\mathbf{y})$ for cases in which the sinusoidal components of the signal are not approximately orthogonal. This will be the case if

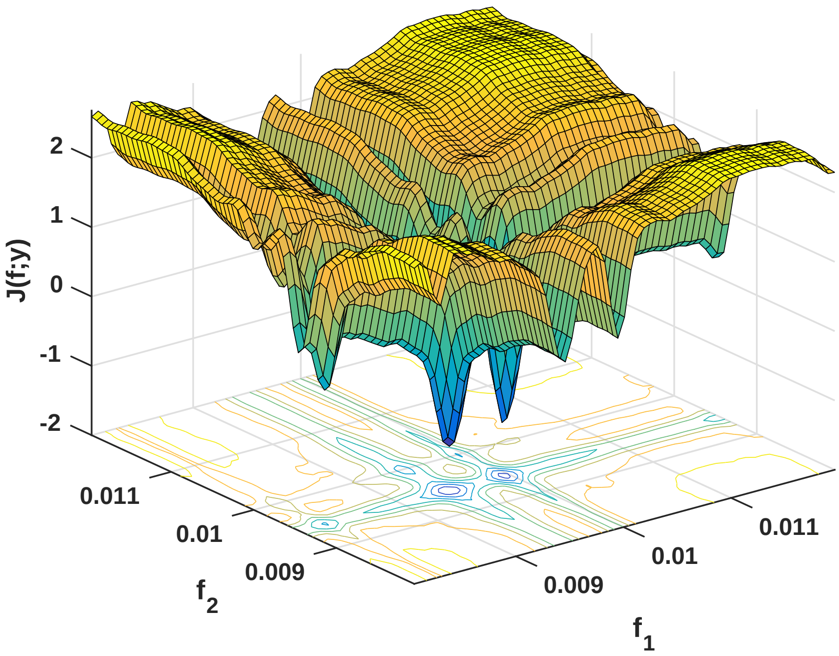

Figure 6: The approximate cost function $\tilde{J}(\mathbf{f};\mathbf{y})$ evaluated for frequencies $\mathbf{f} \in (0.008, 0.012)^2$ for a two-pitch signal with fundamental frequencies $f_1 = 0.01, f_2 = 0.0095$, having $L_1 = 6$ and $L_2 = 4$ harmonics, respectively.

the sample size is small, if the pitches are not well separated, or if the fundamental frequencies are low. The estimator in (28) will be used as a comparison method in this thesis, where it will be referred to as the approximate non-linear least squares, abbreviated ANLS, estimator.

2.2 Sub-octaves and model order knowledge

As noted above, the approximate MLE in (28) is non-robust and may produce erroneous or biased pitch frequency estimates, despite having perfect knowledge of the number of pitches, K, as well as the number of harmonics for each pitch, L_k, constituting the signal. However, it should be noted that also the exact MLE in (21) becomes non-robust if we relax the assumption of having perfect knowledge of the number of harmonics. This is illustrated in Figure 7, showing the

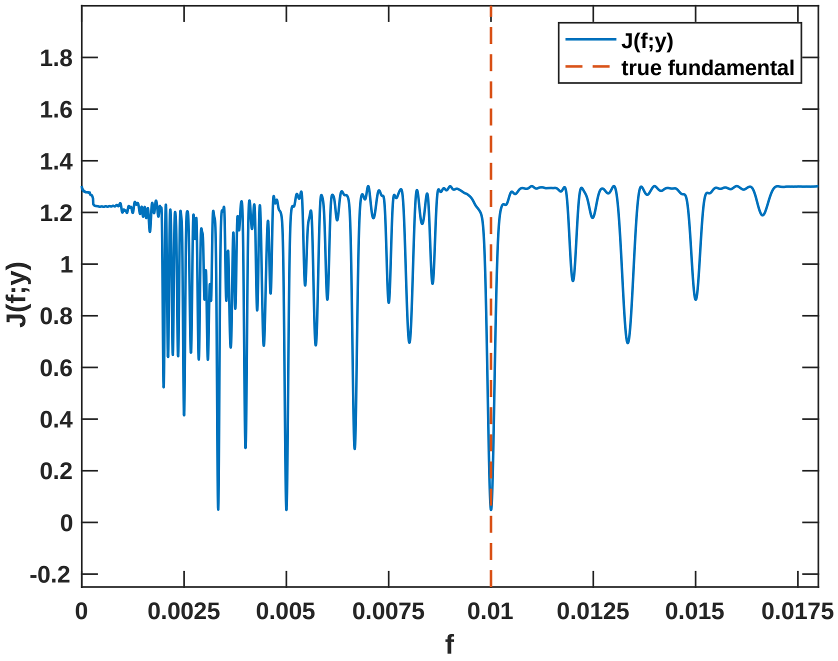

Figure 7: The cost function $J(\mathbf{f};\mathbf{y})$ evaluated for frequencies $f \in (0, 0.018)$ for a single pitch signal with fundamental frequency $f_k = 0.01$ and $L_k = 6$ harmonics. The cost function has been evaluated under the assumption $L_k = 14$.

cost function $J(\mathbf{f};\mathbf{y})$ evaluated evaluated for frequencies $f \in (0, 0.018)$ for the same single-pitch signal as before, i.e., $f_k = 0.01$ and $L_k = 6$. However, this time, we have assumed that we only have imprecise knowledge of the number of harmonics; specifically, we assume that $L_k \leq 14$, and thus evaluate the cost function using 14 harmonics. As can be seen, the cost function $J(\mathbf{f};\mathbf{y})$ no longer has a clear global minimum at $f = 0.01$, but has a similar minimum at $f = 0.005$; in fact, this is the global minimum. Thus, minimizing $J(\mathbf{f};\mathbf{y})$ would in this case yield the erroneous estimate $\hat{f}_k = 0.005 = f_k/2$. This is caused by the set of harmonics of f_k being a subset of the harmonics for all divisors of f_k, if no upper bound on the number of harmonics is set. In practice, if the assumed number of harmonics are more than twice the actual number of harmonics, the estimator in (21) will be prone to mistake these divisors for the true pitch. This is commonly referred to as the *sub-octave* problem due to the relationship between $f_k/2$ and f_k

corresponding to the music theoretical concept of an octave. The general problem of model order selection for multi-pitch signals, i.e., to estimate the number of pitches and harmonics alongside the fundamental frequencies, is, as can be imagined, considerably more difficult than for the single-pitch case, disqualifying approaches such as (21) and (28). The first two papers in this thesis address this problem by forming convex relaxations of (17), coupled with techniques from sparse modeling described briefly below.

3 Sparse modeling

As noted above, the problem in (17) is non-convex in the pitch frequencies, $\mathbf{f}$, whereas it is linear in the amplitudes, $\mathbf{a}$. To arrive at a convex relaxation of (17), let us therefore first create a set $\mathcal{F}$, containing P candidate fundamental frequencies, that are so finely spaced that the true fundamental frequencies are contained in $\mathcal{F}$ (see, e.g., [6]), i.e.,

$$f_k \in \mathcal{F}\ , k = 1, 2, \ldots, K\ . \tag{31}$$

Let $\mathbf{f}^{(K)}$ and $\mathbf{a}^{(K)}$ denote the vectors of the true fundamental frequencies and the true complex amplitudes, respectively, and let $\mathbf{f}^{(P)}$ be the $P \times 1$ vector of candidate fundamental frequencies. Then, letting $L_{\max} \geq \max_k L_k$, we have that there exists an amplitude vector $\mathbf{a}^{(P)}$ such that

$$\mathbf{Z}^{(K)}\big(\mathbf{f}^{(K)}\big)\mathbf{a}^{(K)} = \mathbf{Z}^{(P)}\big(\mathbf{f}^{(P)}\big)\mathbf{a}^{(P)}\ , \tag{32}$$

i.e., it is possible to represent the signal as a linear combination of elements of the dictionary $\mathbf{Z}^{(P)}\big(\mathbf{f}^{(P)}\big)$. Also, as $P \gg K$ in order to have a fine enough grid, we expect this representation to be sparse, i.e., only a few elements of $\mathbf{a}^{(P)}$ will be non-zero. In fact, as $\mathbf{a}^{(P)}$ is a concatenation of the amplitude vectors corresponding to each candidate pitch, we expect the sparsity pattern to have a group structure, i.e., sub-vectors of $\mathbf{a}^{(P)}$ corresponding to non-present pitches are expected to be zero. Having an estimate of $\mathbf{a}^{(P)}$ thus yields an estimate of $\mathbf{f}^{(K)}$ by considering the non-zero sub-vectors of $\mathbf{a}^{(P)}$. However, the problem

$$\underset{\mathbf{a}\in\mathcal{C}^{PL_{\max}}}{\text{minimize}}\ \left\|\mathbf{y} - \mathbf{Z}^{(P)}\big(\mathbf{f}^{(P)}\big)\mathbf{a}\right\|_2^2 \tag{33}$$

is ill-posed due to the dimension of the amplitude vector $\mathbf{a}$ being $PL_{\max} \gg N$, which will cause $\mathbf{Z}^{(P)}\big(\mathbf{f}^{(P)}\big)$ to be rank-deficient and the solution is therefore not unique. On the other hand, the fact that $\mathbf{a}^{(P)}$ is sparse allows for the use of ideas from sparse reconstruction in order to form an estimate of the fundamental frequencies, $\mathbf{f}^{(K)}$. Sparse reconstruction (see, e.g., [7]) refers to the problem of reconstructing a signal using fewer data points than would in general be possible, had the signal not had a sparse representation in some basis. In this framework, we assume that we observe a down-sampled version of the P-sample signal $\boldsymbol{\Psi}\mathbf{x}$, where $\boldsymbol{\Psi}$ is a $P \times P$ matrix of basis functions so that only $K \ll P$ elements of $\mathbf{x}$ are non-zero. The down-sampled signal is then

$$\mathbf{y} = \boldsymbol{\Phi}\boldsymbol{\Psi}\mathbf{x} \tag{34}$$

where $\boldsymbol{\Phi}$ is a $N \times P$ matrix, where $N \ll P$. Assuming that both $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are known, the reconstruction of the signal $\boldsymbol{\Psi}\mathbf{x}$ is then equivalent to estimating $\mathbf{x}$ from $\mathbf{y}$. Under some conditions on $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, and K (see, e.g., [7]) error free reconstruction without knowledge of the sparsity level K can (with high probability) be performed by solving

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & ||\mathbf{x}||_0 \\ \text{subject to} \quad & \mathbf{y} = \mathbf{A}\mathbf{x} \end{aligned} \tag{35}$$

where $\mathbf{A} = \boldsymbol{\Phi}\boldsymbol{\Psi}$ and $||\cdot||_0$ is the ℓ_0 pseudo-norm, counting the non-zero elements of its argument. Due to the ℓ_0-norm, this problem is in general combinatorial, and is therefore often approximated using the convex relaxation

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & ||\mathbf{x}||_1 \\ \text{subject to} \quad & \mathbf{y} = \mathbf{A}\mathbf{x}\,, \end{aligned} \tag{36}$$

referred to as basis pursuit [8]. In some instances, the problems in (35) and (36) are in fact equivalent, in the sense of having the same optimal point [9]. The corresponding problem for the noise-corrupted signal

$$\mathbf{y} = \boldsymbol{\Phi}\boldsymbol{\Psi}\mathbf{x} + \mathbf{e} \tag{37}$$

is

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & ||\mathbf{x}||_1 \\ \text{subject to} \quad & ||\mathbf{y} - \mathbf{A}\mathbf{x}||_2 \leq \delta \end{aligned} \tag{38}$$

where $\delta \geq \|e\|_2$. Also, due to Lagrangian duality, there exists a $\lambda > 0$ such that this problem has the equivalent representation

$$\underset{\mathbf{x}}{\text{minimize}}\ \frac{1}{2}\left\|\mathbf{y}-\mathbf{A}\mathbf{x}\right\|_2^2+\lambda\left\|\mathbf{x}\right\|_1\ . \tag{39}$$

The problem (39) is often referred to as the LASSO [10] and has been used extensively over the years for the estimation of sparse signals from noisy measurements. Notably, (39) was used for spectral estimation in [11] wherein $\mathbf{A}$ constituted an oversampled discrete Fourier transform matrix, yielding higher resolution than the one provided by the FFT. This thesis builds on these foundations, and considers the general formulation

$$\underset{\mathbf{a}\in\mathcal{C}^{PL_{\max}}}{\text{minimize}}\ g\left(\mathbf{y}-\mathbf{Z}^{(P)}\left(\mathbf{f}^{(P)}\right)\mathbf{a}\right)+h\left(\mathbf{a}\right) \tag{40}$$

where $g(\cdot)$ is a measure of fit between the signal and the model, and $h(\cdot)$ is a penalty function inducing structure on the amplitude vector $\mathbf{a}$ and provides regularization as to make the problem well-posed. In particular, $h(\cdot)$ will be chosen as to impose a sparse pitch structure on $\mathbf{a}$, as well as counter the sub-octave problem described earlier. Also, in order to allow for the implementation of efficient solvers, $g(\cdot)$ and $h(\cdot)$ will be chosen so that the objective function is convex.

3.1 Proximal operators

In this thesis, the implementation of efficient solvers of (40) will lead to having the need to solve sub-problems of the form

$$\underset{\mathbf{x}}{\text{minimize}}\ \frac{1}{2}\left\|\mathbf{y}-\mathbf{x}\right\|_2^2+\lambda h(\mathbf{x}) \tag{41}$$

where $\lambda \geq 0$ and $h(\cdot)$ is a convex, non-smooth function. The solution to this problem is often referred to as the *proximal operator* of $\lambda h(\cdot)$ [12], denoted as

$$\text{prox}_{\lambda h}(\mathbf{y})=\arg\underset{\mathbf{x}}{\min}\ \frac{1}{2}\left\|\mathbf{y}-\mathbf{x}\right\|_2^2+\lambda h(\mathbf{x})\ . \tag{42}$$

Often, it is straightforward to interpret the proximal operator as a Euclidian projection onto a convex set defined by $h(\cdot)$, i.e., the proximal operator solves

$$\begin{aligned}&\underset{\mathbf{x}}{\text{minimize}} && \frac{1}{2}\left\|\mathbf{y}-\mathbf{x}\right\|_2^2\\ &\text{subject to} && h(\mathbf{x})\leq\gamma\end{aligned} \tag{43}$$

for some $\gamma > 0$. As an example, consider the problem

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} && \frac{1}{2}\left\|\mathbf{y}-\mathbf{x}\right\|_2^2 \\ &\text{subject to} && \left\|\mathbf{x}\right\|_1 \leq \gamma \end{aligned} \tag{44}$$

where both $\mathbf{x}$ and $\mathbf{y}$ are real vectors of length N, i.e., we want to find the Euclidian projection of $\mathbf{y}$ onto an ℓ_1 unit ball, scaled by γ. We here assume that $\left\|\mathbf{y}\right\|_1 > \gamma$ in order to avoid the trivial solution of projecting $\mathbf{y}$ onto itself. Note that the solution $\mathbf{x}^\star$ will be sparse due to the geometry of the ℓ_1 unit ball; it has protruding corners along the coordinate axes, meaning that the level sets of the ℓ_2-norm will, with high probability, tangent the feasible set at points where one or more of the variables are zero. To verify this, consider the Lagrangian of (44):

$$\begin{aligned} L(\mathbf{x},\lambda) &= \frac{1}{2}\left\|\mathbf{y}-\mathbf{x}\right\|_2^2 + \lambda\left(\left\|\mathbf{x}\right\|_1 - \gamma\right) && (45) \\ &= \sum_{n=0}^{N}\left(\frac{1}{2}\left(y_n - x_n\right)^2 + \lambda\left|x_n\right|\right) - \lambda\gamma && (46) \end{aligned}$$

where $\lambda \geq 0$ is the Lagrange multiplier. Computing the sub-differential for $L(\mathbf{x},\lambda)$ with respect to x_n yields

$$\frac{\partial L(\mathbf{x},\lambda)}{\partial x_n} = -\left(y_n - x_n\right) + \lambda s_n \tag{47}$$

where

$$s_n \in \begin{cases} \operatorname{sign}(x_n) & \text{if } x_n \neq 0 \\ [-1,1] & \text{if } x_n = 0\,. \end{cases} \tag{48}$$

Given this, the Karush-Kuhn-Tucker (KKT) conditions (see, e.g., [13]), which, as the problem is convex, are necessary and sufficient for a point $\mathbf{x}^\star$ to be optimal, are

$$\begin{aligned} -\left(y_n - x_n\right) + \lambda s_n = 0\ ,\ \ n = 1,2,\dots,N && (49) \\ \lambda\left(\left\|\mathbf{x}\right\|_1 - \gamma\right) = 0 && (50) \\ \left\|\mathbf{x}\right\|_1 - \gamma \leq 0 && (51) \\ \lambda \geq 0 && (52) \end{aligned}$$

Noting that the optimal point must lie on the boundary of the feasible set, i.e., $\|\mathbf{x}^\star\|_1 = \gamma$, we arrive at the solution

$$x_n^\star = \begin{cases} \operatorname{sign}\left(y_n\right)\left(|y_n| - \lambda^\star\right) & \text{if } |y_n| > \lambda^\star \\ 0 & \text{if } |y_n| \leq \lambda^\star \end{cases} \tag{53}$$

where $\lambda^\star$ solves the fixed-point equation

$$\lambda = \frac{\sum_{n=0}^{N} |y_n| \, \mathbf{1}_{\{|y_n|>\lambda\}} - \gamma}{\sum_{n=0}^{N} \mathbf{1}_{\{|y_n|>\lambda\}}} \tag{54}$$

where

$$\mathbf{1}_{\{|y_n|>\lambda\}} = \begin{cases} 1 & \text{if } |y_n| > \lambda \\ 0 & \text{if } |y_n| \leq \lambda \end{cases} . \tag{55}$$

As $\lambda^\star$ is the solution to a fixed-point equation, it can be found by fixed-point iteration, initialized at $\lambda = 0$, or by a one-dimensional search. Thus, we see that the projection of $\mathbf{y}$ onto the feasible set amounts to shrinking each element by $\lambda^\star$, and setting an element to zero if its magnitude is smaller than $\lambda^\star$, i.e., $\mathbf{x}^\star$ will be sparser than $\mathbf{y}$. Also, we see that

$$\mathbf{x}^\star = \operatorname{prox}_{\lambda^\star h}(\mathbf{y}) = \arg\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 + \lambda^\star \|\mathbf{x}\|_1 \ , \tag{56}$$

which is the LASSO given in (39), using $\mathbf{A} = \mathbf{I}$, with parameter $\lambda^\star$. When used in this thesis, $h(\cdot)$ will be chosen as to define feasible sets that induce sparsity in the solution in ways analogous to (44).

4 Inharmonicity and optimal transport

In what has been presented so far, we have assumed perfectly harmonic pitches, i.e., that for each pitch k, we have

$$f_{k,\ell} = f_k \ell \ , \ \ell = 1, 2, \ldots, L_k \, . \tag{57}$$

However, in some scenarios, signals may contain sinusoidal components that have an approximate, but not perfectly harmonic structure. This phenomenon is called inharmonicity and is found in, e.g., the sound produced by vibrating stiff strings, such as piano notes. Specifically, the frequencies of the sinusoidal components of an inharmonic pitch are [14]

$$f_{k,\ell} = f_k \ell + \Delta_{k,\ell} \ , \ \ell = 1, 2, \ldots, L_k \tag{58}$$

where $\Delta_{k,\ell}$ is the inharmonicity parameter of the ℓth harmonic. In some cases, there might exist a parametric description of $\Delta_{k,\ell}$, based on the physical properties of the signal source. For example, an often used model for the frequencies produced by a piano is

$$f_{k,\ell} = f_k \ell \sqrt{1 + \beta \ell^2} \tag{59}$$

where $\beta > 0$ is a parameter detailing the string's stiffness. Realistic values of β are on the order of 10^{-3}.

As an illustration, Figure 8 shows the estimated magnitude spectrum of a 30 ms excerpt of a single-pitch signal consisting of the piano note B4, having a fundamental frequency of 493.883 Hz. Also plotted are the frequency locations of the harmonics assuming the perfectly harmonic model in (57) as well as the piano model in (59), respectively. For the piano model, an approximate value of $\beta = 10^{-3}$ was chosen using hand-tuning. As can be seen the piano model better describes the signal, as it accounts for the higher-order harmonics deviating from the perfectly harmonic model.

Importantly, if not taken into account, inharmonicity might have a detrimental effect on the quality of the pitch estimates. To illustrate this, we revisit the earlier example with a single-pitch signal with fundamental frequency $f_k = 0.01$ and $L_k = 6$ harmonics. Using the same complex amplitudes for the harmonics as before, we now alter the frequencies of the harmonics to be detailed by (59), using $\beta = 10^{-3}$, rather than via (57). Figure 9 shows the effect of the inharmonicity when estimating f_k by the MLE (21), which is constructed under the assumption of a perfectly harmonic model. As can be seen, the global minimum of the cost function $J(\mathbf{f}; \mathbf{y})$ has now been shifted upwards in frequency due to the shifted locations of the harmonics, causing an upward bias in the estimate of f_k. There have been estimators proposed to handle this type of deviation, but they are often based on either parametric descriptions of the inharmonicity, similar to (59), narrowing their scope of applicability to certain classes of signals, or only consider the single-pitch case. Although it might be argued that this type of

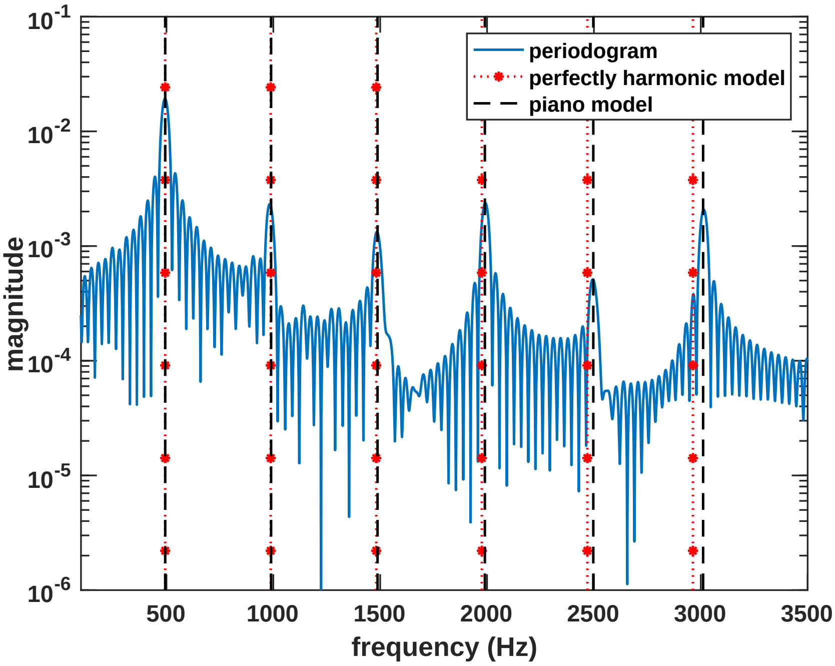

Figure 8: Magnitude spectrum of a signal consisting of a piano note. The frequency locations of the harmonics, assuming the perfectly harmonic model (57), as well as the piano model (59), using $\beta = 10^{-3}$, are indicated by the vertical lines.

bias is only a real concern in cases when highly accurate estimates of the fundamental frequencies are needed, it should be noted that inharmonicity might lead to overestimating the model order, i.e., that an estimator will find more pitches than the ones actually present in the signal. As noted in (23), Fourier vectors corresponding to different frequencies become orthogonal as their lengths increase; however, even for moderate sample sizes, the Fourier vector corresponding to an inharmonic overtone might be almost orthogonal to its perfectly harmonic counterpart. This can be see in Figure 8, where the two highest-order harmonics, as given by a perfectly harmonic model, are located outside the main lobes of the periodogram of the signal. As can be seen in the figure, these two harmonics are located close to the nulls of the periodogram, i.e., they are almost orthogonal to the sinusoidal components of the signal. Thus, inharmonicity, if not taken into

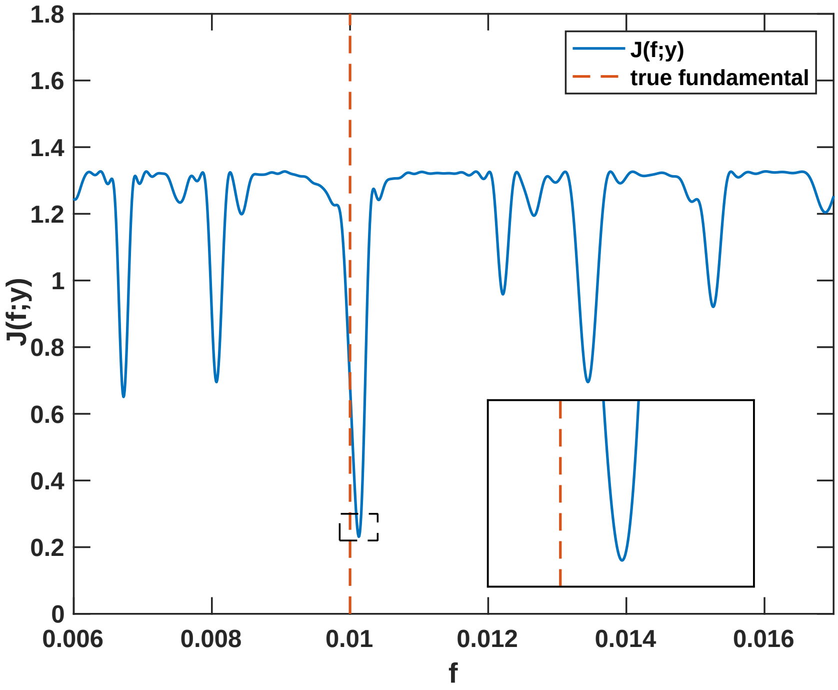

Figure 9: The cost function $J(\mathbf{f};\mathbf{y})$ evaluated for frequencies $f \in (0, 0.018)$ for an inharmonic single-pitch signal with fundamental frequency $f_k = 0.01$ and $L_k = 6$ harmonics. The frequency of each harmonic is detailed by (59), using $\beta = 10^{-3}$.

account, might lead to cases where inharmonic overtones are classified as stray sinusoids not belonging to other found sources. In the third paper of this thesis, we propose a multi-pitch estimator that is designed to be robust to possibly occurring inharmonicity, without requiring specific knowledge of the structure of the inharmonicity. The estimator is based on the concept of optimal transport, briefly described below.

4.1 Optimal transport

Optimal transport is an old subject, first introduced by the French mathematician Gaspard Monge in the 18th century, with the application being how to optimally transport building material from quarries to construction sites; that is, given a known set of locations at which material is deposited, as well as a set of known locations to be supplied, the goal is to find the transportation scheme that minimizes some measure of cost. The modern formulation of this problem was devised by the Soviet mathematician Leonid Kantorovich, and is called the Monge-Kantorovich minimization problem. Formally, let $\mathcal{X}$ and $\mathcal{Y}$ be two spaces, with probability measures μ and ν, respectively; that is, μ describes the distribution of mass on $\mathcal{X}$, and ν describes the distribution of mass on $\mathcal{Y}$. Then, given a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ describing the cost of moving one unit of mass from a point in $\mathcal{X}$ to a point in $\mathcal{Y}$, the optimal transport problem is (see, e.g., [15])

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & \int_{\mathcal{X}\times\mathcal{Y}} c\left(x,y\right) \mathrm{d}w\left(x,y\right) \\ \text{subject to} \quad & \int_{\mathcal{X}} \mathrm{d}w(x,\mathcal{B}) = \nu(\mathcal{B}) \ , \ \forall \mathcal{B} \subset \mathcal{Y} \\ & \int_{\mathcal{Y}} \mathrm{d}w(\mathcal{A},y) = \mu(\mathcal{A}) \ , \ \forall \mathcal{A} \subset \mathcal{X} \ , \end{aligned} \tag{60}$$

where $\mathcal{A}$ and $\mathcal{B}$ are measurable subsets of $\mathcal{X}$ and $\mathcal{Y}$, respectively. Here, the joint probability measure w describes the association between sets in $\mathcal{X}$ and sets in $\mathcal{Y}$, i.e., transportation of mass between the spaces. The constraints of the optimization problem, requiring that μ and ν are marginals of w, ensures the conservation of mass. In a discrete, finite setting, i.e.,

$$\mathcal{X} = \{x_m \mid m = 1, 2, \ldots, M\} \tag{61}$$

$$\mathcal{Y} = \left\{y_p \mid p = 1, 2, \ldots, P\right\} \tag{62}$$

the corresponding problem is

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & \sum_{m=1}^{M} \sum_{p=1}^{P} c\left(x_m, y_n\right) w\left(x_m, y_p\right) \\ \text{subject to} \quad & \sum_{m=1}^{M} w\left(x_m, y_p\right) = \nu\left(y_p\right) \ , \ \forall y_p \in \mathcal{Y} \\ & \sum_{p=1}^{P} w\left(x_m, y_p\right) = \mu\left(x_m\right) \ , \ \forall x_m \in \mathcal{X} \ . \end{aligned} \tag{63}$$

From this, it can be seen that the optimal transportation problem is a convex, linear program, as the objective function and the constraints of (63) are linear in w.

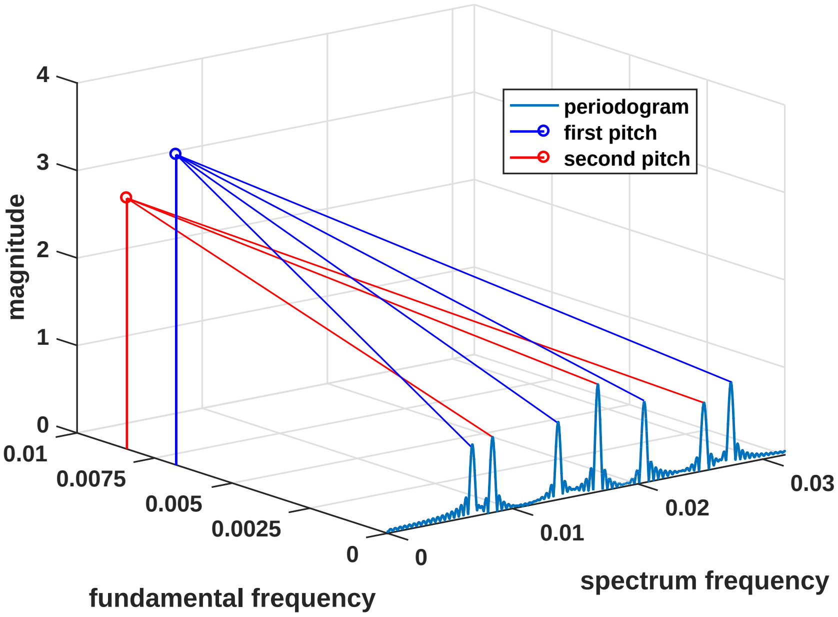

Figure 10: The mapping of seven measured spectral peaks to two fundamental frequencies by solving an optimal transport problem.

