LUND UNIVERSITY

Multi-Pitch Estimation Exploiting Block Sparsity

S. I. ADALBJÖRNSSON, A. JAKOBSSON, AND
M. G. CHRISTENSEN

Published in: Elsevier Signal Processsing
doi:10.1016/j.sigpro.2014.10.014

Lund 2015

Mathematical Statistics
Centre for Mathematical Sciences
Lund University

Multi-Pitch Estimation Exploiting Block Sparsity[☆]

Stefan I. Adalbjörnsson[*,a], Andreas Jakobsson[a], Mads G. Christensen[b]

[a]*Dept. of Mathematical Statistics, Lund University, P.O. Box 118, SE-221 00 Lund, Sweden*

[b]*Audio Analysis Lab, Dept. of Arch., Design & Media Technology, Aalborg University, Denmark*

Abstract

We study the problem of estimating the fundamental frequencies of a signal containing multiple harmonically related sinusoidal components using a novel block sparse signal representation. An efficient algorithm for solving the resulting optimization problem is devised exploiting a novel variable step-size alternating direction method of multipliers (ADMM). The resulting algorithm has guaranteed convergence and shows notable robustness to the f_0 vs $f_0/2$ ambiguity problem. The superiority of the proposed method, as compared to earlier presented estimation techniques, is demonstrated using both simulated and measured audio signals, clearly indicating the preferable performance of the proposed technique.

Key words: Pitch estimation, block sparsity, total variation, spectral smoothness, order estimation.

1. Introduction

Estimating the fundamental frequency of harmonically related signals form an integral part in a wide range of signal processing applications, and perhaps especially so in speech and audio processing. For example, the fundamental frequency, or *pitch*, is necessary when forming the long-term prediction used in linear prediction-based speech codecs [2], and is similarly the

[☆]This work was supported in part by the Swedish Research Council and Carl Trygger's foundation. This work has been presented in part at the ICASSP 2013 conference [1].

[*]Corresponding author. Phone: +4646-2227976. Fax: +4646-2224623.

Email addresses: sia@maths.lth.se (Stefan I. Adalbjörnsson), aj@maths.lth.se (Andreas Jakobsson), mgc@create.aau.dk (Mads G. Christensen)

key component in music information retrieval applications, such as automatic music transcription, and in musical genre classification [3]. The fundamental frequency is also of notable importance in problem such as source separation, enhancement, compression, and classification (see, e.g., [4, 5] and the references therein), as well as in several biomedical, mechanical and acoustic applications, and the topic has for these reasons attracted a notable interest during the recent decades. Commonly, the pitch estimate is formed assuming a single source model, such that only a single fundamental frequency and its harmonics are assumed to be present in the signal, using different kinds of similarity measures, such as the cross-correlation, cepstrum, or the average squared difference function (see, e.g., [6–12]), although notable exceptions treating the multi-pitch problem can be found in, e.g., [4, 13–24]. Regrettably, the problem is hard, and most of these techniques will suffer from not yielding unique estimates even in the ideal case, even for a single source, and/or will typically also require perfect *a priori* knowledge of both the number of sources and the model order of each of these sources. Often, such limitations necessitate notable post-processing or correction steps in order to improve on an initially poor pitch estimate. In this work, we focus on improving the initial pitch estimate, proposing a novel multi-pitch estimation approach making no *a priori* model order assumptions. The method is based on a sparse signal recovery framework, wherein a signal is assumed to consist of only a small number of components from a large set of potential signal vectors. This approach has been found to yield high quality estimates in a wide variety of fields (see, e.g., [25–27]), and has also earlier been exploited in machine learning settings, where sparse modeling of pitch signals is accomplished by learning a dictionary of pitches from a training data set (see, e.g., [17, 22, 23]). For sinusoidal signals, it was early on shown that using a sparse representation technique allowed for high resolution frequency estimates; typical examples include [28, 29], wherein the sparse signal reconstruction from noisy observations was accomplished with the by now well-known sparse least squares (LS) technique. A similar approach may clearly also be applied to the pitch estimation problem, although one is then not fully exploiting the harmonic signal structure. Herein, we instead propose a novel block sparse signal representation, such that each signal source is grouped in one data block for each pitch frequency. By then extending the representation to all considered pitch frequencies, reminiscent to the extended dictionaries used in, e.g., [14, 28, 30], the resulting model will be sparse in the sense that it will be formed from only a few of the possible blocks in the dictionary.

Different from estimates such as the ones presented in [17, 22, 23], the presented method does not exploit any training data, with the method inferring the pitch parameters and the model orders from the spectral content of the signal. The proposed pitch estimation method instead exploits the group sparse structure, without requiring any prior knowledge of either the number of sources present, or their number of harmonics. The presented algorithm, in its presented form, does not take into account for any possible inharmonicity in the pitch structure, such that the higher order frequencies would not occur precisely as a multiple of the fundamental frequency. Such inharmonicities are common in audio signals, and should be taken into account for such signals. As we are here focusing on the general problem, occurring also for numerous other forms of signals, we have here opted to exclude the treatment of inharmonicity, although note that the algorithm may be extended to allow for this along the lines presented in [31, 32], or using a dictionary learning approach such as in [33, 34]. The theoretical study of block sparse signals was initially suggested in [35], where it is shown that including this structure in the estimation procedure has great practical consequences, improving both theoretical recovery limits and numerical results in many cases (see, e.g., [35–38]). Generally, this form of group sparse convex optimization problems are computationally cumbersome; for this reason, we also derive an efficient algorithm to form the estimate based on the alternating directions methods of multipliers (ADMM) (see, e.g., [39, 40]). The resulting algorithm will have a guaranteed convergence as well as exhibit a significant robustness to the common problem of the f_0 vs $f_0/2$ ambiguity, i.e., when a pitch candidate at half the nominal frequency fits the observed signal as well, or possibly even better, than the true pitch frequency. The remainder of this paper is organized as follows: in the next section, we briefly present the data model. Then, in Section 3, we introduce the proposed pitch estimation technique. Section 4 introduces the efficient ADMM-based implementation, and Section 5 includes numerical evaluations of the proposed method as compared to earlier techniques. Finally, Section 6 concludes on the work.

2. Block Sparse Signal Model

Consider a complex-valued signal, $y(n)$, consisting of K harmonically related (signal) sources with fundamental frequencies f_k, for $k = 1, \ldots, K$,

such that (see also [4])

$$y(n) = \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} a_{k,\ell} e^{j2\pi f_k \ell n} + e(n) \tag{1}$$

for $n = 1, \ldots, N$, where $a_{k,\ell}$ and L_k denote the (complex-valued) amplitude of the ℓ:th harmonic of the k:th source, and the number of harmonically related sinusoids for the k:th source, respectively, and where $e(n)$ is an additive noise term, here assumed to be an identically distributed independent circularly symmetric complex Gaussian process with variance σ_e^2. It is worth noting that due to the restriction of the allowed frequency range, the number of harmonics are restricted as a function of the fundamental frequency, such that $L_k < \lfloor 1/f_k \rfloor$, $\forall k$, where $\lfloor \cdot \rfloor$ denotes the round-down to nearest integer operation. Let

$$\mathbf{y} = \left[\begin{array}{ccc} y(1) & \ldots & y(N) \end{array} \right]^T \tag{2}$$

where $(\cdot)^T$ denotes the transpose. Then, (1) may be expressed succinctly as

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{V}_k \mathbf{a}_k + \mathbf{e} \triangleq \mathbf{W}\mathbf{a} + \mathbf{e} \tag{3}$$

where $\mathbf{e}$ is a vector of noise terms constructed in the same manner as $\mathbf{y}$, and

$$\begin{aligned} \mathbf{W} &= \left[\begin{array}{ccc} \mathbf{V}_1 & \ldots & \mathbf{V}_K \end{array} \right] && (4) \\ \mathbf{V}_k &= \left[\begin{array}{cccc} \mathbf{z}_k & \mathbf{z}_k^2 & \ldots & \mathbf{z}_k^{L_k} \end{array} \right] && (5) \\ \mathbf{a} &= \left[\begin{array}{ccc} \mathbf{a}_1^T & \ldots & \mathbf{a}_K^T \end{array} \right]^T && (6) \\ \mathbf{a}_k &= \left[\begin{array}{ccc} a_{k,1} & \ldots & a_{k,L_k} \end{array} \right]^T && (7) \end{aligned}$$

with the vector powers, $\mathbf{z}_k^\ell$, being evaluated element-wise,

$$\mathbf{z}_k^\ell = \left[\begin{array}{ccc} e^{j2\pi f_k \ell} & \ldots & e^{j2\pi f_k N \ell} \end{array} \right]^T \tag{8}$$

Reminiscent to the models considered for line spectra (see, e.g., [14, 28, 30]), the matrix $\mathbf{W}$ may be expanded to be formed instead over a (large) range of possible fundamental frequencies, ν_ℓ, for $\ell = 1, \ldots, P$, where P denotes the total number of considered frequencies, such that the corresponding amplitude vector, $\mathbf{a}$, will have elements different from zero only for those frequencies actually coinciding with the frequencies in the signal. Thus, for

the signal in (1), for each source in the signal, there will be a corresponding non-zero block in the amplitude vector, i.e., if the source has fundamental frequency ν_ℓ, the sub-block $\mathbf{a}_\ell$ will be non-zero. It should be noted that this formulation thus implicitly assumes that P is selected large enough so that the true pitch frequencies lie close to the used grid, such that a sparse solution, that has correctly found the dictionary elements closest to the true frequency, will result in a small approximation error. Practical experience with similar methods, e.g., [28, 29], shows that they are quite robust to this approximation (see also the related discussions in [30, 41, 42]). Given the structure of (3), the resulting approximation of the signal is not only sparse, but thus also *block sparse*, since for each source present, several harmonics will be included in the signal.

3. Pitch Estimation using Block Sparsity

Reminiscent of the block sparse formulations introduced in [35], one may thus form an estimate of the present sources as

$$\underset{\mathbf{a}}{\text{minimize}} \ \frac{1}{2}||\mathbf{y} - \mathbf{W}\mathbf{a}||_2^2 + \alpha \sum_{k=1}^{P} ||\mathbf{a}_k||_2 \tag{9}$$

where $||\cdot||_p$ denotes the ℓ_p norm, and with $\alpha > 0$ denoting a tuning parameter that controls the relative importance of the block sparsity promoting ℓ_2 norm and the squared ℓ_2 norm fitting term, discussed further below. It should be noted that the cost function is clearly convex as it is a sum of a norm and the composition of a norm and an affine function. The second term in (9) is included to promote a block sparse solution, i.e., a solution with the property that most blocks, $\mathbf{a}_i$, are zero (see also Appendix A). As noted, the number of harmonics of each source, L_k, is generally not known, and to be able to use the presented sparse approximation model, one needs to set some maximum allowed number of harmonics for all possible fundamental frequencies, say $L_{\max}$. This implies that the data blocks, $\mathbf{a}_k$, as given in (7), will typically contain some amplitudes that are close to zero, for those harmonics that are not present in the source signal. To allow for this, we introduce a further ℓ_1 penalty term, generally forcing small amplitudes to zero, resulting in the following sparse group lasso (see also [43] and the discussion in Appendix B)

$$\underset{\mathbf{a}}{\text{minimize}} \ \frac{1}{2}||\mathbf{y} - \mathbf{W}\mathbf{a}||_2^2 + \lambda||\mathbf{a}||_1 + \alpha \sum_{k=1}^{P} ||\mathbf{a}_k||_2 \tag{10}$$

where $\alpha > 0$ is a tuning parameter. Using the formulation in (10), this would imply that the (generic) f_0 harmonics will make up a subset of the block detailing the $f_0/2$ harmonics, i.e., the frequencies $\{f_0, 2f_0, 3f_0, ..., L_{f_0} f_0\}$ will be present in both blocks, and thus the minimization in (10) will then in all cases prefer the block corresponding the lower frequency. In order to partly resolve this problem, we introduce a further scaling of the norms in the minimization, such that the blocks are given comparable weights, instead forming the minimization as

$$\underset{\mathbf{a}}{\text{minimize}} \frac{1}{2}||\mathbf{y} - \mathbf{W}\mathbf{a}||_2^2 + \lambda||\mathbf{a}||_1 + \alpha \sum_{k=1}^{P} \sqrt{L_k}||\mathbf{a}_k||_2 \tag{11}$$

