Robust Multi-Pitch Estimation via Optimal Transport Clustering

Anton Björkman
Dept. of Information and Communications Engineering
Aalto University
Espoo, Finland
anton.bjorkman@aalto.fi

Filip Elvander
Dept. of Information and Communications Engineering
Aalto University
Espoo, Finland
filip.elvander@aalto.fi

***Abstract*—In this work, we consider the multi-pitch estimation problem, i.e., to estimate multiple sets of harmonically related sinusoids from noisy measurements. We propose to phrase this as a clustering problem with indirect measurements, where we simultaneously infer the spectral content of the signal and group its power into a small set of harmonic structures. The grouping is enforced using a regularization function building on optimal transport theory. The resulting estimator is formulated in terms of the solution of a convex optimization problem, and we present an efficient algorithm implementing the estimator. In numerical experiments, we show that the proposed estimator displays competitive performance as compared to the state-of-the-art. In particular, the proposed estimator is shown to be highly robust to inharmonicities, i.e., deviations from perfect harmonicity.**

***Index Terms*—Multi-pitch estimation, fundamental frequency, inharmonicity, optimal transport, spectral estimation**

I. Introduction

The task of estimating fundamental frequency, or pitch, is a problem arising in various applications, including speech processing [1], music audio signal processing [2] and biomedical modeling [3]. The problem can typically be separated into two classes: the single-pitch problem considering signal consisting of a single harmonic series [4]–[8], and the multi-pitch problem, in which the signal is modeled as a mixture of several harmonic series with different fundamental frequencies (see [7] for an overview). Currently, the state-of-the-art in multi-pitch estimation consists of parametric methods building on the compressed sensing paradigm and in different ways aim to separate and identify individual pitches by exploiting sparse signal representations (see, e.g., [9]–[11]). In particular, pitches are estimated by constructing dictionaries whose atoms correspond to harmonic series, and the signal components are identified by minimizing sparsity-promoting optimization criteria, often in the form of regularized least-squares problems. One drawback of these methods is that dictionary atoms become highly coherent due to harmonic overlap between separate candidate pitches. Furthermore, the assumption of perfect harmonicity make these methods susceptible to errors in the face of inharmonicity, i.e., deviations of individual sinusoids from the perfectly harmonic model, such as seen in, for example, stringed musical instruments [12]–[14], and to some extent in human speech [15].

In this work, we propose to alleviate the need of pitch-based dictionaries, as well as to induce robustness to potential inharmonicity, by formulating multi-pitch estimation as a problem of clustering from indirect measurements. In particular, we aim to infer the spectral content of the signal, i.e., the power and frequencies of individual sinusoids, and to assign this power to a small number of groups, corresponding to pitches. This grouping, i.e., assignment of sinusoidal power to pitches, is performed based on the Monge-Kantorovich problem of optimal transport (OT) [16]. OT problems have earlier been used for spectral estimation (see, e.g., [17]–[20]), as well as in the context of pitch estimation [21]–[24]. However, these earlier works are either set in the single-pitch scenario, or consider the case when the frequency content of the signal is *a priori* known. To the best of the authors' knowledge, the current work is the first one to propose an OT-based technique for solving multi-pitch estimation as an inverse problem, i.e., using only noisy observations of the signal waveform. As is shown, the proposed method induces considerable robustness to potential inharmonic deviations, and avoids the use of a highly coherent pitch-based dictionary.

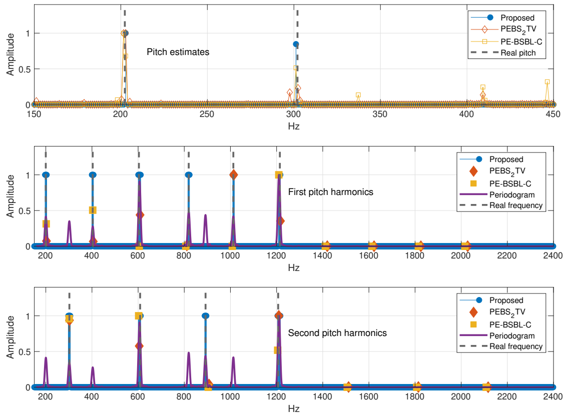

Fig. 1: The top subplot shows the pitch estimates, the middle shows frequency estimates for the first pitch, and the bottom shows frequency estimates of the second pitch, comparing the proposed and two reference methods.