The minimum value of the objective function can be interpreted as a measure of the distance between two probability measures μ and ν. This idea has earlier been used to measure the distance between different power spectra [16], as the power spectral density is non-negative, and has been used for, e.g., the construction of smooth sequences of spectra [17], [18]. This thesis will expand on the idea of the objective function as a similarity measure, in order to use the optimal transportation problem for the estimation of inharmonic pitch signals. In our setting, $\mathcal{X}$ will be the set of observed spectrum frequencies in a signal, and $\mathcal{Y}$ will be a large set of candidate fundamental frequencies, i.e.,

$$\mathcal{X} = \{f_m \mid m = 1, 2, \ldots, M\} \tag{64}$$

$$\mathcal{Y} = \left\{f_p \mid p = 1, 2, \ldots, P\right\} . \tag{65}$$

Also, μ and ν will no longer be proper probability measures; $\mu(f_m)$ will be the spectrum magnitude at frequency f_m and $\nu(f_p)$ will be the sum of all spectrum magnitudes associated with the fundamental frequency f_p. As the fundamental frequencies are not known, ν will also be part of the optimization problem in order to find an optimal association of spectral energy with fundamental frequencies. The cost function $c(\cdot,\cdot)$ will be designed in such a way that this association allows for inharmonicities. Also, the optimization problem (63) will be refined in order to encourage sparse solutions. An illustration of the idea is shown in Figure 10, displaying a measured spectrum containing seven peaks, simulated as a mixture of two pitches, containing $L_1 = 4$ and $L_2 = 3$ harmonics, respectively. The fundamental frequencies are $f_1 = 6.8 \cdot 10^{-3}$ and $f_2 = 8.4 \cdot 10^{-3}$, with the frequencies of the harmonics being detailed by (59) with $\beta = 10^{-3}$. Relating to the problem in (63), the distribution of power, in magnitude, in the observed spectrum is then described by μ. By solving a more elaborate version of (63), as will be described in the third paper of this thesis, the distribution of power among fundamental frequencies, i.e., an estimate of ν, is found. In this example, the estimated ν is only non-zero for two fundamental frequencies, as can be seen in Figure 10.

5 Outline of the papers

Paper A: An Adaptive Penalty Multi-Pitch Estimator with Self-Regularization

In the first paper, multi-pitch estimation for stationary signal frames is considered. Estimates are produced by minimizing a cost function of type (40), where the squared ℓ_2-norm of the difference between the observed signal and the model serves as the measure of fit. The penalty term is designed as to both encourage few harmonics in the solution, as well as smooth spectral envelopes for each pitch. By enforcing spectral smoothness, the proposed estimator specifically targets the sub-octave problem, described earlier, and is shown to yield estimates with sparse pitch representations. The objective function is minimized by solving a series of convex optimization problems, with an alternating direction method of multipliers (ADMM) implementation provided in the paper. Also, as the objective function contains two tuning parameters, determining the trade-off between the fit between the signal and the model and the sparseness of the reconstruction, a scheme for automatic, signal adaptive selection of these parameters is presented. The proposed estimator is extensively evaluated on both simulated and real audio signals, and is shown to outperform comparison methods in high signal-to-noise ratio (SNR) settings. The work in paper A has been published in part as

> Filip Elvander, Stefan Ingi Adalbjörnsson, Ted Kronvall, and
> Andreas Jakobsson, "An Adaptive Penalty Multi-Pitch Estimator with Self-Regularization". *Elsevier Signal Processing*, vol. 127, pp. 56-70, October 2016.

> Ted Kronvall, Filip Elvander, Stefan Ingi Adalbjörnsson, and
> Andreas Jakobsson, "An Adaptive Penalty Approach to Multi-Pitch Estimation", *23rd European Signal Processing Conference*, Nice, France, August 31 - September 4, 2015.

Paper B: Online Estimation of Multiple Harmonic Signals

In the second paper, the assumption of stationary frames made in paper A is relaxed, and an estimator that allows for non-stationarities such as amplitude and frequency modulation is proposed. The estimator is recursive in the signal samples, allowing for smooth signal tracking as it exploits the time correlation between samples. Inspired by sparse recursive least squares, the estimator is the minimizer of a time-weighted least squares criterion augmented by sparsifying penalty terms, where the penalty parameters are signal adaptive as to allow for dynamic changes throughout the signal duration. The objective function is minimized using a proximal gradient algorithm. Also, as the penalty terms induce a magnitude bias towards zero for the harmonics, a de-biasing step is proposed. The paper also presents a scheme for adaptively adjusting the frequencies of the pitch candidates in order to allow for tracking of frequency modulated signals. The estimator is shown to perform comparably to the estimator in paper A, although having a lower computational cost due, in part, to its recursive formulation. Also, experiments on real audio signals suggest that the proposed method is more general in its area of applicability than machine learning-type algorithms specialized on automatic music transcription. The work in paper B has been published in part as

> Filip Elvander, Johan Swärd, and Andreas Jakobsson, "Online Estimation of Multiple Harmonic Signals". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 273-284, February 2017.

> Filip Elvander, Johan Swärd, and Andreas Jakobsson, "Time-Recursive Multi-Pitch Estimation Using Group Sparse Recursive Least Squares", *50th Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, USA, November 6-9, 2016.

Paper C: Using Optimal Transport for Estimating Inharmonic Pitch Signals

In the third paper, an estimator designed to be robust to possibly occurring inharmonicities is proposed. In that respect, it differs from the estimators proposed in papers A and B, as these were derived under the assumption of perfectly harmonic signal components. Building on the concept of optimal transport, the proposed method constructs estimates of the fundamental frequencies by mapping found spectral components to a set of candidate fundamental frequencies. The method achieves robustness to inharmonicity by solving an extended version of the problem in (63) where the cost function $c(\cdot,\cdot)$ is designed as to allow the harmonic frequencies to deviate from perfect integer multiples of the fundamental frequency. Although being an extension of the problem in (63), the resulting optimization problem is still a linear program. When evaluated on simulated inharmonic signals, the proposed method is shown to be robust to the bias in the fundamental frequency affecting standard methods. Also, when applied to a real audio mixture of both harmonic and inharmonic sources, the proposed method is shown to outperform the comparison methods. The work in paper C has been published in part as

> Filip Elvander, Stefan Ingi Adalbjörnsson, Johan Karlsson, and Andreas Jakobsson, "Using Optimal Transport for Estimating Inharmonic Pitch Signals". *42nd IEEE International Conference on Acoustics, Speech, and Signal Processing*, New Orleans, USA, March 5-9, 2017.

References

[1] S-R. Huang, R. M. Lerner, and K. J. Parker, "On estimating the amplitude of harmonic vibration from the Doppler spectrum of reflected signals," *J. Acoust. Soc. Am.*, vol. 88, no. 6, pp. 2702–2712, 1990.

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

[3] 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.

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

[5] J. Cooley and J. Tukey, "An Algorithm for the Machine Calculation of Complex Fourier Series," *Math. Computation*, vol. 19, pp. 297–301, 1965.

[6] P. Stoica and P. Babu, "Sparse Estimation of Spectral Lines: Grid Selection Problems and Their Solutions," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 962–967, Feb. 2012.

[7] M. Elad, *Sparse and Redundant Representations*, Springer, 2010.

[8] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic Decomposition by Basis Pursuit," *SIAM Review*, vol. 43, pp. 129–159, 2001.

[9] E. J. Candes and T. Tao, "Decoding by Linear Programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[10] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society B*, vol. 58, no. 1, pp. 267–288, 1996.

[11] J. J. Fuchs, "On the Use of Sparse Representations in the Identification of Line Spectra," in *17th World Congress IFAC*, Seoul, Jul. 2008, pp. 10225–10229.

[12] N. Parikh and S. Boyd, “Proximal Algorithms,” *Found. Trends Optim.*, vol. 1, pp. 127–239, 2014.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[14] H. Fletcher, “Normal vibration frequencies of stiff piano string,” *Journal of the Acoustical Society of America*, vol. 36, no. 1, 1962.

[15] C. Villani, *Topics in Optimal Transportation*, vol. 58, Graduate studies in Mathematics, AMS, 2003.

[16] T. T. Georgiou, J. Karlsson, and M. S. Takyar, “Metrics for Power Spectra: An Axiomatic Approach,” *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 859–867, Mar. 2009.

[17] X. Jiang, Z-Q. Luo, and T. T. Georgiou, “Geometric Methods for Spectral Analysis,” *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1064–1074, Mar. 2012.

[18] L. Ning, T. T. Georgiou, and A. Tannenbaum, “On Matrix-Valued Monge-Kantorovich Optimal Mass Transport,” *IEEE Trans. Autom. Control*, vol. 60, no. 2, pp. 373–382, Feb. 2015.

A

Paper A

An Adaptive Penalty Multi-Pitch Estimator with Self-Regularization

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

Centre for Mathematical Sciences, Lund University, Lund, Sweden

Abstract

This work treats multi-pitch estimation, and in particular the common misclassification issue wherein the pitch at half the true fundamental frequency, the sub-octave, is chosen instead of the true pitch. Extending on current group LASSO-based methods for pitch estimation, this work introduces an adaptive total variation penalty, which enforces both group- and block sparsity, as well as deals with errors due to sub-octaves. Also presented is a scheme for signal adaptive dictionary construction and automatic selection of the regularization parameters. Used together with this scheme, the proposed method is shown to yield accurate pitch estimates when evaluated on synthetic speech data. The method is shown to perform as good as, or better than, current state-of-the-art sparse methods while requiring fewer tuning parameters than these, as well as several conventional pitch estimation methods, even when these are given oracle model orders. When evaluated on a set of ten musical pieces, the method shows promising results for separating multi-pitch signals.

Key words: Multi-pitch estimation, block sparsity, adaptive sparse penalty, self-regularization, ADMM

1 Introduction

Pitch estimation is a problem arising in a variety of fields, not least in audio processing. It is a fundamental building block in several music information retrieval applications, such as automatic music transcription, i.e., automatic sheet music generation from audio (see, e.g., [1], [2]). Pitch estimation could also be used as a component in methods for cover song detection and music querying, possibly improving currently available services. For example, the popular query service Shazam [3] operates by matching hashed portions of spectrograms of user-provided samples against a large music database. As a change of instrumentation would alter the spectrogram of a song, such algorithms can only identify recordings of a song that are very similar to the actual recording present in the database. Thus, services such as Shazam might fail to identify, e.g., acoustic alternate versions of rock songs. A query algorithm based on pitch estimation could on the other hand correctly match the acoustic version to the original electrified one as it would recognize, e.g., the main melody.

The applicability of pitch estimation to music is due to the fact that the notes produced by many instruments used in Western tonal music, e.g., woodwind instruments such as the clarinet, exhibit a structure that is well modeled using a harmonic sinusoidal structure [4]. However, for some plucked stringed instruments, such as the guitar and the piano, the tension of the string results in the harmonics deviating from perfect integer multiples of the fundamental frequency, a phenomenon called inharmonicity. For some instruments, such as the piano, there are models describing the structure of the inharmonicity based on physical properties of the instrument [5]. Such signals require agile pitch estimation algorithms allowing for this form of deviations (see, e.g., [6–8]). In this work, we will assume such deviations to be small, although noting that one may extend the here presented work along the lines in [6–8].

Estimating the fundamental frequencies of multi-pitch signals is generally a difficult problem. There are many methods available, see, e.g., [9], but most of them require *a priori* model order knowledge, i.e., they require knowledge of the number of pitches present in the signal, as well as the number of active harmonics for each pitch.[1] Three such methods will be used in this work as reference estimators. The first method, here referred to as ORTH, exploits orthogonality

[1] It may be noted that, generally, obtaining correct model order information is a most challenging problem, with the model order estimates strongly affecting the resulting performance of the estimator.

between the signal and noise subspaces to form pitch frequency estimates. The second method is an optimal filtering method based on the Capon estimator, and is therefore here referred to as Capon. The third method is an approximate non-linear least squares method, here referred to as ANLS [10–12] (see also [9] for an overview of these methods). Methods not requiring *a priori* model order knowledge have also been proposed. For example, Adalbjörnsson et al. [13] use a sparse dictionary representation of the signal and regularization penalties to implicitly choose the model order. A similar, but less general, method was introduced in [14], which used a dictionary specifically tailored to piano notes for estimating pitch frequencies generated by pianos. Other source specific methods include [15], [16]. In [17], the author proposes a sparsity-exploiting method, where the dictionary atoms are learned from databases of short-time Fourier transforms of musical notes. A similar idea is used in [18] for pitch-tracking in music. In [16], [19], pitch estimation is based on the assumption of spectral smoothness, i.e., the amplitudes of the harmonics within a pitch are assumed to be of comparable magnitude.

Another field of research is performing multi-pitch estimation, often in the context of automatic music transcription, by decomposing the spectrogram of the signal into two matrices, one that describes the frequency content of the signal and one that describes the time activation of the frequency components. This method makes use of the non-negative matrix factorization, first introduced in this context in [20] and since then widely used, such as in, e.g., [21]. There are also more statistical approaches to multi-pitch estimation, posing the estimation as a Bayesian inference problem (see, e.g., [22]).

The approach to multi-pitch estimation presented in this work is to solve the problem in a group sparse modeling framework, which allows us to avoid making explicit assumptions on the number of pitches, or on the number of harmonics in each pitch. Instead, the number of components in the signal is chosen implicitly, by the setting of some tuning parameters. These tuning parameters determine how appropriate a given pitch candidate is to be present in the signal and may be set using cross-validation, or by using some simple heuristics. The sparse modeling approach has earlier been used for audio (see, e.g., [23]), and specifically for sinusoidal components in [24]. We extend on these works by exploiting the harmonic structure of the signals in a block sparse framework, where each block represents a candidate pitch. A similar method was introduced in [13], where block sparsity was enforced using block-norms, penalizing the number of active pitches.

As the block-norm penalty, under some circumstances, cannot distinguish a true pitch from its sub-octave, i.e., the pitch with half the true fundamental frequency, the method is also complemented by a total variation penalty, which is shown to solve such issues. Total variation penalties are often applied in image analysis to obtain block-wise smooth image reconstructions (see, e.g., [25]). For audio data, one can similarly assume that signals often are block-wise smooth, as the harmonics of a pitch are expected to be of comparable magnitude [19]. Enforcing this feature will specifically deal with octave errors, i.e., the choosing of the sub-octave instead of the true pitch, as, in the noise free case, only every other harmonic of the sub-octave will have non-zero power. In this paper, we show that a total variation penalty, in itself, is enough to enforce a block sparse solution, if utilized efficiently. More specifically, by making the penalty function adaptive, we may improve upon the convex approximation used in [13], allowing us to drop the block-norm penalty altogether, and so reduce the number of tuning parameters. In some estimation scenarios, e.g., when estimating chroma using the approach in [26], this would simplify the tuning procedure significantly.

Furthermore, we show that the proposed method performs comparably to that of [13], albeit with the notable improvement of requiring fewer tuning parameters. The method operates by solving a series of convex optimization problems, and to solve these we present an efficient algorithm based on the alternating direction method of multipliers (ADMM) (see, e.g., [27] for an overview of ADMM in the context of convex optimization). As the proposed method requires two tuning parameters to operate, we also present a scheme for automatic selection of appropriate model orders, thereby avoiding the need of user-supplied parameters.

The remainder of this work is organized as follows; in the following section, we introduce the signal model, followed in Section 3 by the proposed estimation algorithm. Section 4 summarizes the efficient ADMM implementation whereas Section 5 examines how to adaptively choose the regularization parameters. Numerical results illustrating the achieved performance are presented in Section 6. Finally, Section 7 concludes upon the work.

2 Signal model

Consider a complex-valued[2] 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

$$x(t) = \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} a_{k,\ell} \mathrm{e}^{i\omega_k \ell t} \tag{1}$$

for $t = 1, \ldots, N$, where $\omega_k \ell$ is the frequency of the ℓth harmonic in the kth pitch, and with the complex number $a_{k,\ell}$ denoting its magnitude and phase. The occurrence of such harmonic signals is often in combination with non-sinusoidal components, such as, for instance, colored broadband noise or non-stationary impulses. In this work, only the narrowband components of the signal are part of the signal model, such that all other signal structures, including the signal's timbre and the background noise, are treated as part of an additive noise process, $e(t)$.

In general, selecting model orders in (1) may be a daunting task, with both the number of sources, K, and the number of harmonics in each of these sources, L_k, being unknown, as well as often being structured such that different sources may have spectrally overlapping overtones. In order to remedy this, this work proposes a relaxation of the model onto a predefined grid of $P \gg K$ candidate fundamentals, each having $L_{\max} \geq \max_k L_k$ harmonics. Here, $L_{\max}$ should be selected to ensure that the corresponding highest frequency harmonic is limited by the Nyquist frequency, and could thus vary depending on the considered candidate frequency (see also [13]). For notational simplicity, we will hereafter, without loss of generality, use the same $L_{\max}$ for all candidate frequencies. Assume that the candidate fundamentals are chosen so numerous and so closely spaced that the approximation

$$x(t) \approx \sum_{p=1}^{P} \sum_{\ell=1}^{L_{\max}} a_{p,\ell} \mathrm{e}^{i\omega_p \ell t} \tag{2}$$

holds reasonably well. As only K pitches are present in the actual signal, we want to derive an estimator of the amplitudes $a_{p,\ell}$ such that only few, ideally $\sum_{k=1}^{K} L_k$,

[2] For notational simplicity and computational efficiency, we here use the discrete-time analytical signal formed from the measured (real-valued) signal (see, e.g., [9], [28]).

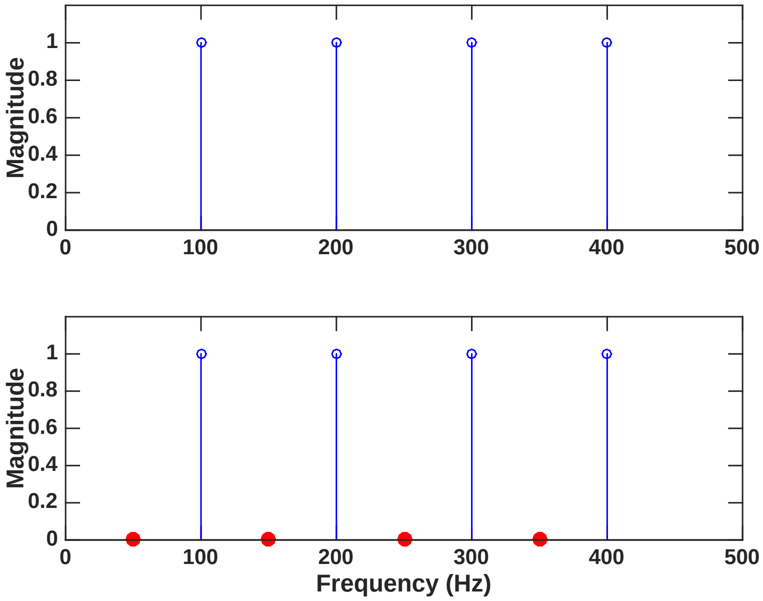

Figure 1: The upper picture depicts a pitch with fundamental frequency 100 Hz and four harmonics. The lower picture depicts a pitch with fundamental frequency 50 Hz and eight harmonics where all odd-numbered harmonics are zero (marked red dots).

of the amplitudes in (2) are non-zero. This approach may be seen as a sparse linear regression problem reminiscent of the one in [24] and has been thoroughly examined in the context of pitch estimation in, e.g., [13, 29, 30]. For notational convenience, define the set of all amplitude parameters to be estimated as

$$\boldsymbol{\Psi} = \{\boldsymbol{\Psi}_{\omega_1}, \ldots, \boldsymbol{\Psi}_{\omega_P}\} \tag{3}$$

$$\boldsymbol{\Psi}_{\omega_p} = \{a_{p,1}, \ldots, a_{p,L_{\max}}\} \tag{4}$$

where, as described above, most of the $a_{p,\ell}$ in $\boldsymbol{\Psi}$ will be zero. Note that $\boldsymbol{\Psi}$ will be sparse, i.e., having few non-zero elements. Also, the pattern of this sparsity will be group wise, meaning that if a pitch with fundamental frequency ω_p is not present, then neither will any of its harmonics, i.e., $\boldsymbol{\Psi}_{\omega_p} = \mathbf{0}$.

Due to the harmonic structure of the signal, candidate pitches having fundamental frequencies at fractions of the present pitches' fundamentals will have a partial fit of their harmonics. This may cause misclassification, i.e., erroneously identifying a present pitch as one or more non-present candidate pitches. This is the cause of the so-called sub-octave problem, which is mistaking the true pitch with fundamental frequency ω_p for the candidate pitch with fundamental frequency $\omega_p/2$. This may occur if the candidate set $\boldsymbol{\Psi}$ is structured such that the sub-octave pitch may perfectly model the true pitch, which is when $L_{\max} \geq 2L_p$. This is illustrated in Figure 1, displaying an extreme case with a pitch with fundamental frequency 100 Hz and four harmonics as well as its sub-octave, i.e., a pitch with fundamental frequency 50 Hz and eight harmonics where only the even-numbered harmonics are non-zero. Relating to music signals, this is the same as mistaking a pitch for the pitch an octave below it. Thus, when estimating the elements of $\boldsymbol{\Psi}$, one also has to take into account the structure of the block sparsity, in order to avoid erroneously selecting sub-octaves.

3 Proposed estimation algorithm

Consider N samples of a noise-corrupted measurement of the signal in (1), $y(t)$, such that it may be well modeled as $y(t) = x(t) + e(t)$, where $e(t)$ is a broadband noise signal. A straightforward approach to estimate $\boldsymbol{\Psi}$ would then be to minimize the residual cost function

$$g_1(\boldsymbol{\Psi}) = \frac{1}{2} \sum_{t=1}^{N} \left| y(t) - \sum_{p=1}^{P} \sum_{\ell=1}^{L_{\max}} a_{p,\ell} e^{i\omega_p \ell t} \right|^2 \tag{5}$$

However, setting

$$\hat{\boldsymbol{\Psi}} = \arg\min_{\boldsymbol{\Psi}} g_1(\boldsymbol{\Psi}) \tag{6}$$

will not yield the desired sparsity structure of $\boldsymbol{\Psi}$ and will be prone to also model the noise, $e(t)$. Also, solutions (6) will not be unique due to the over-completeness of the approximation (2). A remedy for this would be to add terms penalizing solutions $\hat{\boldsymbol{\Psi}}$ that are not sparse, for example as

$$\hat{\boldsymbol{\Psi}} = \arg\min_{\boldsymbol{\Psi}} g_1(\boldsymbol{\Psi}) + \lambda||\boldsymbol{\Psi}||_0 \tag{7}$$

where $||\boldsymbol{\Psi}||_0$ is the pseudo-norm counting the number of non-zero elements in $\boldsymbol{\Psi}$, and λ is a regularization parameter. However, this in general leads to a combinatorial problem whose complexity grows exponentially with the dimension of $\boldsymbol{\Psi}$. To avoid this, one can approximate the ℓ_0 penalty by the convex function

$$g_2(\boldsymbol{\Psi}) = \sum_{p=1}^{P}\sum_{\ell=1}^{L_{\max}} |a_{p,\ell}| \tag{8}$$

The resulting problem

$$\underset{\boldsymbol{\Psi}}{\text{minimize}}\, g_1(\boldsymbol{\Psi}) + \lambda g_2(\boldsymbol{\Psi}) \tag{9}$$

is known as the LASSO [31]. In fact, it can be shown that under some restrictions on the set of frequencies ω (see also [32]), the LASSO is guaranteed to retrieve the non-zero indices of $\boldsymbol{\Psi}$ with high probability, although these conditions are not assumed to be met here. To encourage the group-sparse behavior of $\hat{\boldsymbol{\Psi}}$, one can further introduce

$$g_3(\boldsymbol{\Psi}) = \sum_{p=1}^{P}\sqrt{\sum_{\ell=1}^{L_{\max}} |a_{p,\ell}|^2} \tag{10}$$

which is also a convex function. The inner sum corresponds to the ℓ_2-norm, and does not enforce sparsity within each pitch, whereas instead the outer sum, corresponding to the ℓ_1-norm, enforces sparsity between pitches. Thereby, adding the $g_3(\boldsymbol{\Psi})$ constraint will penalize the number of non-zero pitches. The resulting estimator was in [13] termed the Pitch Estimation using Block Sparsity (PEBS) estimator. However, if we for some p have $2L_p \leq L_{\max}$, the above penalties have no way of discriminating between the correct pitch candidate ω_p and the spurious sub-octave candidate $\omega_p/2$. However, as the candidates will differ in that the sub-octave will only contribute to the harmonic signal at every other frequency in the block, as was seen in Figure 1, one may reduce the risk of such a misclassification by further adding the penalty

$$\breve{g}_4(\mathbf{\Psi}) = \sum_{p=1}^{P} \sum_{\ell=0}^{L_{\max}} \Big| |a_{p,\ell+1}| - |a_{p,\ell}| \Big| \tag{11}$$

where we define

$$a_{p,0} = a_{p,L_{\max}+1} = 0 \quad , \forall p \tag{12}$$

which would add a cost to blocks where there are notable magnitude variations between neighboring harmonics. Unfortunately, (11) is not convex, but a simple convex approximation would be

$$\tilde{g}_4(\mathbf{\Psi}) = \sum_{p=1}^{P} \sum_{\ell=0}^{L_{\max}} |a_{p,\ell+1} - a_{p,\ell}| \tag{13}$$

which would be a good approximation of (11) if all the harmonics had similar phases. This estimator was in [13] termed the PEBS-TV estimator. Clearly, this may not be the case, resulting in that the penalty in (13) would also penalize the correct candidate. An illustration of this is found by considering the worst-case scenario, when all the adjacent harmonics are completely out of phase and have the same magnitudes, i.e., $a_{p,\ell+1} = a_{p,\ell} e^{i\pi}$ with magnitude $|a_{p,\ell}| = r$, for $\ell = 1, \ldots, L_p - 1$. Then, the penalty in (13) will yield a cost of $\tilde{g}_4(\mathbf{\Psi}_{\omega_p}) = 2rL_p$ rather than the desired $\breve{g}_4(\mathbf{\Psi}_{\omega_p}) = 2r$. The cost may also be compared with that of (8), which is $g_2(\mathbf{\Psi}_{\omega_p}) = rL_p$, suggesting that this would add a relatively large penalty. More interestingly, for the sub-octave candidate pitch, the cost will be just as large, i.e., if $\omega_{p'} = \omega_p/2$, then $\tilde{g}_4(\mathbf{\Psi}_{\omega_{p'}}) = 2rL_p$ provided that $L_{\max} \geq 2L_p$, thereby offering no possibility of discriminating between the true pitch and its sub-octave. Such a worst case scenario is just as unlikely as all harmonics having the same phase, if assuming that the phases are uniformly distributed on $[0, 2\pi)$. Instead, the $\tilde{g}_4$ penalty of the true pitch will be slightly smaller than its sub-octave counterpart, on average, and together with (10), the scales tip in favor of the true pitch, as shown in [13]. One may thus conclude that the combination of g_3 and $\tilde{g}_4$ provides a block sparse solution where sub-octaves are usually discouraged. However, it should be noted that such a solution requires the tuning of two functions to control the block sparsity.

This work proposes to simplify the PEBS-TV estimator by improving the approximation in (13), by using an adaptive penalty approach. In order to do so, let $\varphi_{p,\ell}$ denote the phase of the component with frequency $\omega_{p,\ell}$, and collect all the phases in the parameter set

$$\mathbf{\Phi} = \{\mathbf{\Phi}_{\omega_1}, \ldots, \mathbf{\Phi}_{\omega_P}\} \tag{14}$$

$$\mathbf{\Phi}_{\omega_p} = \{\varphi_{p,1}, \ldots, \varphi_{p,L_{\max}}\} . \tag{15}$$

The penalty function in (11) may then instead be approximated as

$$g_4(\mathbf{\Psi}, \mathbf{\Phi}) = \sum_{p=1}^{P} \sum_{\ell=0}^{L_{\max}} |a_{p,\ell+1} \mathrm{e}^{-i\varphi_{p,\ell+1}} - a_{p,\ell} \mathrm{e}^{-i\varphi_{p,\ell}}| \tag{16}$$

thus penalizing only differences in magnitude, given that the phases $\varphi_{p,\ell+1}$ have been chosen as to offset phase differences between the harmonics. In order to do so, the phases $\varphi_{p,\ell}$ need to be estimated as the arguments of the latest available amplitude estimates $a_{p,\ell}$. As a result, (16) yields an improved approximation of (11), avoiding the issues of (13) described above, and also promotes a block sparse solution. The block sparsity is promoted due to the introduction of zero amplitudes in (12). In effect, this introduces a penalty for activating a pitch block. As a result, the block-norm penalty function g_3 may be omitted, which simplifies the algorithm noticeably. Thus, we form the parameter estimates by solving

$$\hat{\mathbf{\Psi}} = \arg\min_{\mathbf{\Psi}} \; g_1(\mathbf{\Psi}) + \lambda_2 g_2(\mathbf{\Psi}) + \lambda_4 g_4(\mathbf{\Psi}, \mathbf{\Phi}) \tag{17}$$

where λ_2 and λ_4 are user-defined regularization parameters that weigh the importance of each penalty function with that of the residual cost. To form the convex criteria and to facilitate the implementation, consider the signal expressed in matrix notation as

$$\mathbf{y} = \left[\begin{array}{ccc} y(1) & \ldots & y(N) \end{array} \right]^T = \sum_{p=1}^{P} \mathbf{W}_p \, \mathbf{a}_p + \mathbf{e} \triangleq \mathbf{W}\mathbf{a} + \mathbf{e} \tag{18}$$

where

$$\mathbf{W} = \left[\begin{array}{ccc} \mathbf{W}_1 & \ldots & \mathbf{W}_P \end{array} \right] \tag{19}$$

$$\mathbf{W}_p = \left[\begin{array}{ccc} \mathbf{z}_p^1 & \ldots & \mathbf{z}_p^{L_{\max}} \end{array} \right] \tag{20}$$

$$\mathbf{z}_p = \begin{bmatrix} e^{i\omega_p 1} & \dots & e^{i\omega_p N} \end{bmatrix}^T \tag{21}$$

$$\mathbf{a} = \begin{bmatrix} \mathbf{a}_1^T & \dots & \mathbf{a}_P^T \end{bmatrix}^T \tag{22}$$

$$\mathbf{a}_p = \begin{bmatrix} a_{p,1} & \dots & a_{p,L_{\max}} \end{bmatrix}^T \tag{23}$$

where the powers in the vectors $\mathbf{z}_p^k$ are taken element-wise. The dictionary matrix $\mathbf{W}$ is constructed by P horizontally stacked blocks, or dictionary atoms $\mathbf{W}_p$, where each is a matrix with $L_{\max}$ columns and N rows. In order to obtain an acceptable approximation of (11), the problem must be solved iteratively, where the last solution is used to improve the next. To pursue an even sparser solution, a re-weighting procedure is simultaneously used for $g_2(\mathbf{\Psi})$, similar to the one used in [33]. Redefining the functions g_j to operate on matrices, the solution is thus found at the kth iteration as

$$\hat{\mathbf{a}}^{(k)} = \arg\min_{\mathbf{a}} \frac{1}{2} \left\| \mathbf{y} - \mathbf{H}_1^{(k)} \mathbf{a} \right\|_2^2 + \lambda_2 \left\| \mathbf{H}_2^{(k)} \mathbf{a} \right\|_1 + \lambda_4 \left\| \mathbf{H}_4^{(k)} \mathbf{a} \right\|_1 \tag{24}$$

where

$$\mathbf{H}_1^{(k)} = \mathbf{W} \tag{25}$$

$$\mathbf{H}_2^{(k)} = \text{diag}\left(1/\left(\left|\hat{\mathbf{a}}^{(k-1)}\right| + \varepsilon\right)\right) \tag{26}$$

$$\mathbf{H}_4^{(k)} = \mathbf{F}\,\text{diag}\left(\arg\left(\hat{\mathbf{a}}^{(k-1)}\right)\right)^{-1} \tag{27}$$

where diag$(\cdot)$ denotes a diagonal matrix formed with the given vector along its diagonal, $|\cdot|$ is element-wise absolute value, arg$(\cdot)$ is the element-wise complex argument, and $\varepsilon \ll 1$. If the magnitude of a certain component of $\hat{\mathbf{a}}^{(k-1)}$ is small, the construction of $\mathbf{H}_2^{(k)}$ will ensure that the magnitude of the corresponding component of $\hat{\mathbf{a}}^{(k)}$ will be penalized harder. This iterative re-weighting procedure will then be a sequence of convex approximations of a non-convex logarithmic penalty on the ℓ_1 norm of $\mathbf{a}$. The inclusion of ε is made to ensure that a division by zero is avoided. Also, $\mathbf{I}$ denotes the identity matrix, and $\mathbf{F}$ is a $P(L_{\max}+1) \times PL_{\max}$ matrix $\mathbf{F} = \text{diag}(\mathbf{F}_1, \dots, \mathbf{F}_P)$, where each block $\mathbf{F}_p$ is a $(L_{\max}+1) \times L_{\max}$ matrix with elements

$$f_{k,\ell} = \begin{cases} 1 & \text{if } k = \ell = 1 \\ -1 & \text{if } k = \ell, \ell \neq 1 \\ 1 & \text{if } k = \ell + 1 \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

As intended, the minimization in (24) is convex, and may be solved using one of many publicly available convex solvers, such as, for instance, the interior point methods SeDuMi [34] or SDPT3 [27]. However, these methods are quite computationally burdensome and will scale poorly with increased data length and larger grids. Instead, we here propose an efficient implementation using ADMM. The problem in (24) may be implemented in a similar manner as was done in [25], requiring only two tuning parameters, λ_2 and λ_4. The proposed method compares to the PEBS and PEBS-TV algorithms as improving upon the former, and requiring fewer tuning parameters than the latter. The proposed method is therefore termed a light and improved version of PEBS, here denoted the PEBSI-Lite algorithm.