However, this does not completely remove the ambiguity from the model since one might well consider, in certain scenarios, restricting the maximum number of allowed harmonics such that the sub-vectors corresponding to some f_0 and $f_0/2$ could have the same number of elements. Thus, a signal composed of a fundamental frequency f_0, with L_{f_0} harmonics, can be written interchangeably using the first L_{f_0} elements of the sub-vector corresponding to the fundamental frequency f_0, or every other element of the first $2L_{f_0}$ elements of the sub-vector corresponding to $f_0/2$. By instead including a *total variation* penalty function

$$\text{Tv}(\mathbf{a}_k) = \sum_{i=1}^{L_{k-1}} |a_{k,i} - a_{k,i+1}|$$

in the cost function, blocks with constant amplitudes will not be penalized, whereas $f_0/2$ vectors, such as $\mathbf{a}_{f_0/2}$ mentioned above, will incur a large penalty. Note that even for signals that contain only odd harmonics, such as the clarinet, or other audio signals created with a closed cylindrical pipe [44], the total variation penalty will resolve halvling ambiguity. The resulting spectral smoothness is similar to often imposed assumption in the modeling of audio signals, see e.g., [15]. Adding the total variation function to the suggested criteria will still result in a convex problem as the total variation function is convex since it may be written as composition of an affine function, say $\mathbf{F}$, and the ℓ_1 norm, i.e.,

$$\sum_{k=1}^{P} \text{Tv}(\mathbf{a}_k) = ||\mathbf{F}\mathbf{a}||_1 \tag{12}$$

where $\mathbf{F} \in \mathbb{R}^{\sum_{k=1}^{P} L_k \times \sum_{k=1}^{P} L_k}$ is created such that the rows corresponding to the first $L_k - 1$ elements of each block have a one on the diagonal and minus one on the first super-diagonal, and the row corresponding to element L_k is zero, or equivalently, a difference operator with rows $L_1, L_1 + L_2 \ldots, \sum_{k=1}^{P} L_k$ set to zero. Thus, we propose forming the pitch estimate via the minimization

$$\underset{\mathbf{a}}{\text{minimize}} \ \frac{1}{2}||\mathbf{y} - \mathbf{W}\mathbf{a}||_2^2 + \lambda||\mathbf{a}||_1 + \alpha \sum_{k=1}^{P} ||\mathbf{a}_k||_2 + \gamma \sum_{k=1}^{P} \mathrm{Tv}(\mathbf{a}_k) \qquad (13)$$

where $\gamma > 0$ is a tuning parameter, which should be set small enough such that the total effect of adding the TV term is only to resolve the f_0 and $f_0/2$ ambiguity in a consistent and correct manner; in the numerical section it was set to 0.01 for all simulations.The tuning parameters, λ and α may, for instance, be estimated for example with a cross validation approach. However, in our experience, if the signal to noise ratio (SNR) is high enough, they may preferably be set by simply inspecting the amplitudes in the zero padded discrete Fourier transform, as is shown in Appendix B, i.e., by setting α as the smallest significant amplitude above the noise floor, and by setting λ similarly, but for each pitch. It is worth noting that an alternative formulation may be obtained by instead using a covariance fitting formulation; as recently shown in [45, 46], the sparse SPICE covariance fitting algorithm [47] may be equivalently expressed using an weighted penalized ℓ_1 formulation, for a particular choice of λ. One may similarly form a covariance fitting style minimization of the here proposed minimization by replacing the squared ℓ_2 fitting term in (11) or (13) with a corresponding ℓ_1 fitting term; we will below examine what such a choice would imply. Reminiscent of the work in [29, 48–50], another approach would be to instead consider other penalties, e.g., the ℓ_q penalties with $0 < q < 1$, or the reweighted ℓ_1, which would both lead to non-convex optimization problems, that can nevertheless often be efficiently solved with the benefit of, in many cases, sparser solutions, with less biased amplitude estimates, although with local minima being a recurring problem and without the global optimality conditions of convex optimization problems. Herein, given that our main objective is the estimation of the non-linear fundamental frequency parameters, we restrict our attention to convex criteria, but note that especially the re-weighted ℓ_1 algorithm and the ℓ_q-like criteria suggested in [48] are easily adapted to the algorithm and the here presented criteria.

Considering that the signals of interest are only approximately sparse in $\mathbf{W}$,

and as two closely spaced fundamental frequencies will result in that the corresponding matrices, $\mathbf{W}_s$ and $\mathbf{W}_r$, will be rather similar, one cannot expect the resulting (block) pseudo spectral solution, formed over the peaks of the 2-norm of the estimated amplitudes, $||\hat{\mathbf{a}}_k||_2$, to have exactly as many non-zero blocks as there are sources present in the signal. In order to determine the number of sources present, we therefore introduce a novel BIC-style criterion, such that the number of sources are selected as (cf. [30, 51])

$$\hat{K} = \underset{k\in[1,K_{\max}]}{\operatorname{argmin}} \operatorname{BIC}_k(\lambda,\alpha) \tag{14}$$

where $K_{\max}$ denotes the maximum number of considered sources, here selected as the number of peaks present in the initially obtained (block) pseudo-spectra, and where the (λ,α)-dependent BIC cost function is formed as

$$\operatorname{BIC}_k(\lambda,\alpha) = 2N\ln(\hat{\sigma}_k^2) + (2H_k+1)\ln(N) \tag{15}$$

with $\hat{\sigma}_k^2$ denoting the variance of the estimation residual when modeling the pitch signal using

$$H_k = \sum_{\ell=1}^{k} \hat{L}_{k_\ell} \tag{16}$$

(dependent) sinusoidal components (see also [4]), where $\hat{L}_{k_\ell}$ is the number of frequencies corresponding to the non-zero elements of $\hat{\mathbf{a}}_{k_\ell}$. It should be stressed that the $\hat{L}_{k_\ell}$ considered harmonics are not necessarily consecutive, thereby allowing for the case of missing harmonics (including the possibility that the signal lacks the fundamental frequency component), which is a case commonly occurring in many form of acoustic signals.

4. An Efficient ADMM Implementation

As the minimizations in (11) and (13) are composed of simple convex functions, they may be solved using one of the freely available interior point based solvers, such as SeDuMi [52] and SDPT3 [53], although such solvers will scale badly both with increased data length and with the use of a finer grid size for the fundamental frequency. As a result, such a solution will in many cases be too computationally intensive to be practically useful. In order to form a more efficient implementation, we therefore reformulate the minimization in (11) using an ADMM formulation, which may be used to

solve convex optimization problems which are the sum of two convex functions by decomposing the optimization into two simpler problems, which are then solved in an iterative fashion (see, e.g., [39]). For completeness and to introduce our notation, we here include a brief outline of the main steps involved. Consider the convex optimization problem

$$\underset{\mathbf{z}}{\text{minimize}} \quad f_1(\mathbf{z}) + f_2(\mathbf{G}\mathbf{z}) \tag{17}$$

where $\mathbf{z} \in \mathbb{R}^p$ is the optimization variable, $f_1(\cdot)$ and $f_2(\cdot)$ are convex functions, and $\mathbf{G} \in \mathbb{R}^{N \times p}$ is a known matrix. If one introduces an auxiliary variable, $\mathbf{u}$, then (17) may be equivalently be expressed as

$$\begin{aligned} \underset{\mathbf{z},\ \mathbf{u}}{\text{minimize}} \quad & f_1(\mathbf{z}) + f_2(\mathbf{u}) + \frac{\mu}{2}||\mathbf{G}\mathbf{z} - \mathbf{u}||_2^2 \\ \text{subject to} \quad & \mathbf{G}\mathbf{z} - \mathbf{u} = \mathbf{0} \end{aligned} \tag{18}$$

Under the assumption that there is no duality gap, which is true for all the optimization problems considered herein, one can solve the optimization problem via the dual function defined as the infimum with respect to $\mathbf{u}$ and $\mathbf{z}$ of the augmented Lagrangian [39]

$$L_\mu(\mathbf{z}, \mathbf{u}, \mathbf{d}) = f_1(\mathbf{z}) + f_2(\mathbf{u}) + \mathbf{d}^T(\mathbf{G}\mathbf{z} - \mathbf{u}) + \frac{\mu}{2}||\mathbf{G}\mathbf{z} - \mathbf{u}||_2^2 \tag{19}$$

which holds for all μ, since at any feasible point $||\mathbf{G}\mathbf{z} - \mathbf{u}||_2^2 = 0$. The ADMM does this by iteratively maximizing the dual function, such that at step ℓ+1, one minimizes the Lagrangian for one of the variables, while holding the other fixed at its most recent value, i.e.,

$$\mathbf{z}(\ell + 1) = \underset{\mathbf{z}}{\text{argmin}}\, L_\mu(\mathbf{z}, \mathbf{u}(\ell), \mathbf{d}(\ell)) \tag{20}$$

$$\mathbf{u}(\ell + 1) = \underset{\mathbf{u}}{\text{argmin}}\, L_\mu(\mathbf{z}(\ell + 1), \mathbf{u}, \mathbf{d}(\ell)) \tag{21}$$

where the notation $\mathbf{x}(\ell)$ denotes the vector $\mathbf{x}$ at iteration ℓ. Finally one updates the dual variable by taking a gradient ascent step to maximize the dual function, resulting in

$$\tilde{\mathbf{d}}(\ell + 1) = \tilde{\mathbf{d}}(\ell) - \mu(\mathbf{G}\mathbf{z}(\ell + 1) - \mathbf{u}(\ell + 1)) \tag{22}$$

from which the interpretation of μ as the dual variable step size may be seen (see also [39] for further details). The general ADMM steps are outlined in

Algorithm 1, using the scaled version of the dual variable $\mathbf{d}(\ell) = \tilde{\mathbf{d}}(\ell)/\mu$, which is more convenient for implementation. As a stopping criterion, it is shown in [39] that by studying the necessary and sufficient conditions for the optimality of a solution, say $\mathbf{z}^*, \mathbf{u}^*$, and $\mathbf{d}^*$, of the minimization in (18), i.e., the primal feasibility

$$\mathbf{G}\mathbf{z}^* - \mathbf{u}^* = \mathbf{0} \tag{23}$$

and the dual feasibility

$$\mathbf{0} \in \ \partial f_1(\mathbf{z}^*) + \mathbf{G}^T\mathbf{d}^* \tag{24}$$

$$\mathbf{0} \in \ \partial f_2(\mathbf{u}^*) - \mathbf{d}^* \tag{25}$$

where ∂ is the sub-differential operator, imply that the so-called primal and dual residuals, which are defined as $\mathbf{r}(\ell) = \mathbf{G}\mathbf{z}(\ell) - \mathbf{u}(\ell)$ and $\mathbf{s}(\ell) = \mu\mathbf{G}^T(\mathbf{u}(\ell) - \mathbf{u}(\ell-1))$, respectively, will converge to zero. Thus, as a stopping criterion, one may use that the norm of the primal and dual residuals are small enough. Clearly, the ADMM is only relevant when the optimizations in steps 3 and 4 in Algorithm 1 can be carried out easily as compared to the original problem. We begin by examining the implementation of (11), and then proceed to extending this to form (13). One possibility to reformulate (11) in this fashion would be to choose $f_1(\cdot)$ as the 2-norm fitting term and $f_2(\cdot)$ as the sum of the sparse regularization term, i.e., with $\mathbf{G} = \mathbf{I}$ and

$$f_1(\mathbf{z}) = \frac{1}{2}||\mathbf{y} - \mathbf{W}\mathbf{z}||_2^2 \tag{26}$$

$$f_2(\mathbf{u}) = \lambda||\mathbf{u}||_1 + \alpha\sum_{k=1}^{P}\sqrt{\Delta_k}||\mathbf{u}_k||_2 \tag{27}$$

which yields

$$\mathbf{z}(\ell+1) = \underset{\mathbf{z}}{\operatorname{argmin}}\ \frac{1}{2}||\mathbf{y} - \mathbf{W}\mathbf{z}||_2^2 + \frac{\mu}{2}||\mathbf{z} - \mathbf{u}(\ell) - \mathbf{d}(\ell)||_2^2 \tag{28}$$

$$= \left(\mathbf{W}^H\mathbf{W} + \mu\mathbf{I}\right)^{-1}\left(\mathbf{W}^H\mathbf{y} - \mathbf{u}(\ell) - \mathbf{d}(\ell)\right) \tag{29}$$

where $(\cdot)^H$ denotes the Hermitian (conjugate) transpose. It should be noted that the matrix inversion lemma can be used such that the solution can be calculated by solving an $N \times N$ system corresponding to the matrix $\mathbf{W}\mathbf{W}^H + \mathbf{I}/\mu$, i.e.,

$$\left(\mathbf{W}^H\mathbf{W} + \mu I\right)^{-1}\boldsymbol{\kappa} = \frac{\mathbf{y}}{\mu} + 1/\mu\mathbf{W}^H\left(\mathbf{I}/\mu + \mathbf{W}\mathbf{W}^H\right)^{-1}\mathbf{W}\boldsymbol{\kappa} \tag{30}$$

for some vector $\boldsymbol{\kappa} \in \mathbb{C}^p$, thus transforming the $P \times P$ matrix inversion into that of an $N \times N$ matrix inversion. Moreover,

$$\mathbf{u}(\ell+1) = \underset{\mathbf{u}}{\operatorname{argmin}}\, \lambda||\mathbf{u}||_1 + \alpha \sum_{k=1}^{P} \sqrt{\Delta_k}||\mathbf{u}_k||_2 + \frac{\mu}{2}||\mathbf{z}(\ell+1) - \mathbf{u} - \mathbf{d}(\ell)||_2^2 \tag{31}$$

which decouples into P optimization problems as

$$\mathbf{u}_k(\ell+1) = \underset{\mathbf{u_k}}{\operatorname{argmin}}\, \lambda||\mathbf{u}_k||_1 + \alpha\sqrt{\Delta_k}||\mathbf{u}_k||_2 + \frac{\mu}{2}||\mathbf{z}_k(\ell+1) - \mathbf{u}_k - \mathbf{d}_k(\ell)||_2^2 \tag{32}$$