An illustrative example is presented in Figure 1, with a signal containing two pitches, both being subject to inharmonicities (see [24] for a detailed discussion on the concept of pitch for inharmonic signals). The top panel displays estimates of the pitches (i.e., the power assigned to fundamental frequencies) by the herein proposed method, as well as for the sparsity-based methods PEBS_2TV [9] and PE-BSBL-C [11]. As can be seen, the proposed method concentrates all signal energy into two distinct pitches, whereas the references methods suffer from spurious estimates caused by the inharmonic deviations that cannot be represented by only two dictionary atoms. The middle and bottom panels show the obtained frequency estimates for the two pitches, separated into the two panels. As a consequence of the perfectly harmonic model for the reference methods, they are unable to assign energy content to all of the harmonics of the identified pitches. While this energy content is missing for the identified pitches, it is still assigned to other pitches, corresponding to the spurious peaks in the top panel. The proposed method on the other hand distributes the energy freely between frequency and pitch, and is able to classify a set of frequencies centered around the harmonics of both pitches.

II. Signal model

Consider the observed signal[1] y_t, $t = 0, 1, \ldots, N-1$,

$$y_t = x_t + v_t \ , \ v_t \in \mathbb{C}\mathcal{N}(0, \sigma^2), \tag{1}$$

$$x_t = \sum_{p=1}^{P} x_t^{(p)} \ \text{, with } \ x_t^{(p)} = \sum_{h}^{H^{(p)}} \alpha_h^{(p)} e^{i\omega_h^{(p)} t}, \tag{2}$$

where $\omega_h^{(p)} \in (0, \pi]$, $\alpha_h^{(p)} \in \mathbb{C}$, and $\mathbb{C}\mathcal{N}(0, \sigma^2)$ denotes the circularly symmetric complex distribution with variance σ^2. Here, each constituent wave-form $x_t^{(p)}$ is assumed to be a (approximately) harmonic series with (nominal) fundamental frequency $\omega_0^{(p)}$ and $H^{(p)}$ harmonics. In particular, the set of frequencies $\{\omega_h^{(p)}\}_{h=1}^{H^{(p)}}$ is assumed to satisfy

$$\omega_h^{(p)} = h\omega_0^{(p)} + \Delta_h^{(p)}, \ h \in [\![H^{(p)}]\!] \triangleq \{1, \ldots, H^{(p)}\} \tag{3}$$

where $\Delta_h^{(p)}$ are so-called inharmonicities assumed to be small in the sense $\left|\Delta_h^{(p)}\right| \ll \omega_0^{(p)}$. Note that for the case of all $\Delta_h^{(p)}$ being zero, y_t is a noisy multi-pitch signal with a set of fundamental frequencies $\boldsymbol{\omega}_0 = \{\omega_0^{(p)}, p \in [\![P]\!]\}$.

In this work, we consider the problem of identifying the components $x_t^{(p)}$ from noisy observations of y_t in (1) and to estimate the fundamental frequencies $\boldsymbol{\omega}_0$. In particular, motivated by the example in Figure 1, we aim to propose a multi-pitch estimation method displaying robustness to potential inharmonicity.

It may be noted that several works have considered estimation of $\boldsymbol{\omega}_0$ in the perfectly harmonic case (see, e.g., [7], [9]–[11]). For the inharmonic signals, the single-pitch case, i.e., $P = 1$, has been considered in [23], [26]. For the multi-pitch case, the only existing works (to the best of the authors' knowledge) consider a scenario where estimates of frequencies and amplitudes of the sinusoidal components are already available, and multi-pitch estimation boils down to finding suitable groupings of these components [21], [22]. The works [21]–[23] all consider using the concept of optimal transport for performing pitch estimating. Building on these efforts, we herein propose solving an OT-regularized inverse problem allowing for directly estimating $\boldsymbol{\omega}_0$ from the noisy observations (1). In particular, the regularizing function is constructed from an OT problem considering the spectral representation of the noise-free wave-form x_t.

III. Spectral representation

It may be noted that the spectrum of x_t in (2) is the non-negative distribution on $(0, \pi]$ given by

$$\Phi(\omega) = \sum_{p=1}^{P} \sum_{h=1}^{H^{(p)}} |\alpha_h^{(p)}|^2 \delta(\omega - \omega_h^{(p)}), \tag{4}$$

where $\delta(\cdot)$ is the Dirac delta. Herein, we model this as a *perturbation in frequency* of a perfectly harmonic spectrum of the form

$$\Phi_{harm}(\omega) = \sum_{p=1}^{P} \sum_{h=1}^{H^{(p)}} |\alpha_h^{(p)}|^2 \delta(\omega - h\omega_0^{(p)}). \tag{5}$$

As may be noted, the parameters of x_t, i.e., the frequencies and amplitudes, uniquely determine (4). With this, we aim to infer the parameters of x_t promoting estimates that are close-to-harmonic, in the sense of being ”close” to having a perfectly harmonic spectrum (5). Building on the works [20], [22], [24] we here measure closeness by means of optimal transport problems.