4 ADMM implementation

In order to solve (24), we proceed to introduce an efficient ADMM implementation. To this end, let $\mathbf{z} \in \mathbb{C}^{PL_{\max}}$ be the primal optimization variable and introduce the auxiliary variables $\mathbf{u}_1 \in \mathbb{C}^N$, $\mathbf{u}_2 \in \mathbb{C}^{PL_{\max}}$, and $\mathbf{u}_4 \in \mathbb{C}^{P(L_{\max}+1)}$ and let

$$\mathbf{G}^{(k)} = \begin{bmatrix} \mathbf{H}_1^{(k)T} & \mathbf{H}_2^{(k)T} & \mathbf{H}_4^{(k)T} \end{bmatrix}^T \tag{29}$$

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_1^T & \mathbf{u}_2^T & \mathbf{u}_4^T \end{bmatrix}^T . \tag{30}$$

Thus, we want to solve

$$\underset{\mathbf{z}}{\text{minimize}} f\left(\mathbf{G}^{(k)}\mathbf{z}\right) \tag{31}$$

where

$$f\left(\mathbf{G}^{(k)}\mathbf{z}\right) = \frac{1}{2}\left\|\mathbf{y} - \mathbf{H}_1^{(k)}\mathbf{z}\right\|_2^2 + \lambda_2\left\|\mathbf{H}_2^{(k)}\mathbf{z}\right\|_1 + \lambda_4\left\|\mathbf{H}_4^{(k)}\mathbf{z}\right\|_1 . \tag{32}$$

Using the auxiliary variabel $\mathbf{u}$, one may equivalently solve

$$\begin{aligned} &\underset{\mathbf{z},\mathbf{u}}{\text{minimize}} f(\mathbf{u}) + \frac{\mu}{2}\left\|\mathbf{G}^{(k)}\mathbf{z} - \mathbf{u}\right\|_2^2 \\ &\text{subject to } \mathbf{G}^{(k)}\mathbf{z} - \mathbf{u} = \mathbf{0} \end{aligned} \tag{33}$$

where μ is a positive scalar, as the added term is zero for any feasible point. The Lagrangian can be succinctly expressed using the (scaled) dual variable

$$\mathbf{d} = \begin{bmatrix} \mathbf{d}_1^T & \mathbf{d}_2^T & \mathbf{d}_4^T \end{bmatrix}^T \tag{34}$$

where $\mathbf{d}_1 \in \mathbb{C}^N$, $\mathbf{d}_2 \in \mathbb{C}^{PL_{\max}}$, and $\mathbf{d}_4 \in \mathbb{C}^{P(L_{\max}+1)}$. By completing the square, the Lagrangian of the problem can be equivalently expressed as

$$L_\mu(\mathbf{z}, \mathbf{u}, \mathbf{d}) = f(\mathbf{u}) + \frac{\mu}{2}\left\|\mathbf{G}^{(k)}\mathbf{z} - \mathbf{u} - \mathbf{d}\right\|_2^2 - \frac{\mu}{2}\|\mathbf{d}\|_2^2 \ . \tag{35}$$

Also, define

$$\boldsymbol{\zeta}(j) = \begin{bmatrix} \boldsymbol{\zeta}_1^T(j) & \boldsymbol{\zeta}_2^T(j) & \boldsymbol{\zeta}_4^T(j) \end{bmatrix}^T \tag{36}$$

where

$$\boldsymbol{\zeta}_\ell(j) = \mathbf{H}_\ell^{(k)}\mathbf{z}(j+1) - \mathbf{d}_\ell(j) \ , \ \ell = 1, 2, 4 \ . \tag{37}$$

The Lagrangian (35) is separable in the variables $\mathbf{z}$, $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_4$, and one may thus form an updating scheme similar to that in [25], as

$$\mathbf{z}(j+1) = \arg\min_{\mathbf{z}} \ \left\|\mathbf{G}^{(k)}\mathbf{z} - \mathbf{u}(j) - \mathbf{d}(j)\right\|_2^2 \tag{38}$$

$$\mathbf{u}_1(j+1) = \arg\min_{\mathbf{u}_1} \frac{1}{2}\|\mathbf{y} - \mathbf{u}_1\|_2^2 + \frac{\mu}{2}\|\boldsymbol{\zeta}_1(j) - \mathbf{u}_1\|_2^2 \tag{39}$$

$$\mathbf{u}_2(j+1) = \arg\min_{\mathbf{u}_2} \lambda_2\|\mathbf{u}_2\|_1 + \frac{\mu}{2}\|\boldsymbol{\zeta}_2(j) - \mathbf{u}_2\|_2^2 \tag{40}$$

$$\mathbf{u}_4(j+1) = \arg\min_{\mathbf{u}_4} \lambda_4\|\mathbf{u}_4\|_1 + \frac{\mu}{2}\|\boldsymbol{\zeta}_4(j) - \mathbf{u}_4\|_2^2 \tag{41}$$

$$\mathbf{d}(j+1) = \mathbf{u}(j+1) - \boldsymbol{\zeta}(j) \ . \tag{42}$$

The updates of $\mathbf{z}$ and $\mathbf{u}_1$ are given by

$$\mathbf{z}(j+1) = \left(\mathbf{G}^{(k)H}\mathbf{G}^{(k)}\right)^{-1}\mathbf{G}^{(k)H}\left(\mathbf{u}(j) + \mathbf{d}(j)\right) \tag{43}$$

and

$$\mathbf{u}_1(j+1) = \frac{\mathbf{y} + \mu\boldsymbol{\zeta}_1(j)}{1+\mu} \tag{44}$$

respectively.

Algorithm 1 The proposed PEBSI-Lite algorithm

```
initiate $k := 0$, $\mathbf{H}_1^{(0)} = \mathbf{I}$, $\mathbf{H}_4^{(0)} = \mathbf{F}$, and
    $\hat{\mathbf{a}}^{(0)} = \mathbf{z}_{\text{save}} = \mathbf{d}_{\text{save}} = \mathbf{0}^{PL_{\max} \times 1}$
repeat {adaptive penalty scheme}
    initiate $j := 0$, $\mathbf{u}_2(0) = \hat{\mathbf{a}}^{(k)}$,
        $\mathbf{z}(0) = \mathbf{z}_{\text{save}}$, and $\mathbf{d}(0) = \mathbf{d}_{\text{save}}$
    repeat {ADMM scheme}
        $\mathbf{z}(j) = \left(\mathbf{G}^{(k)H}\mathbf{G}^{(k)}\right)^{-1} \mathbf{G}^{(k)H}\left(\mathbf{u}(j) + \mathbf{d}(j)\right)$
        $\mathbf{u}_1(j+1) = \frac{\mathbf{y} + \mu \boldsymbol{\zeta}_1(j)}{1+\mu}$
        $\mathbf{u}_2(j+1) = \mathbf{T}\left(\boldsymbol{\zeta}_2(j), \frac{\lambda_2}{\mu}\right)$
        $\mathbf{u}_4(j+1) = \mathbf{T}\left(\boldsymbol{\zeta}_4(j), \frac{\lambda_4}{\mu}\right)$
        $\mathbf{d}(j+1) = \mathbf{u}(j+1) - \boldsymbol{\zeta}(j)$
        $j \leftarrow j + 1$
    until convergence
    store $\hat{\mathbf{a}}^{(k)} = \mathbf{u}_2(\text{end})$, $\mathbf{z}_{\text{save}} = \mathbf{z}(\text{end})$, and $\mathbf{d}_{\text{save}} = \mathbf{d}(\text{end})$
    update $\mathbf{H}_2^{(k+1)} = \mathrm{diag}\left(1/|\hat{\mathbf{a}}^{(k)}| + \varepsilon\right)$, $\mathbf{H}_4^{(k+1)} = \mathbf{F}\,\mathrm{diag}\left(\arg\left(\hat{\mathbf{a}}^{(k)}\right)\right)^{-1}$
    $k \leftarrow k + 1$
until convergence
```

Using the element-wise shrinkage function,

$$\mathbf{T}\left(\mathbf{x}, \xi\right) = \frac{\max(|\mathbf{x}| - \xi, 0)}{\max(|\mathbf{x}| - \xi, 0) + \xi} \odot \mathbf{x} \tag{45}$$

where the max function operates on each element in the vector $\mathbf{x}$ separately and $\odot$ denotes element-wise multiplication, one may update $\mathbf{u}_2$ and $\mathbf{u}_4$ as

$$\mathbf{u}_2(j+1) = \mathbf{T}\left(\boldsymbol{\zeta}_2(j), \frac{\lambda_2}{\mu}\right) \tag{46}$$

and

$$\mathbf{u}_4(j+1) = \mathbf{T}\left(\boldsymbol{\zeta}_4(j), \frac{\lambda_4}{\mu}\right) \tag{47}$$

respectively. The resulting PEBSI-Lite algorithm is summarized in Algorithm 1, where the solution is given as $\hat{\mathbf{a}} = \hat{\mathbf{a}}^{(k_{\text{end}})}$ with k_{end} denoting the last iteration index of the outer loop. The complexity of the resulting algorithm will be dominated by the computation of step 5 in Algorithm 1. This system of equations can be solved efficiently by storing the Cholesky factorization of the matrix to be inverted, with a one-time cost of $\mathcal{O}\left(p^3\right)$ operations, where p denotes the number of variables (here, assumed to be larger than the number of data points). Furthermore, at each iteration, one needs to perform a back solve costing $\mathcal{O}\left(p^2\right)$ operations.

5 Self-regularization

The quality of the pitch estimates produced by the PEBSI-Lite algorithm depends on the values of the regularization parameters λ_2 and λ_4. In general, large values of λ_2 encourage sparse solutions, whereas large values of λ_4 encourage solutions that are smooth within blocks. As the model order is unknown, it is generally hard to determine how sparse the solution should be in order to be considered the desired one. Therefore, one often determines the values of the regularization parameters using cross-validation schemes, making the performance of the methods user dependent. Instead, one would like to have a systematic and preferable automatic method for choosing λ_2 and λ_4, and thereby the model order.

A common approach to solving model order problems is to use information criteria such as AIC or BIC [35], which measure the fit of the model to the data, while penalizing high model orders, resulting in a trade-off criterion that should take its optimal value for the correct model order. For the LASSO problem, there have been suggestions of appropriate model order criteria [36], [37]. In [13], the authors suggest a BIC-style criterion for multi-pitch estimation for given regularization parameters. However, this criterion can only be used to determine which of the found pitches are true and which are spurious, and not to determine the appropriate regularization parameters. Thus, even if one has an efficient criterion for choosing between different models, one first has to form a set of candidate models, in effect running Algorithm 1 for different values of λ_2 and λ_4. For the simpler case of the LASSO, the analog is to solve (9) for all $\lambda \in \mathbb{R}_+$, for which there are algorithms such as LARS [38]. There have also been methods suggested for solving the LASSO for only a finite number of values λ, i.e., only values of the regularization parameter where the number of active components of the solution change (see, e.g., [37]). For our problem, the analog is to find solutions for the set of parameter values

$$\{(\lambda_2, \lambda_4) | (\lambda_2, \lambda_4) \in \mathbf{R}_+ \times \mathbf{R}_+\} . \tag{48}$$

For the real-variable counterpart of the here considered pitch estimation problem, known as the Sparse Fused LASSO [39], there have been algorithms suggested for computing the whole solution surface. In [40], the authors present an elegant way of finding a solution path for the case of the dictionary $\mathbf{W}$ being the identity matrix, meaning that the estimated amplitude vector is just a smoothed version of the signal $\mathbf{y}$. The algorithm can be used for general matrices $\mathbf{W}$, under the condition that $\mathbf{W}$ has full column rank, something that is not true for dictionaries in high-resolution spectral estimation applications such as the one considered here. In [41], the authors present an approach to find the solution path of

$$\underset{\beta}{\text{minimize}}\ \frac{1}{2}\left\|\mathbf{y}-\mathbf{W}\beta\right\|_2^2+\lambda\left\|\mathbf{D}\beta\right\|_1 \tag{49}$$

for the real-variable case with a general penalty matrix $\mathbf{D}$ by considering the solution paths of the dual variable. Unfortunately, this is only for the one-dimensional case, i.e., for the case when the minimization has only a single regularization parameter.

Despite the above efficient ADMM implementation, it is computationally cumbersome to conduct a search on (48) in order to find an appropriate model order, with the computation complexity increasing both in the case of longer signals, and when using more elements in the dictionary. Instead of constructing a fully general path algorithm for PEBSI-Lite, we therefore proceed to propose a scheme for constructing a reduced size signal adapted dictionary that combined with a parametrization of the regularization parameters (λ_2, λ_4) will allow us to form good pitch estimates without having to predefine values of the regularization parameters, by means of a simple line search instead of searching through (48). The proposed dictionary construction begins by estimating the frequency content of the signal without imposing any harmonic structure. This estimation may be performed by any standard method, such as ESPRIT (see, e.g., [42]). As the number of sinusoidal components is unknown, estimates corresponding to different model orders can be evaluated using, for instance, the BIC criterion (see, e.g., [35])

$$\text{BIC}_k = 2N\log\hat{\sigma}_k^2+(5k+1)\log N \tag{50}$$

where $\hat{\sigma}_k^2$ is the maximum likelihood estimate of the residual variance corresponding to the model constituted by k estimated sinusoids, in order to choose a suitable model order. The accuracy of the frequency estimates produced by ESPRIT will suffer if a too low model order is determined, whereas it is less sensitive to cases when the model order is moderately overestimated. Thus, we propose to increase the robustness of the frequency estimates by using $k + \delta$, $\delta \geq 1$, estimated sinusoids for the case when order k is determined optimal by the BIC. As the only interesting pitch candidates are those having at least one harmonic corresponding to a present sinusoidal component, we can then design a considerably reduced dictionary, containing only pitches with such matching harmonics. If one has some prior knowledge of the nature of the signal, one could impose stronger assumptions on the candidate pitches in order to reduce the dictionary further, e.g., by allowing only pitches whose first harmonic is found in the set of estimated sinusoids. Using the obtained dictionary, one could then proceed to conduct a search for λ_2 and λ_4.

Although considerably cheaper as compared to when performed using a full dictionary, a complete evaluation of the $\lambda_2\lambda_4$-plane is still somewhat expensive. To avoid a full grid search, the following heuristic concerning the connection between λ_2 and λ_4 can be used. Assume that we have a single-pitch signal where all L_k harmonics have equal magnitude r. Further, assume that when setting $\lambda_4 = 0$, λ' is the largest value of λ_2 resulting in a nonzero solution, where each harmonic amplitude is estimated to r_0. If we would instead set $\lambda_2 = 0$, and consider which value of λ_4 that should result in the same solution, this value should be

$$\lambda_4 = \frac{L_k}{2}\lambda' \tag{51}$$

as this would result in precisely the same penalty as with $\lambda_4 = 0$, $\lambda_2 = \lambda'$. More compactly, we have that

$$\lambda_2 = \alpha\lambda' \, , \, \lambda_4 = \big(1 - \alpha\big)\frac{L_k}{2}\lambda' \tag{52}$$

yields the penalty $\lambda' \, L_k r_0$ for all $\alpha \in [0, 1]$. If we assume (52) to be true, we should, for spectrally smooth signals, expect to see ridges in the solution surface where the number of pitches present in the solution changes, and the shapes of the ridges in the $\lambda_2\lambda_4$-plane should be described by lines similar to (52).

This is illustrated in Figure 2, presenting a plot of the number of pitches present in the solution for different values (λ_2, λ_4) for a signal consisting of three pitches with fundamental frequencies 400, 550 and 700 Hz, and with 4, 8, and 12 harmonics, respectively. The magnitude of each harmonic amplitude has

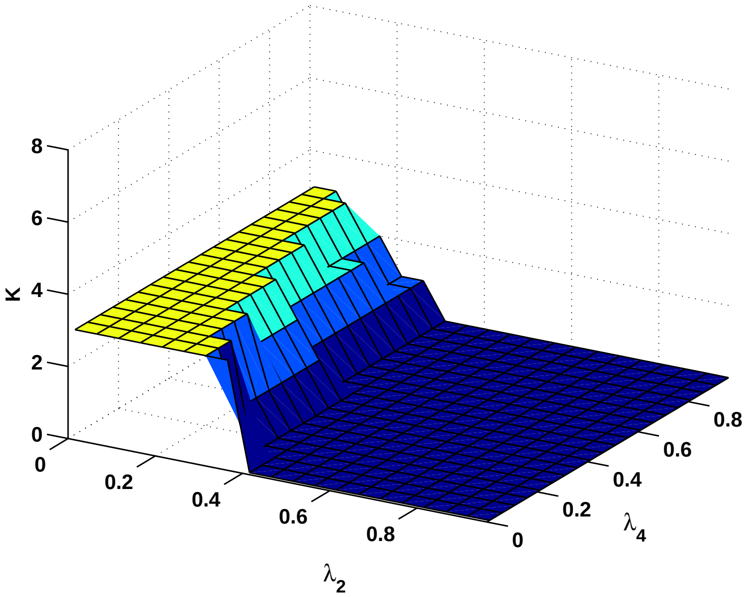

Figure 2: Number of pitches, K, present in the solution of PEBSI-Lite for different values (λ_2, λ_4) when applied to a three pitch signal with 4, 8, and 12 harmonics, respectively.

been drawn uniformly on $(0.9, 1.1)$ and each phase has been drawn uniformly on $(0, 2\pi)$. The signal was sampled at frequency 20 kHz in a time frame of length 40 ms, generating 800 samples of the signal. The signal-to-noise ratio (SNR), as defined in (55), was 20 dB. On the plateau with two pitches, the pitch with four harmonics have been forced to zero, whereas on the plateau with one pitch present, only the pitch with 12 harmonics is present. Note the shape of the different plateaus: seen in the $\lambda_2\lambda_4$-plane, the slopes of the ridges seem to be well described by (52) where $L_k = 4, 8$, and 12, for the three ridges corresponding to changes from three to two, from two to one, and from one to zero pitches, respectively. The signal corresponding to Figure 2 has a relatively low level of noise. Increasing the noise level, the least regularized solutions, i.e., with λ_2 and λ_4 close to zero, results in more than three non-zero pitches. Guided by this observation, one could reduce the search for (λ_2, λ_4) from a 2-D to a 1-D search by using a

Algorithm 2 Self-Regularized PEBSI-Lite

```
initiate $\ell = 1$
repeat {sinusoidal component estimation}
    $\hat{\boldsymbol{\omega}}_\ell \leftarrow \ell$ sinusoidal components from ESPRIT
    $\mathrm{BIC}_\ell \leftarrow 2N \log \hat{\sigma}^2(\hat{\boldsymbol{\omega}}_\ell) + (5\ell + 1) \log N$
until $\mathrm{BIC}_\ell > \mathrm{BIC}_{\ell-1}$
$\hat{\boldsymbol{\omega}}_{\ell+\delta} \leftarrow \ell + \delta$ sinusoidal components from ESPRIT, where $\delta \geq 1$ is a safety margin
construct dictionary $\mathbf{W}$ from $\hat{\boldsymbol{\omega}}_{\ell+\delta}$
$L \leftarrow$ largest number of active harmonics among candidate pitches in $\mathbf{W}$
initiate $\lambda = \varepsilon$, $k = 1$
$\hat{\sigma}_y^2 \leftarrow \mathrm{Var}(y)$
$\hat{\sigma}_{\mathrm{MLE}}^2 \leftarrow$ maximum likelihood (least squares) estimate of noise power
repeat {regularization parameter line search}
    $\lambda_2 \leftarrow \lambda$, $\lambda_4 \leftarrow \frac{L}{2}\lambda$
    form amplitude estimate $\hat{\mathbf{a}}^{(k)}$ from Algorithm 1
    estimate the power of the model residual $\hat{\sigma}^2(\lambda_2, \lambda_4)$
    $\lambda \leftarrow \lambda + \varepsilon$
    $k \leftarrow k + 1$
until $\left(\hat{\sigma}^2(\lambda_2, \lambda_4) - \hat{\sigma}_{\mathrm{MLE}}^2\right) > \tau \hat{\sigma}_y^2$
$\hat{\mathbf{a}} \leftarrow \hat{\mathbf{a}}^{(k-1)}$
```

re-parametrization. Keeping the plateaus in Figure 2 and our assumption of spectral smoothness in mind, we should expect a desirable solution to correspond to a (λ_2, λ_4)-pair with $\lambda_2 \leq \lambda_4$. In order to get solutions regularized with respect to spectral smoothness, while keeping the risk of getting only zero solutions low, the following parametrization can be used. Let λ denote the only free parameter and set

$$\lambda_2 = \lambda \tag{53}$$

$$\lambda_4 = \frac{L}{2}\lambda \tag{54}$$

where L is the largest number of harmonics among the pitches present in the signal. Although L is unknown, it can be estimated during the dictionary construction phase using the BIC and ESPRIT estimates, permitting us to conduct a line search for the value of λ. Having obtained a solution with PEBSI-Lite using

Estimator	SNR (dB)	-5	0	5	10	15	20
PEBS-TV	λ_2	0.2	0.2	0.2	0.15	0.1	0.1
	λ_3	0.3	0.3	0.3	0.2	0.2	0.15
	λ_4	0.1	0.1	0.1	0.75	0.75	0.05
PEBS	λ_2	0.2	0.2	0.2	0.15	0.15	0.1
	λ_3	0.4	0.4	0.4	0.3	0.3	0.2

Table 1: Regularization parameter values for PEBS-TV and PEBS.

the regularization parameter λ, the residual power σ_λ^2 can be estimated by least squares. It is worth noting that in low noise environments, it can be expected that false pitches modeling noise will not contribute much to the signal power. Thus, the first significant rise in residual power is expected to occur when one of the true pitches are set to zero. Therefore, we propose keeping only models that correspond to lower values of σ_λ^2 and then choosing the optimal model as the one having the least number of active pitches. The complete algorithm for the dictionary construction, line search, and pitch estimation is outlined in Algorithm 2, where ε denotes the step size of the line search and $\tau \in (0, 1)$ is a threshold for detecting an increase in model residual power. The step size ε can be chosen based on afforded estimation time, as small values of ε will result in more steps for the line search. τ can be chosen based on estimates of the noise power, if available.

6 Numerical results

We proceed to examine the performance of the proposed algorithm using signals simulated from the pitch model (1) as well as synthetic audio signals generated from MIDI, and measured audio signals.

6.1 Two-pitch signal

We initially examine a simulated dual-pitch signal, measured in white Gaussian noise at different SNRs ranging from -5 dB to 20 dB in steps of 5 dB. The SNR is here defined as

$$\text{SNR} = 10 \log_{10} \frac{\sigma_x^2}{\sigma_e^2} \tag{55}$$

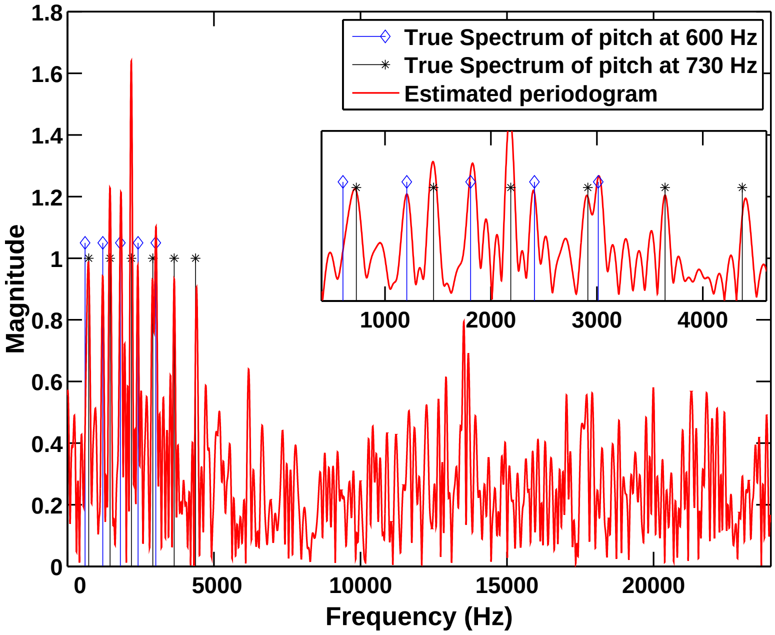

Figure 3: The periodogram estimate and the true signal studied in Figure 4.

where σ_x^2 and σ_e^2 are the powers of the signal and the noise, respectively. For a pitch signal generated by (1), under the simplifying assumption of distinct sinusoidal components, the power of the signal is given by

$$\sigma_x^2 = \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} \frac{|a_{k,\ell}|^2}{2} \,. \tag{56}$$

At each SNR, 200 Monte Carlo simulations were performed, each simulation generating a signal with fundamental frequencies of 600 and 730 Hz. As PEBS and PEBS-TV rely on a predefined frequency grid, the fundamental frequencies were randomly chosen at each simulation uniformly on $600 \pm d/2$ and $730 \pm d/2$, where d is the grid point spacing, to reflect performance in present of off-grid effects. The phases of the harmonics in each pitch were chosen uniformly on $[0, 2\pi)$, whereas all had unit magnitude. The signal was sampled at $f_s = 48$ kHz on a time frame of 10 ms, yielding $N = 480$ samples per frame. As a result, the

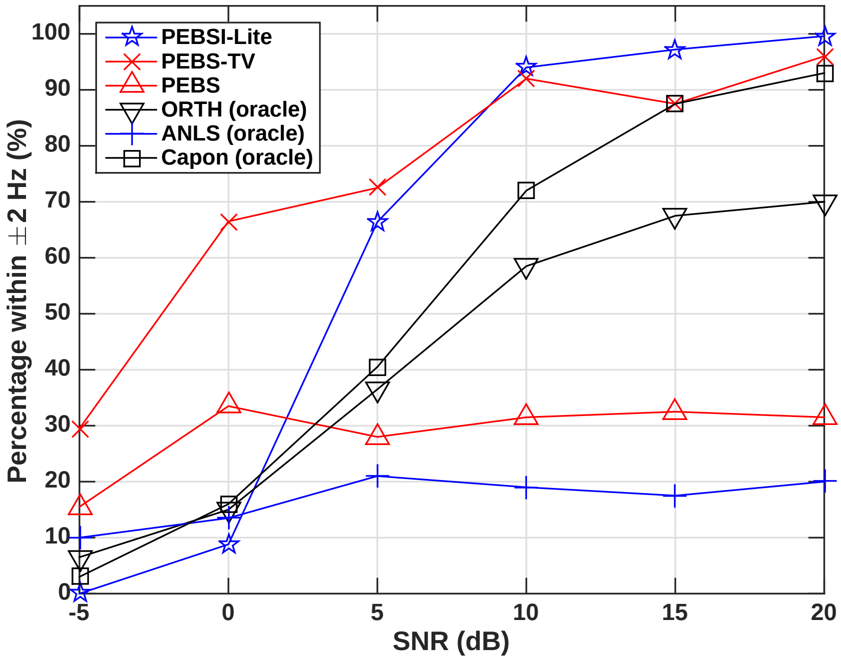

Figure 4: Percentage of estimated pitches where both fundamental frequencies lie at most 2 Hz, or $d/5 = 1/50N$, from the ground truth, plotted as a function of SNR. Here, the pitches have $[5, 6]$ harmonics, respectively, and $L_{\max} = 10$.

pitches were spaced by approximately f_s/N Hz, which is the resolution limit of the periodogram. This is also seen in Figure 3, illustrating the resolution of the periodogram as well as the frequencies of the harmonics, at SNR $= -5$ dB. From the figure, it may be concluded that the signal contains more than one harmonic source, as the observed peaks are not harmonically related. Furthermore, it is clear that the fundamental frequencies are not separated by the periodogram, indicating that any pitch estimation algorithm based on the periodogram would suffer notable difficulties. For PEBSI-Lite, the estimates are formed using Algorithm 2 with $\tau = 0.1$ and $\varepsilon = 0.05$. The safety margin for the sinusoidal model order is $\delta = 1$. For PEBS and PEBS-TV, the estimation procedure is initiated using a coarse dictionary, with candidate pitches uniformly distributed on the interval $[280, 1500]$ Hz, thus also including $\omega_p/2$ and $2\omega_p$ for both pitches. The coarse resolution was $d = 10$ Hz, i.e., still a super-resolution of $f_s/10N$. After estima-

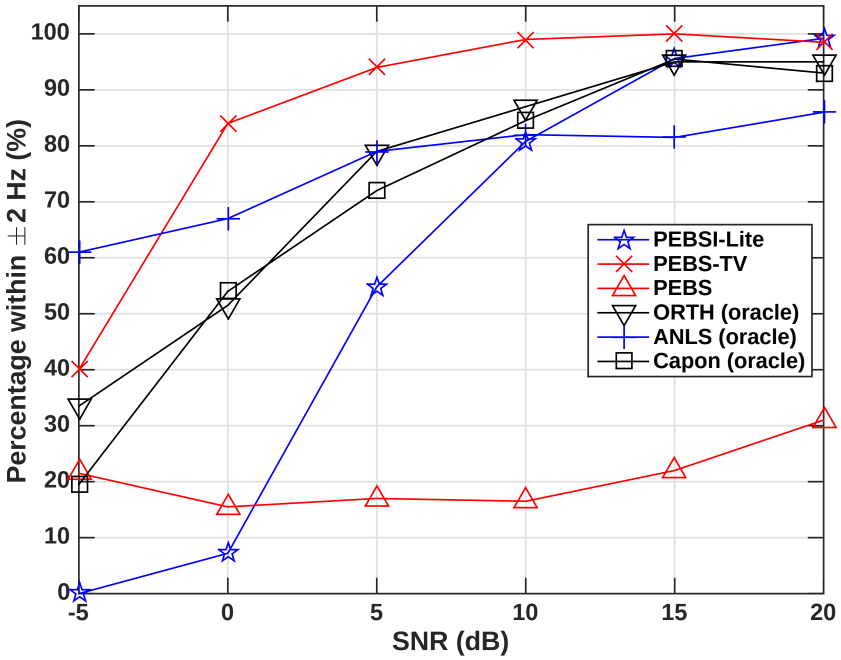

Figure 5: Percentage of estimated pitches where both fundamental frequencies lie at most 2 Hz, or $d/5 = 1/50N$, from the ground truth, plotted as a function of SNR. Here, the pitches have [10, 11] harmonics, respectively, and $L_{\max} = 20$.

tion on this grid, a zooming step was taken where a new grid with spacing $d/10$ was laid $\pm 2d$ around each pitch having non-zero power. The regularization parameter values used for PEBS-TV and PEBS are presented in Table 1. The values where selected using *manual cross-validation* for similar signals. Comparisons were also made with the ANLS, ORTH, and the harmonic Capon estimators, which had been given the *oracle* model orders (see [9] for more details on these methods). The simulation and estimation procedure was performed for two cases; one where the number of harmonics L_k were set to 5 and 6, and one where L_k were set to 10 and 11. In the former case, $L_{\max} = 10$ and in the latter, $L_{\max} = 20$, i.e., well above the true number of harmonics.