One can easily solve the sub-differential equations

$$\lambda\mathbf{r} + \alpha\sqrt{\Delta_k}\mathbf{s} + \mu(\tilde{\mathbf{z}}(\ell+1) - \tilde{\mathbf{u}}_k - \tilde{\mathbf{d}}(\ell)) = 0 \tag{33}$$

where the notation $\tilde{\mathbf{x}}$ denotes the real valued version of the complex vector $\mathbf{x}$, created as specified in Appendix A, and the vectors $\mathbf{s}$ and $\mathbf{r}$ are real-valued and are defined such that

$$\mathbf{s} = \begin{cases} \frac{\tilde{\mathbf{u}}_k}{||\tilde{\mathbf{u}}_k||_2} & \text{if } \tilde{\mathbf{u}}_k \neq 0 \\ \mathbf{v} & \text{otherwise} \end{cases} \tag{34}$$

with $||\mathbf{v}||_2 \leq 1$, and

$$\begin{bmatrix} r_i \\ r_{i+L_k} \end{bmatrix} = \begin{cases} \frac{\left[\tilde{\mathbf{u}}_{k,i},\, \tilde{\mathbf{u}}_{k,i+L_k}\right]^T}{\left|\left|\left[\tilde{\mathbf{u}}_{k,i},\, \tilde{\mathbf{u}}_{k,i+L_k}\right]\right|\right|_2} & \text{if } \left[\tilde{\mathbf{u}}_{k,i},\, \tilde{\mathbf{u}}_{k,i+L_k}\right]^T \neq 0 \\ \mathbf{p} & \text{otherwise} \end{cases} \tag{35}$$

with $||\mathbf{p}_i||_2 \leq 1$, for $i = 1, \ldots, L_k$, where $\mathbf{a}_{i,j}$ denotes element j of sub-vector i and $[a,\, b]$ denoting a vector with two scalars a and b, and

$$\mathbf{r} = \begin{bmatrix} r_1 & \ldots & r_{2L_k} \end{bmatrix}^T \tag{36}$$

This leads to

$$\mathbf{u}_k(\ell+1) = \bar{\Psi}\left(\Psi\left(\mathbf{z}_k(\ell+1) - \mathbf{d}_k(\ell), \frac{\lambda}{\mu}\right), \frac{\alpha\sqrt{\Delta_k}}{\mu}\right) \tag{37}$$

where $\Psi(\cdot)$ is an element-wise shrinkage function, defined as

$$\Psi(\mathbf{a}, \gamma) = \frac{\max(|\mathbf{a}| - \gamma, 0)}{\max(|\mathbf{a}| - \gamma, 0) + \gamma} \odot \mathbf{a} \tag{38}$$

where the max function acts element-wise on the vector, and $\odot$ denotes the element-wise multiplication of two vectors. Similarly, $\bar{\Psi}(\cdot)$ is a vector shrinkage functions formed as

$$\bar{\Psi}(\mathbf{a}, \gamma) = \frac{\max(||\mathbf{a}||_2 - \gamma, 0)}{\max(||\mathbf{a}||_2 - \gamma, 0) + \gamma}\mathbf{a}$$

The resulting ADMM algorithm for (11), here termed the Pitch Estimation using ℓ_2 norm and Block Sparsity (PEBS$_2$), is summarized in Algorithm 2. For (13), one could similarly define $f_1(\cdot)$ as the sum of all the regularization terms. However, the subdifferential equations can then unfortunately not be solved as easily as before. Instead, we exploit the recent idea introduced in [40], where, by a clever choice of functions the $f_1(\cdot)$ and $f_2(\cdot)$, one may extend (17) to a minimization of a sum of B convex functions, i.e.,

$$\underset{\mathbf{z}}{\text{minimize}} \quad \sum_{k=1}^{B} g_k(\mathbf{H}\mathbf{z}) \tag{39}$$

where $\mathbf{H}_k \in \mathbf{R}^{N \times p}$ are known matrices, and $g_k(\cdot)$ convex functions. This is accomplished by setting $f_1(\mathbf{z}) = 0$, and

$$f_2(\mathbf{G}\mathbf{u}) = \sum_{k=1}^{B} g_k(\mathbf{G}\mathbf{u}) = \sum_{k=1}^{B} g_k(\mathbf{H}_k\mathbf{u}^{(k)}) \tag{40}$$

where

$$\mathbf{G} = \begin{bmatrix} \mathbf{H}_1^T & \dots & \mathbf{H}_K^T \end{bmatrix}^T \tag{41}$$

$$\mathbf{u} = \begin{bmatrix} (\mathbf{u}^{(1)})^T & \dots & (\mathbf{u}^{(K)})^T \end{bmatrix}^T \tag{42}$$

Thereby step 4 in Algorithm 1 is allowed to be decomposed into B independent optimization problems. Rewriting (13) on the form in (39), noting that for this case, $B = 3$, and

$$f_2(\mathbf{G}\mathbf{u}) = \frac{1}{2}||\mathbf{u}^{(1)} - \mathbf{y}|| + \lambda||\mathbf{u}^{(2)}||_1 + \alpha \sum_{k=1}^{P} \sqrt{\Delta_k}||\mathbf{u}_k^{(2)}||_2 + \gamma||\mathbf{u}^{(3)}||_1 \tag{43}$$

where $\mathbf{G} = \begin{bmatrix} \mathbf{A}^T & \mathbf{I} & \mathbf{F}^T \end{bmatrix}^T$, and

$$\mathbf{u} = \begin{bmatrix} (\mathbf{u}^{(1)})^T & (\mathbf{u}^{(2)})^T & (\mathbf{u}^{(3)})^T \end{bmatrix}^T \tag{44}$$

This implies that step 3 in Algorithm 1 can be solved as

$$\mathbf{z}(\ell+1) = \underset{\mathbf{z}}{\operatorname{argmin}} \, ||\mathbf{G}\mathbf{z} - \mathbf{u}(\ell) - \mathbf{d}(\ell)||_2^2 \tag{45}$$

$$= \left[\mathbf{A}^H\mathbf{A} + \mathbf{F}^H\mathbf{F} + \mathbf{I}\right]^{-1} \left(\mathbf{A}^H\boldsymbol{\xi}^{(1)}(\ell) + \boldsymbol{\xi}^{(2)}(\ell) + \mathbf{F}^H\boldsymbol{\xi}^{(3)}(\ell)\right) \tag{46}$$

where $\mathbf{d}$ is decomposed in the same manner as $\mathbf{u}$, and

$$\boldsymbol{\xi}^{(m)}(\ell) \triangleq \mathbf{u}^{(m)}(\ell) + \mathbf{d}^{(m)}(\ell) \tag{47}$$

for $m = 1, 2, 3$. Here, we are mostly interested in situations where the number of parameters far outnumbers the number of measurements, i.e., $N \ll p$. Thus, since (45) needs to be solved at each iteration, one may solve it efficiently using the matrix inversion lemma, i.e.,

$$\mathbf{z}(\ell+1) = \boldsymbol{\chi}(\ell) - \left(\mathbf{F}^H\mathbf{F} + \mathbf{I}\right)^{-1}\mathbf{A}^H \left(\mathbf{I} + \mathbf{A}\left(\mathbf{F}^H\mathbf{F} + \mathbf{I}\right)^{-1}\mathbf{A}^H\right)^{-1}\mathbf{A}\boldsymbol{\chi}(\ell) \tag{48}$$

with

$$\boldsymbol{\chi}(\ell) = \left(\mathbf{F}^H\mathbf{F} + \mathbf{I}\right)^{-1} \left(\mathbf{A}^H\boldsymbol{\xi}^{(1)}(\ell) + \boldsymbol{\xi}^{(2)}(\ell) + \mathbf{F}^H\boldsymbol{\xi}^{(3)}(\ell)\right) \tag{49}$$

where we instead of solving one full $p \times p$ system of equations solve two tridiagonal systems of equations, which may be solved using $\mathcal{O}(p)$ operations [54, p. 153] and one $N \times N$ system of equations. Furthermore, since

$$\left(\mathbf{I} + \mathbf{A}\left(\mathbf{F}^H\mathbf{F} + \mathbf{I}\right)^{-1}\mathbf{A}^H\right)^{-1}\mathbf{A}\boldsymbol{\chi}_k \tag{50}$$

needs to be calculated at each step, the computational complexity can be decreased even further by calculating the Cholesky factor, and at each stop solving two triangular systems of equations. Thus, for a one time cost of $\mathcal{O}(N^3)$ operations, one can at each step solve two triangular systems of equations at cost of $\mathcal{O}(N^2)$ operations. Step 4 in Algorithm 1 thereby decomposes

into three different and decoupled optimization problems; firstly, for the first block,

$$\begin{aligned}\mathbf{u}^{(1)}(\ell+1) &= \underset{\mathbf{u}}{\operatorname{argmin}}\, \frac{1}{2}\left\|\mathbf{u}-\mathbf{y}\right\|_2^2 + \frac{\mu}{2}\left\|\mathbf{A}\mathbf{z}(\ell+1)-\mathbf{u}-\mathbf{d}^{(1)}(\ell)\right\|_2^2 \\ &= \frac{\mathbf{y}-\mu\left(\mathbf{A}\mathbf{z}(\ell+1)-\mathbf{d}^{(1)}(\ell)\right)}{1+\mu} \end{aligned} \tag{51}$$

Secondly, the optimization problem for the second block is equivalent to (31), leading again to

$$\mathbf{u}_k^{(2)}(\ell+1) = \underset{\mathbf{u_k}}{\operatorname{argmin}}\, \lambda||\mathbf{u}||_1 + \alpha\sum_{k=1}^{P}\sqrt{\Delta_k}||\mathbf{u}_k||_2 + \frac{\mu}{2}||\mathbf{z}(\ell+1)-\mathbf{u}-\mathbf{d}^{(2)}(\ell)||_2^2 \tag{52}$$

$$= \bar{\Psi}\left(\Psi\left(\mathbf{z}_k(\ell+1)-\mathbf{d}_k^{(2)}(\ell), \frac{\lambda}{\mu}\right), \frac{\alpha\sqrt{\Delta_k}}{\mu}\right) \tag{53}$$

Finally, the third block can be similarly updated to

$$\mathbf{u}^{(3)}(\ell+1) = \underset{\mathbf{u}}{\operatorname{argmin}}\, \gamma||\mathbf{u}||_1 + \frac{\mu}{2}||\mathbf{F}\mathbf{z}_{k+1}-\mathbf{u}-\mathbf{d}_k^{(3)}||_2^2 \tag{54}$$

$$= \Psi\left(\mathbf{F}\mathbf{z}(\ell+1)-\mathbf{d}^{(3)}(\ell), \frac{\gamma}{\mu}\right) \tag{55}$$

The resulting ADMM algorithm for the block sparse pitch estimation problem, including the TV penalty (PEBS$_2$TV), is summarized in Algorithm 3. Alternatively, if one wish to use a covariance fitting formulation, as discussed above, one may simply change the appropriate step, e.g., the update for $\mathbf{u}_{k+1}^{(1)}$ in Algorithm 3 leads to

$$\begin{aligned}\mathbf{u}^{(1)}(\ell+1) &= \underset{\mathbf{u}}{\operatorname{argmin}}\, \frac{1}{2}\left\|\mathbf{u}-\mathbf{y}\right\|_1 + \frac{\mu}{2}\left\|\mathbf{A}\mathbf{z}(\ell+1)-\mathbf{u}-\mathbf{d}^{(1)}(\ell)\right\|_2^2 \\ &= \mathbf{y} + \Psi\left(\mathbf{A}\mathbf{z}^{(1)}(\ell+1)-\mathbf{d}^{(1)}(\ell), \frac{1}{\mu}\right)\end{aligned}$$