In particular, we replace (5) with a distribution over only the fundamental frequencies according to

$$\Phi_{pitch}(\omega_0) = \sum_{p=1}^{P} \left(\sum_{h=1}^{H^{(p)}} |\alpha_h^{(p)}|^2 \right) \delta(\omega_0 - \omega_0^{(p)}). \tag{6}$$

Note here that the power of the full set of harmonics is allocated to the corresponding fundamental frequency. Then, letting $\mathbf{\Phi} \in \mathbb{R}^F$, with $F = \sum_{p=1}^{P} H^{(p)}$, and $\mathbf{\Phi}_{pitch} \in \mathbb{R}^P$ be vector representations of the distributions Φ and Φ_{pitch}, one may measure the distance between the two using the Monge-Kantorovich problem of optimal transport [16], [27],

$$\begin{aligned} \mathcal{Q}(\mathbf{\Phi}, \mathbf{\Phi}_{pitch}) \triangleq & \underset{\mathbf{M} \in \mathbb{R}_+^{F \times P}}{\text{minimize}} \ \ \langle \mathbf{C}, \mathbf{M} \rangle \\ & \text{s.t.} \quad \mathbf{M}\mathbf{1}_P = \mathbf{\Phi}, \quad \mathbf{M}^T \mathbf{1}_F = \mathbf{\Phi}_{pitch}, \end{aligned} \tag{7}$$

where $\langle \mathbf{C}, \mathbf{M} \rangle \triangleq \sum_{f,p} [\mathbf{C}]_{f,p} [\mathbf{M}]_{f,p}$. Here, the transport plan $\mathbf{M}$ describes the amount of mass (i.e., power) at frequencies in Φ that is assigned to fundamental frequencies in Φ_{pitch}, an example of which can be seen in Figure 2. The corresponding elements of the cost matrix $\mathbf{C}$ describes the cost of that assignment. Herein, we will consider two so-called ground-cost functions defining the elements of $\mathbf{C}$. We will use

$$c_H(\omega, \omega_0) = \min_{h \in [\![H]\!]} (\omega/\omega_0 - h)^2, \tag{8}$$

i.e., the normalized squared distance of ω to the closest integer multiple of ω_0, where the maximal allowed harmonic order is H, as well as the case when only the Nyquist rate limits the order, denoted as

$$c_\infty(\omega, \omega_0) = \min_{h \in \mathbb{Z}_+} (\omega/\omega_0 - h)^2. \tag{9}$$

In fact, the minimal objective $\langle \mathbf{C}, \mathbf{M} \rangle$ is zero if and only if all components of Φ are harmonic, and strictly positive otherwise.

It may be noted that for the setting of this paper, we only have access to the noisy measurements y_t from (1) and not to the spectrum Φ and pitch spectrum Φ_{pitch} which in effect are the quantities that we want to estimate. However, we in the next section show how the problem in (7) can be extended as to yield a regularizing amenable to our inverse setting.

IV. Sparse entropic harmonic clustering

As noted earlier, the spectrum $\Phi(\omega)$ is not available and we only have access to the noisy measurement $\mathbf{y} = [y_0, \ldots, y_{N-1}]^T$. However, with a fine-enough gridding of the frequency axis $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_F\} \subset [0, \pi)$ for some $F \in \mathbb{N}$, we assume that the noise-free samples $\mathbf{x} = [x_0, \ldots, x_{N-1}]^T$ can be approximated as

$$\mathbf{x} \approx \mathbf{A}(\boldsymbol{\omega})\boldsymbol{\alpha},$$

where $\mathbf{A}(\boldsymbol{\omega}) = \begin{bmatrix} \mathbf{a}(\omega_1) & \ldots & \mathbf{a}(\omega_F) \end{bmatrix}$ is a dictionary matrix whose columns $\mathbf{a}(\omega_f) \in \mathbb{C}^N$ are Fourier vectors corresponding to frequencies ω_f, $f \in [\![F]\!]$. The vector of complex amplitudes $\boldsymbol{\alpha} \in \mathbb{C}^F$ here corresponds to a discretization of $\Phi(\omega)$, i.e., $|\boldsymbol{\alpha}|^2 \approx \mathbf{\Phi}$, where $|\cdot|^2$ is to be interpreted elementwise. With this, we are ready to formulate

[1] Here, for generality, we will consider the complex-valued representation, noting that this can easily be formed as the discrete-time analytical version of a real-valued signal [25].