Figures 4 and 5 show the percentage of pitch estimates where both lie within ±2 Hz from the true values for the six compared methods, for the case of 5 and 6 as well as 10 and 11 harmonics, respectively. In this setting, PEBS performs

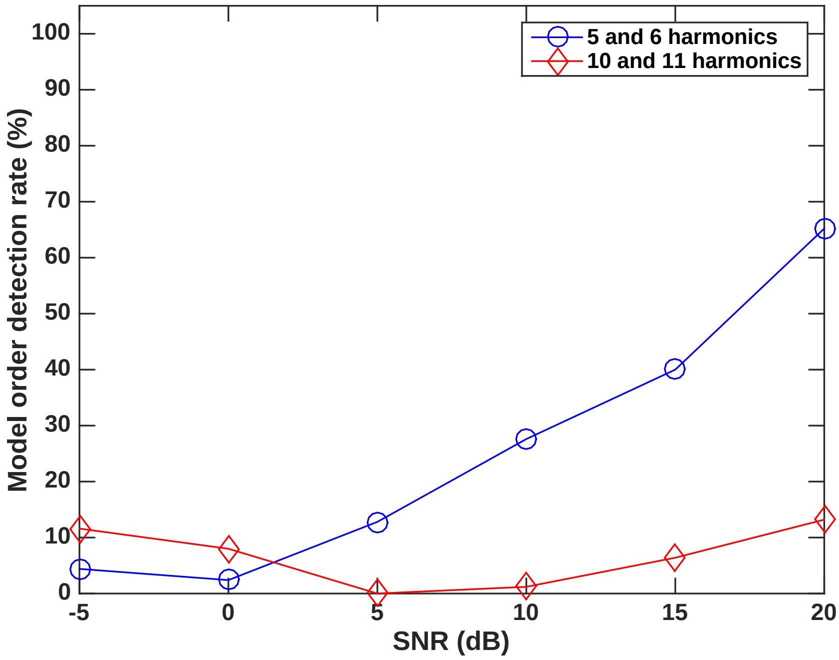

Figure 6: The percentage of the estimates in which the model order choice criterion (50) correctly determines the number of sinusoidal components in the two-pitch signal, for the case of 5 and 6 harmonics, and 10 and 11 harmonics, respectively.

poorly, as the generous choices of L_{max} allow it to pick the sub-octave, as predicted. As can be seen in Figure 4, PEBSI-Lite performs better than all reference methods for SNRs above and including 10 dB despite not having the model order information given to ORTH, ANLS, and Capon, nor having the supervised regularization parameter choices of PEBS and PEBS-TV. Though, in higher noise settings, the performance of PEBSI-Lite degrades and its pitch frequency estimates are worse than those produced by the reference methods for SNRs below 10 dB. For the case of 10 and 11 harmonics, PEBSI-Lite performs on par with the reference methods for SNRs above and including 15 dB, while performing worse in

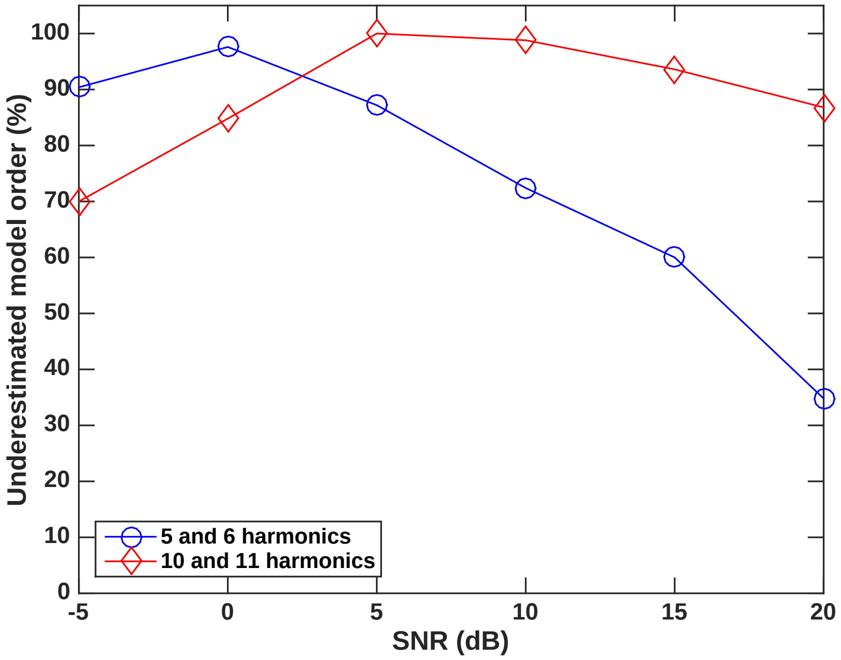

Figure 7: The percentage of the estimates in which the model order choice criterion (50) selects a model with too few sinusoidal components for the two-pitch signal, for the case of 5 and 6 harmonics, and 10 and 11 harmonics, respectively.

higher noise settings. As shown in Figures 6 and 7, the drop in performance for lower SNRs results from the difficulty of accurately estimating the total number of sinusoids, as used by the ESPRIT step, for such signals. In Figure 6, the percentage of the estimates in which the the BIC criterion (50) correctly determines the number of sinusoidal components in the signal is presented, whereas Figure 7 shows the percentage of the estimates in which the BIC criterion (50) determines a too low model order. As is clear from the figures, the model order estimates strongly degrade for lower SNRs, thus causing the PEBSI-Lite dictionary to be inaccurate. Clearly, all the other methods here shown using oracle model order information would suffer drastically from such inaccuracies, although it should be

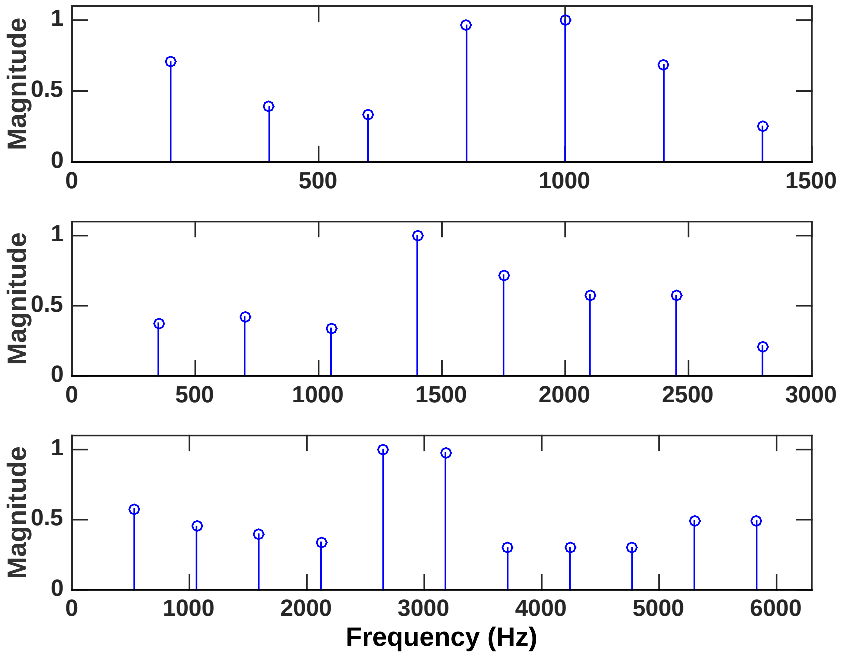

Figure 8: Magnitudes for the harmonics of the three pitches constituting the test signal for the Monte Carlo simulations.

stressed that one may expect these methods to suffer further, as they also need to perform an exhaustive combinatorial search to determine the number of pitches given the found number of sinusoids.

6.2 Three-pitch signal

To further examine the performance of Algorithm 2, it was evaluated using a simulated triple-pitch signal, measured in white Gaussian noise at different SNR levels, ranging from 0 dB to 25 dB, in steps of 5 dB. Instead of using unit magnitudes of the harmonics, as was the case for the above presented two-pitch setting, the spectral envelopes of the three pitch components were constructed from periodograms of three different speech recordings. The formants of the three pitches are displayed in Figure 8. The pitches had fundamental frequencies 200,

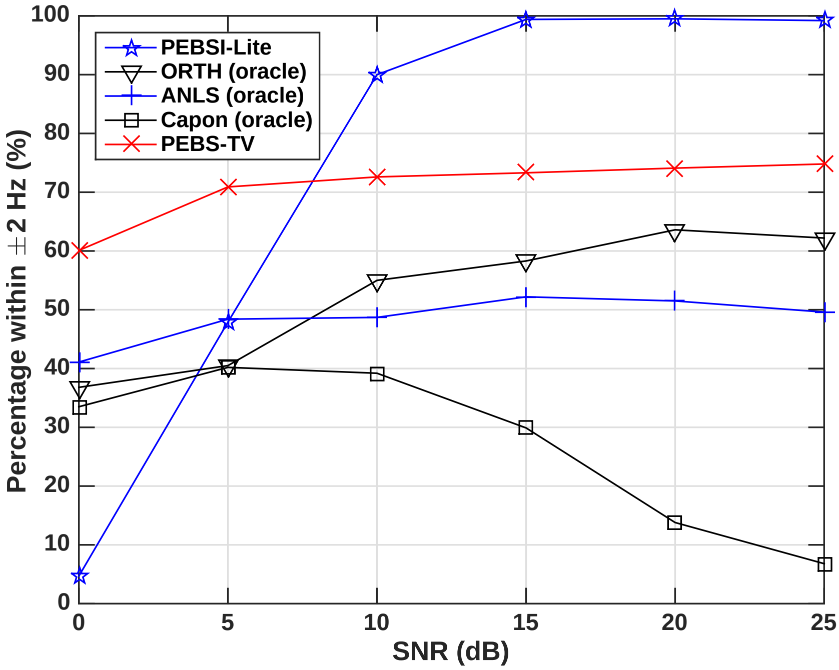

Figure 9: Percentage of estimated pitches where all three fundamental frequencies lie at most 2 Hz from the ground truth.

350, and 530 Hz, and 7, 8, and 11 harmonics, respectively. At each level of SNR, 1000 Monte Carlo simulations were performed, where the fundamental frequencies were chosen uniformly on 200 ± 2.5, 350 ± 2.5, and 530 ± 2.5 Hz, respectively, and the phase of each harmonic was chosen uniformly on $[0, 2\pi)$. The signal was sampled in a 40 ms window at a sampling frequency of 20 kHz, generating 800 samples of the signal. The algorithm settings were $\tau = 0.1$, $\varepsilon = 0.05$, and $\delta = 1$. Here, Algorithm 2 was compared to the ANLS, ORTH, harmonic Capon, as well as PEBS-TV estimators. The three first comparison methods were given the oracle model orders.

To illustrate the fact that the choice of regularization parameter values is not universal, the values found using cross-validation for the two-pitch case (see Table 1) were used for PEBS-TV initially. However, this resulted in such poor performance that the parameter values had to be slightly altered in order to make PEBS-TV an interesting reference method. As a compromise, the parameter val-

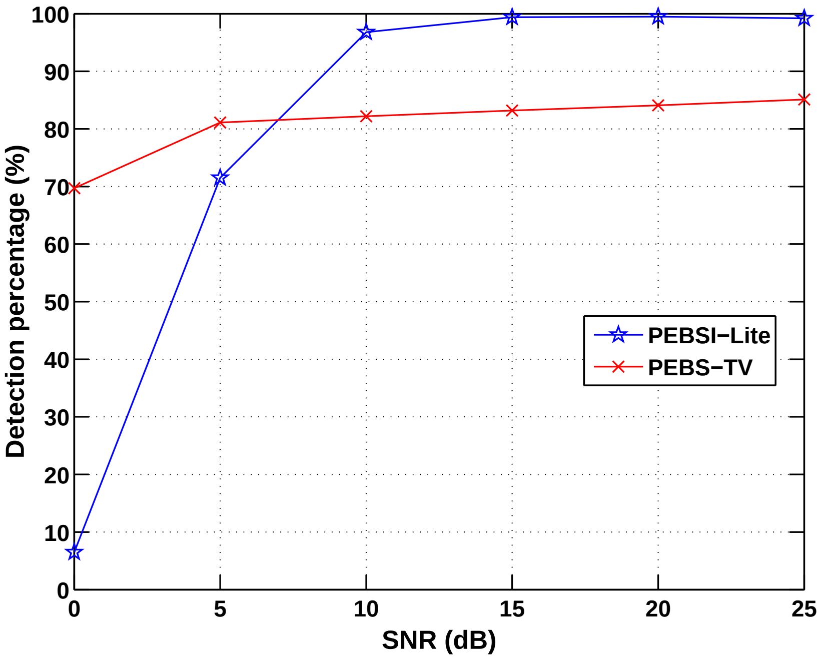

Figure 10: Estimated probability of PEBSI-Lite determining the correct number of pitches for the triple pitch test signal.

ues corresponding to SNR 20 dB in Table 1 were used for all SNRs in this simulation setting. For the dictionaries of PEBSI-Lite and PEBS-TV, $L_{\max} = 16$ was used, well above the true model orders. Figure 9 shows the percentage of the pitch estimates where all three pitch estimates lie within ±2 Hz of the true values for the five different methods. As can be seen, the performance of PEBSI-Lite is again poor for low SNRs while improving considerably for lower noise levels. The low scoring for PEBSI-Lite for low SNRs is mainly due to the selection of wrong model orders. This is illustrated in Figure 10, which shows the percentage of the estimates in which PEBSI-Lite and PEBS-TV select the correct number of pitches. As can be seen, for an SNR of 0 dB, PEBSI-Lite selects the true model order in less than 10% of the simulations. Mostly, a too high model order is selected, which is to be expected as the model order choice is based on the power of the model residual and that the pitch estimates depend on the accuracy of the initial ESPRIT estimates. Arguably, one could improve on these results by either

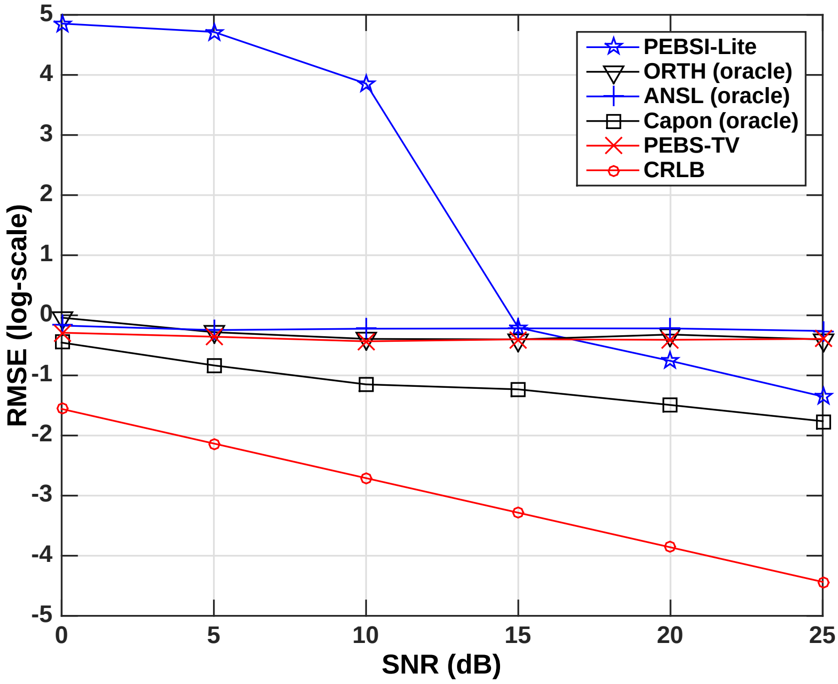

Figure 11: The RMSE for the fundamental frequency estimates for the triple pitch test signal, as compared to the (root) CRLB. For PEBSI-Lite and PEBS-TV, only estimates where the number of pitches is found are considered. For the reference methods ORTH, ANLS, Capon, and PEBS-TV only estimates where all estimated pitch frequencies lie within 2 Hz of the true pitch frequencies are considered.

using prior knowledge of the noise level or by estimating it, and based on this make the model order selection scheme more robust. Figure 11 shows the root mean squared error (RMSE) for the estimated fundamental frequencies. Instead of presenting three separate RMSE plots, Figure 11 shows an aggregate version where the MSE for the three pitches have been summed. In order to compute relevant RMSE values for PEBSI-Lite and PEBS-TV, estimates where the model order has not been correctly determined have been discarded. Thus, for an SNR level of 0 dB, the RMSE values for PEBSI-Lite are based on quite few samples. However, as PEBSI-Lite finds the correct model order for high SNR levels with high probability, the corresponding RMSE values are more trustworthy in these regions. For the reference methods ORTH, ANLS, Capon, and PEBS-TV, some of the estimates deviate from the true pitch frequencies with as much as 100 Hz, resulting in very large RMSE values should all estimates be used in their computation. Thus, in order to obtain RMSE values comparable to that of the PEBSI-Lite estimates, only estimates found within 2 Hz of the true pitch frequencies are used when computing RMSE for the reference methods. With this, as can be seen in Figure 11, PEBSI-Lite performs worse than the reference methods for SNRs below and including 10 dB, while outperforming all reference methods except Capon for SNRs above and including 20 dB. Though, one should bear in mind that the RMSE values for Capon for these SNRs are based on only 15% respectively 8% of the available pitch estimates, as can be seen in Figure 9, and that the Capon method has been allowed oracle model order knowledge. Also presented in Figure 11 is the root Cramér-Rao lower bound (CRLB) for the estimates of the pitch frequencies. As the frequencies of the harmonics in this case are distinct and the additive noise is white Gaussian, the lower limit for the variance of an unbiased pitch frequency estimate $\hat{f}_k$ is given by [9]

$$\mathrm{Var}\left(\hat{f}_k\right) \geq \frac{6\sigma^2\left(f_s/2\pi\right)^2}{N(N^2-1)\sum_{\ell=1}^{L_k}|a_{k,\ell}|^2\ell^2} \tag{57}$$

where σ^2 is the power of the additive noise, $a_{k,\ell}$ is the amplitude of harmonic ℓ of pitch k, N is the number of data samples, and f_s is the sampling frequency. In analog with the summed MSE values for the pitch estimates, the root CRLB curve presented here is the sum of the three separate limits, i.e.,

$$\mathrm{CRLB} = \sum_{k=1}^{3} \frac{6\sigma^2\left(f_s/2\pi\right)^2}{N(N^2-1)\sum_{\ell=1}^{L_k}|a_{k,\ell}|^2\ell^2}\,. \tag{58}$$

As can bee seen in Figure 11, PEBSI-Lite, as well as the other methods, fails to reach the CRLB. In an attempt to improve the PEBSI-Lite estimates for SNR levels above and including 15 dB, a non-linear least squares (NLS) search was performed, using the presented algorithm estimate as an initial estimate of all the unknown parameters, including the model orders. This means that we obtain refined estimates of the pitch frequencies f_k contained in the vector $\mathbf{f}$ as (see, e.g, [42])

$$\mathbf{f} = \arg\max_{\mathbf{f}} \mathbf{y}^{\mathbf{H}}\mathbf{B}\left(\mathbf{B}^{\mathbf{H}}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathbf{H}}\mathbf{y} \tag{59}$$

where **B** is a block matrix consisting of K blocks,

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_1 & \ldots & \mathbf{B}_K \end{bmatrix} \tag{60}$$

where each block $\mathbf{B}_j$ corresponds to a separate pitch and is constructed as

$$\mathbf{B}_j = \begin{bmatrix} e^{i2\pi f_j/f_s t_1} & \ldots & e^{i2\pi L_j f_j/f_s t_1} \\ \vdots & & \vdots \\ e^{i2\pi f_j/f_s t_N} & \ldots & e^{i2\pi L_j f_j/f_s t_N} \end{bmatrix} . \tag{61}$$

Given that the PEBSI-Lite estimates are fairly close to the true pitch frequencies, we expect the NLS scheme to converge if we solve (59) using routines like MATLAB's *fminsearch* initialized with the PEBSI-Lite estimates. However, the success of such a scheme is not only dependent on good initial frequency estimates, we also need the true number of harmonics L_j for each pitch.

Figure 12 presents a plot of the average absolute error in the number of detected harmonics for each pitch for the test signal when using PEBSI-Lite. As can be seen, the number of detected harmonics is only correct for the third pitch even for the largest SNRs. The errors in number of harmonics for the first and second pitches are due to the relatively small amplitudes of both pitches highest order harmonics, as shown in Figure 8, making these harmonics prone to occasionally being cancelled out by the PEBSI-Lite regularization penalties. Using erroneous harmonic orders as input to the NLS search, we expect the resulting pitch frequency estimates to be somewhat biased. Indeed, this is what happens. Figure 13 presents a plot of the RMSE of the pitch frequency estimates when the PEBSI-Lite estimates for SNRs above and including 15 dB have been post-processed using NLS. As can be seen, the estimator still fails to reach the CRLB, although the estimation errors have become smaller. Note also that the slopes of the RMSE curve for PEBSI-Lite and CRLB are now somewhat different, which is due to that the erroneous harmonic orders induces varying degrees of bias in the estimates. Considering computational complexity, ANLS and ORTH are by far the fastest methods, with average running times of 0.03 and 1.6 seconds per estimation cycle on a regular PC, respectively. For Capon and PEBS-TV, the corresponding running times are 6.1 and 6.4 seconds for the considered example, respectively, while running PEBSI-Lite using Algorithm 2 requires on average 40.1 seconds per estimation cycle. As a comparison, it may be noted that if one replaces Algorithm 1 in Algorithm 2 to instead use SeDuMi or SDPT3, the computation time for this

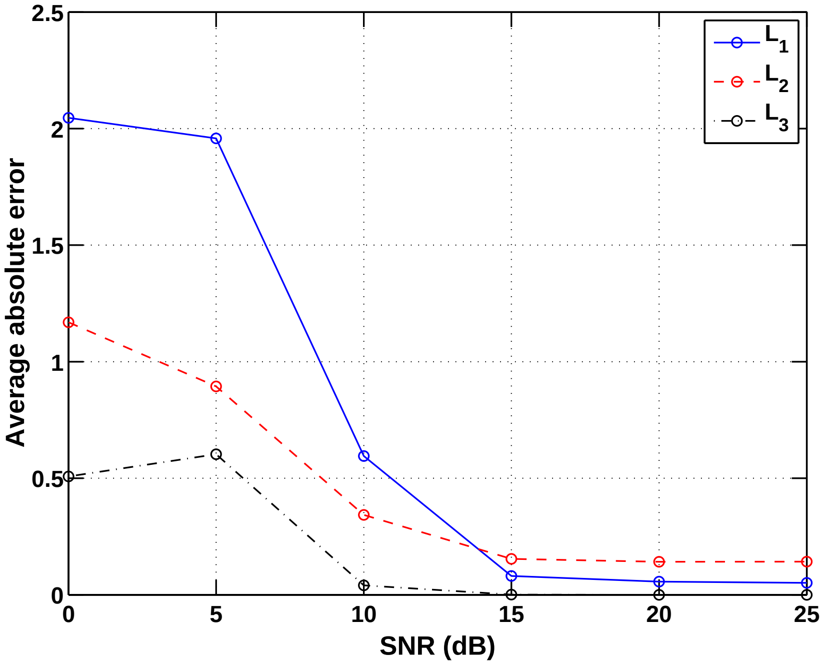

Figure 12: The average absolute error in the number of detected harmonics (L_1, L_2, L_3) for the three pitches of the test signal when using PEBSI-Lite. Only estimates where the correct number of pitches is found are considered.

step of Algorithm 2 increases almost tenfold[3]. Although Algorithm 2 is considerably more expensive to run than the reference methods, it should be noted that the method does not require any user input in terms of regularization parameter values. PEBS-TV could arguably be tuned to perform on par with PEBSI-Lite if one is allowed to change the values of its regularization parameters. However, PEBS-TV needs the setting of three parameter values and after trying only seven such triplets, the computational time is the same as running Algorithm 2 in its entirety.

[3] For all algorithms, the given execution times are those of direct implementations of the corresponding methods. Clearly, these methods can be more efficiently implemented by fully exploiting their inherent structures.

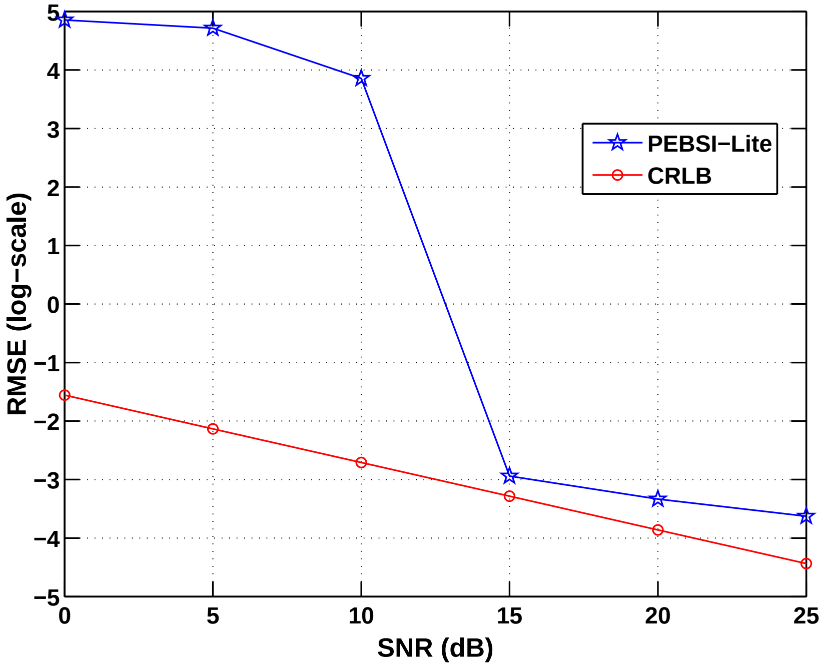

Figure 13: The RMSE for the fundamental frequency estimates where the estimates obtained using PEBSI-Lite have been improved using NLS for SNR levels 15, 20, and 25 dB, as compared to the (root) CRLB. Only estimates where the number of pitches is found are considered.

6.3 MIDI and measured audio signals

Figure 14 shows a plot of the spectrogram of a signal consisting of three MIDI-saxophones playing notes with fundamental frequencies 311, 277, and 440 Hz. The signal was sampled initially at 44 kHz and then down sampled to 20 kHz. The 311 Hz saxophone starts out alone and is after 0.45 seconds joined by the 277 Hz saxophone and after 0.95 seconds by the 440 Hz saxophone. The image is quite blurred for the later parts of the signal, but for the first half second, one can clearly see the harmonic structure of the saxophone pitch. It is worth noting that a large number of harmonics is present. Figure 15 shows pitch estimates produced by Algorithm 2, using $\tau = 0.1$ and $L_{\max} = 15$, when applied to the same signal, using windows of lengths 40 ms. As can be seen, the estimates are quite

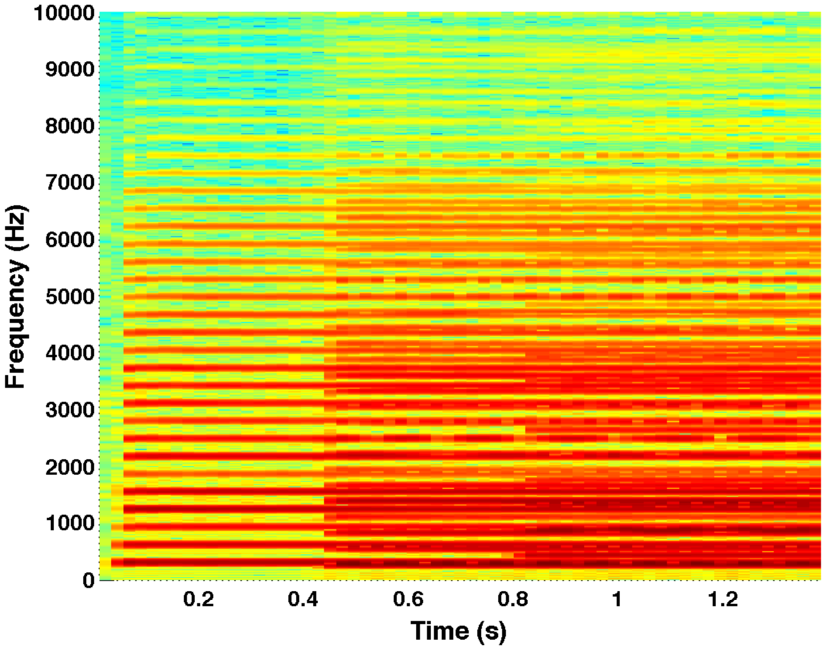

Figure 14: Spectrogram for a signal consisting of one, two and lastly three MIDI-saxophones playing notes with fundamental frequencies 311, 277, and 440 Hz, respectively.

accurate, with the exception of the beginning of the first tone and for a single frame where the 440 Hz pitch is mistaken for a 220 Hz pitch. It is worth noting that such errors may be avoided using the information resulting from earlier frames, for instance using an approach similar to [22]. The figure also shows the estimated pitch tracks obtained using the ESACF estimator [43]; this estimator requires *a priori* knowledge of the number of sources in the signal, but is, given this information, able to estimate the number of harmonics of each source. Here, ESACF has thus been provided oracle knowledge of the number of sources, with each source given the same maximum harmonic order as used by PEBSI-Lite (as before, the latter also has to estimate the number of sources). As can be seen from the figure, the ESACF estimator fails to track the pitches in several of the frames. In particular, it fails to estimate the pitch with fundamental frequency 440 Hz

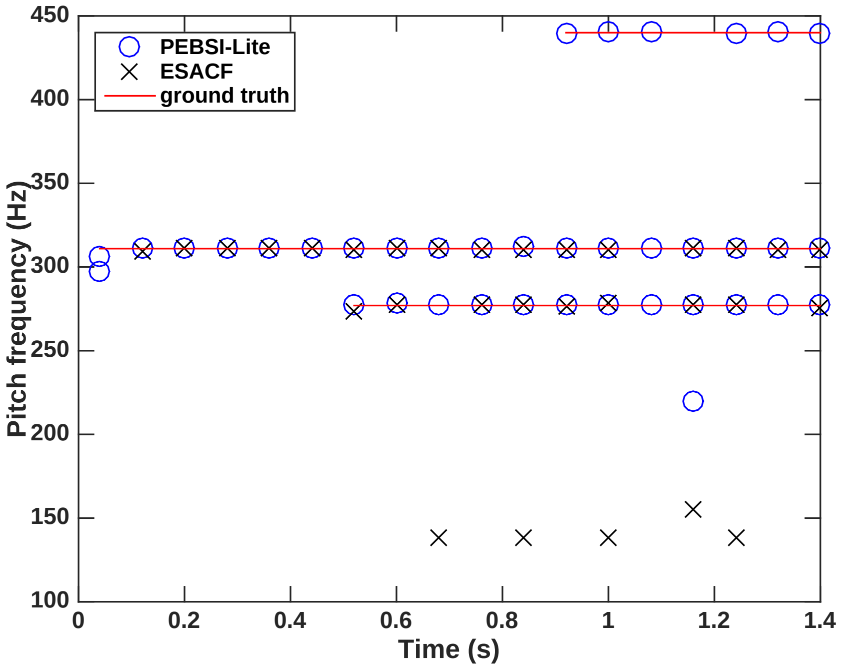

Figure 15: Pitch tracks for a signal consisting of one, two, and lastly three MIDI-saxophones playing notes with fundamental frequencies 311, 277, and 440 Hz, respectively.

altogether. Furthermore, Figure 16 examines the performance of the PEBSI-Lite estimator when applied to a measured audio signal. The considered signal consists of three trumpets playing the notes A4, B4, and C♯4, which, using concert tuning, corresponds to the fundamental frequencies 440, 493.883, and 554.365 Hz, respectively. However, it should be noted, that as the musicians play with vibrato, the fundamental frequencies are not constant across the frames, which may also be seen in the resulting estimates. To facilitate for a comparison, the ground truth estimates of the fundamental frequencies have been obtained using the joint order and (single) pitch estimation algorithm ANLS, presented in [11], when applied to each individual trumpet separately. As a comparison, the figure also shows the three fundamental frequencies obtained using the ESACF estimator (which has here, again, been allowed oracle knowledge of the number of sources, but using the same maximum number of harmonics as used by PEBSI-Lite). As can

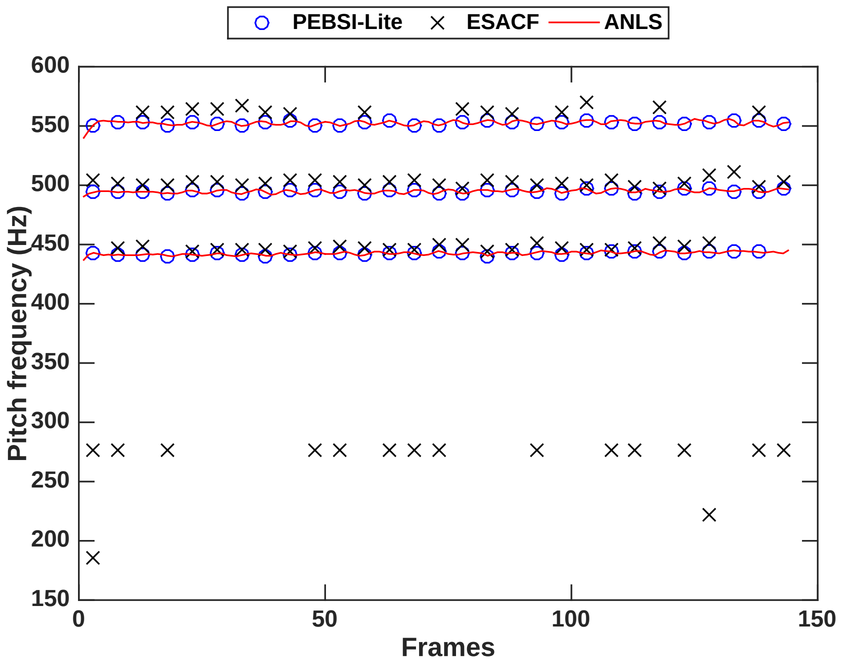

Figure 16: Pitch tracks produced by PEBSI-Lite as well as ESACF when applied to a triple-pitch signal consisting of three trumpets. The ground truth has been obtained using ANLS applied to the single source signals.

be seen, PEBSI-Lite accurately tracks each of the three pitches, even catching the pitch variations caused by the vibrato. As before, it may be noted that the estimates produced by ESACF have lower accuracy as compared to PEBSI-Lite, with the ESACF estimator here erroneously picking one of the sub-octaves in some of the frames. The trumpet signal was sampled at 8 kHz. The pitch estimates where formed in non-overlapping frames of length 30ms.