We denote the thus resulting estimators the PEBS$_1$ and PEBS$_1$TV, where the latter includes the TV penalty. The computational cost of each iteration of Algorithm 2 and 3 is, for typical problem dimensions, dominated by calculating $\mathbf{A}\mathbf{x}$ and $\mathbf{A}^H\mathbf{y}$, for various vectors $\mathbf{x}$ and $\mathbf{y}$, and requires considerably

less operations than the $\mathcal{O}(p^3)$ needed for the solvers mentioned earlier. It is worth noting that the cost of the PEBS algorithms may be significantly reduced for signals sampled at equidistant time-points by using fast Fourier transform (FFT) techniques. Further improvements are possible by addressing the choice of the dual variable step size, μ. Instead of tuning it for each problem depending on the typical sizes of the various inputs and outputs, an adaptive approach is possible using the following heuristic [39]: considering the fact that μ can be seen as controlling the relative importance of the dual and primal feasibility condition suggests an adaptive choice by comparing the norms of the primal and dual residuals and adjusting μ appropriately, i.e., after step 9 in Algorithm 3, one may update μ according to

$$\mu(\ell+1) = \begin{cases} \mu(\ell)\tau & \text{if } ||\mathbf{r}(\ell)||_2 > \rho||\mathbf{s}(\ell)||_2 \\ \mu(\ell)/\tau & \text{if } ||\mathbf{s}(\ell)||_2 > \rho||\mathbf{r}(\ell)||_2 \\ \mu(\ell) & \text{otherwise} \end{cases} \tag{56}$$

where τ is the multiplicative change in the step size, and μ set such that the step size is changed to keep the ratio between the norms of the primal and dual residuals within a factor μ. In our experience, setting $\tau = 2$ and $\rho = 10$ results in about an order of magnitude fewer steps being needed. Note that changing μ here does not cause any additional computational cost in any of the above steps, except for the negligible cost of rescaling the dual variables, i.e., $\tilde{d}(\ell+1) = \mu(\ell)/\mu(\ell+1)d(\ell+1)$.

5. Numerical Results

We proceed to examine the robustness and performance of the proposed estimators, using both simulated and real audio signals, comparing with the optimal filtering (Capon), approximative nonlinear least squares (ANLS), and multi-pitch estimator based on subspace orthogonality (ORTH) algorithms [10, 55]. These estimators have in several studies been found to offer state-of-the-art performance, and have freely available implementations, allowing for easily reproducible comparisons in future studies[1]. Initially, examining simulated signals, the performance of the estimates for the different algorithms are computed using 250 Monte-Carlo simulations and

[1]The implementation of the proposed PEBS$_2$TV estimator will be provided online in case of publication.

$N = 160$ samples, wherein the number of harmonics are selected uniformly over $[3, \min(\text{floor}(1/f), 10)]$ in each simulation, where f denotes the fundamental frequency, in order to ensure that all frequencies are below the Nyquist limit. Here, frequencies are given as normalized frequencies with unit cycles/sample, in the interval [0, 1], unless otherwise specified. The signal to noise ratio (SNR), defined as $10log_{10}(||\mathbf{y}||_2/||\mathbf{w}||_2)$, is set to 18 dB, unless otherwise stated. To ensure the best possible performance, the reference methods are allowed perfect *a priori* knowledge of both the number of present sources and their respective number of harmonics, whereas the proposed estimators are only given that the maximum number of harmonics for any present source is 10. All methods are given the same grid size, equivalent to 1000 equally spaced points in [0.025, 0.1].

We begin by examining the performance of the estimators in a case with one source when random harmonics are allowed to be missing. As shown in earlier studies (see, e.g., [10]), the reference methods are well able to estimate the pitch of a single source, but can be expected to suffer somewhat of a loss of performance when the number of assumed harmonics differ from the actual number present in the signal. To illustrate this, we simulate a signal with the fundamental frequency drawn uniformly on $[0.025, 0.05]$, with $L_1 = 10$ with $2 - 8$ harmonics missing at random, with all the amplitudes set to 1 with uniformly distributed phases. The results are shown in Figure 1, illustrating the ratio of estimates for which the estimated pitch is within ± 0.0002, i.e., approximately within two grid points from the true value, for a varying number of missing harmonics. As seen in the figure, it is clear that the PEBS estimators are performing as well as, or even better, than the reference methods. Of the methods, only ORTH is seen to suffer noticeably by the missing harmonics, which is natural due to the resulting loss of orthogonality between the subspaces. It is worth noting that the fundamental frequency is here allowed to be one of the randomly missing harmonics. We have here used $\alpha = c\chi$, $\lambda = (1-c)\chi$, for $c = 0.5$ and $\chi = 0.2$.

Next, we illustrate how the TV penalty influences the performance of the estimate. Figure 2 shows the results for a single pitch signal with fundamental frequency chosen uniformly in [0.04, 0.0625], with four harmonics, where, as before, all the amplitudes are set to 1 with random phases, and the dictionary for both methods is chosen such that a maximum of 8 harmonics are allowed for the frequency range [0.02, 0.1]. The result of this choice of signal and dictionary is that the cost function for PEBS_2 will not be able to distinguish between the block corresponding to f_0 and $f_0/2$ in a consistent manner.

This is clearly visible in the figure, where one can see that the fundamental frequency is only correctly identified in roughly 60 % of the simulation for the PEBS_2 estimator, with noise in the spectrum basically deciding if f_0 or $f_0/2$ is chosen, whereas the PEBS_2TV estimate yields consistent performance for all SNRs. Here, and in all other simulations, γ was set to 0.01.

We proceed with the more interesting case of more than one signal source, forming a signal consisting of two sources with the fundamental frequencies, f_k, drawn uniformly on $[0.025, 0.1]$, where we have ensured that the minimum difference between the frequencies is at least $1/25$ of the frequency range. To illustrate the effect of non-equal amplitudes, the amplitudes are here drawn such that both pitches have equal power, with $a_{i,k} \sim \mathbf{N}(1, 1)$, i.e., Gaussian with expected value one and variance one, with uniformly distributed phase, which also means that no harmonics will be missing, but some might have small amplitudes. Figure 3 shows the ratio of estimates where the estimated pitches are both within two grid points from the true value, for varying SNR, clearly showing the preferable performance of the proposed PEBS algorithms. As seen from the figure, the PEBS_2 estimates achieve almost perfect performance for SNRs greater than 5 dB, whereas the other examined estimators fail to do so, even for larger SNRs. The reference methods thus fail to properly identify the pitches for the two sources, even though being provided perfect *a priori* information of the number of sources and harmonics. This can to some extent be explained by the fact that, being random variables, some of the amplitudes may well be quite small, mimicking the missing harmonics case previously studied. Also, as the fundamental frequency decreases, the harmonics become more closely spaced, implying a more difficult estimation problem.

To examine the effects of closely spaced fundamental frequencies, we proceed to consider the pitches $f_1 = 0.02 + \xi$, where the random variable ξ, uniformly distributed on $[0,\ 0.00005]$ and redrawn for each Monte-Carlo simulation, is added to make sure that the signal is not lying exactly on the grid of proposed fundamental frequencies, and with $f_2 = f_1 + \Delta f$. Here, to clarify the effects of the source separation, $L_1 = 4$ and $L_2 = 4$, $\alpha_{k,l} = 1$, $\forall k, l$, with the amplitudes having a uniformly distributed phase. Figure 4 shows the resulting performance as a function of Δf, again confirming the preferable performance of the proposed estimators. In particular, it is worth noting how the Capon and ORTH estimators suffers loss in performance as frequencies corresponding to the overtones of the fundamental frequencies. Here, the performance of the reference methods can be largely explained

by the difficulty of estimating lower fundamental frequencies. To illustrate this, Figure 5 shows the ratio when selecting larger fundamental frequencies, $f_0 = 0.05$ instead of 0.025 in the previous example. As can be seen in the figure, the more well separated pitches are easier for the reference methods to resolve. As is clear from both figures, the proposed estimator does not suffer this shortcoming, and offer a uniformly preferable performance.

We continue on to examine the robustness to the selection of the user parameters. Figure 6 illustrates the resulting performance as a function of χ for different values of c, for SNR=15 dB, while the other signal parameters are the same as for the signals used for Figure 3. To increase clarity, the results are here only compared to the ORTH estimator, which exhibited the best performance of the reference methods. As shown in the figure, the performance of the PEBS estimate is quite insensitive to the choice of the user parameters, although their relative ratio, typically estimated using a modified cross validation approach, where the prediction of the estimated model is done with a re-estimated LS solution using only the non-zero blocks chosen (see, e.g., [56]), does make some difference in performance. The figure illustrates that a better results was obtained by including the ℓ_1 penalty ($c \neq 0$), as compared to using only the block penalty ($c = 0$).

Turning our attention to actual audio recordings, we consider a real audio signal[2] using a recorded guitar playing in succession three chords, first a single note, then a 2-note chord, and, finally, a 3-note chord. Figures 7-9 show the spectrogram of the recorded signal as well as the resulting PEBS$_2$TV and ORTH estimates, respectively. For this signal, where one may expect a fundamental frequency in the range 80 to 1600 Hz, and with varying number of pitches and harmonics, the f_0 vs $f_0/2$ ambiguity should be expected. As can be seen in the figures, the PEBS$_2$TV method estimates the fundamental frequencies consistently with the actual number of sources, as well as the fundamental frequencies of the underlying notes. Figure 8 also shows the (estimated) scaled standard deviation of the signal, clearly illustrating the initial uncertainty in the measurement when the chord is struck. The dictionary is chosen using the entire span of the fundamental frequency range of a guitar, and the number of harmonics is chosen to be at a maximum 8, c was set to 0.3 and χ was set to equal the standard deviation of the signal. Overall, PEBS$_2$TV manages to find the correct number of pitches and the

[2]The authors are grateful to Mr Tommy Nilsson for this recording.

true fundamental frequency. Since the estimator is not given the number of pitches, artificial fundamental frequency estimates appear when string is struck or damped. This shows the importance of better preprocessing or modeling for music signal applications. Furthermore, the frequency estimate at around 990 Hz might be due to the inharmonicity in the guitar (see, e.g., [44]). For comparison, we in Figure 9 show, the resulting estimates for the ORTH estimator, which was best performing of the reference methods for this signal. The model order was here set using oracle information of the number of pitches and manually tuning the number of pitches to give the best results. As can be seen, the ORTH estimator manages to do reasonably well, with the most troublesome region being between 1 and 1.5 seconds, where several cases of $f_0/2$ or $2f_0$ being chosen instead of the correct fundamental frequency.

Finally, we examine a signal obtained by superimposing two recordings from the SQAM database [57], being a viola and the voice of a female speaker. The viola has a single fundamental frequency of about 131 Hz with roughly 15 overtones, although it may be noted that both the first and fifth harmonics are missing, and several other harmonics are quite small. For the speech signal, we have selected a part of the phrase "to administer", analyzing the two vowels "o" and "a", corresponding to the first third of the spectrogram in Figure 10. To allow the speech signal to be reasonably stationary, we use (non-overlapping and un-windowed) 20 ms time windows. During the examined time period, the voice varies considerable, and the number of harmonics can be seen to vary over the segments from one to eight with a fundamental frequency varying between 180 and 220 Hz. The spectrogram of the resulting signal is shown in Figure 10. To allow for the range of possible pitch frequencies a viola and a female voice may be expected to span, the dictionary was selected to cover the frequency range 130–1200 Hz, using 500 grid points, with the maximum number of harmonics set to $L_{\max} = 15$. Figures 11 and 12 show the resulting pitch estimates for PEBS$_2$TV and the ORTH estimator, respectively. Here, ORTH has been allowed oracle knowledge of the number of harmonics of each source, as well as the number of sources. As can be seen from the figures, the PEBS$_2$TV estimator is able to correctly identify the two pitch signals throughout, except in the transition period when the speech signal is too weak to be detected, whereas the ORTH estimate gives poor pitch estimates for the latter part of the signal, where it yields pitch estimates which are multiples of the correct pitch, corresponding to the higher order overtones. As the PEBS$_2$TV estimator does not assume prior knowledge of

the number of sources, it may yield spurious pitches. This may be seen, for instance, at time 0.15 s, where a (weak) third pitch appears. By tuning the estimator better, or by allowing for information from previous frames, for instance via pitch tracking (see, e.g., [58]), this may easily be remedied.