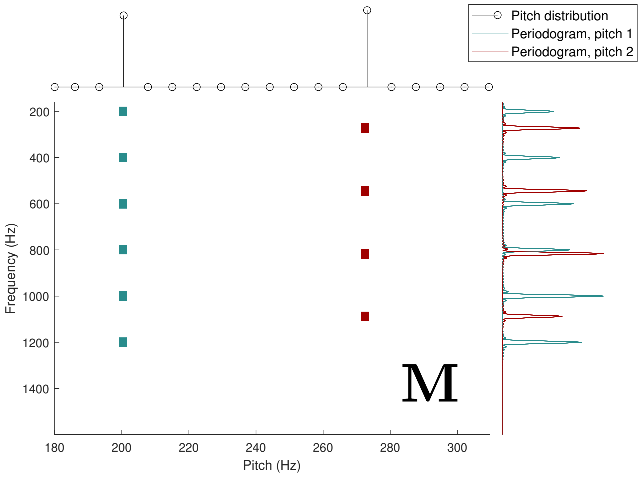

Fig. 2: Illustrative example of a transport plan between a distribution on the frequency grid and a distribution on the pitch grid. The upper plot is the pitch distribution, while the rotated plot on the right-hand side displays the periodogram for the two pitches present in the signal.

our inverse estimation problem. In particular, in order to estimate $\mathbf{\Phi}$, we solve the regularized least-squares problem

$$\underset{\boldsymbol{\alpha}\in\mathbb{C}^F}{\text{minimize}} \quad \frac{1}{N}\left\|\mathbf{y}-\mathbf{A}(\boldsymbol{\omega})\boldsymbol{\alpha}\right\|_2^2+\mathcal{R}(\boldsymbol{\alpha}), \tag{10}$$

where we want the regularizing function $\mathcal{R}$ to allow us to estimate and retrieve the pitch structure of x_t, as well as allow for efficient solution algorithms. We note that the power in $|\boldsymbol{\alpha}|^2$ should be concentrated in just a few pitches, and argue the mechanism of assigning power over frequency to pitches can be formalized using transport problems similar to (7).

To this end, let $\Gamma = \{\omega_0^{(1)}, \ldots, \omega_0^{(G)}\}$, $G \in \mathbb{N}$, be a set of candidate pitch frequencies. Similar to the approximation of $\mathbf{x}$, we assume that for large enough Γ, the (nominal) fundamental frequencies of x_t in (2) can be found close to elements of Γ. Relating to the transport problem in (7), we let $\mathbf{\Phi}_{pitch} \in \mathbb{R}^G$ be the (unknown) power distribution over the pitch grid Γ. As $G \gg 1$, we expect $\mathbf{\Phi}_{pitch}$ to be sparse. Building on our earlier works on OT-based clustering [22]–[24], [28], [29], we propose the regularizer

$$\mathcal{R}(\boldsymbol{\alpha}) = \beta\left\|\boldsymbol{\alpha}\right\|_1 + \zeta\mathcal{S}(\boldsymbol{\alpha}),$$

where $\beta, \zeta > 0$, the penalty on $\left\|\boldsymbol{\alpha}\right\|_1$ reflects the assumption that $\mathbf{\Phi}$ is sparse, and $\mathcal{S}(\boldsymbol{\alpha})$ is defined as

$$\begin{aligned} \mathcal{S}(\boldsymbol{\alpha}) \triangleq \min_{\mathbf{M}\in\mathbb{R}_+^{F\times G}} \quad & \langle \mathbf{C}, \mathbf{M}\rangle + \epsilon D(\mathbf{M}) + \eta\left\|\mathbf{M}\right\|_{\infty,1} \\ \text{s.t.} \quad & \mathbf{M}\mathbf{1}_G \geq |\boldsymbol{\alpha}|^2, \end{aligned} \tag{11}$$