The performance of PEBSI-Lite and ESACF on real audio was also evaluated on the Bach10 dataset [44]. This dataset consists of ten chorales 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 seconds long. Each piece was sampled at 44.1 kHz, then downsampled to 22.05 kHz, and divided into non-overlapping frames of length 30 ms. Estimates of the ground truth fundamental frequencies in each frame were obtained by applying YIN [45] to

Performance measure	PEBSI-Lite	ESACF
Accuracy	0.499	0.269
Precision	0.631	0.471
Recall	0.609	0.386

Table 2: Performance measures for PEBSI-Lite and ESACF when evaluated on the Bach10 dataset.

each individual channel. Obvious errors in the YIN estimates were then corrected manually.

As before, to yield its best possible performance, ESACF was given oracle knowledge of the number of present pitches and both methods were given a maximum harmonic order of 15. For PEBSI-Lite, $\tau = 0.1$ was used. Table 2 presents the resulting measures of the *accuracy*, *precision*, and *recall* for the dataset, defined as

$$\text{Accuracy} = \frac{\sum_{i=1}^{I}\sum_{t=1}^{T_i}\text{TP}(t,i)}{\sum_{i=1}^{I}\sum_{t=1}^{T_i}\text{TP}(t,i)+\text{FP}(t,i)+\text{FN}(t,i)} \tag{62}$$

$$\text{Precision} = \frac{\sum_{i=1}^{I}\sum_{t=1}^{T_i}\text{TP}(t,i)}{\sum_{i=1}^{I}\sum_{t=1}^{T_i}\text{TP}(t,i)+\text{FP}(t,i)} \tag{63}$$

$$\text{Recall} = \frac{\sum_{i=1}^{I}\sum_{t=1}^{T_i}\text{TP}(t,i)}{\sum_{i=1}^{I}\sum_{t=1}^{T_i}\text{TP}(t,i)+\text{FN}(t,i)} \tag{64}$$

where $\text{TP}(t,i)$, $\text{FP}(t,i)$, and $\text{FN}(t,i)$ denote the number of true positive, false positive, and false negative pitch estimates, respectively, for frame t in music piece i. Furthermore, T_i is the number of frames for music piece i, whereas I is the number of music pieces. Here, an estimated pitch is associated with a ground truth pitch only if its fundamental frequency lies within a quarter tone, or 3%, of the ground truth pitch (see also, e.g., [46]). To avoid the most non-stationary frames, where we cannot expect the estimates produced by PEBSI-Lite and ESACF, nor the ground truth, to be reliable, frames containing note onsets, defined as frames where one of the ground truth pitches change with more than a semitone, have been excluded when computing the measures. As can be seen from the table, PEBSI-Lite performs better than ESACF for all of the three considered measures *accuracy*, *precision*, and *recall*. As PEBSI-Lite does, for now, not incor-

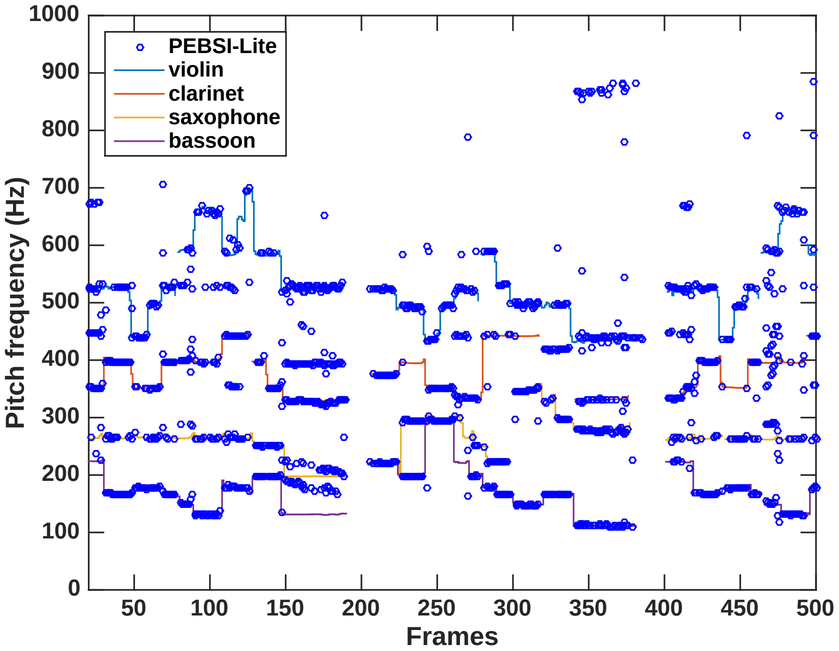

Figure 17: Pitch tracks produced by PEBSI-Lite when applied to first 15 seconds of J. S. Bach's *Ach, lieben Christen*, performed by a violin, a clarinet, a saxophone, and a bassoon. The ground truth has been obtained using YIN applied to the single source signals.

porate information between adjacent frames, these results are most promising for what might be achievable when extended to include such information.

As an illustration of the performance, Figures 17 and 18 present pitch tracks produced by PEBSI-Lite and ESACF when applied to the first 15 seconds of one of the pieces in the dataset, namely *Ach, lieben Christen*. As can be seen from the figures, PEBSI-Lite tracks the fundamental frequencies of the violin, the saxophone, and the bassoon fairly well, while having trouble with the clarinet. This problem is caused by the shape of the spectral envelope of the clarinet, as it is dominated by a large peak at the fundamental frequency, with very weak overtones, and thus deviates from the here used model assumption of spectral smoothness. It may also be noted that PEBSI-Lite has better performance at the stationary parts of the signal, while producing more erroneous estimates at note

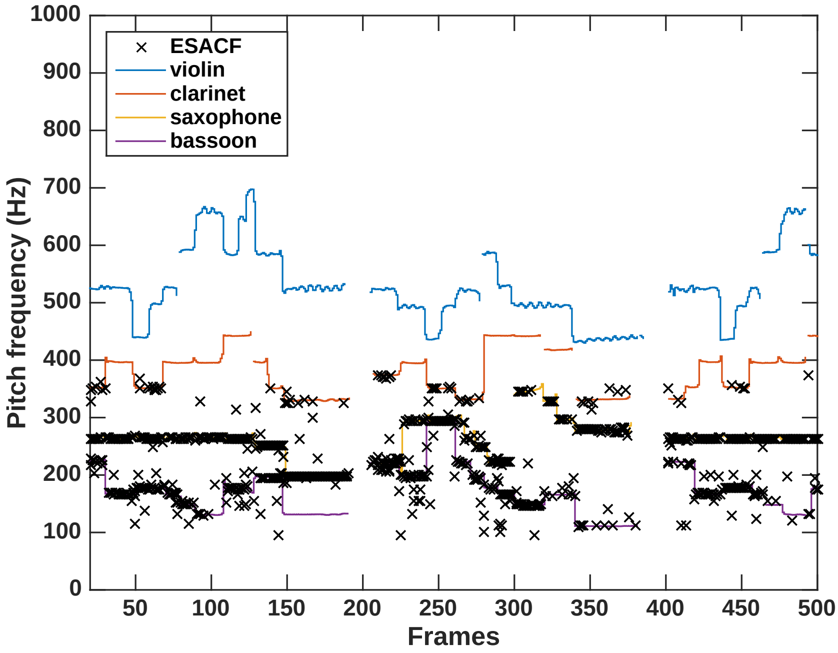

Figure 18: Pitch tracks produced by ESACF when applied to the first 15 seconds of J. S. Bach's *Ach, lieben Christen*, performed by a violin, a clarinet, a saxophone, and a bassoon. The ground truth has been obtained using YIN applied to the single source signals.

on- and offsets due to quickly changing spectral content. The ESACF estimator on the other hand has serious problems tracking the violin and clarinet, often picking sub-octaves estimates instead of the correct pitch, although being able to track the saxophone and bassoon fairly well.

7 Conclusions

The proposed algorithm PEBSI-Lite has been shown to be an accurate method for multi-pitch estimation. The method was shown to perform as good as, or better than, state-of-the-art methods. As compared to related methods, the presented algorithm requires fewer regularization parameters, simplifying the calibration of the method. Furthermore, the work introduces an adaptive dictionary scheme for determining suitable regularization parameters. Combined with this scheme, PEBSI-Lite was shown to outperform other multi-pitch estimation methods for high levels of SNR, while breaking down in too noisy settings. However, even if this scheme would fail to select the correct model order, the obtained efficient dictionary facilitates a more rigorous grid search in terms of computational complexity. Such a grid search could also exploit information about the solution surface obtained from the line search. Using an additional refinement step, the proposed algorithm is found to yield estimates reasonably close to being efficient, if considering that the method has not been allowed any knowledge of the model order of the signal.

References

[1] 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.

[2] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic Music Transcription: Challenges and Future Directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, Dec. 2013.

[3] A. Wang, "An Industrial Strength Audio Search Algorithm," in *4th International Conference on Music Information Retrieval*, Baltimore, Maryland, USA, Oct. 26-30 2003.

[4] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, Springer-Verlag, New York, NY, 1988.

[5] H. Fletcher, "Normal vibration frequencies of stiff piano string," *Journal of the Acoustical Society of America*, vol. 36, no. 1, 1962.

[6] N. R. Butt, S. I. Adalbjörnsson, S. D. Somasundaram, and A. Jakobsson, "Robust Fundamental Frequency Estimation in the Presence of Inharmonicities," in *38th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancouver, May 26–31, 2013.

[7] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "On the Influence of Inharmonicities in Model-Based Speech Enhancement," in *European Signal Processing Conference*, Marrakesh, Sept. 10-13 2013.

[8] T. Nilsson, S. I. Adalbjörnsson, N. R. Butt, and A. Jakobsson, "Multi-Pitch Estimation of Inharmonic Signals," in *European Signal Processing Conference*, Marrakech, Sept. 9-13, 2013.

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

[10] M. G. Christensen, S. H. Jensen, S. V. Andersen, and A. Jakobsson, “Subspace-based Fundamental Frequency Estimation,” in *European Signal Processing Conference*, Vienna, Sept. 7-10 2004.

[11] M. G. Christensen, A. Jakobsson, and S. H. Jensen, “Joint High-Resolution Fundamental Frequency and Order Estimation,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 15, no. 5, pp. 1635–1644, Jul. 2007.

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

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

[14] M. Genussov and I. Cohen, “Multiple fundamental frequency estimation based on sparse representations in a structured dictionary,” *Digit. Signal Process.*, vol. 23, no. 1, pp. 390–400, Jan. 2013.

[15] C. Kim, W. Chang, S-H. Oh, and S-Y. Lee, “Joint Estimation of Multiple Notes and Inharmoncity Coefficient Based on f0-Triplet for Automatic Piano Transcription,” *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1536–1540, Dec. 2014.

[16] V. Emiya, R. Badeau, and B. David, “Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1643–1654, Aug. 2010.

[17] K. O’Hanlon, “Structured Sparsity for Automatic Music Transcription,” in *37th IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Kyoto, 25-30 Mar. 2012.

[18] M. Bay, A.F. Ehmann, J.W. Beauchamp, P. Smaragdis, and J.S. Downie, “Second Fiddle is Important Too: Pitch Tracking Individual Voices in Polyphonic music,” in *13th Annual Conference of the International Speech Communication Association*, Portland, Sept. 2012, pp. 319–324.

[19] A. Klapuri, “Multiple fundamental frequency estimation based on harmonicity and spectral smoothness,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 804–816, 2003.

[20] P. Smaragdis and J.C. Brown, "Non-Negative Matrix Factorization for Polyphonic Music Transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 177–180.

[21] N. Bertin, R. Badeau, and E. Vincent, "Enforcing Harmonicity and Smoothness in Bayesian Non-Negative Matrix Factorization Applied to Polyphonic Music Transcription," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 18, no. 3, pp. 538–549, 2010.

[22] S. Karimian-Azari, A. Jakobsson, J. R. Jensen, and M. G. Christensen, "Multi-Pitch Estimation and Tracking using Bayesian Inference in Block Sparsity," in *23rd European Signal Processing Conference*, Nice, Aug. 31-Sept. 4 2015.

[23] 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.

[24] J. J. Fuchs, "On the Use of Sparse Representations in the Identification of Line Spectra," in *17th World Congress IFAC*, Seoul, Jul. 2008, pp. 10225–10229.

[25] M. A. T. Figueiredo and J. M. Bioucas-Dias, "Algorithms for imaging inverse problems under sparsity regularization," in *Proc. 3rd Int. Workshop on Cognitive Information Processing*, May 2012, pp. 1–6.

[26] T. Kronvall, M. Juhlin, S. I. Adalbjörnsson, and A. Jakobsson, "Sparse Chroma Estimation for Harmonic Audio," in *40th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Brisbane, Apr. 19-24 2015.

[27] 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.

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

[29] T. Kronvall, S. I. Adalbjörnsson, and A. Jakobsson, "Joint DOA and Multi-Pitch Estimation Using Block Sparsity," in *39th IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Florence, May 4-9 2014.

[30] T. Kronvall, S. I. Adalbjörnsson, and A. Jakobsson, "Joint DOA and Multi-pitch estimation via Block Sparse Dictionary Learning," in *22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, Sept. 1-5 2014.

[31] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society B*, vol. 58, no. 1, pp. 267–288, 1996.

[32] E. J. Candès, J. Romberg, and T. Tao, "Robust Uncertainty Principles: Exact Signal Reconstruction From Highly Incomplete Frequency Information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[33] E. J. Candès, M. B. Wakin, and S. Boyd, "Enhancing Sparsity by Reweighted l_1 Minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, Dec. 2008.

[34] 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.

[35] 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, Jul. 2004.

[36] C. D. Austin, R. L. Moses, J. N. Ash, and E. Ertin, "On the Relation Between Sparse Reconstruction and Parameter Estimation With Model Order Selection," *IEEE J. Sel. Topics Signal Process.*, vol. 4, pp. 560–570, 2010.

[37] A. Panahi and M. Viberg, "Fast Candidate Point Selection in the LASSO Path," *IEEE Signal Processing Letters*, vol. 19, no. 2, pp. 79–82, Feb. 2012.

[38] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, Apr. 2004.

[39] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society B*, vol. 67, no. 1, pp. 91–108, Jan. 2005.

[40] H. Hoefling, "A Path Algorithm for the Fused Lasso Signal Approximator," *Journal of Computational and Graphical Statistics*, vol. 19, no. 4, pp. 984–1006, Dec. 2010.

[41] R.J. Tibshirani and J. Taylor, “The Solution Path of the Generalized Lasso,” *The Annals of Statistics*, vol. 39, no. 3, pp. 1335–1371, Jun. 2011.

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

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

[44] Z. Duan and B. Pardo, “Bach10 dataset,” [Online]. Available: http://music.cs.northwestern.edu/data/Bach10.html, Accessed on: Dec. 2015.

[45] 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.

[46] 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, Oct. 2009.

B

Paper B

Online Estimation of Multiple Harmonic Signals

Filip Elvander, Johan Swärd, and Andreas Jakobsson

Centre for Mathematical Sciences, Lund University, Lund, Sweden

Abstract

In this paper, we propose a time-recursive multi-pitch estimation algorithm using a sparse reconstruction framework, assuming that only a few pitches from a large set of candidates are active at each time instant. The proposed algorithm does not require any training data, and instead utilizes a sparse recursive least squares formulation augmented by an adaptive penalty term specifically designed to enforce a pitch structure on the solution. The amplitudes of the active pitches are also recursively updated, allowing for a smooth and more accurate representation. When evaluated on a set of ten music pieces, the proposed method is shown to outperform other general purpose multi-pitch estimators in either accuracy or computational speed, although not being able to yield performance as good as the state-of-the art methods, which are being optimally tuned and specifically trained on the present instruments. However, the method is able to outperform such a technique when used without optimal tuning, or when applied to instruments not included in the training data.

Key words: Adaptive signal processing, dictionary learning, group sparsity, multi-pitch estimation, sparse recursive least squares

1 Introduction

The problem of estimating the fundamental frequency, or pitch, arises in a variety of fields, such as in speech and audio processing, non-destructive testing, and biomedical modeling (see, e.g., [1–6], and the references therein). In such applications, the measured signal may often result from several partly simultaneous sources, meaning that both the number of pitches, and the number of overtones of each such pitch, may be expected to vary over the signal. Such would be the case, for instance, in most forms of audio signals. The resulting multi-pitch estimation problem is in general difficult, with one of the most notorious issues being the so-called sub-octave problem, i.e., distinguishing between pitches whose fundamental frequencies are related by powers of two. Both non-parametric, such as methods based on autocorrelation (see, e.g., [7] and references therein), and parametric multi-pitch estimators (see, e.g., [2]) have been suggested, where the latter are often more robust to the sub-octave problem, 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.

Regrettably, the need for accurate model order information is a significant drawback, as such information is typically difficult to obtain and may vary rapidly over the signal. In order to alleviate this, several sparse reconstruction algorithms tailored for multi-pitch estimation have recently been proposed, allowing for estimators that do not require explicit knowledge of the number of sources or their harmonics; for example, in [8], the so-called PEBS estimator was introduced, exploiting the block-sparse structure of the pitch signal. This estimator was then further developed in [9], such that the likelihood of erroneously selecting a sub-octave in place of the true pitch was lowered, while also introducing a self-regularization technique for selecting the penalty parameters. Both these estimators form implicit model order decisions based on one or more tuning parameters that dictate the relative weight of various penalties. As shown in the above cited works, the resulting estimators are able to allow for (rapidly) varying model orders, without significant loss of performance. Earlier works based on sparse representations of signals also include works such as [10], which considers atomic decomposition of audio signals in both the time and the frequency domains.

There have also been methods proposed for multi-pitch estimation and tracking that are source specific, i.e., tailored specifically to sources, e.g., musical instruments, that are known to be present in the signal. In [11], the authors perform multi-pitch estimation on music mixtures by, via a probabilistic framework, matching the signal to a pre-learned dictionary of spectral basis vectors that correspond to instruments known to be present in the signal. A similar source specific idea was used in [12], where pitch estimation was performed by matching the signal to spectral templates learned from individual piano keys. Other methods specifically designed to handle multi-pitch estimation for pianos include [13–15]. Another field of research is designing multi-pitch estimators based on a two-matrix factorisation of the short-time Fourier transform, i.e., a non-negative matrix factorization (see, e.g., [16–18]). The method has also been used in the sparse reconstruction framework, for instance to learn atoms in order to decompose the signal [19]. A common assumption is also that of spectral smoothness within each pitch, which may also be exploited in order to improve the estimation performance (see, e.g., [13, 17, 20, 21]).

In many audio processing applications, pitch tracking is of great interest and despite being a problem that has been studied for a long time, it still attracts a lot of attention. Over the years, there have been many different approaches for tracking pitches; some of the more recent include particle filters [22], neural networks [23], and Bayesian filtering [24]. Many of these methods require *a priori* model order information, and/or are limited to the single pitch case. The sparse pitch estimators in [8], [9] are robust to these model assumptions, and allow for multiple pitches. However, these estimators process each data frame separately, treating each as an isolated and stationary measurement, without exploiting the information obtained from earlier data frames when forming the estimates. To allow for such correlation over time, the PEBS estimator introduced in [8] was recently extended to exploit the previous pitch estimates, as well as the power distribution of the following frame, when processing the current data frame [25]. In this work, we extend on this effort, but instead propose a fully time-recursive problem formulation using the sparse recursive least squares (RLS) estimator. The resulting estimator does not only allow for more stable pitch estimates as compared to earlier sparse multi-pitch estimators, as more information is used at each time-point, but also decreases the computational burden of each update, as new estimates are formed by updating already available ones.

On the other hand, sparse adaptive filtering is a field attracting steadily increasing attention, with, for instance, the sparse RLS algorithm being explored for adaptive filtering in, e.g., [26–28]. Other related studies include [29], wherein the authors use a projection approach to solve a recursive LASSO-type problem, and [30], which introduced an online recursive method allowing for an underlying dynamical signal model and the use of sparsity-inducing penalties. Recursive algorithms designed for group-sparse systems have also been introduced, such as the ones presented in [31–33], but to the best of our knowledge, no such technique has so-far been applied to the multi-pitch estimation problem. This is the problem we strive to address in this paper. It should be noted that the here presented work differs from many other multi-pitch estimators in that it only exploits the assumption that the signal of interest is generated by a harmonic sinusoidal model. Recently, quite a few methods for multi-pitch estimation adhering to the machine learning paradigm have been proposed (see, e.g., [34], [35]). In these methods, a model is trained on labeled signals, such as, e.g., notes played by individual music instruments, extracting features from the training data that are then used for classification in the estimation stage. As opposed to this, the method presented here is not dependent on being trained on any dataset prior to the estimation.

Our earlier efforts on multi-pitch estimation based on sparse modeling, such as the PEBS [8] and PEBSI-Lite [9] algorithms, have focused on frame-based multi-pitch estimation techniques, with PEBS introducing the use of block sparsity to form the pitch estimates, and PEBSI-Lite refining these ideas and introducing a self-regularization technique to select the required user parameters. In this work, we build on the insights from these algorithms, and expand these ideas by introducing a method that allows for a sample-by-sample updating, in the form of an RLS-like sparse estimator, thereby allowing the estimates to also exploit information available in earlier data samples. The sub-octave problems experienced by PEBS and later alleviated by PEBSI-Lite, with the use of a total-variation penalty enforcing spectral smoothness, is here addressed using an adaptively re-weighted block penalty. Furthermore, we introduce a signal-adaptive updating scheme for the dictionary frequency atoms that allows the proposed method to, e.g., track frequency modulated signals, and alleviates grid mismatches otherwise commonly experienced by dictionary based methods.

The remainder of this paper is organized as follows; in the next section, we introduce the multi-pitch signal model and its corresponding dictionary formulation. Then, in Section 3, we introduce the group sparse RLS formulation for multi-pitch estimation, followed by a scheme for decreasing the bias of the harmonic amplitude estimates in Section 4. Section 5 presents a discussion about various algorithmic considerations. Section 6 contains numerical examples illustrating the performance of the proposed estimator on various audio signals.

Finally, Section 7 concludes upon the work.

1.1 Notation

In this work, we use lower case non-bold letters such as x to denote scalars and lower case boldface letter such as $\mathbf{x}$ to denote vectors. Upper case bold face letters such as $\mathbf{X}$ are used for matrices. We let $\operatorname{diag}(\mathbf{x})$ denote a diagonal matrix formed with the vector $\mathbf{x}$ along its diagonal. Sets are denoted using upper case calligraphic letters such as $\mathcal{A}$. If $\mathcal{A}$ and $\mathcal{B}$ are sets of integers, then $\mathbf{x}_{\mathcal{A}}$ denotes the sub-vector of $\mathbf{x}$ indexed by $\mathcal{A}$. For matrices, $\mathbf{X}_{\mathcal{A},\mathcal{B}}$ denotes the matrix constructed using the rows indexed by $\mathcal{A}$ and columns indexed by $\mathcal{B}$. We use the shorthand $\mathbf{X}_{\mathcal{A}}$ to denote $\mathbf{X}_{\mathcal{A},\mathcal{A}}$. Furthermore, $[\bar{\cdot}]$, $[\cdot]^H$, and $[\cdot]^T$ denotes complex conjugation, conjugate transpose, and transpose, respectively. Also, $|\mathcal{A}|$ is the cardinality of the set $\mathcal{A}$, and $|\mathbf{x}|$ denotes the number of elements in the vector $\mathbf{x}$, unless otherwise stated. Finally, we for vectors $\mathbf{x} \in \mathbb{C}^n$ let $||\mathbf{x}||_\ell$ denote the ℓ-norm, defined as

$$||\mathbf{x}||_\ell = \left(\sum_{j=1}^{n} |x_j|^\ell \right)^{1/\ell} \tag{1}$$

and use $i = \sqrt{-1}$.

2 Signal model

Consider a measured signal[1], $y(t)$, that is generated according to the model
$y(t) = x(t) + e(t)$, where

$$x(t) = \sum_{k=1}^{K(t)} \sum_{\ell=1}^{L_k(t)} w_{k,\ell}(t) e^{i2\pi f_k(t)\ell t} \tag{2}$$

with $K(t)$ denoting the number of pitches at time t, with fundamental frequencies $f_k(t)$, having $L_k(t)$ harmonics, $w_{k,\ell}(t)$ the complex-valued amplitude of the ℓth harmonic of the kth pitch, and where $e(t)$ denotes a broad-band additive noise. It should be stressed that the number of pitches, as well as their fundamental frequencies, and the number of harmonics for each source, may vary over time.

[1] For notational and computational simplicity, we here consider the discrete-time analytic signal of any real-valued measured signal.

It is worth noting that we here assume a harmonic signal, such as detailed in (2); however, as shown in the numerical section, the proposed method does also work well for somewhat inharmonic signals, such as, e.g., those resulting from a piano.

We here attempt to approximate the measured signal using a sparse representation in an over-complete harmonic basis, see, e.g., [36]. Specifically, as in [8], [9], the signal sources are approximated using a sparse modeling framework containing P candidate pitches, each allowed to have up to $L_{\max}$ harmonics, such that

$$x(t) \approx \sum_{p=1}^{P} \sum_{\ell=1}^{L_{\max}} w_{p,\ell}(t) e^{i2\pi f_p(t)\ell t} \tag{3}$$

where the dictionary is selected large enough so that (at least) $K(t)$ candidate pitches, $f_p(t)$, reasonably well approximate the true pitch frequencies (see also, e.g., [37], [38]), i.e., such that $P \gg \max_t K(t)$ and $L_{\max} \gg \max_{t,k} L_k(t)$. It should be noted that as the signal is assumed to contain relatively few pitches at each time instance, the resulting amplitude vector will be sparse, although with a harmonic structure reflecting the overtones of the pitches. Furthermore, it may be noted that the frequency grid-points, $f_p(t)$, are allowed to vary with time, which will here be implemented using an adaptive dictionary learning scheme. Using this framework, the pitches present in the signal at time t may be implicitly estimated by identifying the non-zero amplitude coefficients, $w_{p,\ell}(t)$.

3 Group-sparse RLS for pitches

Exploiting the structure of the signal, we introduce the group-sparse adaptive filter, $\mathbf{w}(t)$, which at time t is divided into P groups according to

$$\mathbf{w}(t) = \begin{bmatrix} \mathbf{w}_1^T(t) & \dots & \mathbf{w}_P^T(t) \end{bmatrix}^T \tag{4}$$

$$\mathbf{w}_p(t) = \begin{bmatrix} w_{p,1}(t) & \dots & w_{p,L_{\max}}(t) \end{bmatrix}^T \tag{5}$$

implying that, ideally, only $K(t)$ sub-vectors $\mathbf{w}_p(t)$ will be non-zeros at time t. In order to achieve this, the filter is formed as

$$\hat{\mathbf{w}}(t) = \arg\min_{\mathbf{w}} g_t(\mathbf{w}) + h_t(\mathbf{w}) \tag{6}$$

where $\hat{\mathbf{w}}(t)$ denotes the solution of (6), $g_t(\mathbf{w})$ the regular RLS criterion, (see, e.g., [39]), formed as

$$g_t(\mathbf{w}) = \frac{1}{2}\sum_{\tau=1}^{t} \lambda^{t-\tau} \left| y(\tau) - \mathbf{w}^T \mathbf{a}(\tau) \right|^2 \tag{7}$$

and $h_t(\mathbf{w})$ a sparsity inducing penalty function. Note that a similar adaptive filter formulation for estimating sparse data structures was introduced in [27]. However, whereas [27] considered sparse signals, we in this work expand this approach to also consider block sparsity, and specifically the pitch structure. As a result, the dictionary is here formed as

$$\mathbf{a}(t) = \left[\begin{array}{ccc} \mathbf{a}_1^T(t) & \ldots & \mathbf{a}_P^T(t) \end{array}\right]^T \tag{8}$$

$$\mathbf{a}_p(t) = \left[\begin{array}{ccc} e^{i2\pi f_p(t)t} & \ldots & e^{i2\pi f_p(t)L_{\max}t} \end{array}\right]^T \tag{9}$$

and $\lambda \in (0, 1)$ being a user-determined forgetting factor. The choice of the forgetting factor λ will reflect assumptions on the variability of the spectral content of the signal, with λ close to 1 implying an almost stationary signal, whereas a smaller value will allow for a quicker adaption to changes in the spectral content. The sparsity inducing function, $h_t(\mathbf{w})$, should be selected as to encourage a pitch-structure in the solution; in [9], which considered multi-pitch estimation on isolated time frames, this function, which then was not a function of time, was selected as

$$h(\mathbf{w}) = \gamma_1 \left\|\mathbf{w}\right\|_1 + \gamma_2 \sum_{p=1}^{P} \left\|\mathbf{F}\mathbf{w}_{\mathcal{G}_p}\right\|_1 \tag{10}$$

where $\mathbf{F}$ is the first difference matrix and $\mathcal{G}_p$ is the set of indices corresponding to the harmonics of the candidate pitch p. The second term of this penalty function is the ℓ_1-norm of the differences between consecutive harmonics and acts as a total variation penalty on the spectral envelope of each pitch. Often referred to as the sparse fused LASSO [40], this penalty was in [9] used to promote solutions with spectral smoothness in each pitch, although requiring some additional refinements to achieve this. To allow for a fast implementation, we will here instead consider the time-varying penalty function

$$h_t(\mathbf{w}) = \gamma_1(t) \left\|\mathbf{w}\right\|_1 + \sum_{p=1}^{P} \gamma_{2,p}(t) \left\|\mathbf{w}_{\mathcal{G}_p}\right\|_2 \tag{11}$$

where $\gamma_1(t)$ and $\gamma_{2,p}(t)$ are non-negative regularization parameters. This penalty, often called the sparse group LASSO [41] when combined with a squared ℓ_2-norm model fit term, is reminiscent of the one used in the PEBS method introduced in [8], and belongs to the class of methods utilizing mixed norms for sparse signal estimation (see, e.g., [42]). The second term of this penalty function, the pitch-wise ℓ_2-norm, has a group-sparsifying effect, encouraging solutions where active harmonics are grouped together into a few number of pitches. As the frequency content of different pitches may be quite similar due to overlapping, or close to overlapping, harmonics, the group penalty thus prevents erroneous activation of isolated harmonics, while still allowing the different groups to retain harmonics shared by different sources (see also [8], [9]). In the case of overlapping harmonics in the signal, i.e., the presence of two pitches which share at least one harmonic, the ℓ_2-norm will favor solutions of the optimization problem (6) in which the powers of these harmonics are shared among the two pitches. The precise level of sharing is decided by the relative powers of the unique harmonics of each pitch so that the pitch having unique harmonics with more power will also be assigned a larger share of the power corresponding to the overlapping harmonics. In the case of the the two pitches having unique harmonics with equal combined power, the power of the overlapping harmonics will also be shared equally. However, when, as in [8], using fixed penalty parameters $\gamma_1(t)$ and $\gamma_{2,p}(t)$, the resulting estimate has been shown to be prone to mistaking a pitch for its sub-octave (see also [9]). In order to discourage this type of erroneous solutions, we will herein introduce a way of adaptively choosing the group sparsity parameter, $\gamma_{2,p}(t)$, as further discussed below.