6. Conclusions

In this work, we introduced the idea of using block sparsity in the estimation of the fundamental frequencies of a multi-pitch signal. Formulating the estimation as a sum of a fitting term and convex sparsity inducing norms, ensuring a block sparse solution, the proposed algorithm is shown to offer significantly improved performance as compared to a range of state-of-the-art multi-pitch estimators. Furthermore, by including a total variation penalty on each block, the algorithm avoids the f_0 vs $f_0/2$ ambiguity that many estimators suffer from. The algorithm is shown to be capable of handling issues such as missing harmonics as well as closely spaced fundamental frequencies. Furthermore, novel ADMM algorithms are devised for the entailing optimizations, resulting in a iterative dual ascent method, where each step has a simple closed form expression that scales well with the problem dimensions.

A.

Insight into how the penalty term in (9) induces a block sparse solution can be gained by studying the sub-differential equations of the equivalent real-valued cost function (see also [43]), which may be expressed as

$$\tilde{\mathbf{W}}_\ell^T \left(\tilde{\mathbf{y}} - \sum_{k=1}^{P} \tilde{\mathbf{W}}_k \tilde{\mathbf{a}}_k \right) + \alpha \mathbf{s}_\ell = 0 \tag{57}$$

for $\ell = 1, 2, \ldots, P$, where $\mathbf{s}_\ell$ is either a vector such that $||\mathbf{s}_\ell||_2 \leq 1$, or equal to $\tilde{\mathbf{a}}_\ell / ||\tilde{\mathbf{a}}_\ell||$, depending on if $\tilde{\mathbf{a}}_\ell = 0$ or not, $\tilde{\mathbf{W}}$ is the real counterpart of $\mathbf{W}$, created such that

$$\tilde{\mathbf{W}}_\ell = \begin{bmatrix} \Re\mathfrak{e}\{\mathbf{W}_\ell\} & -\Im\mathfrak{m}\{\mathbf{W}_\ell\} \\ \Im\mathfrak{m}\{\mathbf{W}_\ell\} & \Re\mathfrak{e}\{\mathbf{W}_\ell\} \end{bmatrix}$$

where $\Re\mathfrak{e}\{\cdot\}$ and $\Im\mathfrak{m}\{\cdot\}$ denote the real and imaginary part of a matrix, and $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{a}}$ are formed similarly, i.e.,

$$\begin{aligned} \tilde{\mathbf{y}} &= \begin{bmatrix} \Re\mathfrak{e}\{\mathbf{y}\} & \Im\mathfrak{m}\{\mathbf{y}\} \end{bmatrix} \\ \tilde{\mathbf{a}}_\ell &= \begin{bmatrix} \Re\mathfrak{e}\{\mathbf{a}_\ell\} & \Im\mathfrak{m}\{\mathbf{a}_\ell\} \end{bmatrix} \end{aligned}$$

Thus, for any minimizing vector $\breve{\mathbf{a}}$, a necessary and sufficient condition for a sub-vector, or block, $\breve{\mathbf{a}}_\ell$, to be zero is that [43]

$$\left\| \tilde{\mathbf{W}}_\ell^T \left(\tilde{\mathbf{y}} - \sum_{k=1}^{P} \tilde{\mathbf{W}}_k \breve{\breve{\mathbf{a}}}_k \right) \right\|_2 < \alpha \tag{58}$$

which shows the (block) sparsifying effect of the (block) 2-norm. Note further that if the inequality does not hold, $\tilde{\mathbf{a}}_\ell$ could have been found by solving

$$\tilde{\mathbf{a}}_\ell = \left(\tilde{\mathbf{W}}_\ell^T \tilde{\mathbf{W}}_\ell + \alpha / ||\tilde{\mathbf{a}}_\ell|| \right)^{-1} \tilde{\mathbf{W}}_\ell^T \left(y - \sum_{k \neq \ell} \tilde{\mathbf{W}}_k \tilde{\mathbf{a}}_k \right) \tag{59}$$

This can be recognized as being similar to the solution of a Tikhonov regularized LS, or ridge regression, solution which is known to lack a sparsifying effect. Thus, if the block is non-zero, one may expect each element in the block to be non-zero.

B.

Similarily as in Appendix A, the sparsity of the solution of (11) may be understood by studying the subdifferential equations for the equivalent real-valued problem, which are given by

$$\tilde{\mathbf{W}}_\ell^T \left(\tilde{\mathbf{y}} - \sum_{k=1}^{P} \tilde{\mathbf{W}}_k \tilde{\mathbf{a}}_k \right) + \alpha \mathbf{s}_\ell + \lambda \mathbf{r}_\ell = 0 \tag{60}$$

for $\ell = 1, \ldots, P$, where $\mathbf{s}_\ell$ and $\mathbf{r}_\ell$ are real-valued vectors defined such that

$$\mathbf{s}_\ell = \begin{cases} \frac{\tilde{\mathbf{a}}_\ell}{||\tilde{\mathbf{a}}_\ell||_2} & \text{if } \tilde{\mathbf{a}}_\ell \neq 0 \\ \mathbf{v} & \text{otherwise} \end{cases} \tag{61}$$

where $||\mathbf{v}||_2 \leq 1$, and

$$\begin{bmatrix} r_{\ell,i} \\ r_{\ell,i+L_\ell} \end{bmatrix} = \begin{cases} \frac{\left[\tilde{\mathbf{a}}_{\ell,i}, \tilde{\mathbf{a}}_{k,i+L_\ell}\right]^T}{\left\| \left[\tilde{\mathbf{a}}_{\ell,i}, \tilde{\mathbf{a}}_{\ell,i+L_\ell}\right] \right\|_2} & \text{if } \left[\tilde{\mathbf{a}}_{\ell,i},\ \tilde{\mathbf{a}}_{\ell,i+L_\ell}\right]^T \neq 0 \\ \mathbf{p}_i & \text{otherwise} \end{cases} \tag{62}$$

with $||\mathbf{p}_i||_2 \leq 1$, for $i = 1, \ldots, L_k$, where $\mathbf{a}_{i,j}$ denotes element j of sub-vector i, $[a,\ b]$ a vector with two scalars a and b, and

$$\mathbf{r}_\ell = \begin{bmatrix} r_{\ell,1} & \ldots & r_{\ell,2L_k} \end{bmatrix}^T \tag{63}$$

This implies that for any minimizing vector $\breve{\mathbf{a}}$, it holds that $\breve{\mathbf{a}}_\ell = 0$ if

$$\left\| \tilde{\mathbf{W}}_\ell^T \left(\tilde{\mathbf{y}} - \sum_{k=1}^{P} \tilde{\mathbf{W}}_k \breve{\breve{\mathbf{a}}}_k \right) - \lambda \mathbf{r} \right\|_2 \leq \alpha \tag{64}$$

or, equivalently, if

$$\sum_{k=1}^{L_\ell} \left\| \mathbf{z}_k (\|\mathbf{z}_k\|_2 - \lambda)^+ \right\|_2^2 \leq \alpha^2 \tag{65}$$

where $\mathbf{z}_k$ is a vector composed of the elements k and $k + L_\ell$ of the vector

$$\mathbf{z} = \tilde{\mathbf{W}}_\ell^T \left(\tilde{\mathbf{y}} - \sum_{k=1}^{P} \tilde{\mathbf{W}}_k \breve{\breve{\mathbf{a}}}_k \right) \tag{66}$$

Interestingly, but perhaps not surprisingly, this is a similar solution as one would obtain from the analysis of the real-valued version of (10) analyzed in [43]. However, in this case, the analysis holds for any kind of non-overlapping sub-division of the sub-vectors, not only into the two variables corresponding to the same complex variables. This insight was used in [59] to generalize the above results to the case of multiple measurements vectors (array) case.

References

[1] S. I. Adalbjörnsson, A. Jakobsson, M. G. Christensen, Estimating Multiple Pitches Using Block Sparsity, in: 38th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Vancouver, 2013.

[2] P. Kroon, W. B. Kleijn, Linear-prediction based analysis-by-synthesis coding, in: W. B. Kleijn, K. K. Paliwal (Eds.), Speech Coding and Synthesis, Elsevier, Berlin, Germany, 1995, Ch. 3, pp. 79–119.

[3] G. Tzanetakis, P. Cook, Musical genre classification of audio signals, IEEE Trans. Acoust., Speech, Signal Process. 10 (5) (2002) 293–302.

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

[5] M. Müller, D. P. W. Ellis, A. Klapuri, G. Richard, Signal Processing for Music Analysis, IEEE Journal of Selected Topics in Signal Processing 5 (6) (2011) 1088–1110.

[6] W. Hess, Pitch Determination of Speech Signals, Springer, Berlin, 1983.

[7] H. Li, P. Stoica, J. Li, Computationally Efficient Parameter Estimation for Harmonic Sinusoidal Signals, Signal Processing 80 (2000) 1937–1944.

[8] A. de Cheveigné, H. Kawahara, YIN, a fundamental frequency estimator for speech and music, J. Acoust. Soc. Amer. 111 (4) (2002) 1917–1930.

[9] K. W. Chan, H. C. So, Accurate frequency estimation for real harmonic sinusoids, IEEE Signal Processing Letters 11 (7) (2004) 609–612.

[10] M. G. Christensen, P. Stoica, A. Jakobsson, S. H. Jensen, Multi-pitch estimation, Signal Processing 88 (4) (2008) 972–983.

[11] J. X. Zhang, M. G. Christensen, S. H. Jensen, M. Moonen, A Robust and Computationally Efficient Subspace-Based Fundamental Frequency Estimator 18 (3) (2010) 487–497.

[12] Z. Zhou, H. C. So, F. K. W. Chan, Optimally Weighted Music Algorithm for Frequency Estimation of Real Harmonic Sinusoids, in: 37th IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Kyoto, 2012.

[13] T. Tolonen, M. Karjalainen, A computationally efficient multipitch analysis model 8 (6) (2000) 708–716.

[14] R. Gribonval, E. Bacry, Harmonic decomposition of audio signals with matching pursuit, IEEE Transactions on Signal Processing 51 (1) (2003) 101–111.

[15] A. Klapuri, Multiple fundamental frequency estimation based on harmonicity and spectral smoothness, IEEE Trans. Acoust., Speech, Signal Process. 11 (6) (2003) 804–816.

[16] S. S. Abeysekera, Multiple pitch estimation of poly-phonic audio signals in a frequency-lag domain using the bispectrum, in: Proc. IEEE International Symposium on Circuits and Systems, Vol. 14, 2004, pp. 469–472.

[17] M. D. Plumbley, S. A. Abdallah, T. Blumensath, M. E. Davies, Sparse representations of polyphonic music, Signal Processing 86 (3) (2006) 417–431.

[18] J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigne, S. Sagayama, Single and Multiple Contour Estimation Through Parametric Spectrogram Modeling of Speech in Noisy Environments 15 (4) (2007) 1135–1145.

[19] V. Emiya, R. Badeau, B. David, Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle 18 (6) (2010) 1643–1654.

[20] E. Benetos, S. Dixon, Joint Multi-Pitch Detection Using Harmonic Envelope Estimation for Polyphonic Music Transcription, IEEE Journal of Selected Topics in Signal Processing 5 (6) (2011) 1111–1123.

[21] A. Koretz, J. Tabrikian, Maximum A Posteriori Probability Multiple-Pitch Tracking Using the Harmonic Model 19 (7) (2011) 2210–2221.

[22] C. Lee, Y. Yang, H. H. Chen, Multipitch Estimation of Piano Music by Exemplar-Based Sparse Representation, IEEE Transactions on Multimedia 14 (3) (2012) 608–618.

[23] M. Genussov, I. Cohen, Multiple fundamental frequency estimation based on sparse representations in a structured dictionary, Digit. Signal Process. 23 (1) (2013) 390–400.

[24] F. Huang, T. Lee, Pitch Estimation in Noisy Speech Using Accumulated Peak Spectrum and Sparse Estimation Technique 21 (1) (2013) 99–109.

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

[26] D. Donoho, Compressed Sensing, IEEE Transactions on Information Theory 52 (2006) 1289 –1306.

[27] R. Tibshirani, Regression shrinkage and selection via the Lasso, Journal of the Royal Statistical Society B 58 (1) (1996) 267–288.

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

[29] I. F. Gorodnitsky, B. D. Rao, Sparse Signal Reconstruction from Limited Data Using FOCUSS: A Re-weighted Minimum Norm Algorithm, IEEE Transactions on Signal Processing 45 (3) (1997) 600–616.