with $\epsilon, \eta > 0$. Although maybe not immediately apparent, the regularizer $\mathcal{S}(\boldsymbol{\alpha})$ in fact performs OT clustering that concentrates the power in $|\boldsymbol{\alpha}|^2$ in a small number of pitches on the grid Γ. Furthermore, the pitch "spectrum" can be retrieved as $\mathbf{\Phi}_{pitch} = \mathbf{M}^T\mathbf{1}_F$ from the optimal $\mathbf{M}$. It may be noted that (11) differs from (7) in the following respects. First of all, only one margin of the transport problem, i.e., $\mathbf{M}\mathbf{1}_G$, is specified, whereas the margin $\mathbf{M}^T\mathbf{1}_F$ is free. The only assumption put on this margin is that it is sparse, which is enforced be the penalty $\left\|\mathbf{M}\right\|_{\infty,1} \triangleq \sum_g \max_f |[\mathbf{M}]_{f,g}|$, i.e., the transport plan is promoted to be concentrated to just a few columns, as illustrated in Figure 2. Secondly, $D(\mathbf{M}) = \sum_{f,g}[\mathbf{M}]_{f,g}\log[\mathbf{M}]_{f,g} - [\mathbf{M}]_{f,g} + 1$ is an entropic regularization term, added as to make the objective in (11) strictly convex. For the case of known and fixed $|\boldsymbol{\alpha}|^2$, problems of this type have earlier been used successfully to perform OT-based clustering (see [29] for details), where the constraint is in the form of an equality, as in the standard problem (7). As we in the setting considered herein want to estimate $\boldsymbol{\alpha}$, the constraint is relaxed to an inequality, resulting in $\mathcal{S}$ being convex in $\boldsymbol{\alpha}$. For the cost matrix $\mathbf{C}$, we in the numerical experiments consider defining its elements as $[\mathbf{C}]_{f,g} = c_H(\omega_f, \omega_0^{(g)})$ as well as $[\mathbf{C}]_{f,g} = c_\infty(\omega_f, \omega_0^{(g)})$, where c_H and c_∞ are defined in (8) and (9), respectively.

Taken together, we have arrived at our proposed estimator. In particular we propose to estimate the pitch spectrum as $\mathbf{\Phi}_{pitch} = \mathbf{M}^T\mathbf{1}_F$, where $\mathbf{M}$ is found by solving

$$\underset{\boldsymbol{\alpha}\in\mathbb{C}^F}{\text{minimize}} \quad \frac{1}{N}\left\|\mathbf{y}-\mathbf{A}(\boldsymbol{\omega})\boldsymbol{\alpha}\right\|_2^2 + \beta\left\|\boldsymbol{\alpha}\right\|_1 + \zeta\mathcal{S}(\boldsymbol{\alpha}), \tag{12}$$

The fundamental frequencies are then found by inspecting the (sparse) support of $\mathbf{\Phi}_{pitch}$. It may be noted that the herein proposed methods differs substantially from earlier sparsity-based multi-pitch estimator, not only in how the estimation criterion is formulated, but also how the signal is approximated. In particular, the dictionary atoms used herein are simply Fourier vectors, whereas the "atoms" of dictionaries in, e.g., [9]–[11] correspond to entire harmonic series. The latter approach leads to a highly coherent dictionary due to overlap of the harmonic frequencies. The herein proposed method alleviates this by, essentially, formulating pitch estimation as a clustering problem: the signal x_t is approximated using sinusoidal components, and the pitch structure is retrieved using the transport regularization.

As may be noted, the problem in (12) is convex, and we next present an efficient algorithm implementing the estimator.

V. Efficient implementation

We herein propose to solve (12) using a proximal gradient scheme, alternating between taking gradient steps with respect to the data-fit term and computing the proximal mapping of the regularizer $\beta\left\|\cdot\right\|_1 + \zeta\mathcal{S}(\cdot)$. We use the constant stepsize $1/L$, where $L = \left\|\mathbf{A}(\boldsymbol{\omega})\right\|^2/N$, and $\left\|\cdot\right\|$ is the operator norm. Then, iterates are given by

$$\boldsymbol{\alpha}^{(j+1)} = \text{prox}_{\frac{1}{L}\beta\|\cdot\|_1 + \frac{1}{L}\zeta\mathcal{S}(\cdot)}\left(\boldsymbol{\alpha}^{(j)} - \frac{1}{L}\nabla_\alpha\frac{1}{N}\left\|\mathbf{y}-\mathbf{A}(\boldsymbol{\omega})\boldsymbol{\alpha}^{(j)}\right\|_2^2\right)$$

where $\nabla_\alpha(\cdot)$ denotes the (Wirtinger) gradient with respect to $\boldsymbol{\alpha}$. For the proximal operator, the following proposition holds.