We note that $g_t(\mathbf{w})$, as defined in (7), may be expressed in matrix form as

$$g_t(\mathbf{w}) = \frac{1}{2} \left\| \mathbf{\Lambda}_{1:t}^{1/2} \mathbf{y}_{1:t} - \mathbf{\Lambda}_{1:t}^{1/2} \mathbf{A}_{1:t} \mathbf{w} \right\|_2^2 \tag{12}$$

where

$$\mathbf{y}_{\tau:t} = \begin{bmatrix} y(\tau) & \dots & y(t) \end{bmatrix}^T \tag{13}$$

$$\mathbf{A}_{\tau:t} = \begin{bmatrix} \mathbf{a}(\tau) & \dots & \mathbf{a}(t) \end{bmatrix}^T \tag{14}$$

and with $\mathbf{\Lambda}_{1:t} = \text{diag}\left(\begin{bmatrix} \lambda^{t-1} & \lambda^{t-2} & \dots & 1 \end{bmatrix}\right)$. To simplify notation, define

$$\mathbf{R}(t) \triangleq \mathbf{A}_{1:t}^H \mathbf{\Lambda}_{1:t} \mathbf{A}_{1:t} \tag{15}$$

$$\mathbf{r}(t) \triangleq \mathbf{A}_{1:t}^H \mathbf{\Lambda}_{1:t} \mathbf{y}_{1:t} \,. \tag{16}$$

With these definitions, the minimization in (6) may be formed using proximal gradient iterations, (see, e.g., [43]), such that the jth iteration may be expressed as

$$\hat{\mathbf{w}}^{(j+1)}(t) = \arg\min_{\mathbf{w}} \frac{1}{2s(t)} \left\| \boldsymbol{\nu}^{(j)} - \mathbf{w} \right\|_2^2 + h_t(\mathbf{w}) \tag{17}$$

where

$$\boldsymbol{\nu}^{(j)} = \hat{\mathbf{w}}^{(j)}(t) + s(t) \left[\mathbf{r}(t) - \mathbf{R}(t)\hat{\mathbf{w}}^{(j)}(t) \right] \tag{18}$$

with $s(t)$ denoting the step-size. We note that this update is reminiscent of the one presented in [27], which considers the problem of ℓ_1-regularized recursive least squares, although it should be noted that the ℓ_1-norm for complex vectors in [27] is defined to be the sum of the absolute values of the real and imaginary parts separately, whereas we here use the more common definition, as given by (1). In [27], the authors motivate their minimization algorithm by casting it as an EM-algorithm using reasoning from [44], as well as some further assumptions about properties of the signal. By studying the zero sub-differential equations for (17), it can be shown that the closed form solution for each group p can be computed separately as (see, e.g., equations (54)-(55) and (32)-(38) in [8]; for further details, see also [41])

$$\tilde{\boldsymbol{\nu}}_{\mathcal{G}_p}^{(j)} = S_1 \left(\boldsymbol{\nu}_{\mathcal{G}_p}^{(j)}, s(t)\gamma_1(t) \right) \tag{19}$$

$$\hat{\mathbf{w}}_{\mathcal{G}_p}^{(j+1)}(t) = S_2 \left(\tilde{\boldsymbol{\nu}}_{\mathcal{G}_p}^{(j)}, s(t)\gamma_{2,p}(t) \right) \tag{20}$$

where $S_1(\cdot)$ and $S_2(\cdot)$ are the soft thresholding operators corresponding to the ℓ_1- and ℓ_2-norms, respectively, i.e.,

$$S_1(\mathbf{z}, \alpha) = \frac{\max\left(|\mathbf{z}| - \alpha, 0 \right)}{\max\left(|\mathbf{z}| - \alpha, 0 \right) + \alpha} \odot \mathbf{z} \tag{21}$$

$$S_2(\mathbf{z}, \alpha) = \frac{\max\left(\|\mathbf{z}\|_2 - \alpha, 0 \right)}{\max\left(\|\mathbf{z}\|_2 - \alpha, 0 \right) + \alpha} \mathbf{z} \tag{22}$$

where, in (21), $|\mathbf{z}|$ denotes the vector obtained by taking the absolute value of each element of the vector $\mathbf{z}$, the max function operates element-wise on the vector $\mathbf{z}$, and $\odot$ denotes element-wise multiplication. Furthermore, as $\mathbf{R}(t)$ and $\mathbf{r}(t)$ can be expressed as

$$\mathbf{R}(t) = \sum_{\tau=1}^{t} \lambda^{t-\tau} \mathbf{a}(\tau)\mathbf{a}^H(\tau) \tag{23}$$

$$\mathbf{r}(t) = \sum_{\tau=1}^{t} \lambda^{t-\tau} y(\tau)\bar{\mathbf{a}}(\tau) \tag{24}$$

these entities can be updated according to

$$\mathbf{R}(t) = \lambda \mathbf{R}(t-1) + \mathbf{a}(t)\mathbf{a}^H(t) \tag{25}$$

$$\mathbf{r}(t) = \lambda \mathbf{r}(t-1) + y(t)\bar{\mathbf{a}}(t) \ , \tag{26}$$

when new samples become available. Here, $\bar{(\cdot)}$ denotes complex conjugation.

4 Refined amplitude estimates

In general, the sparsity promoting penalty function $h_t(\mathbf{w})$ will introduce a downward bias on the magnitude of the amplitude estimates formed by (6). However, as the support of $\hat{\mathbf{w}}(t)$ will reflect the fundamental frequencies present in the signal, we can refine the amplitude estimates by minimizing a least squares criterion. As this problem only considers amplitudes of harmonics of pitches that are believed to be in the signal, we do not need to use any sparsity inducing penalties and can therefore avoid the magnitude bias. This will be analogous to estimating the amplitudes of each harmonic using recursive least squares assuming that the support of the filter is known. To this end, let

$$\mathcal{S}(t) = \bigcup_{p \in \mathcal{A}(t)} \mathcal{G}_p \tag{27}$$

$$\mathcal{A}(t) = \left\{ p \mid \left\| \hat{\mathbf{w}}_{\mathcal{G}_p}(t) \right\|_2 > 0 \right\} \ , \tag{28}$$

i.e., $\mathcal{A}(t)$ is the set of active pitches determined by the sparse filter $\hat{\mathbf{w}}(t)$, at time t, and $\mathcal{S}(t)$ is the index set corresponding to the harmonics of these pitches. Let $\breve{\mathbf{w}}(t)$ denote the refined amplitude estimates at time t. Given $\hat{\mathbf{w}}(t)$, and thereby $\mathcal{S}(t)$, we update this filter according to

$$\breve{\mathbf{w}}_k(t) = 0 \;, k \notin \mathcal{A}(t) \tag{29}$$

$$\begin{aligned}\breve{\mathbf{w}}_{\mathcal{S}(t)}(t) = \; & \underset{\mathbf{w}\in\mathbb{C}^{|\mathcal{S}(t)|}}{\arg\min}\; \mathbf{w}^H \mathbf{R}_{\mathcal{S}(t)} \mathbf{w} - \mathbf{w}^H \mathbf{r}_{\mathcal{S}(t)} - \mathbf{r}_{\mathcal{S}(t)}^H \mathbf{w} \\ & + \xi \|\mathbf{w} - \breve{\mathbf{w}}_{\mathcal{S}(t)}(t-1)\|_2^2 \end{aligned} \tag{30}$$

where $\mathbf{R}_{\mathcal{S}(t)}(t)$ is the $|\mathcal{S}(t)| \times |\mathcal{S}(t)|$ matrix constructed by the rows and columns of $\mathbf{R}(t)$ indexed by $\mathcal{S}(t)$ and $\mathbf{r}_{\mathcal{S}(t)}(t)$ is the $|\mathcal{S}(t)|$ dimensional vector constructed by the elements of $\mathbf{r}(t)$, indexed by $\mathcal{S}(t)$. The second term of (30) is a proximal term that will promote a smooth trajectory for the magnitude of the filter coefficients, where the parameter $\xi > 0$ controls the smoothness. This type of smoothness-promoting penalty has earlier been used, for instance, to enforce temporal continuity in NMF applications [45]. To avoid inverting large matrices, we split the solving of (30) into $\mathcal{A}(t)$ problems of size $L_{\max}$ using a cyclic coordinate descent scheme (see also, e.g., [26]). To this end, define the index sets

$$\mathcal{Q}_p = \mathcal{S}(t) \setminus \mathcal{G}_p \;, p \in \mathcal{A}(t) \;, \tag{31}$$

i.e., the indices corresponding to harmonics that are not part of pitch p. Considering only terms in the cost function in (30) that depend on harmonics of the pth pitch, we can form an update of the corresponding filter coefficients according to

$$\begin{aligned}\breve{\mathbf{w}}_{\mathcal{G}_p}(t) = & \underset{\mathbf{w}\in\mathbb{C}^{L_{\max}}}{\arg\min}\; \mathbf{w}^H \mathbf{R}_{\mathcal{G}_p} \mathbf{w} - \mathbf{w}^H \mathbf{r}^{(p)} - \mathbf{r}^{(p)H} \mathbf{w} \\ & + \xi \left\|\mathbf{w} - \breve{\mathbf{w}}_{\mathcal{G}_p}(t-1)\right\|_2^2 \end{aligned} \tag{32}$$

where

$$\mathbf{r}^{(p)} = \mathbf{r}_{\mathcal{G}_p} - \mathbf{R}_{\mathcal{G}_p,\mathcal{Q}_p} \tilde{\mathbf{w}}_{\mathcal{Q}_p} \;. \tag{33}$$

The vector $\tilde{\mathbf{w}}_{\mathcal{Q}_p} \in \mathbb{C}^{|\mathcal{Q}_p|}$ contains the (partially updated) filter coefficients that correspond to other pitches than p, i.e.,

$$\tilde{\mathbf{w}}_{\mathcal{G}_q} = \begin{cases} \breve{\mathbf{w}}_{\mathcal{G}_q}(t) & \text{if updated} \\ \breve{\mathbf{w}}_{\mathcal{G}_q}(t-1) & \text{if not updated} \end{cases} \tag{34}$$

for $q \neq p$. By setting the gradient of (32) with respect to $\mathbf{w}$ to zero, we find the update of $\breve{\mathbf{w}}_{\mathcal{G}_p}(t)$ to be

$$\breve{\mathbf{w}}_{\mathcal{G}_p}(t) = \left(\mathbf{R}_{\mathcal{G}_p} + \xi \mathbf{I}\right)^{-1} \left(\mathbf{r}^{(p)} + \xi \breve{\mathbf{w}}_{\mathcal{G}_p}(t-1)\right) \;. \tag{35}$$

5 Algorithmic considerations

We proceed to examine some implementation aspects of the presented algorithm, first discussing the appropriate choice of the penalty parameters, then possible computational speed-ups, as well as ways of adaptively updating the used pitch dictionary.

5.1 Parameter choices

In order to discourage solutions containing erroneous sub-octaves, we here propose to update the group penalty parameter, in iteration j of the filter update (17), as

$$\gamma_{2,p}(t) = \gamma_2(t) \max \left(1, \frac{1}{\left| \hat{w}_{p,1}^{j-1}(t) \right| + \varepsilon} \right) \tag{36}$$

where $\left| \hat{w}_{p,1}^{j-1}(t) \right|$ is the estimated amplitude of the first harmonic of group p, obtained in iteration $j-1$, with $\varepsilon \ll 1$ being a user-specified parameter selected to avoid a division by zero. In this paper, we use $\varepsilon = 10^{-5}$. As sub-octaves will typically have missing first harmonics, such a choice will encourage shifting power from the sub-octave to the proper pitch. Similar types of re-weighted penalties have earlier been used to enhance sparsity in the estimated signal (see, e.g., [46], [47]). Studies using many different kinds of pitch signals indicate that the overall performance of the algorithm is relatively insensitive to the choice of the parameter $s(t)$, which may typically be selected in the range $s(t) \in \left[10^{-5}, 10^{-3}\right]$. Here, we use $s(t) = 10^{-4}$. The choice of the penalty parameters $\gamma_1(t)$ and $\gamma_2(t)$ can be made using inner-products between the dictionary and the signal. Letting Δ denote the time-lag, define

$$\eta(t, \mu) = \mu \left\| \mathbf{\Lambda}_{1:\Delta} \mathbf{A}_{t-\Delta:t}^{H} \mathbf{y}_{t-\Delta:t} \right\|_{\infty} \tag{37}$$

where $\mu \in (0, 1)$. A good rule of thumb is choosing $\gamma_1(t)$ in the neighborhood of (37) with $\mu = 0.1$, whereas a corresponding reasonable value for $\gamma_2(t)$ is $\mu = 1$. Empirically, the performance of the algorithm has been seen to be robust to variations of these choices of μ. This method emulates choosing the values of the penalty parameters based on the correlation between the signal and the dictionary in a finite window. Here, the window length, Δ, is determined by the forgetting factor, λ, and by how much correlation one is willing to lose as a result from the truncation. For example, selecting

$$\Delta = \frac{\log(0.01)}{\log \lambda} \tag{38}$$

will yield a window such that the excluded samples will contribute to less than 0.01 of the correlation. It should be noted that for smoothly varying signals, $\gamma_1(t)$ and $\gamma_2(t)$ only need to be updated infrequently.

5.2 Iteration speed-up

As the signal is assumed to have a sparse representation in the dictionary $\mathbf{a}(t)$, one may expect updates of the coefficients of many groups, here indexed by q, to result in zero amplitude estimates. As such groups do not contribute to the pitch estimates, these groups would preferably be excluded from the updates in (17)-(18). If assuming the support of $\mathbf{w}(t)$ to be constant for all t, one could thus sequentially discard such groups from the updating step, and thereby decrease computation time. However, as generally pitches may disappear and then reappear, as well as drift in frequency over time, we will here only exclude the groups q from the updating steps temporarily. That is, if at time τ, we have $\left\|\hat{\mathbf{w}}_{\mathcal{G}_q}\right\|_2 < \tilde{\varepsilon}$, where $\tilde{\varepsilon} \ll 1$, the group q is considered not to be present in the signal and is therefore excluded from the updating steps for a waiting period, T. After that period, it is again included in the updates, allowing it to again appear in the signal. Defining the set $\mathcal{U}$, indexing the groups that are considered active, the group q is adaptively included and excluded from $\mathcal{U}$ depending on the size of $\left\|\hat{\mathbf{w}}_{\mathcal{G}_q}\right\|_2$. If the signal can be assumed to have slowly varying spectral content, meaning that the support of $\mathbf{w}(t)$ is also varying slowly, the waiting period T may be chosen to be quite long, as to improve the computational efficiency. In general, choosing T as to correspond to a few milliseconds allows for a speed-up of the algorithm while at the same time enabling it to track the time evolution of $\mathbf{w}(t)$.

5.3 Dictionary learning

In general, a signal's pitch frequencies may vary over time, for instance, due to vibrato. Applying the filter updating scheme using fixed grid-points will therefore result in rapidly changing support of the filter or energy leakage between adjacent blocks of the filter, here indexed by p. In order to overcome this problem, and to

Algorithm 1 The PEARLS algorithm

```
Initialise $\hat{\mathbf{w}}(0) \leftarrow \mathbf{0}$, $\mathbf{R}(0) \leftarrow \mathbf{0}$ , $\mathbf{r}(0) \leftarrow \mathbf{0}$
$t \leftarrow 1$
repeat {Recursive update scheme}
    $\mathbf{R}(t) \leftarrow \lambda\mathbf{R}(t-1) + \mathbf{a}(t)\mathbf{a}^H(t)$
    $\mathbf{r}(t) \leftarrow \lambda\mathbf{r}(t-1) + y(t)\bar{\mathbf{a}}(t)$
    $j \leftarrow 0$
    $\hat{\mathbf{w}}^{(j)}(t) \leftarrow \hat{\mathbf{w}}(t-1)$
    repeat {Proximal gradient update}
        $\boldsymbol{\nu}^{(j)} \leftarrow \hat{\mathbf{w}}^{(j)}(t) + s(t)\left[\mathbf{r}(t) - \mathbf{R}(t)\hat{\mathbf{w}}^{(j)}(t)\right]$
        $\hat{\mathbf{w}}^{(j+1)}(t) \leftarrow \arg\min_{\mathbf{w}} \frac{1}{2s(t)} \left\|\boldsymbol{\nu}^{(j)} - \mathbf{w}\right\|_2^2 + h_t(\mathbf{w})$
        $j \leftarrow j + 1$
    until convergence
    $\hat{\mathbf{w}}(t) \leftarrow \hat{\mathbf{w}}^{(j)}(t)$
    Determine $\mathcal{A}(t)$ and $\mathcal{S}(t)$
    $\breve{\mathbf{w}}_k(t) \leftarrow 0$ , $k \notin \mathcal{A}(t)$
    $\breve{\mathbf{w}}_{\mathcal{S}(t)}(t) = \arg\min_{\mathbf{w} \in \mathbb{C}^{|\mathcal{S}(t)|}} \mathbf{w}^H\mathbf{R}_{\mathcal{S}(t)}\mathbf{w} - \mathbf{w}^H\mathbf{r}_{\mathcal{S}(t)} - \mathbf{r}_{\mathcal{S}(t)}^H\mathbf{w}$
                $+\xi \left\|\mathbf{w} - \breve{\mathbf{w}}_{\mathcal{S}(t)}(t-1)\right\|_2^2$
    Update active set $\mathcal{U}$
    if $t \in \mathcal{T}$ then
        Update dictionary
    end if
    $t \leftarrow t + 1$
until end of signal
```

allow for smooth tracking of pitches over time, we propose a scheme for adaptively updating the dictionary of candidate pitches. This adaptive adjustment scheme also allows for the use of a grid with coarser resolution than would otherwise be possible. Let $\mathcal{T} = \{\tau_k\}_k$ be the set of time points in which the dictionary is updated. As only groups $\hat{\mathbf{w}}_{\mathcal{G}_p}(\tau_k)$ with non-zero power are considered to be present in the signal, one only has to adjust the fundamental frequencies of these. Assuming that the current estimate of such a candidate pitch frequency is $f_p(\tau_{k-1})$, one only needs to consider adjusting it on the interval $f_p(\tau_{k-1}) \pm \frac{1}{2}\delta_{f,k}(t)$, where

$\delta_{f,k}(t)$ denotes the current grid-point spacing. The update can be formed using the approximate non-linear least squares method in [48], [2], where, instead of $L_{\max}$, one uses the harmonic order corresponding to the non-zero components of $\hat{\mathbf{w}}_{\mathcal{G}_p}(\tau_k)$. This refined estimate is obtained by first forming the residual, and adding back the current group of harmonics, whereafter the approximate non-linear least squares method is applied to update the frequencies. The adjusted frequency $f_p(\tau_k)$ is then used to update the dictionary on the time interval $\left[\tau_k, \tau_{k+1}\right)$. After updating the dictionary, the filter coefficient estimates will, due to the recursive nature of the method, be partly based on the old dictionary and partly on the updated one. It is thus very likely that after the dictionary update the phase component of the two filter coefficient parts will differ. To avoid this, we instead incorporate the phase into the dictionary, thus obtaining a filter coefficient with zero phase. This is accomplished by estimating the phases at the same time as the frequencies are updated in the dictionary updating step. Each estimated phase is then multiplied with the corresponding column of the dictionary, thus including the phases into the dictionary. This update corresponds to changing (8) and (9) to

$$\mathbf{a}(t, \boldsymbol{\varphi}) = \left[\begin{array}{ccc} \mathbf{a}_1^T(t, \boldsymbol{\varphi}_1) & \ldots & \mathbf{a}_P^T(t, \boldsymbol{\varphi}_P) \end{array} \right]^T \tag{39}$$

$$\mathbf{a}_p(t, \boldsymbol{\varphi}_p) = \left[\begin{array}{ccc} \mathrm{e}^{i2\pi f_p(t)t + i\pi\varphi_{p_1}} & \ldots & \mathrm{e}^{i2\pi f_p(t)L_{\max}t + i\pi\varphi_{p_{L_{\max}}}} \end{array} \right]^T \tag{40}$$

where

$$\boldsymbol{\varphi} = \left[\begin{array}{ccc} \boldsymbol{\varphi}_1^T & \cdots & \boldsymbol{\varphi}_P^T \end{array} \right]^T \tag{41}$$

$$\boldsymbol{\varphi}_p = \left[\begin{array}{ccc} \varphi_{p_1}^T & \cdots & \boldsymbol{\varphi}_{p_{L_{\max}}} \end{array} \right]^T \tag{42}$$

with φ_{p_ℓ} denoting the phase of the ℓth harmonic of the pth pitch. With this formulation the phases are incorporated into the dictionary, thus rendering the amplitudes real valued.

Together with the discussed algorithmic considerations, the presented time-recursive multi-pitch estimator is detailed in Algorithm 1. The algorithm is termed the Pitch Estimation using dictionary-Adaptive Recursive Least Squares (PEARLS) method[2].

[2] An implementation in MATLAB may be found at `http://www.maths.lu.se/staff/andreas-jakobsson/publications/`.

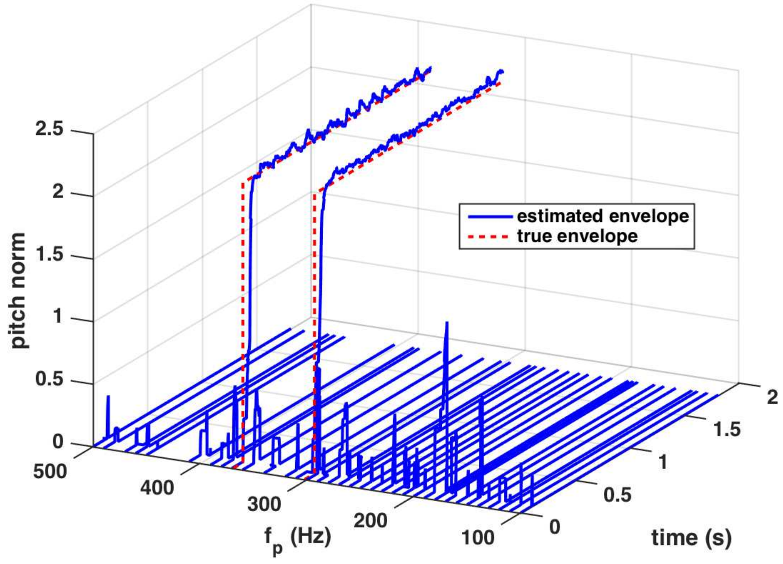

Figure 1: Pitch frequency and pitch norm estimates, i.e., estimates of $f_p(t)$ and $\left\|\breve{\mathbf{w}}_{\mathcal{G}_p}(t)\right\|_2$ as produced by PEARLS when applied to a simulated two-pitch signal with fundamental frequencies 302 and 369 Hz, respectively, deviating from the original dictionary grid points by 2 and 1 Hz respectively.

6 Numerical results

In this section, we evaluate the performance of the proposed PEARLS algorithm using both simulated signals and real audio recordings.

6.1 Simulated signals

To demonstrate the effect of the smoothing parameter, ξ, as well as the ability of PEARLS to smoothly track the amplitudes of pitches, we first consider an illustrative example with a two-pitch signal. Figure 1 shows the time evolution of the pitch frequency and pitch norm estimates, i.e., estimates of $f_p(t)$ and $\left\|\breve{\mathbf{w}}_{\mathcal{G}_p}(t)\right\|_2$, as produced by PEARLS when applied to a two-pitch signal with fundamental frequencies 302 and 369 Hz, respectively, where both pitches are constituted

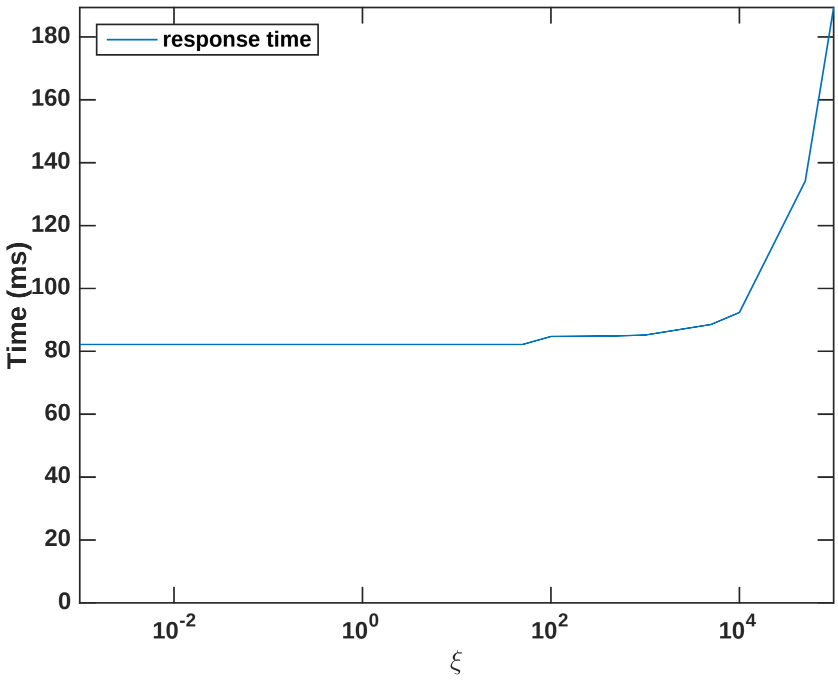

Figure 2: Respone time for different values of the smoothing parameter ξ.

by 5 harmonics each. Both pitches enter the signal after 90 ms, reaching their maximum amplitudes momentarily and keeping them for the rest of the signal duration. The signal was sampled at 11 kHz. The settings for PEARLS was $L_{\max} = 10$, $\lambda = 0.995$, and the smoothing parameter was $\xi = 10^4$. The original pitch frequency grid was chosen so that the true pitch frequencies deviated from the closest grid points by 2 and 1 Hz, respectively. As can be seen from the figure, the estimate initially, before the pitch signals appear, contains several spurious pitch estimates, but then quickly finds the pitch signals when these appear in the data. At this point, the spurious peaks are suppressed and the estimates are seen to well follow the true pitch envelopes. It is worth noting that both the response time and the steady state variance of the estimates will be influenced by the choice of the smoothing parameter, ξ. Figures 2 and 3 illustrate this effect by considering the response time, defined as the time required for the PEARLS amplitude estimate to reach 95% of its peak value, and the steady state amplitude variance,

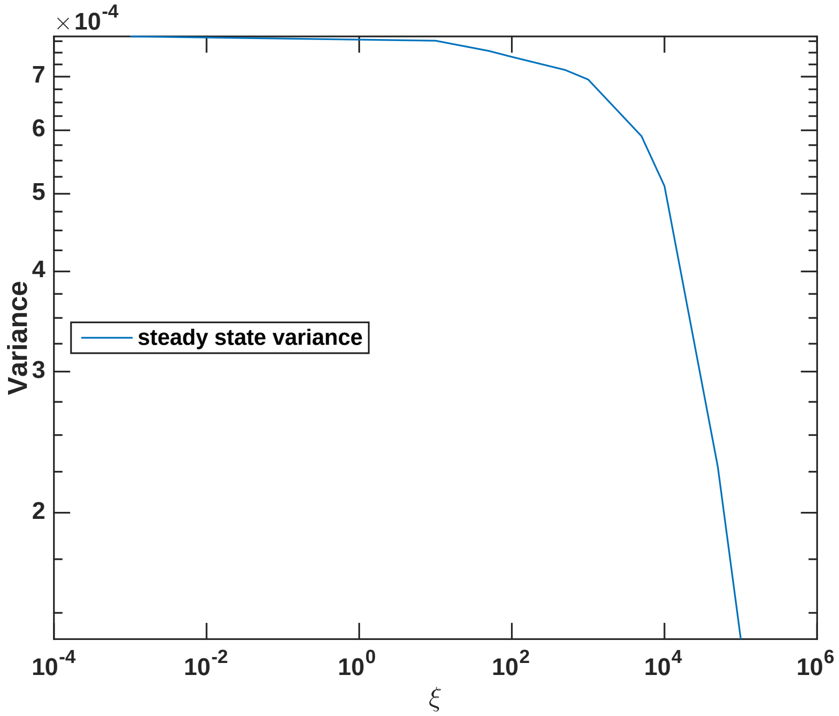

Figure 3: Steady state variance of the pitch norm estimate for different values of the smoothing parameter ξ.

respectively. The signal considered is the same as in Figure 1. As can be seen from the figures, a higher value of ξ implies a longer response time for PEARLS, while at the same time promoting a more smooth pitch norm trajectory, just as could be expected.