[30] P. Stoica, P. Babu, Sparse Estimation of Spectral Lines: Grid Selection Problems and Their Solutions, IEEE Transactions on Signal Processing 60 (2) (2012) 962–967.

[31] T. Nilsson, S. I. Adalbjörnsson, N. R. Butt, A. Jakobsson, Multi-Pitch Estimation of Inharmonic Signals, in: European Signal Processing Conference, Marrakech, 2013.

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

[33] C. D. Austin, J. N. Ash, R. L. Moses, Dynamic Dictionary Algorithms for Model Order and Parameter Estimation, IEEE Transactions on Signal Processing 61 (20) (2013) 5117–5130.

[34] J. Swärd, S. I. Adalbjörnsson, A. Jakobsson, High Resolution Sparse Estimation of Exponentially Decaying Signals, in: 39th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Florence, Italy, 2014.

[35] M. Yuan, Y. Lin, Model Selection and Estimation in Regression with Grouped Variables, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (1) (2006) 49–67.

[36] Y. V. Eldar, P. Kuppinger, H. Bolcskei, Block-Sparse Signals: Uncertainty Relations and Efficient Recovery, Signal Processing, IEEE Transactions on 58 (6) (2010) 3042–3054.

[37] X. Lv, G. Bi, C. Wan, The Group Lasso for Stable Recovery of Block-Sparse Signal Representations, IEEE Transactions on Signal Processing 59 (4) (2011) 1371–1382.

[38] A. Juditsky, F. Karzan, A. Nemirovski, B. Polyak, Accuracy guaranties for ℓ_1 recovery of block-sparse signals, Annals of Statistics 40 (2012) 3077–3107.

[39] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers, Found. Trends Mach. Learn. 3 (1) (2011) 1–122.

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

[41] Y. Chi, L. L. Scharf, A. Pezeshki, A. R. Calderbank, Sensitivity to Basis Mismatch in Compressed Sensing, IEEE Transactions on Signal Processing 59 (5) (2011) 2182 –2195.

[42] J. Fang, J. Li, Y. Shen, H. Li, S. Li, Super-resolution compressed sensing: An iterative reweighted algorithm for joint parameter learning and sparse signal recovery, IEEE Signal Processing Letters 21 (6) (2014) 761–765.

[43] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, A Sparse-Group Lasso, Journal of Computational and Graphical Statistics 22 (2) (2013) 231–245.

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

[45] P. Babu, Spectral Analysis of Nonuniformly Sampled Data and Applications, Ph.D. thesis, Uppsala University (2012).

[46] C. R. Rojas, D. Katselis, H. Hjalmarsson, A Note on the SPICE Method, IEEE Transactions on Signal Processing 61 (18) (2013) 4545–4551.

[47] P. Stoica, P. Babu, J. Li, SPICE : a novel covariance-based sparse estimation method for array processing, IEEE Transactions on Signal Processing 59 (2) (2011) 629 –638.

[48] R. Chartrand, B. Wohlberg, A Nonconvex ADMM Algorithm for Group Sparsity with Sparse Groups, in: 38th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Vancouver, Canada, 2013.

[49] E. J. Candes, M. B. Wakin, S. Boyd, Enhancing Sparsity by Reweighted l_1 Minimization, Journal of Fourier Analysis and Applications 14 (5) (2008) 877–905.

[50] X. Tan, W. Roberts, J. Li, P. Stoica, Sparse Learning via Iterative Minimization With Application to MIMO Radar Imaging, IEEE Transactions on Signal Processing 59 (3) (2011) 1088–1101.

[51] P. Stoica, Y. Selén, Model-order Selection — A Review of Information Criterion Rules, IEEE Signal Processing Magazine 21 (4) (2004) 36–47.

[52] J. F. Sturm, Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones, Optimization Methods and Software 11-12 (1999) 625–653.

[53] R. H. Tutuncu, K. C. Toh, M. J. Todd, Solving semidefinite-quadratic-linear programs using SDPT3, Mathematical Programming Ser. B 95 (2003) 189–217.

[54] G. H. Golub, C. F. V. Loan, Matrix Computations, 3^{rd} Edition, The John Hopkins University Press, 1996.

[55] M. Christensen, P. Stoica, A. Jakobsson, S. Jensen, The Multi-Pitch Estimation Problem: some New Solutions, in: 32nd IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vol. 3, 2007, pp. III–1221–III–1224.

[56] N. Meinshausen, Relaxed lasso, Computational Statistics and Data Analysis (2007) 374–393.

[57] Sound Quality Assessment Materical Recodings for Subjective Tests, Tech. rep., European Broadcasting Union (1988).

[58] M. G. Christensen, A Method for Low-Delay Pitch Tracking and Smoothing, in: 37th IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2012, pp. 345–348.

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

Algorithm 1 The general ADMM algorithm

1: Initiate $\mathbf{z} = \mathbf{z}(0), \mathbf{u} = \mathbf{u}(0)$, and $\ell = 0$
2: **repeat**
3: $\quad \mathbf{z}(\ell+1) = \underset{\mathbf{z}}{\text{argmin}}\, f_1(\mathbf{z}) + \frac{\mu}{2}||\mathbf{G}\mathbf{z} - \mathbf{u}(\ell) - \mathbf{d}(\ell)||_2^2$
4: $\quad \mathbf{u}(\ell+1) = \underset{\mathbf{u}}{\text{argmin}}\, f_2(\mathbf{u}) + \frac{\mu}{2}||\mathbf{G}\mathbf{z}(\ell+1) - \mathbf{u} - \mathbf{d}(\ell)||_2^2$
5: $\quad \mathbf{d}(\ell+1) = \mathbf{d}(\ell) - (\mathbf{G}\mathbf{z}(\ell+1) - \mathbf{u}(\ell+1))$
6: $\quad \ell \leftarrow \ell + 1$
7: **until** convergence

Algorithm 2 PEBS_2 via ADMM

Initiate $\mathbf{z} = \mathbf{z}(0), \mathbf{u} = \mathbf{u}(0)$, and $\ell := 0$

repeat

$\quad \mathbf{z}(\ell+1) = \left(\mathbf{W}^H\mathbf{W} + \mu\mathbf{I}\right)^{-1}(\mathbf{y} - \mathbf{u}(\ell) - \mathbf{d}(\ell))$

$\quad \mathbf{u}_k(\ell+1) = \bar{\Psi}\left(\Psi\left(\mathbf{z}_k(\ell+1) - \mathbf{d}_k(\ell+1), \frac{\lambda}{\mu}\right), \frac{\alpha\sqrt{\Delta_k}}{\mu}\right)$ for $k = 1, \ldots, P$

$\quad \mathbf{d}(\ell+1) = \mathbf{d}(\ell) - (\mathbf{z}(\ell+1) - \mathbf{u}(\ell+1))$

$\quad \ell \leftarrow \ell + 1$

until convergence

Algorithm 3 PEBS$_2$TV via ADMM

Initiate $\mathbf{z} = \mathbf{z}(0), \mathbf{u} = \mathbf{u}(0)$, and $\ell := 0$

repeat

$\mathbf{z}(\ell) = \left[\mathbf{A}^H\mathbf{A} + \mathbf{F}^H\mathbf{F} + \mathbf{I}\right]^{-1}\left(\mathbf{A}^H\boldsymbol{\xi}^{(1)}(\ell) + \boldsymbol{\xi}^{(2)}(\ell) + \mathbf{F}^H\boldsymbol{\xi}^{(3)}(\ell)\right)$

$\mathbf{u}^{(1)}(\ell+1) = \frac{\mathbf{y}-\mu\left(\mathbf{A}\mathbf{z}(\ell+1)-\mathbf{d}^{(1)}(\ell)\right)}{1+\mu}$

$\mathbf{u}_k^{(2)}(\ell+1) = \bar{\Psi}\left(\Psi\left(\mathbf{z}_k(\ell+1) - \mathbf{d}_k^{(2)}(\ell), \frac{\lambda}{\mu}\right), \frac{\alpha\sqrt{\Delta_k}}{\mu}\right)$ for $k = 1, \ldots, P$

$\mathbf{u}^{(3)}(\ell+1) = \Psi\left(\mathbf{F}\mathbf{z}(\ell+1) - \mathbf{d}^{(3)}(\ell), \frac{\gamma}{\mu}\right)$

$\mathbf{d}(\ell+1) = \mathbf{d}(\ell) - (\mathbf{G}\mathbf{z}(\ell+1) - \mathbf{u}(\ell+1))$

$\ell \leftarrow \ell + 1$

until convergence

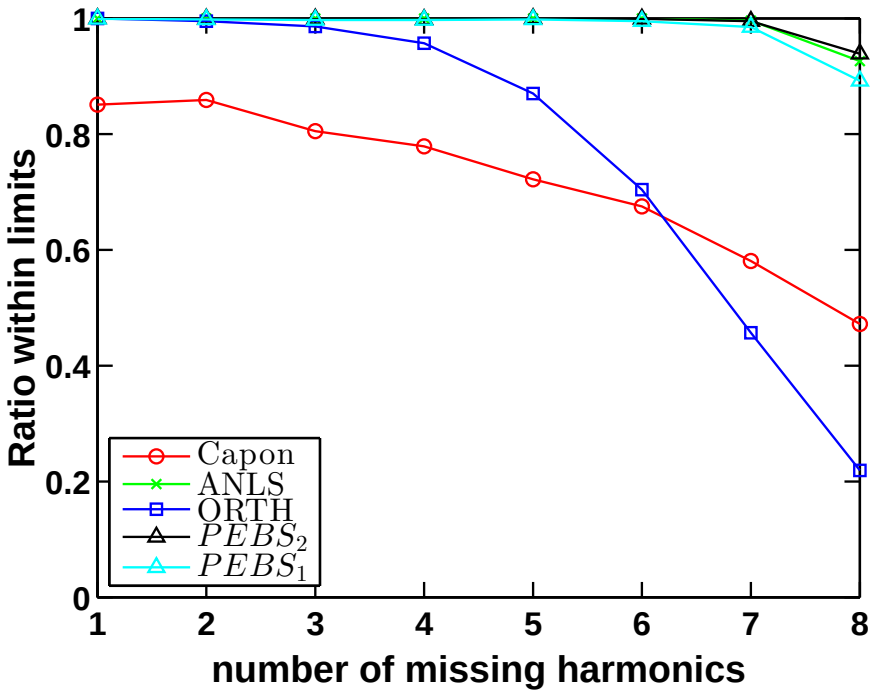

Figure 1: Ratio of estimated pitches where the fundamental frequency lies at most 0.0002 from the ground truth, plotted as a function of the number of harmonics that are missing for $\alpha = \lambda = 0.5\chi$ and $\chi = 0.2$. The fundamental frequency is uniformly distributed on $[0.025, 0.05]$.

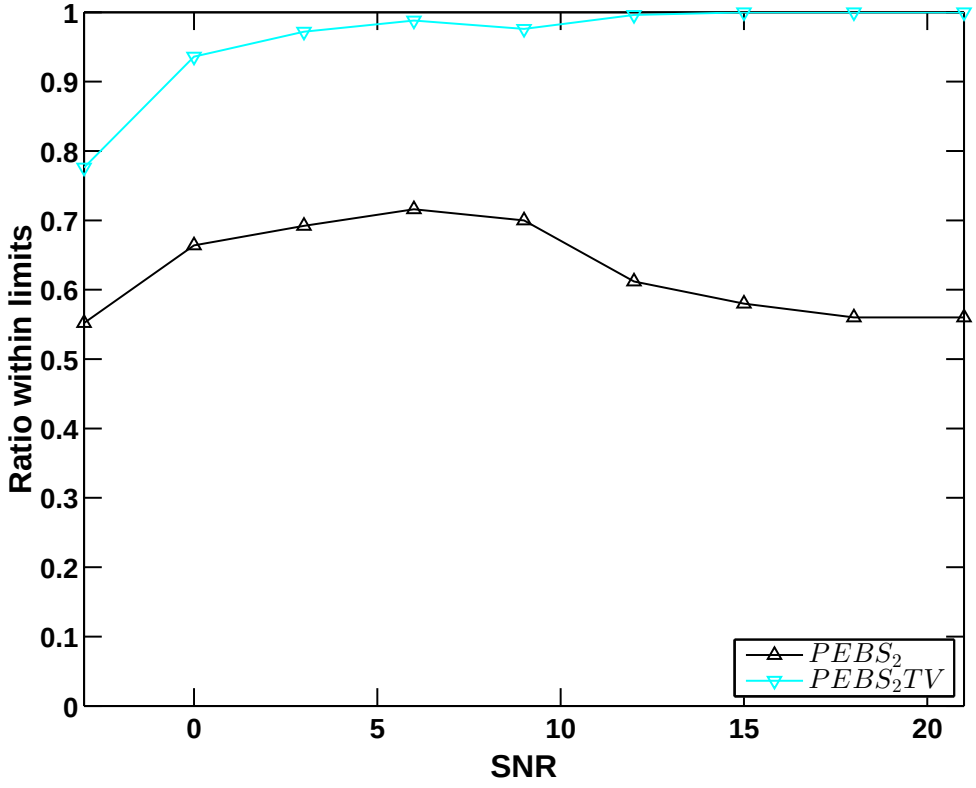

Figure 2: Ratio of estimated pitches where both fundamental frequencies lie at most two grid points from the ground truth, plotted as a function of SNR. The dictionary and signal are chosen such that there is ambiguity in the choice of f_0 vs $f_0/2$.