Proposition 1. *The proximal operator for $\frac{1}{L}\beta\left\|\cdot\right\|_1 + \frac{1}{L}\zeta\mathcal{S}(\cdot)$ is unique and given by*

$$\text{prox}_{\frac{\beta}{L}\|\cdot\|_1 + \frac{\zeta}{L}\mathcal{S}(\cdot)}(\mathbf{u}) = e^{i\angle\mathbf{u}} \odot \left((\rho)_+ \oslash (\mathbf{1}_F + 2\boldsymbol{\lambda})\right)$$

where $\oslash$ and $\odot$ denotes elementwise division and multiplication, respectively, $\angle\mathbf{u}$ denotes the phase angle of $\mathbf{u}$, $(\cdot)_+$ denotes the truncation of negative values to 0, $\rho = (|\boldsymbol{\alpha}| - \frac{\beta}{L}\mathbf{1}_F)$ and $\boldsymbol{\lambda} \in \mathbb{R}_+^F$ solves

$$\begin{aligned} \underset{\boldsymbol{\lambda}\in\mathbb{R}_+^F,\ \mathbf{\Psi}\in\mathbb{R}^{F\times G}:\|\mathbf{\Psi}\|_{1,\infty}\leq\frac{1}{L}\zeta\eta}{\text{minimize}} \quad & \frac{1}{L}\zeta\epsilon\mathbf{v}^T(\mathbf{K}\odot\mathbf{W})\mathbf{1}_G \\ & + \frac{1}{2}\left\langle \mathbf{1}_F \oslash (\mathbf{1}_F + 2\boldsymbol{\lambda}), (\rho)_+^2\right\rangle \end{aligned} \tag{13}$$

where $\mathbf{v} = \exp\left(\frac{L}{\zeta\epsilon}\boldsymbol{\lambda}\right)$, $\mathbf{K} = \exp\left(-\frac{1}{\epsilon}\mathbf{C}\right)$, $\mathbf{W} = \exp\left(\frac{L}{\zeta\epsilon}\mathbf{\Psi}\right)$, $\|\mathbf{\Psi}\|_{1,\infty} = \max_g \sum_f |\Psi_{f,g}|$ is the dual norm of $\|\cdot\|_{\infty,1}$, and all exponentiation and powers are evaluated elementwise.

Proof. See appendix. □

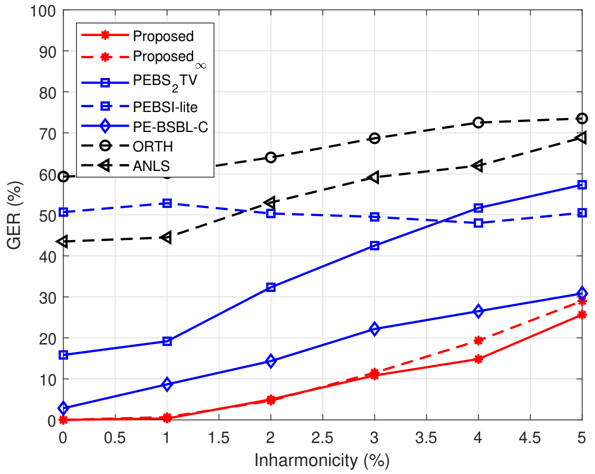

Fig. 3: The gross error rate for 200 Monte Carlo simulations of a simulated signal with three pitches of $197, 240$ and 272 Hz, which in each simulation are perturbed as to avoid biasing due to grid effects, for varying levels of inharmonicity.

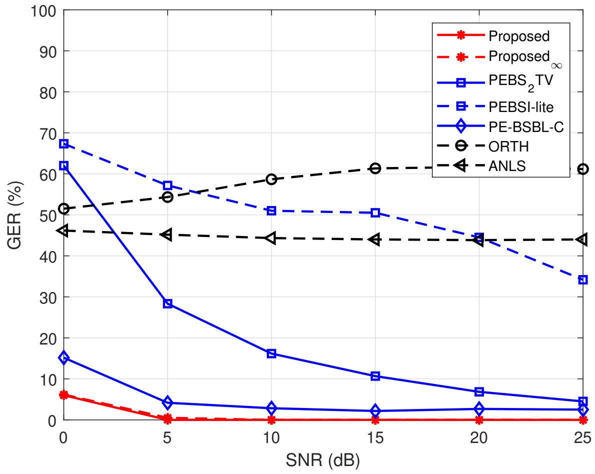

Fig. 4: The gross error rate for 200 Monte Carlo simulations of a simulated signal with three pitches of $197, 240$ and 272 Hz, which in each simulation are perturbed as to avoid biasing due to grid effects, for varying levels of SNR.