The PEARLS algorithm is not restricted to form estimates of stationary pitches; it is also able to cope with amplitude and frequency modulated signals. In Figure 4, PEARLS has been applied to a two-pitch signal with fundamental frequencies that oscillate according to sine waves with frequencies 2 and 3 Hz on the intervals 327 ± 2 Hz and 394 ± 3 Hz, respectively. Also, the pitch norms are not constant, but are amplitude modulated according to a Hamming window. As can be seen, PEARLS is able to track the two pitches smoothly both in frequency and in pitch norm. Here, the pitches consisted of 5 and 7 harmonics, respect-

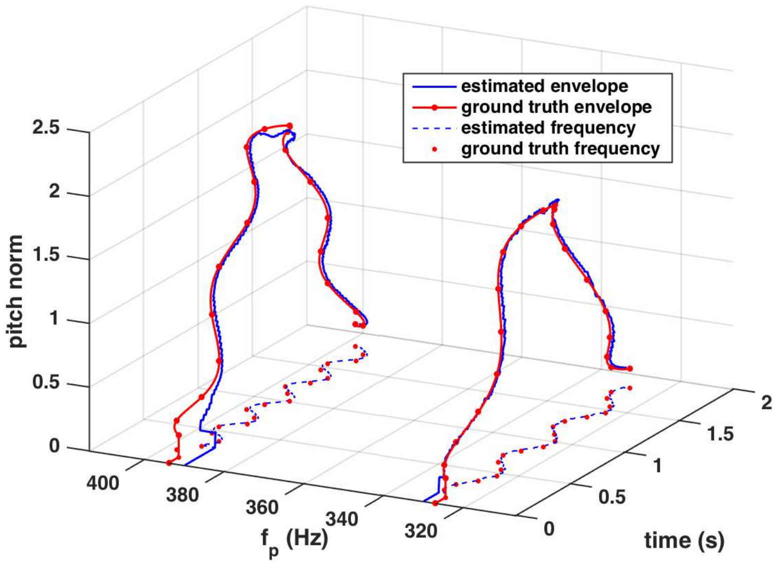

Figure 4: Pitch frequency and pitch norm estimates, i.e., estimates of $f_p(t)$ and $\left\|\breve{\mathbf{w}}_{\mathcal{G}_p}(t)\right\|_2$, as produced by PEARLS when applied to a simulated two-pitch signal with fundamental frequencies that oscillate according to sine waves.

ively. The signal was sampled at 11 kHz, with PEARLS using the same settings as above. As comparison, Figure 5 presents a corresponding plot for the multi-pitch estimator ESACF [7], using recommended settings. As ESACF only estimates pitch frequencies, pitch norm estimates have been obtained using least squares, assuming known harmonic orders. ESACF is a frame based estimator and the signal was therefore here subdivided into 30ms windows. As can be seen, the ESACF estimates deviate from the true pitch frequencies, causing the amplitude estimates to degrade. Figure 6 demonstrates the usefulness of using the dictionary learning procedure. In this figure, PEARLS is again applied to the signal with two frequency modulated pitches, but this time the dictionary learning scheme is excluded from Algorithm 1. As can be seen in the figure, PEARLS is still able to estimate the frequency content, as well as the pitch norms, but the tracking is now performed by different elements of $\breve{\mathbf{w}}(t)$, as the frequency modulation causes the different candidate pitches to become activated and then deactivated, with the

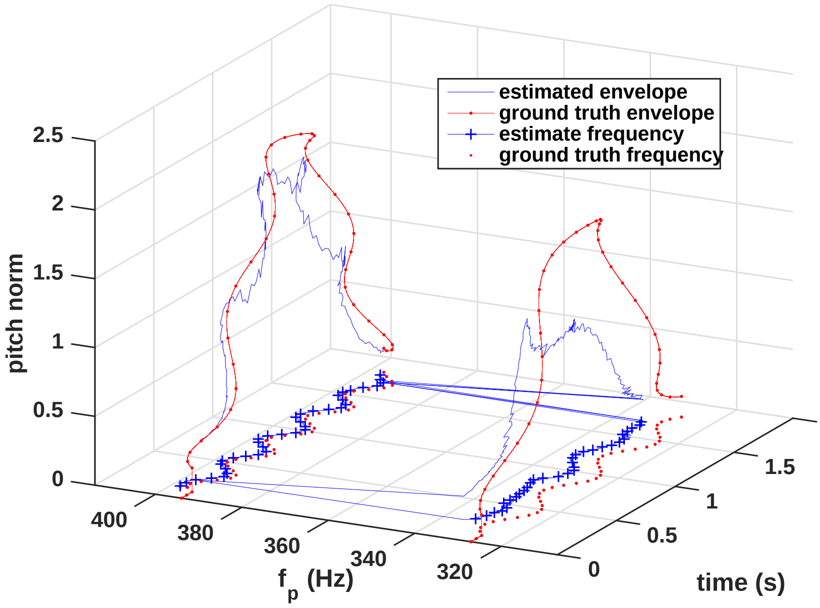

Figure 5: Pitch frequency, i.e., estimates of $f_p(t)$, as produced by ESACF when applied to a simulated two-pitch signal with fundamental frequencies that oscillate according to sine waves. The pitch norms, i.e., $\left\|\breve{\mathbf{w}}_{\mathcal{G}_p}(t)\right\|_2$, have been estimated by applying least squares to the ESACF pitch frequency estimates using oracle harmonic orders.

activation-deactivation cycles following the periods of the frequency modulation. Also, there is some power-sharing between adjacent pitch groups of $\breve{\mathbf{w}}(t)$ at time points where the frequency modulating sinusoids change sign. In contrast, the dictionary learning scheme allows for a much smoother tracking as the movable dictionary elements counters the activation-deactivation phenomenon, which can be observed in Figure 4.

6.2 Real audio

We proceed to evaluate the performance of PEARLS on the Bach10 dataset [49]. This dataset consists of ten excerpts from chorals composed by J. S. Bach, and have been arranged to be performed by an ensemble consisting of a violin, a clarinet, a saxophone, and a bassoon, with each excerpt being 25-42 seconds long.

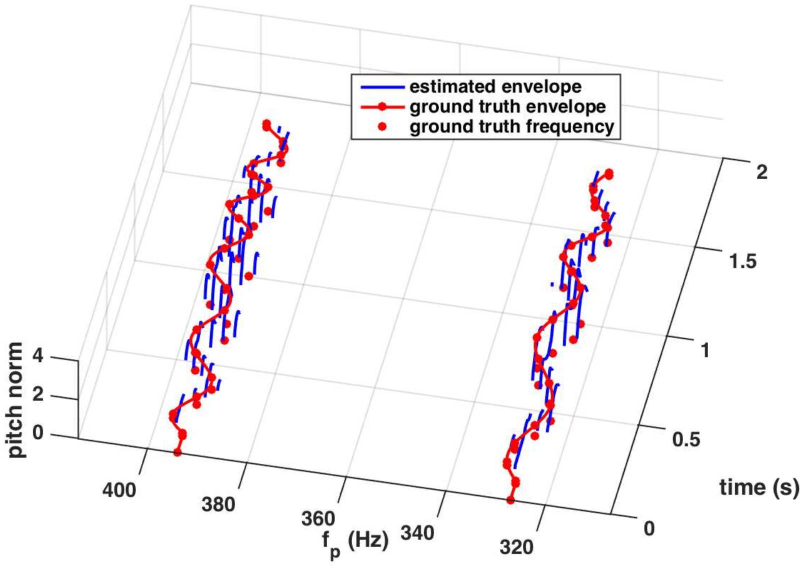

Figure 6: Pitch frequency and pitch norm estimates, i.e., estimates of $f_p(t)$ and $\left\|\check{\mathbf{w}}_{\mathcal{G}_p}(t)\right\|_2$, as produced by PEARLS when applied to a simulated two-pitch signal with fundamental frequencies that oscillate according to sine waves. Here, the dictionary learning scheme is excluded from Algorithm 1.

The algorithm settings for PEARLS were $\lambda = 0.985$, $\xi = 10^3$, $L_{\max} = 6$, and the dictionary was updated every 10 ms using 45 ms of past signal samples. Each music piece, originally sampled at 44.1 kHz, was down-sampled to 11.025 kHz. The PEARLS estimates were compared to ground truth values with a time-resolution of one reference point every 30 ms. The ground truth fundamental frequencies were obtained by applying the single-pitch estimator YIN [50] to each separate channel with manual correction of obvious errors. The results are presented in Table 1, presenting values of the performance measures *Accuracy*, *Precision*, and *Recall*, as defined in [51]. As in [51], an estimated fundamental frequency is associated with a ground truth fundamental frequency if it lies within a quarter-tone, or 3%, of the ground truth fundamental frequency. For comparison, Table 1 also includes corresponding performance measures for the PEBSI-Lite [9] and ESACF algorithms. The values for PEBSI-Lite and ESACF were originally presen-

	PEARLS	PEBSI-Lite	BW15	ESACF
Accuracy	0.437	0.449	0.515	0.269
Precision	0.683	0.631	0.684	0.471
Recall	0.548	0.609	0.675	0.386

Table 1: Performance measures for the PEARLS, PEBSI-Lite, BW15, and ESACF algorithms, when evaluated on the Bach10 dataset.

ted in [9], and the settings for these algorithms are the same as is presented there. Also presented in Table 1 are performance measures obtained when applying the method presented in [35], hereafter referred to as BW15, after the authors and year of publication, to the same dataset. Being trained on databases of music instrument, this method uses probabilistic latent component analysis to produce pitch estimates and is specifically tailored to estimate pitches in music signals. The frequency resolution of the obtained estimates corresponds to that of the Western chromatic scale, i.e., to the keys of the piano.

As can be seen, PEARLS clearly outperforms ESACF and performs on par with PEBSI-Lite when considering these measures, although it should be stressed that PEARLS has significantly lower computational complexity than PEBSI-Lite. The BW15 methods performs better than the other presented methods, including PEARLS, for this dataset. This is as the performance of the BW15 estimate was formed when using an *a posteriori* thresholding of the obtained estimate, optimally selecting the threshold level as to maximize the performance measures; this in order to illustrate the best possible performance achievable for BW15. However, several other choices of possible threshold levels resulted in BW15 performing worse than both PEARLS and PEBSI-Lite. Furthermore, the BW15 estimator is sensitive to mismatches between the examined signal and the training dataset used to construct its priors. This is illustrated by applying the BW15 and PEARLS estimators to a signal consisting of two (harmonic) trumpet notes and two (inharmonic) piano notes. The trumpets are playing the notes A4 and D♭5, corresponding to the fundamental frequencies 440 and 554.37 Hz, whereas the pianos are playing the notes E4 and G♯4, corresponding to the fundamental frequencies 329.65 and 415.3 Hz. The signal was sampled at 11.025 kHz. The ground truth pitches can be seen in Figure 7. Here, the amplitude, i.e., the pitch norm, of each pitch is illustrated by the color of each track. The amplitude has been normalized

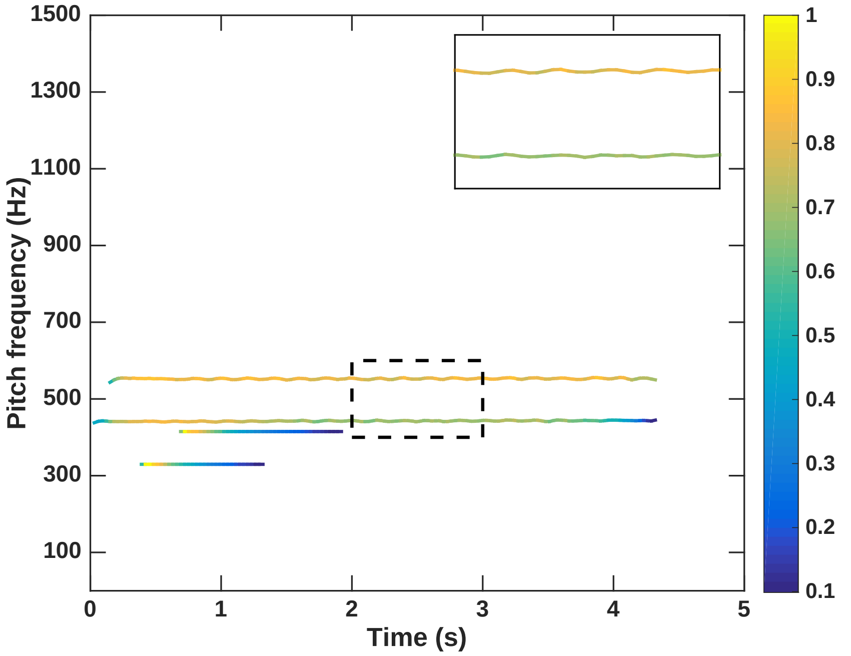

Figure 7: Ground truth for a signal consisting of two trumpets and two pianos. The amplitude of each pitch, i.e., the pitch norm, is illustrated by the color of each track. The amplitudes have been normalized so that the maximal amplitude is 1.

so that the maximum amplitude is equal to one. The corresponding estimates produced by PEARLS (using the same settings as for the Bach10 dataset) and BW15 are presented in Figures 8 and 9, respectively.

As can be seen from Figure 8, PEARLS is able to correctly identify both the trumpet and the piano pitches, despite the pianos being inharmonic and thereby differing from the assumed signal model, as given in (2). Note that PEARLS is also able to smoothly track the frequency modulation caused by that trumpets are playing with vibrato, which can be more clearly seen from the zoomed-in portions of Figures 7 and 8. In contrast, as seen in Figure 9, BW15 is able to correctly identify the piano pitches (note that pianos were included in the training dataset used by the authors of [35]), but instead of identifying the sinusoidal content corresponding to the trumpets (which are not in the training dataset) as

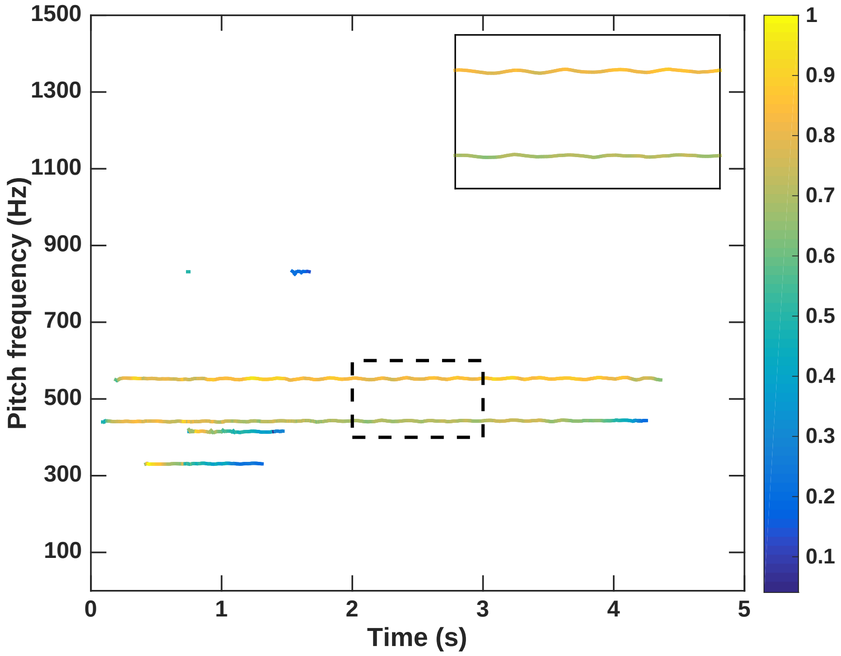

Figure 8: Estimates produced by PEARLS when applied to a signal with two trumpets as well as two pianos. The amplitude of each pitch, i.e., the pitch norm, is illustrated by the color of each track. The amplitudes have been normalized so that the maximal amplitude is 1.

originating from only two pitches, several of the individual harmonics are instead being assigned individual pitches.

It may be noted that the method does not accurately represent the vibratos; this as the estimates of BW15 are restricted to correspond to the keys of the piano. It should further be noted that the pitches indicated as being the most significant by BW15 are not those corresponding to the true fundamental frequencies, but instead higher order harmonics. This problem is arguably due to the mismatch between the content of the signal and the database used to train the method. Thus, for this example, it is not possible to recover the true pitches by thresholding the solution of BW15, as the thresholding would eliminate true pitch candidates before getting rid of the erroneous ones. Although the estimates produced by BW15 could arguably be improved by extending its training data to also include

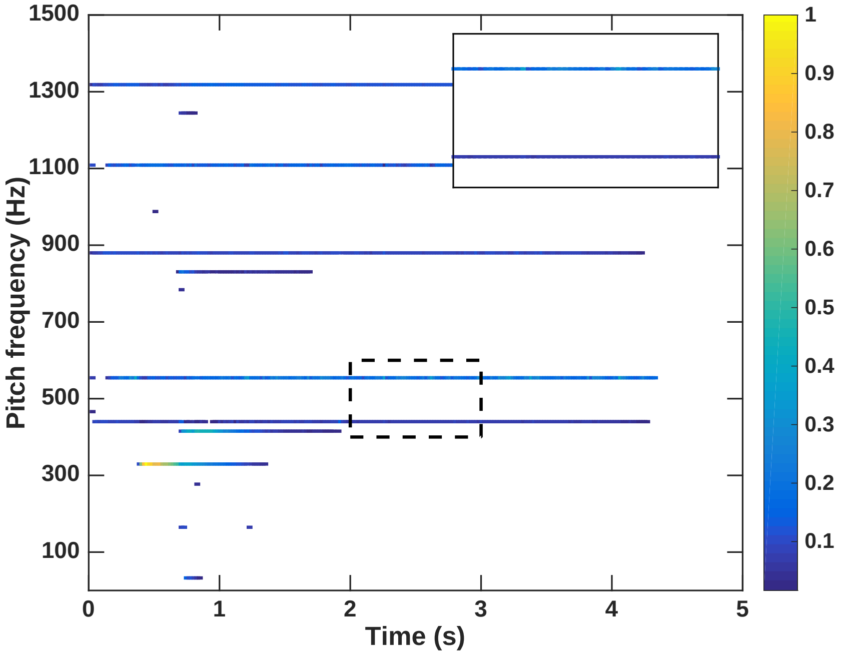

Figure 9: Estimates produced by BW15 when applied to a signal with two trumpets as well as two pianos. The magnitudes of the estimates are illustrated by the color of the pitch tracks. The magnitudes have been normalised so that the maximal magnitude is 1.

trumpets, this example illustrates that basing estimation on exploiting the features of a signal model, as PEARLS does, can be beneficial in terms of the generality of the estimator, even in the face of slight deviations from the assumed signal model, which in this case takes the form of inharmonicity for the pianos. It can be noted that an interesting future development would be to combine the benefits from training a hidden Markov model, as is done in BW15, with the more robust approach in PEARLS.

Another recent method that would be of interest to consider in this respect would be the one presented in [21], which also exhibits some conceptual similarities with the herein presented algorithm. Notably, the sparsifying role played by the ℓ_1-norm herein is in [21] formed by instead determining the significant spectral peaks using an estimate of the noise floor. The pitch selection, herein formed

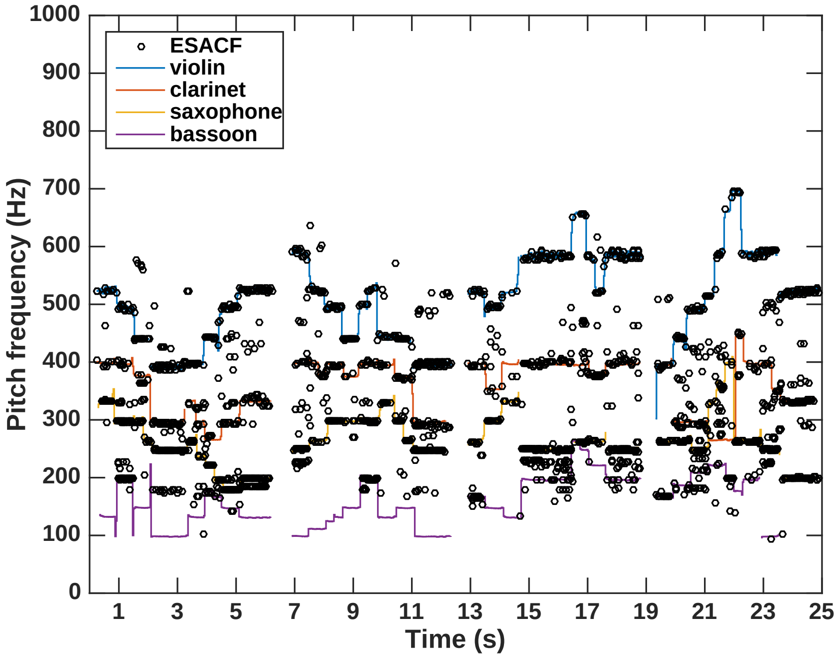

Figure 10: Pitch tracks produced by ESACF when applied to a 25 seconds excerpt of J. S. Bach's *Ach, Gott und Herr* performed by a violin, a clarinet, a saxophone, and a bassoon.

using the group-wise ℓ_2-norm, is in [21] made by matching spectral content with that of components in a large training data set, which is also used to measure the power concentration for low-order harmonics, as well as a synchronicity measure. The relative weighting of these components is selected using training data. Using a greedy approach, the method in [21] then iteratively adds candidate pitches to the estimate; the power allocation between pitches that have overlapping harmonics is resolved using an interpolation scheme utilizing the power of harmonics unique to each candidate pitch. In contrast, the number of active pitches is herein decided by the optimal point of (6), where candidate pitches not contained in the signal should be assigned zero power. It can also be noted that the optimization problem presented here does not favor spectral smoothness; rather, the ℓ_2-norm will favor collecting as much power as possible into a few candidate pitches. The power of overlapping harmonics will therefore tend to be allocated to pitches with more

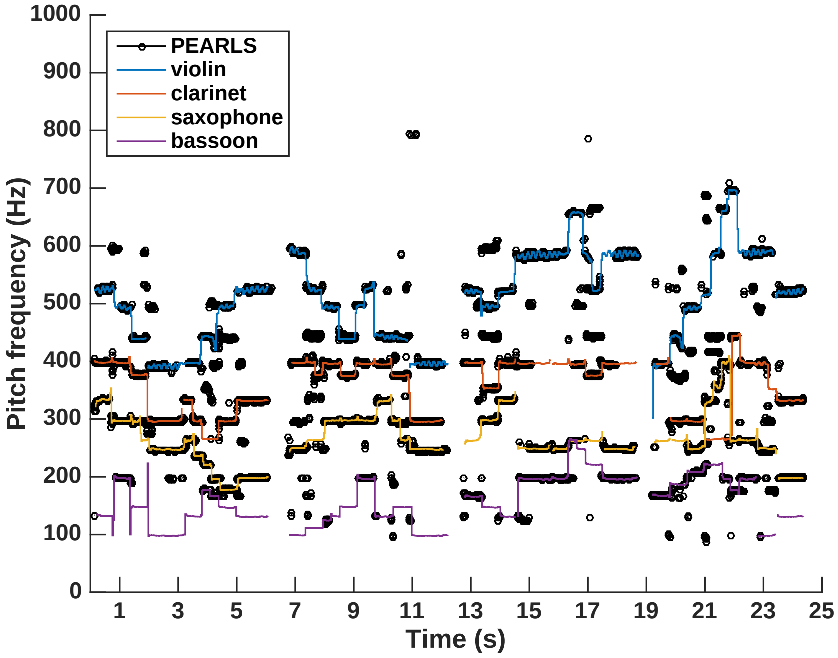

Figure 11: Pitch tracks produced by PEARLS when applied to a 25 seconds excerpt of J. S. Bach's *Ach, Gott und Herr* performed by a violin, a clarinet, a saxophone, and a bassoon.

prominent unique harmonics. Using a MATLAB implementation of PEARLS on a 2.68 GHz PC, the average running time for the Bach pieces was 20 minutes. The Bach pieces were on average 33 seconds long[3]. For PEBSI-Lite, the average running time was 54 minutes, with the signal being divided into non-overlapping frames of length 30 ms.

As an illustration of the performance of PEARLS on the Bach10 dataset, Figures 10 and 11 present the estimated fundamental frequencies obtained using ESACF and PEARLS, respectively, for the piece *Ach, Gott und Herr*, as compared to the ground truth for each instrument. Here, in order to make a fair comparison of the computational complexities of the estimators, the ESACF estimate was com-

[3] We note that the current implementation has not exploited that the filter updating step (17) can be done for all P candidate pitches in parallel. Similarly, the computations for PEBSI-Lite can also be parallelized, as each time frame can be processed in isolation.

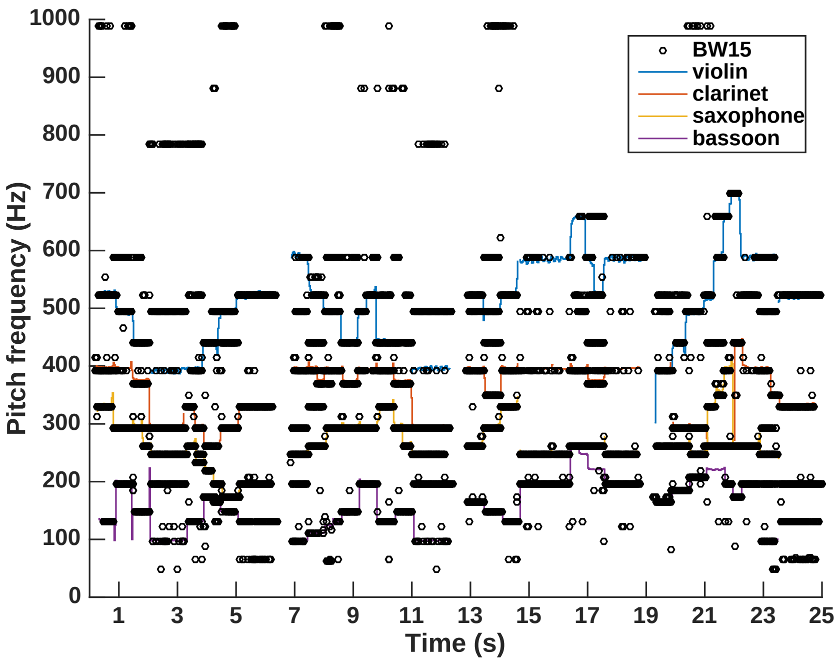

Figure 12: Pitch tracks produced by BW15 when applied to a 25 seconds excerpt of J. S. Bach's *Ach, Gott und Herr* performed by a violin, a clarinet, a saxophone, and a bassoon.

puted on windows of length 30 ms, where two consecutive windows overlapped in all but one sample. Although ESACF can arguably be applied to windows with smaller overlap, this setup meant that ESACF would produce pitch tracks with the same time resolution as PEARLS. This resulted in an average running time of 11 minutes per music piece, that is, about half that of PEARLS. As can be seen from the figures, PEARLS is considerably better at tracking the instruments than ESACF. In Figure 12, the corresponding results for BW15 are shown. The figure has been truncated at 1000 Hz to simplify inspection, although pitch estimates with fundamental frequencies higher than 1000 Hz did occur repeatedly. From the figure, it is clear that BW15 is better able to track the bassoon (which is included in the method's training data) than either PEARLS or ESACF. It can also be noted that the discrete nature of the BW15 estimator prevents it from tracking smaller frequency variations, such as vibratos.

7 Conclusions

In this work, we have presented a time-recursive multi-pitch estimation algorithm, based on a both sparse and group-sparse reconstruction technique. The method has been shown to be able to accurately track multiple pitches over time, in fundamental frequency as well as in amplitude, without requiring prior knowledge of the number of pitches nor the number of harmonics present in the signal. Furthermore, we have presented a scheme for adaptively changing the signal dictionary, thereby providing robustness against grid mismatch, as well as allowing for smooth tracking of frequency modulated signals. We have shown that the proposed method yields accurate results when applied to real data, outperforming other general purpose multi-pitch estimators in either estimation accuracy and/or computational speed. The method has further been shown to be robust to deviations from the assumed signal model, although it is not able to yield performance as good as that achievable by a state-of-the art method being optimally tuned and specifically trained on the present instruments. However, the method is able to outperform such a technique when used without optimal tuning, or when applied to instruments not included in the training data.

References

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

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

[3] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*, Springer International Publishing, 2015.

[4] R. B. Randall, *Vibration-Based Condition Monitoring: Industrial, Aerospace and Automotive Applications*, John Wiley & Sons, Chichester, UK, 2011.

[5] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1015–102, Apr. 2009.

[6] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric Representations of Bird Sounds for Automatic Species Recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2252–2263, Nov. 2006.

[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, Apr. 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, Oct. 2016.

[10] M. D. Plumbley, S. A. Abdallah, T. Blumensath, and M. E. Davies, “Sparse representations of polyphonic music,” *Signal Processing*, vol. 86, no. 3, pp. 417–431, Mar. 2006.

[11] M. Bay, A.F. Ehmann, J.W. Beauchamp, P. Smaragdis, and J.S. Downie, “Second Fiddle is Important Too: Pitch Tracking Individual Voices in Polyphonic music,” in *13th Annual Conference of the International Speech Communication Association*, Portland, Sept. 2012, pp. 319–324.

[12] A. Dessein, A. Cont, and G. Lemaitre, “Real-Time Polyphonic Music Transcription With Non-Negative Matrix Factorisation and Beta-Divergence,” in *Proceedings of the 11th International Society for Music Information Retrieval Conference*, Utrecht, NL, Aug. 2010, pp. 489–494.

[13] V. Emiya, R. Badeau, and B. David, “Multipitch Estimation of Piano Sounds Using a New Probabilistic Spectral Smoothness Principle,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 18, no. 6, pp. 1643–1654, Aug. 2010.

[14] C. Kim, W. Chang, S-H. Oh, and S-Y. Lee, “Joint Estimation of Multiple Notes and Inharmoncity Coefficient Based on f0-Triplet for Automatic Piano Transcription,” *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1536–1540, Dec. 2014.

[15] A. Cogliati, Z. Duan, and B. Wohlberg, “Piano Music Transcription with Fast Convolutional Sparse Coding,” in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing*, Boston, MA, Sept. 2015, pp. 1–6.

[16] P. Smaragdis and J.C. Brown, “Non-Negative Matrix Factorization for Polyphonic Music Transcription,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 177–180.

[17] E. Vincent, N. Bertin, and R. Badeau, “Adaptive Harmonic Spectral Decomposition for Multiple Pitch Estimation,” *IEEE Trans. Acoust., Speech, Language Process.*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

[18] N. Bertin, R. Badeau, and E. Vincent, “Enforcing Harmonicity and Smoothness in Bayesian Non-Negative Matrix Factorization Applied to

Polyphonic Music Transcription," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 18, no. 3, pp. 538–549, 2010.

[19] M. Genussov and I. Cohen, "Multiple fundamental frequency estimation based on sparse representations in a structured dictionary," *Digit. Signal Process.*, vol. 23, no. 1, pp. 390–400, Jan. 2013.

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

[21] C. Yeh, A. Roebel, and X. Rodet, "Multiple Fundamental Frequency Estimation and Polyphony Inference of Polyphonic Music Signals," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 18, no. 6, pp. 1116–1126, Aug. 2010.

[22] G. Zhang and S. Godsill, "Tracking Pitch Period Using Particle Filters," in *IEEE Workhop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, Oct. 2013.

[23] K. Han and D. Wang, "Neural Networks For Supervised Pitch Tracking in Noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[24] H. Hajimolahoseini, R. Amirfattahi, S. Gazor, and H. Soltanian-Zadeh, "Robust Estimation and Tracking of Pitch Period Using an Efficient Bayesian Filter," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1219–1229, Jul. 2016.

[25] S. Karimian-Azari, A. Jakobsson, J. R. Jensen, and M. G. Christensen, "Multi-Pitch Estimation and Tracking using Bayesian Inference in Block Sparsity," in *23rd European Signal Processing Conference*, Nice, Aug. 31-Sept. 4 2015.

[26] D. Angelosante, J. A. Bazerque, and G. B. Giannakis, "Online Adaptive Estimation of Sparse Signals: Where RLS meets the ℓ_1-Norm," *IEEE Trans. Signal Process.*, vol. 58, 2010.

[27] B. Babadi, N. Kalouptsidis, and V. Tarokh, "SPARLS: The Sparse RLS Algorithm," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4013–4025, Aug. 2010.

[28] N. Vaswani and J. Zhan, "Recursive Rrecovery of Sparse Signal Sequences From Compressive Measurements: A Reveiew," *IEEE Trans. Signal Process.*, vol. 64, no. 13, pp. 3523–3549, Jul. 2016.

[29] Y. Kopsinis, K. Slavakis, and S. Theodoridis, "Online Sparse System Identification and Signal Reconstruction Using Projections Onto Weighted ℓ_1 Balls," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 936–952, Mar. 2011.

[30] E. C. Hall and R. M. Willett, "Online Convex Optimization in Dynamic Environments," *IEEE J. Sel. Topics in Signal Processing*, vol. 9, no. 4, pp. 647–662, Jun. 2015.

[31] Y. Chen and A. O. Hero, "Recursive $\ell_{1,\infty}$ Group Lasso," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 3978–3987, Aug. 2012.