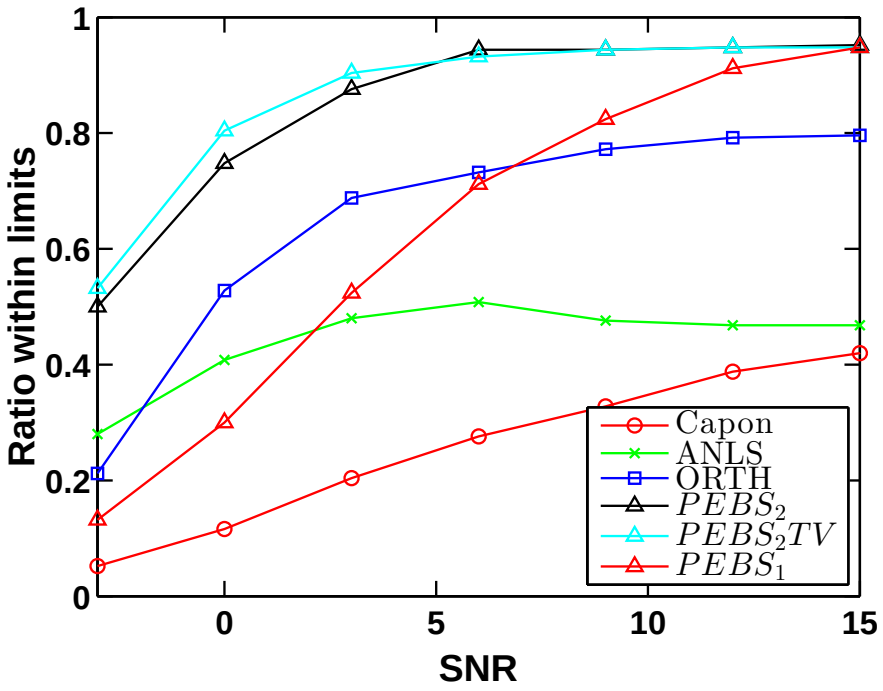

Figure 3: Ratio of estimated pitches where both fundamental frequencies lie at most two gridpoints from the ground truth, plotted as a function of SNR for $\alpha = \lambda = 0.5\chi$ and $\chi = 2.1\sigma_e$. The fundamental frequency is uniformly distributed on $[0.025, 0.1]$.

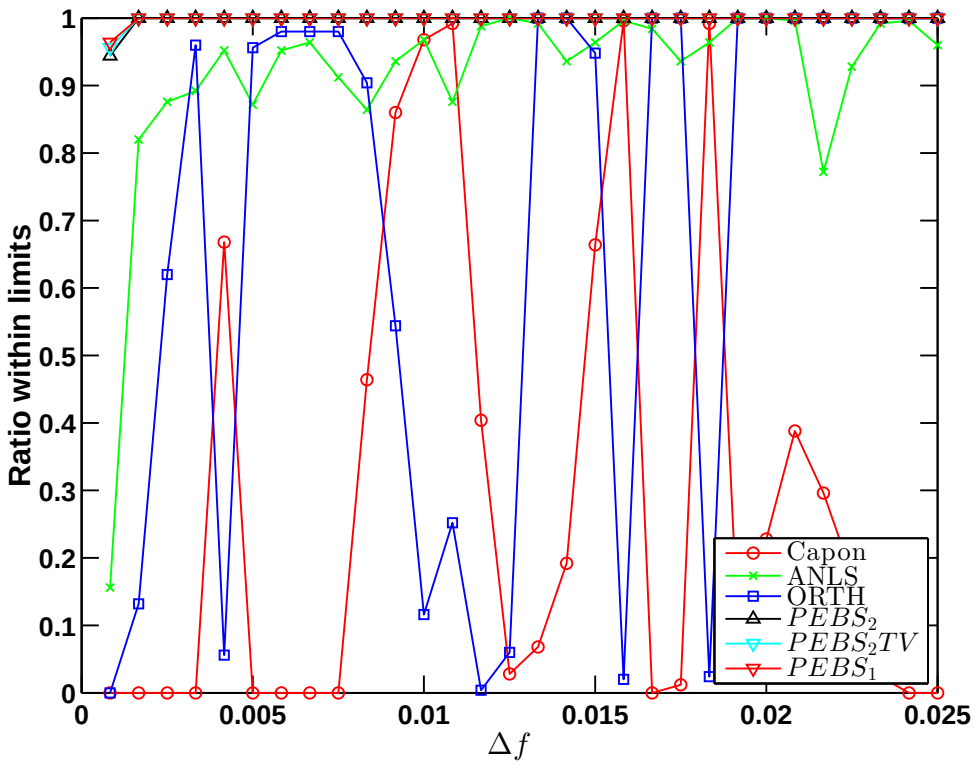

Figure 4: Ratio of estimated pitches where both fundamental frequencies lie at most two grid points from the ground truth, plotted as a function of Δf, for $f_0 = 0.025$, $\alpha = \lambda = 0.5\chi$, $L_1 = 7, L_2 = 5$ and $\chi = 0.2$.

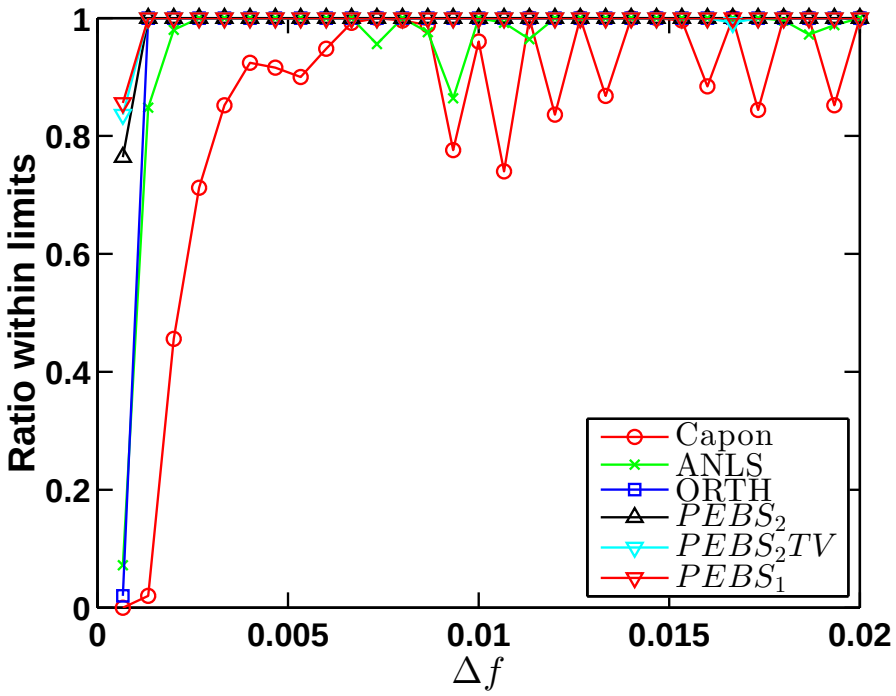

Figure 5: Ratio of estimated pitches where both fundamental frequencies lie at most two grid points from the ground truth, plotted as a function of Δf, for $f_0 = 0.05$, $\alpha = \lambda = 0.5\chi$, $L_1 = 7, L_2 = 5$ and $\chi = 0.2$.

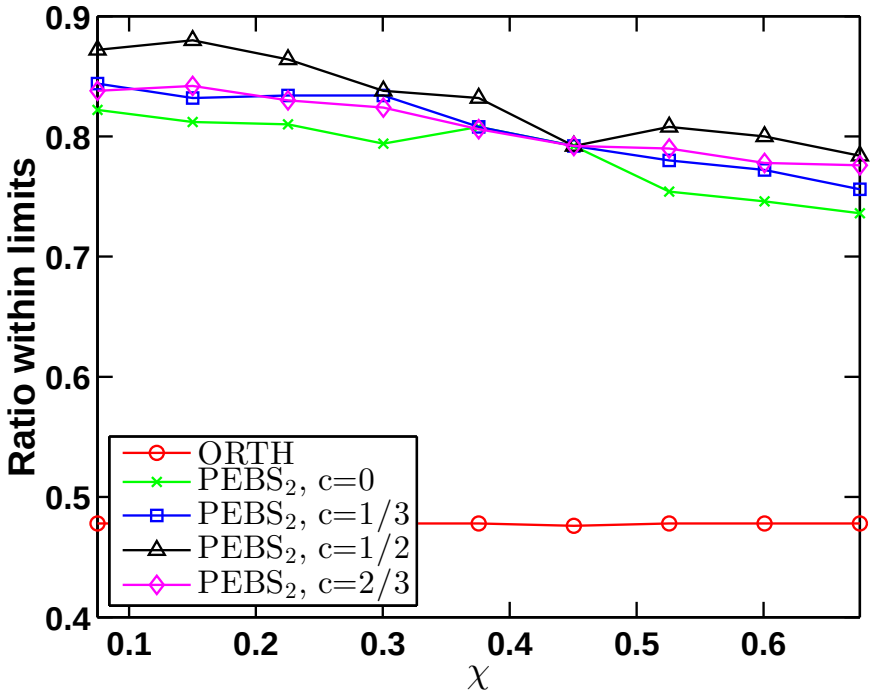

Figure 6: Ratio of estimated pitches where both fundamental frequencies lie at most two grid points from the ground truth, plotted as a function of χ for $\alpha = c\chi$, $\lambda = (1-c)\chi$, for $c \in \{0, 1/2, 1/3, 2/3\}$.

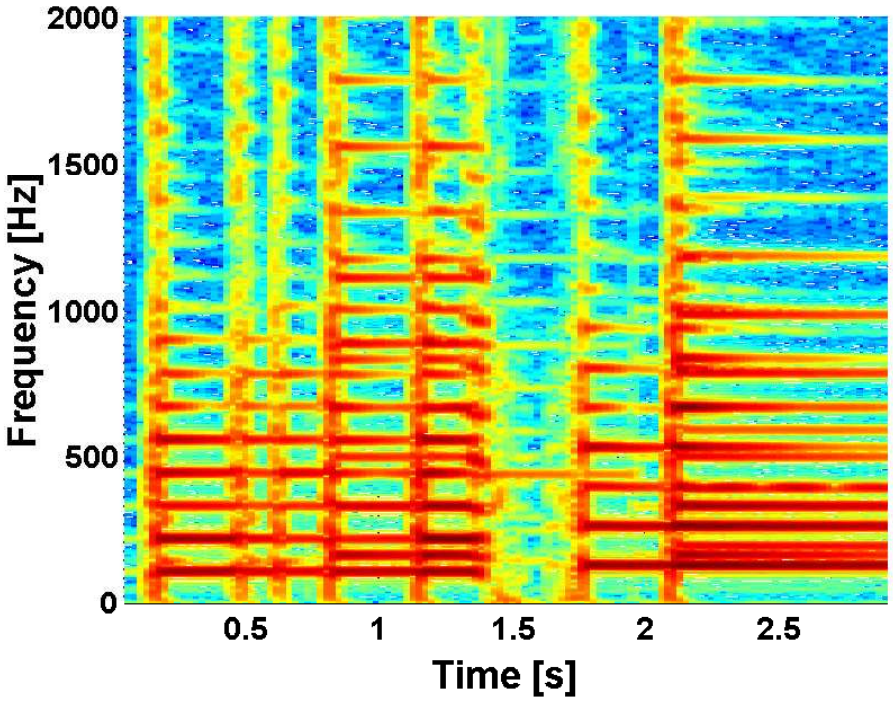

Figure 7: Spectrogram of recorded guitar sound.