For solving the problem (13), as to evaluate the proximal operator, we propose a block-coordinate ascent scheme (see, e.g., [19]). The variable updates in iteration k are given by (see appendix)

$$\boldsymbol{\lambda}^{(k)} = 2\frac{1}{L}\zeta\epsilon\left(\Omega\left(\boldsymbol{\xi}^{(k-1)}\right) - \frac{\mathbf{1}_F}{4\frac{1}{L}\zeta\epsilon}\right)_+,$$

$$\boldsymbol{\Psi}^{(k)} = \underset{\boldsymbol{\Psi}:\|\boldsymbol{\Psi}\|_{1,\infty}\leq\frac{1}{L}\zeta\eta}{\operatorname{argmin}} \quad \frac{1}{L}\zeta\epsilon(\mathbf{v}^{(k)}\odot\mathbf{K})^T\mathbf{W},$$

where

$$\boldsymbol{\xi}^{(k)} = \left(\frac{1}{4\frac{1}{L}\zeta\epsilon} - \log(4\frac{1}{L}\zeta\epsilon)\right)\mathbf{1}_F - \frac{1}{2}\log\left((\mathbf{K}\odot\mathbf{W}^{(k)})\mathbf{1}_G\right) + \frac{1}{2}\log\left((\rho)_+^2\right),$$

and $\Omega(\cdot)$ denotes the (elementwise) Wright omega function [30]. The variable $\boldsymbol{\Psi}^{(k)}$ is computed according to Theorem 2 in [20]. As noted earlier, we retrieve $\boldsymbol{\Phi}_{pitch}$ as $\boldsymbol{\Phi}_{pitch} = \mathbf{M}^T\mathbf{1}_F$.

VI. Numerical experiments

Herein, we evaluate the proposed methods using a Monte Carlo simulation study. Throughout, we consider a signal consisting of 3 pitches, sampled at 8000 Hz, and observed at $N = 250$ samples. In each simulation, the number of harmonics for each pitch is randomized on the interval $[3, 10]$, with magnitudes randomized uniformly on $[0.7, 1]$ and uniform random initial phase. The (nominal) fundamental frequencies are 197, 240, and 272 Hz, respectively, which in each simulation are perturbed as to avoid biasing due to grid effects.

With this setup, we study the performance of the proposed method in the face of varying signal-to-noise ratio (SNR) defined as SNR $= 10\log_{10}\left((\sum_{p=1}^{P}\sum_{h=1}^{H^{(p)}}|\alpha_h^{(p)}|^2)/\sigma^2\right)$, as well as inharmonicity. Performance is evaluated in terms of the gross error rate (GER) defined as the percentage of estimated fundamental frequencies deviating less than 5% from the ground-truth values. The proposed method is compared to the state-of-the-art sparsity-based methods PEBS$_2$TV, PEBSI-lite and PE-BSBL-C [9]–[11], as well as the methods ORTH and ANLS from [7].

For all methods, a uniform grid of fundamental frequencies with grid spacing 2 Hz from 50 to 500 Hz is used, such that $G = 226$, and the highest assumed harmonic order is set to $H = 10$. For the proposed method, we also include the version where the harmonic order is allowed to span up to the Nyquist frequency. For the proposed method, the grid for the sinusoidal frequencies is uniform from 50 to 4000 Hz, corresponding to a grid from the lowest evaluated fundamental frequency to the Nyquist frequency, with $F = 2260$ grid points to match the size of the reference methods' dictionary matrices. For the proposed method, we use the parameter values $\epsilon = 10^{-9}$, $\beta = 4 \cdot 10^{-1}$, $\eta = 5 \cdot 10^{-1}$ and $\zeta = 5 \cdot 10^2$, whereas the hyper parameters of the comparison methods are set according to recommendations in their respective papers.

Figure 4 shows the results for varying SNR for the perfectly harmonic case, i.e., with no inharmonic deviations, where the labels "Proposed" and "Proposed$_\infty$" denote the version with limited and unlimited harmonic order, respectively. As can be seen, the two versions of the proposed method provide more accurate estimates than the comparisons,

For a fixed SNR = 10 dB, Figure 3 evaluates the effect of inharmonicity. Here, each nominal harmonic frequency ω is in each simulation perturbed randomly uniformly on the interval $\omega \pm \kappa\omega$, where $\kappa \in [0, 1]$ corresponds to the x-axis of the figure. It may be noted that the proposed method displays higher robustness to inharmonic deviations as compared to the reference methods.