[32] E. Eksioglu, "Group sparse RLS algorithms," *International Journal of Adaptive Control and Signal Processing*, vol. 28, pp. 1398–1412, 2014.

[33] S. Jiang and Y. Gu, "Block-Sparsity-Induced Adaptive Filter for Multi-Clustering System Identification," *IEEE Trans. Signal Process.*, vol. 63, no. 20, pp. 5318–5330, Oct. 2015.

[34] B. Fuentes, R. Badeau, and G. Richard, "Harmonic Adaptive Latent Component Analysis of Audio and Application to Music Transcription," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 21, no. 9, pp. 1854–1866, Sept. 2013.

[35] E. Benetos and T. Weyde, "An Efficient Temporally-Constrained Probabilistic Model for Multiple-Instrument Music Transcription," in *Proceedings of the 16th International Society for Music Information Retrieval Conference*, Malaga, Spain, Oct. 2015.

[36] 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.

[37] Y. Chi, L. L. Scharf, A. Pezeshki, and A. R. Calderbank, "Sensitivity to Basis Mismatch in Compressed Sensing," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2182 –2195, May 2011.

[38] P. Stoica and P. Babu, "Sparse Estimation of Spectral Lines: Grid Selection Problems and Their Solutions," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 962–967, Feb. 2012.

[39] S. Haykin, *Adaptive Filter Theory (4th edition)*, Prentice Hall, Inc., Englewood Cliffs, N.J., 2002.

[40] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society B*, vol. 67, no. 1, pp. 91–108, Jan. 2005.

[41] 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.

[42] M. Kowalski, "Sparse Regression Using Mixed Norms," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 303 – 324, 2009.

[43] P. Combettes and V. Wajs, "Signal recovery by proximal forward-backward splitting," *SIAM Jour. Multiscale Modeling & Simulation*, vol. 4, pp. 1168–1200, 2005.

[44] M. A. T. Figueiredo and R. D. Nowak, "An EM Algorithm for Wavelet-Based Image Restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, Aug. 2003.

[45] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[46] I. F. Gorodnitsky and B. D. Rao, "Sparse Signal Reconstruction from Limited Data Using FOCUSS: A Re-weighted Minimum Norm Algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[47] E. J. Candès, M. B. Wakin, and S. Boyd, "Enhancing Sparsity by Reweighted l_1 Minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, Dec. 2008.

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

[49] Z. Duan and B. Pardo, "Bach10 dataset," [Online]. Available: http://music.cs.northwestern.edu/data/Bach10.html, Accessed on: Dec. 2015.

[50] 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.

[51] 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, Oct. 2009.

C

Paper C

Using Optimal Transport for Estimating Inharmonic Pitch Signals

Filip Elvander[1], Stefan Ingi Adalbjörnsson[1], Johan Karlsson[2], and Andreas Jakobsson[1]

[1] *Centre for Mathematical Sciences, Lund University, Lund, Sweden*
[2] *Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden*

Abstract

In this work, we propose a novel multi-pitch estimation technique that is robust with respect to the inharmonicity commonly occurring in many applications. The method does not require any *a priori* knowledge of the number of signal sources, the number of harmonics of each source, nor the structure or scope of any possibly occurring inharmonicity. Formulated as a minimum transport distance problem, the proposed method finds an estimate of the present pitches by mapping any found spectral line to the closest harmonic structure. The resulting optimization is a convex and highly tractable linear programming problem. The preferable performance of the proposed method is illustrated using both simulated and real audio signals.

Key words: Multi-pitch estimation, frequency clustering, inharmonicity, optimal transport distance, convex optimization.

1 Introduction

The problem of estimating the fundamental frequency, or pitch, of a harmonic, or close-to-harmonic, signal occurs in a wide range of applications [1–9]. Often, the problem is complicated by the number of sources being unknown, as is the number of components detailing each source. Furthermore, some sources, such as, e.g., audio signals resulting from stringed instruments, exhibit inharmonicity, implying that higher order components may deviate from the harmonic model, often with increasing deviation for the higher harmonics [10–12]. In such scenarios, a naive approach exploiting the sinusoidal frequency model in the time domain results in a cumbersome high dimensional optimization problem, as the uncertainty due to the inharmonicity will occur in the nonlinear frequency parameter. Previously, this problem has been approached by approximate optimization in the time domain [12], [13], approximating the frequency uncertainty with an uncertainty in the functional form of the sinusoid [10], or via a subspace-based framework robust to such deviations [14]. For certain applications, there also exists source specific pitch estimators that rely on the inharmonicity following a parametric model, see, e.g., [15]. However, such estimators are generally unable to resolve cases when harmonics from different sources overlap, as commonly occurs, for instance, in Western music playing in harmony.

In order to handle such situations, while still allowing for an unknown number of sources, we here formulate the multi-pitch problem such that the estimated pitches are obtained as the ones minimizing a particular (convex) Monge-Kantorovich optimal transportation problem. These methods have also earlier been shown useful for problems in signal analysis, e.g., for clustering, tracking, registration, and robust identification [16–19]. Transport problems have a rich history going back to questions concerning how to most efficiently transport soil from one location to another, and has since attracted attention in various fields (see [20] and references therein). An example of this is the facility localization problem, where for a set of customers one seeks to determine locations of facilities that minimize the sum of the distances from each customer to its closest facility. As we will see, the multi-pitch estimation problem can be reformulated as a facility location problem [20].

In this setting, the harmonic model (facilities) should be selected so that the spectral components (customers) can be transported to the closest harmonic model with minimal total cost. In this case, the mass to be moved constitutes the amplitude of the observed spectral component at a given frequency; as this amplitude may originate from two or more sources which have overlapping harmonics at the given frequency, we should allow the optimization to transport parts of the observed amplitude to different harmonic candidates. We further wish to introduce restrictions on the allowed mass transport problem such that ambiguity with different sub-octaves are avoided, promoting spectrally smooth solutions similar to those proposed in [2, 3, 21, 22]. As we show in the following, the desired optimization problem can be formulated as a linear programming (LP) problem, for which powerful solvers are available, even for big data applications [23]. In the numerical section, we illustrate the preferable performance of the proposed method as compared to several previously suggested methods, for both simulated and real audio signals.

2 Signal model

Consider N samples of a (reasonably) stationary signal, $y(t)$, that may be well described as a sum of close-to-harmonic sources, $x(t)$, corrupted by an additive broadband noise, $e(t)$, such that $y(t) = x(t) + e(t)$, where[1]

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

Here, K denotes the number of sources, each containing L_k close-to-harmonic signal components. The constant f_k denotes the pitch of the kth source, and the constants $a_{k,\ell}$ and $\Delta_{k,\ell}$ denote the complex amplitude and frequency deviation, respectively, of the ℓth harmonic of the kth source. The deviation will thus be zero for fully harmonic sources, whereas $\Delta_{k,\ell}$ otherwise details the inharmonicity. Depending on the source, one may have models for such inharmonicities, such as the model used for pianos (see, e.g., [11]). In the frequency domain, the assumed signal may thus be represented as

$$X(f) = \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} a_{k,\ell} \delta(f - f_k\ell - \Delta_{k,\ell}) \tag{2}$$

where $\delta(\cdot)$ denotes the Dirac delta function. In this work, we aim at estimating both the number of sources, K, and their pitches, f_k, while allowing for unknown

[1] For computational and notational simplicity, we here use the time-discrete analytical version of the measured data.

frequency deviations, $\Delta_{k,\ell}$. In order to do so, we consider the transport cost (see, e.g., [20]) associated with assigning each spectral component to a set of candidate pitches, i.e., the transport cost of moving the component onto the assumed harmonic structure related to each candidate pitch.

In order to introduce notation, let $\mathcal{F}$ denote the set of observed spectral components in the signal of interest, whereas $\mathbf{\Omega}$ denotes the set of all considered candidate pitches. Furthermore, let M and P denote the number of elements of the sets $\mathcal{F}$ and $\mathbf{\Omega}$, respectively. Here, the number of candidate pitches are assumed to be much larger than the number of sources, such that $P \gg K$. Finally, each candidate pitch is assumed to have at most $L_{\max} \geq \max_k L_k$ harmonics.

3 Optimal transport

In order to find an optimal assignment of the amplitudes corresponding to the observed line spectrum frequencies to the set of pitch candidates, one needs to define a function describing the cost of a certain assignment and then minimize this function over all possible assignments. In order to do this, let the function $c(f, f_p)$ describe the cost of moving one unit of amplitude from the line spectral frequency f to the pitch candidate f_p. For example, the cost of assigning all amplitudes in the line spectrum $Y(f)$, defined as

$$Y(f) = \sum_{f_m \in \mathcal{F}} a_{f_m} \delta(f - f_m), \tag{3}$$

where a_{f_m} denotes the amplitude of the spectral line at frequency f_m, to the candidate pitch f_p is

$$\sum_{f_m \in \mathcal{F}} |a_{f_m}| \, c(f_m, f_p) \ . \tag{4}$$

To describe the cost of a general assignment, let $\mathbf{C}$ be the $P \times M$ matrix whose (p, m)th element is equal to $c(f_m, f_p)$. Also, let $\mathbf{W}$ be the $P \times M$ matrix describing the amplitude assignment, i.e., the (p, m)th element of $\mathbf{W}$ describes how much of the magnitude $|a_m|$ that is assigned to candidate pitch f_p. Thus, to ensure that all the estimated spectral content is mapped to some pitch, the sum of the mth column of $\mathbf{W}$ must be equal to $|a_m|$. With this, the cost of an assignment described by $\mathbf{W}$ may be expressed as tr $(\mathbf{C}^T\mathbf{W})$, where $(\cdot)^T$ denotes the transpose, and tr$(\cdot)$ denotes the trace of a matrix. Defining the $M \times 1$ vector $\mathbf{a} = \begin{bmatrix} |a_1| & \dots & |a_M| \end{bmatrix}^T$, and letting $\mathbf{1}_P$ be a $P \times 1$ vector of ones, one may formulate the desired optimal transport problem as

$$\begin{aligned} \underset{\mathbf{W},\mathbf{x}}{\text{minimize}} \quad & \text{tr}\left(\mathbf{C}^T\mathbf{W}\right) \\ \text{subject to} \quad & \mathbf{W}^T\mathbf{1}_P = \mathbf{a}, \quad \mathbf{x}^T\mathbf{1}_P = K \\ & \mathbf{W} \leq \mathbf{x}\mathbf{a}^T, \quad \mathbf{W} \geq \mathbf{0} \\ & x_i \in \{0,1\} \ , \ i = 1,\dots,P \end{aligned} \tag{5}$$

where the inequalities for matrices and vectors should be interpreted elementwise. The binary vector $\mathbf{x}$ here controls whether a pitch candidate f_p is present in the solution or not, i.e., if $x_p = 1$, then f_p is present and if $x_p = 0$, then it is not. However, as x_i are binary variables, this problem is not convex. Furthermore, this formulation assumes precise knowledge of the number of sources, K, which in general is unknown. In order to remedy this, we consider the convex relaxation (cf. [16])

$$\begin{aligned} \underset{\mathbf{W},\mathbf{x}}{\text{minimize}} \quad & \text{tr}\left(\mathbf{C}^T\mathbf{W}\right) + \lambda\mathbf{1}_P^T\mathbf{x} \\ \text{subject to} \quad & \mathbf{W}^T\mathbf{1}_P = \mathbf{a}, \quad \mathbf{W} \leq \mathbf{x}\mathbf{a}^T \\ & \mathbf{x} \geq 0, \quad \mathbf{W} \geq \mathbf{0} \end{aligned} \tag{6}$$

with $\lambda > 0$. The second term of the objective function in (6) allows for an implicit choice of the sparsity of $\mathbf{x}$ via the regularization parameter λ. However, using the relaxation in (6), the cost function is unable to distinguish between sub-octaves, i.e., the row of $\mathbf{C}$ corresponding to some f_0 that may be greater or equal to the row corresponding to $f_0/2$. Fortunately, this may be included in the modeling by considering the structure of the amplitude assignment. Specifically, for each candidate pitch f_p, define an $L_{\text{max}} \times M$ matrix $\mathbf{L}^{(p)}$ that describes the mapping between the line spectral frequencies and the harmonics corresponding to f_p. That is, the (ℓ, m)th element of $\mathbf{L}^{(p)}$ is equal to one if $f_p\ell$ is the harmonic of pitch f_p that is closest in frequency to the line spectral frequency f_m, and zero otherwise. As each spectral line is mapped to precisely one harmonic, each column of $\mathbf{L}^{(p)}$ has exactly one element equal to one, whereas all the rest are zero. This linear mapping thus allows for the inclusion of constraints on the relative amplitudes of each pitch. For example, it may be used to promote spectral smoothness in each pitch. In this work, we restrict our attention to only requiring active pitches to have non-zero amplitude in the first harmonic. As this constraint is then convex it can easily be included in (6), yielding

$$\begin{aligned}
\underset{\mathbf{W},\mathbf{x}}{\text{minimize}} \quad & \operatorname{tr}\left(\mathbf{C}^T\mathbf{W}\right) + \lambda \mathbf{1}_P^T \mathbf{x} \\
\text{subject to} \quad & \mathbf{W}^T \mathbf{1}_P = \mathbf{a}, \quad \mathbf{W} \leq \mathbf{x}\mathbf{a}^T \\
& \mathbf{x} \geq 0, \quad \mathbf{W} \geq \mathbf{0} \\
& \left(\mathbf{1}_M - (Q+1)\,\mathbf{e}_1\right)^T \mathbf{L}^{(p)} \left[\mathbf{W}\right]_{p\cdot}^T \leq 0
\end{aligned} \tag{7}$$

for $p = 1, \ldots, P$. Here, $Q > 1$ assures that a scaled version of the amplitude assigned to the first harmonic dominates the amplitude assigned to the rest of the harmonics, thus enforcing solutions where active pitches have non-zero amplitude assigned to their first harmonics. In our simulations, we use $Q = 3L_{\text{max}}$. Here, $\mathbf{e}_1$ denotes the $M \times 1$-vector with its first element equal to one, and the rest zero, with $\left[\mathbf{W}\right]_{p\cdot}$ denoting row p of $\mathbf{W}$. It is worth noting that the resulting problem is an LP, which may thus be solved using standard convex solvers.

4 Choice of transport cost function

To model the amplitude distribution of a pitch, the transport cost function $c(\cdot,\cdot)$ should assign the cost of associating amplitude at a frequency f_m to a candidate pitch f_p depending on the distance between f_m and the closest harmonic of f_p, e.g.,

$$c\left(f_m, f_p\right) = \min_{\ell \in \mathbb{N}} \left| f_m - f_p \ell \right|^2 . \tag{8}$$

However, this function would too harshly penalize inharmonicity, as the higher harmonics of inharmonic pitches could typically deviate significantly from integer multiples of the pitch. We therefore propose to only have harsh penalties for the pitch, while allowing subsequent harmonics to deviate somewhat more. Specifically, for the first harmonic, let

$$c_1\left(f_m, f_p\right) = \rho s_+ \left(\left| f_p - f_m \right|, \frac{\Delta_f}{2} \right)^{\nu} \tag{9}$$

where $s_+(\cdot)$ is the soft threshold function defined as

$$s_+\left(x, \frac{\Delta_f}{2}\right) = \left| \max\left(x - \frac{\Delta_f}{2}, 0\right) \right| \tag{10}$$

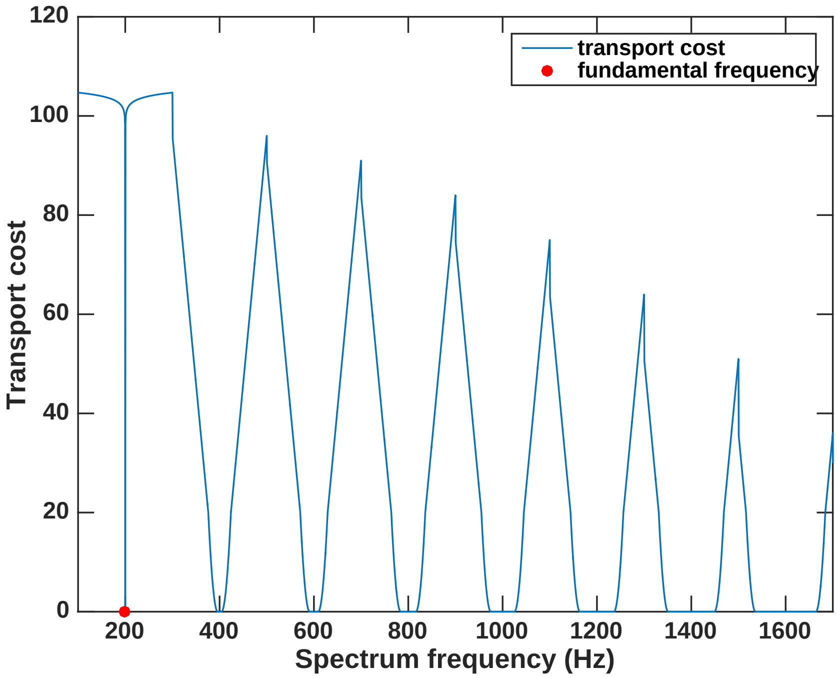

Figure 1: Transportation cost for candidate pitch with fundamental frequency 200 Hz.

and Δ_f is the spacing of the candidate pitch grid. Thus, we allow for a deadzone corresponding to the grid resolution, while penalizing larger deviances according to a scaled, highly non-convex, pseudo-norm. To allow for increasing deviations with higher harmonics, we instead use

$$c_\ell(f_m, f_p) = \min\left(\varepsilon_\ell(f_m, f_p), \xi \varepsilon_\ell(f_m, f_p)^2\right) \tag{11}$$

where

$$\varepsilon_\ell(f_m, f_p) = s_+\left(\left|f_p \ell - f_m\right|, \psi f_p \ell^2\right). \tag{12}$$

Thus, the width of the deadzone is dependent on the harmonic order ℓ as well as being scaled by a small number, ψ. In our simulations, $\rho = 100$, $\xi = 0.01$, $\nu = 0.05$, and $\psi = 0.005$. An illustration of the transport cost function is shown in Figure 1, where the cost of assigning frequencies on the interval (100, 1700) Hz

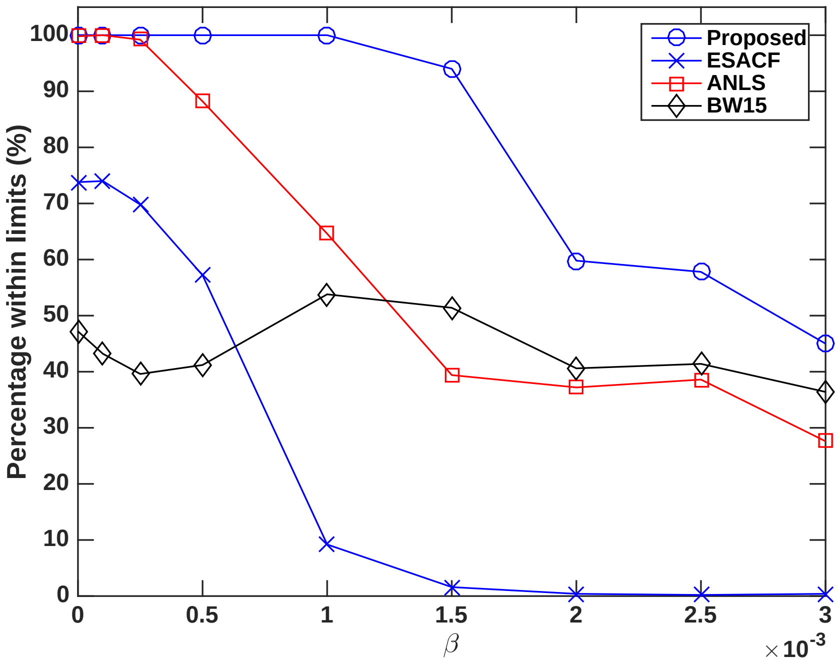

Figure 2: Percentage of pitch estimates found within ±3% of the ground truth in the simulated data case.

is shown for a pitch of 200 Hz. Here, the width of the deadzone scales quadratically with the harmonic order.

5 Numerical results

We proceed to examine the performance of the proposed method using both simulated and measured audio signals. In both settings, the line spectrum is estimated using the MUSIC estimator [24], with $M \gg \sum_k L_k$. The amplitudes a_m are then estimated using least squares. Initially, we examine a simulated signal consisting of two pitches, with pitches f_1 and f_2, with varying degrees of inharmonicity. The harmonics of the pitches are modelled using the piano model (see, e.g., [11]), i.e., $f_{k,\ell} = f_k \ell \sqrt{1+\beta\ell^2}$, for $\ell = 1, \dots, L_k$ and $k = 1, 2$, where the parameter $\beta \ll 1$ controls the level of inharmonicity. The frequencies f_1 and f_2 are drawn

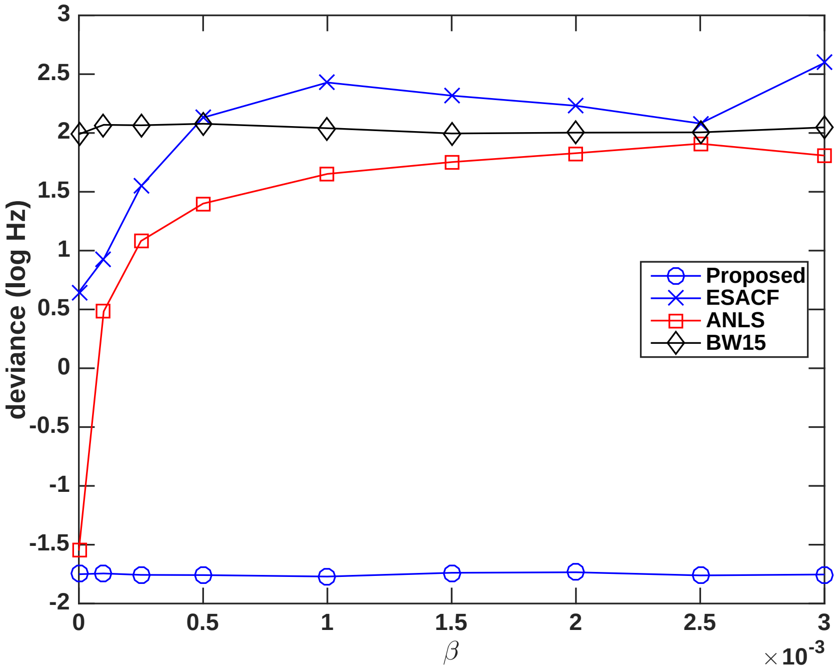

Figure 3: Expected maximal absolute deviation of pitch estimates from the ground truth in the simulated data case.

uniformly on the intervals $(300, 390)$ Hz and $(400, 540)$ Hz, respectively. The harmonic orders L_k are drawn uniformly on $[8, 12]$, whereas the magnitude of each harmonic is drawn uniformly on $(0.75, 1.25)$, with phases drawn uniformly on $[0, 2\pi)$. We thereafter add an additive white Gaussian noise to the signal, resulting in a signal-to-noise-ratio of 30 dB. The signal is then sampled for 30 ms at 40 kHz. This is done for 500 Monte Carlo simulations and for varying values of β. Performance is then measured as the percentage of the simulations in which both pitch estimates are found within $\pm 3\%$ of their respective ground truths and where no erroneous extra pitch estimates are produced. For the proposed method, we set $L_{\max} = 20$ and $\lambda = 15$. As comparison, we include three other types of pitch estimators; the approximate non-linear least squares estimator (ANLS) (see, e.g., [5]); the autocorrelation-based enhanced summary autocorrelation (ESACF) estimator [25]; and the method presented in [9], which is based on probabilistic latent component analysis. The latter method, hereafter

	Proposed	ESACF	BW15
Accuracy	0.928	0.691	0.366
Precision	0.974	0.984	0.391
Recall	0.952	0.699	0.849

Table 1: Performance measures for the proposed method as well as the ESACF and BW15 methods.

referred to as BW15, is specifically designed for multi-pitch estimation for music signals, with pitch estimates restricted to the chromatic Western scale, i.e., to the keys of the piano. This frequency resolution corresponds precisely to the chosen accuracy limit of ±3% of the ground truth pitches. The method is based on extensive training on a database of various forms of signals[2]. As ANLS requires knowledge of both the number of sources and the number of harmonics for each source, it is here provided with oracle model order knowledge. For all methods, the algorithm settings recommended by their respective authors have been used. As shown in Figure 2, the proposed method outperforms the other methods for all considered levels of inharmonicity. It may be noted that the performance of the BW15 method is not strictly decreasing with the inharmonicity parameter β; rather, the best performance is achieved for the value $\beta = 10^{-3}$, arguably due to this being the best match to the method's training library. We also evaluate the accuracy of the pitch estimates, measured as the maximum absolute deviation of each estimate from its corresponding ground truth, conditioned on that the estimates are found within ±3% of their respective ground truths. The results are shown in Figure 3, with deviation shown in log-scale. Again, the proposed method outperforms all comparison methods.

In Figure 4, we study a real audio signal consisting of two harmonic trumpet signals and two piano signals with some inharmonicity. Specifically, the signal is composed of two trumpet signals, with pitches 440 and 554.37 Hz, corresponding to the notes A4 and D♭5, and of two piano notes, with pitches 329.65 and 415.3 Hz, corresponding to the notes E4 and G♯4. Ground truth estimates for the trumpet pitches have been obtained by applying the YIN estimator [26] to the single channel recordings. Ground truths for the pianos are known as the signals are simulated using software synthesizers. As can be seen in Figure 4, the proposed

[2] The implementation used was provided online by the authors of [9].

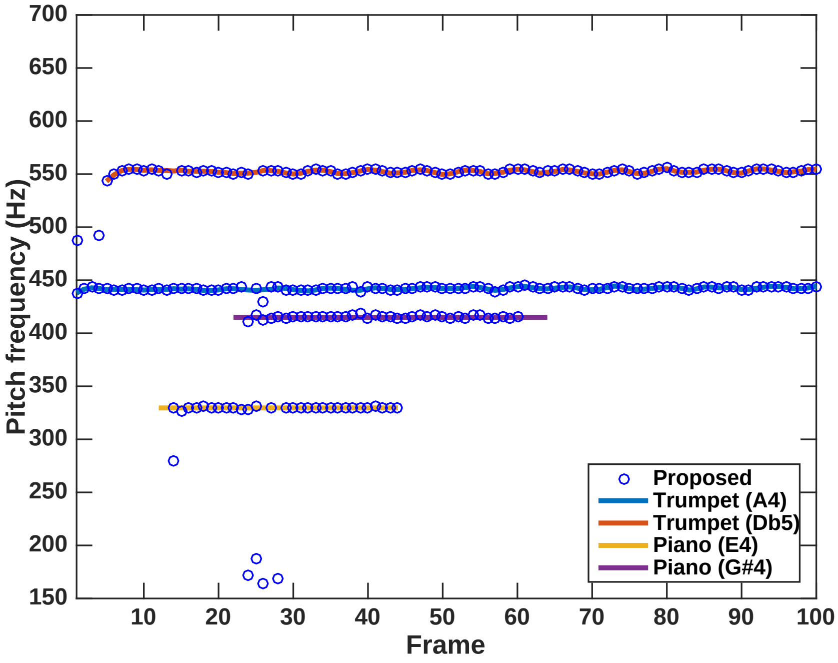

Figure 4: Estimated fundamental frequencies for a signal containing two trumpet notes as well as two piano notes.

method is able to correctly group the frequencies into the correct pitches, with only small errors during the onset phase, where the frequency content is highly transient and non-sparse. The recording was sampled at 44.1 kHz and was subdivided into non-overlapping estimation frames of length 30 ms. The settings for the proposed method was $L_{\max} = 10$ and $\lambda = 15$. Table 1 compares the proposed method to the ESACF and BW15 methods, while excluding ANLS as exact model order information of the number of harmonics of each source is unavailable. The table presents the performance measures *Accuracy*, *Precision*, and *Recall* [27]. As can be seen, the performance of the proposed method is clearly better than that of the comparison methods; likely, this results from ESACF having problems with estimating the pitches of the inharmonic pianos, whereas BW15 suffers from not being able to accurately estimate the trumpets, perhaps caused by bad match to its training data set.

References

[1] R. B. Randall, *Vibration-Based Condition Monitoring: Industrial, Aerospace and Automotive Applications*, John Wiley & Sons, Chichester, UK, 2011.

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

[3] 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, Oct. 2016.

[4] M. S. Reza, M. Ciobotaru, and V. G. Agelidis, "Robust Technique for Accurate Estimation of Single-Phase Grid Voltage Fundamental Frequency and Amplitude," *IET Generation, Transmission Distribution*, vol. 9, no. 2, pp. 183–192, 2015.

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

[6] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1015–102, Apr. 2009.

[7] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric Representations of Bird Sounds for Automatic Species Recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2252–2263, Nov. 2006.

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

[9] E. Benetos and T. Weyde, "An Efficient Temporally-Constrained Probabilistic Model for Multiple-Instrument Music Transcription," in *Proceedings*

of the 16th International Society for Music Information Retrieval Conference, Malaga, Spain, Oct. 2015.

[10] N. R. Butt, S. I. Adalbjörnsson, S. D. Somasundaram, and A. Jakobsson, "Robust Fundamental Frequency Estimation in the Presence of Inharmonicities," in *38th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancouver, May 26–31, 2013.

[11] H. Fletcher, "Normal vibration frequencies of stiff piano string," *Journal of the Acoustical Society of America*, vol. 36, no. 1, 1962.

[12] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "On the Influence of Inharmonicities in Model-Based Speech Enhancement," in *European Signal Processing Conference*, Marrakesh, Sept. 10-13 2013.

[13] T. Nilsson, S. I. Adalbjörnsson, N. R. Butt, and A. Jakobsson, "Multi-Pitch Estimation of Inharmonic Signals," in *European Signal Processing Conference*, Marrakech, Sept. 9-13, 2013.

[14] M. G. Christensen, P. Vera-Candeas, S. D. Somasundaram, and A. Jakobsson, "Robust Subspace-based Fundamental Frequency Estimation," in *33rd IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Las Vegas, Mar. 30-Apr. 4, 2008.

[15] C. Kim, W. Chang, S-H. Oh, and S-Y. Lee, "Joint Estimation of Multiple Notes and Inharmoncity Coefficient Based on f0-Triplet for Automatic Piano Transcription," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1536–1540, Dec. 2014.

[16] F. Carli, L. Ning, and T. Georgiou, "Convex Clustering via Optimal Mass Transport," *arXiv preprint arXiv:1307.5459*, 2013.

[17] Xianhua Jiang, Zhi-Quan Luo, and Tryphon T Georgiou, "Geometric methods for spectral analysis," *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1064–1074, 2012.

[18] S. Haker, L. Zhu, A. Tannenbaum, and S. Angenent, "Optimal mass transport for registration and warping," *International Journal of Computer Vision*, vol. 60, no. 3, pp. 225–240, 2004.

[19] J. Karlsson and L. Ning, “On robustness of ℓ_1-regularization methods for spectral estimation,” in *IEEE 53nd Annual Conference on Decision and Control*, Dec. 2014.

[20] C. Villani, *Topics in Optimal Transportation*, vol. 58, Graduate studies in Mathematics, AMS, 2003.

[21] A. Klapuri, “Multiple fundamental frequency estimation based on harmonicity and spectral smoothness,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 804–816, 2003.

[22] V. Emiya, R. Badeau, and B. David, “Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1643–1654, Aug. 2010.

[23] R. E. Bixby, J. W. Gregory, R. E. Marsten, and D. F. Shanno, “Very Large-Scale Linear Programming: A Case Study in Combining Interior Point and Simplex Methods,” *Operations Research*, vol. 40, no. 5, 885-897 1992.

[24] R. O. Schmidt, *A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation*, Ph.D. thesis, Stanford University, Stanford, C.A., 1981.

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

[26] 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.

[27] 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, Oct. 2009.