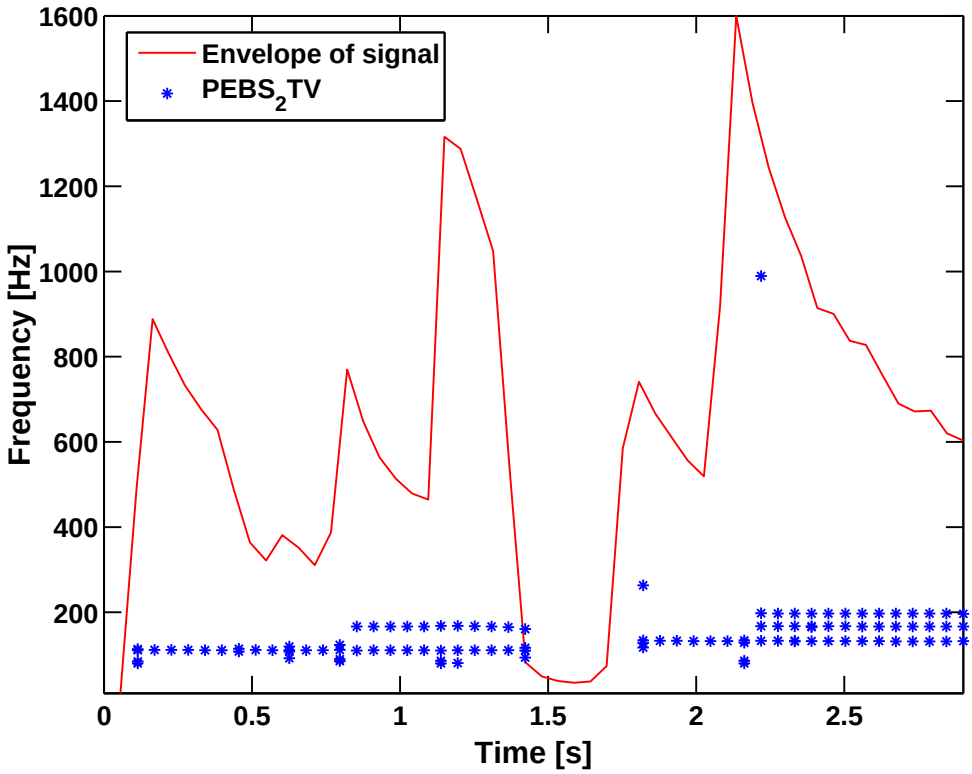

Figure 8: The PEBS estimate of the guitar recording, showing that the correct number of pitches and their corresponding frequencies are revealed. The scaled standard deviation of the signal is superimposed to illustrate at what time points the notes are struck or muted.

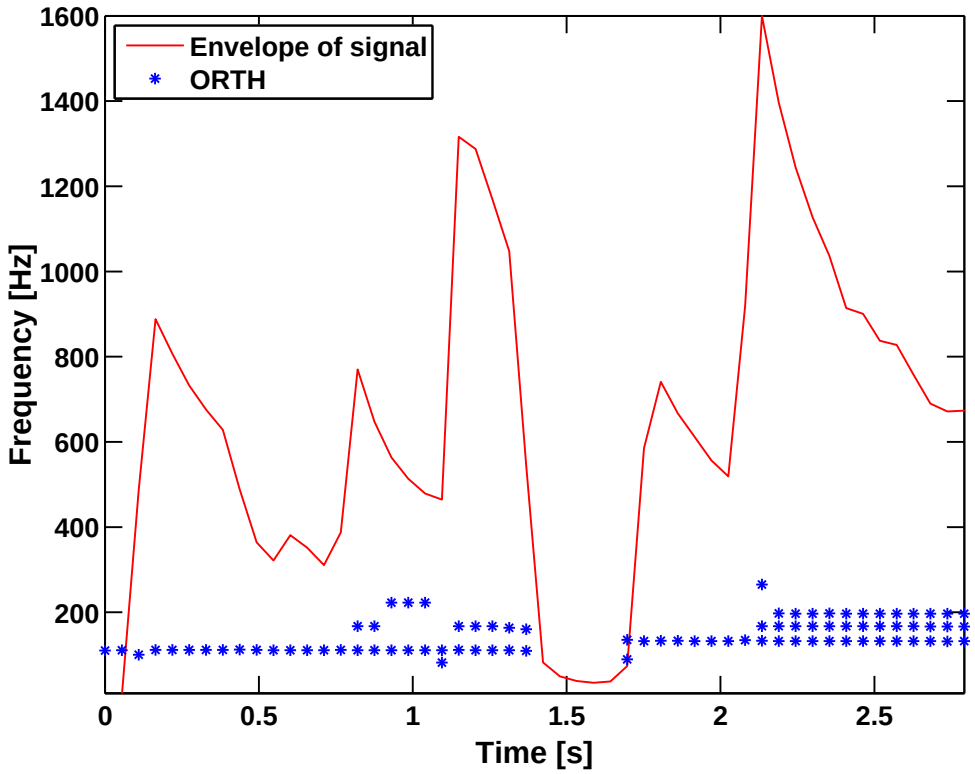

Figure 9: The ORTH estimate of the guitar recording, using oracle information of the model-orders. The scaled standard deviation of the signal is superimposed to illustrate at what time points the notes are struck or muted.

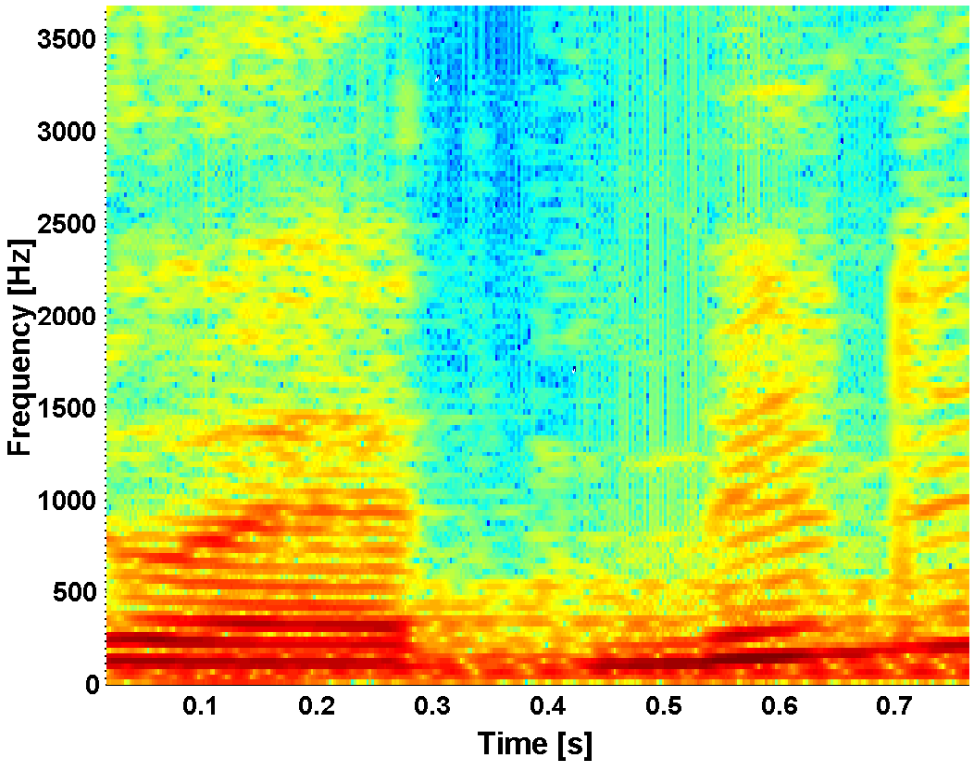

Figure 10: Spectrogram of recorded speech and viola.

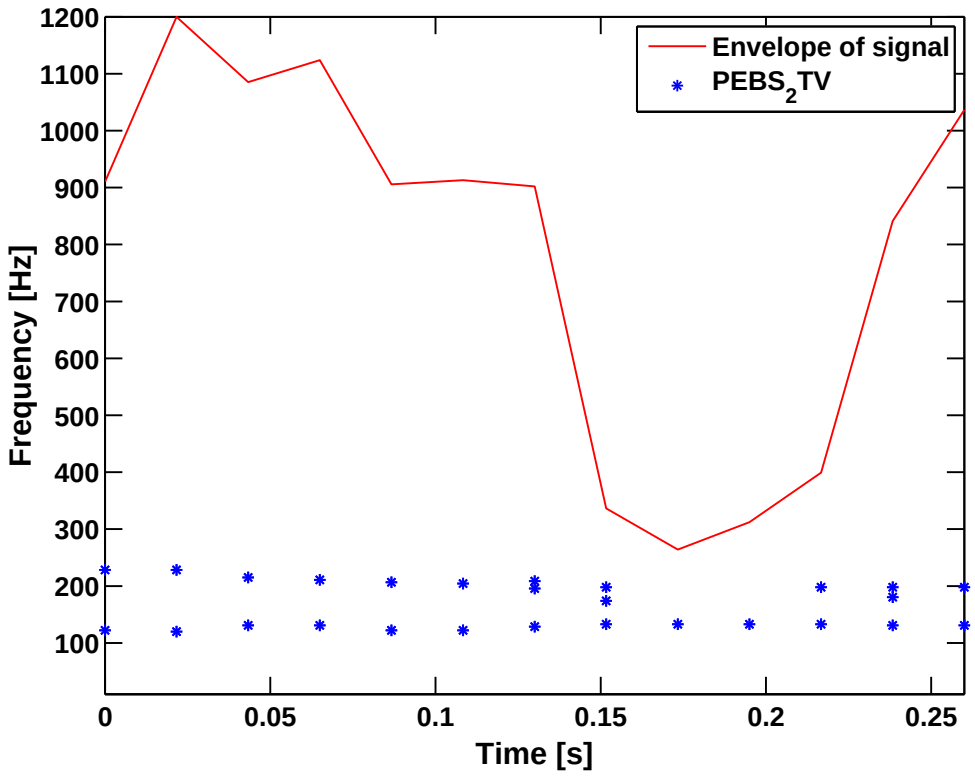

Figure 11: The PEBS_2TV estimate of the speech and viola recording. The scaled standard deviation of the signal is superimposed to illustrate at what time points the voice is silent.

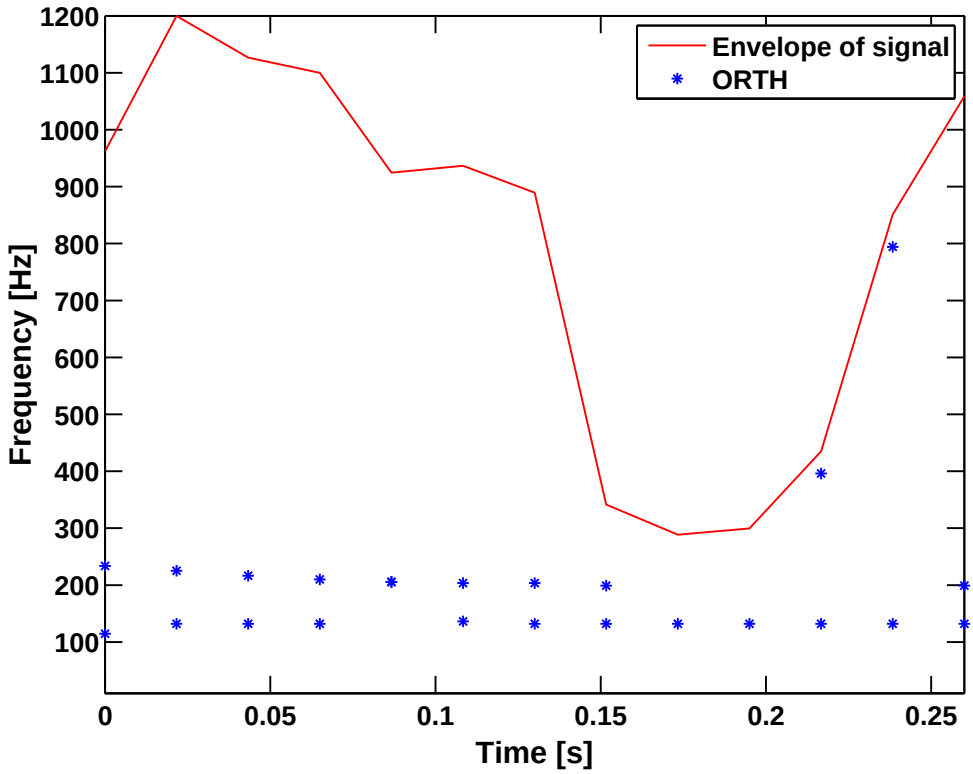

Figure 12: The ORTH estimate of the speech and viola recording, using oracle information of the model-orders. The scaled standard deviation of the signal is superimposed to illustrate at what time points the voice is silent.