As may be noted from Figures 4 and Figure 3, allowing for an unlimited (up to Nyquist) harmonic order only results in a slight decrease of performance as compared to fixing a maximum order.

VII. Conclusions

In this work, we have proposed a technique for multi-pitch estimation based on ideas from optimal transport theory. Phrased as an inverse problem, the proposed methods estimate the spectral content of a signal while simultaneously grouping the power into a small number of pitches, where the grouping is enforced using an OT-based regularizer. The resulting inverse problem is a convex program, and we formulate an efficient solver implementing the solver. As shown in simulations, the proposed methods display considerable robustness to noise, as well as to inharmonic deviations.

Appendix

Below we show the proof of Proposition 1.

Proof. The proximal operator of $\frac{1}{L}\beta\|\cdot\|_1 + \frac{1}{L}\zeta\mathcal{S}(\cdot)$ is defined as

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\arg\min}\, \frac{1}{L}\beta\|\boldsymbol{\alpha}\|_1 + \frac{1}{L}\zeta\mathcal{S}(\boldsymbol{\alpha}) + \frac{1}{2}\|\boldsymbol{\alpha} - \mathbf{u}\|_2^2.$$

We note that $\mathcal{S}(\cdot)$ and $\|\cdot\|_1$ are both invariant to the phase, and that the term $\frac{1}{2}\|\boldsymbol{\alpha} - \mathbf{u}\|$ is minimized when the phase of $\boldsymbol{\alpha}$ aligns with the phase of $\mathbf{u}$. Thus, by introducing the polar form of $\boldsymbol{\alpha}, \mathbf{u}$, such that $\mathbf{u} = \mathbf{r} \odot e^{i\angle\mathbf{u}}$, $\boldsymbol{\alpha} = \hat{\mathbf{r}} \odot e^{i\angle\boldsymbol{\alpha}}$, we simply let $\angle\boldsymbol{\alpha} = \angle\mathbf{u}$, and the proximal operator becomes

$$e^{i\angle\mathbf{u}} \odot \left(\underset{\hat{\mathbf{r}}\in\mathbb{R}_+^N}{\arg\min}\, \frac{1}{L}\beta\|\hat{\mathbf{r}}\|_1 + \frac{1}{L}\zeta\mathcal{S}(\hat{\mathbf{r}}) + \frac{1}{2}\|\hat{\mathbf{r}} - \mathbf{r}\|_2^2 \right).$$

By introducing the auxiliary variable $\mathbf{Q}$, such that $\mathbf{Q} = \mathbf{M}$, $\hat{\mathbf{r}}$ can be obtained through solving

$$\begin{aligned} \min_{\hat{\mathbf{r}},\mathbf{M},\mathbf{Q}} \quad & \frac{1}{L}\beta\|\hat{\mathbf{r}}\|_1 + \frac{1}{L}\zeta\langle\mathbf{C},\mathbf{M}\rangle + \frac{1}{L}\zeta\epsilon D(\mathbf{M}) \\ & + \frac{1}{L}\zeta\eta\|\mathbf{Q}\|_{\infty,1} + \frac{1}{2}\|\hat{\mathbf{r}} - \mathbf{r}\|_2^2, \\ \text{s.t.} \quad & \hat{\mathbf{r}}^2 \leq \mathbf{M}\mathbf{1}_F, \quad \mathbf{Q} = \mathbf{M}. \end{aligned} \tag{14}$$

The Lagrangian of (14) is

$$\begin{aligned} &\mathcal{L}(\hat{\mathbf{r}}, \mathbf{M}, \mathbf{Q}, \boldsymbol{\lambda}, \boldsymbol{\Psi}) = \\ &\theta\left\langle\mathbf{C},\mathbf{M}\right\rangle + \theta\epsilon D(\mathbf{M}) + \theta\eta\|\mathbf{Q}\|_{\infty,1} + \frac{1}{L}\beta\|\hat{\mathbf{r}}\|_1 \\ &+ \frac{1}{2}\|\hat{\mathbf{r}} - \mathbf{r}\|_2^2 + \left\langle\boldsymbol{\lambda}, \hat{\mathbf{r}}^2 - \mathbf{M}\mathbf{1}_F\right\rangle + \left\langle\boldsymbol{\Psi}, \mathbf{Q} - \mathbf{M}\right\rangle, \end{aligned}$$

It may be readily verified that the Lagrangian is strongly convex in $\hat{\mathbf{r}}, \mathbf{M}$ and $\mathbf{Q}$, with unique minimizer for $\hat{\mathbf{r}}$ and $\mathbf{M}$

$$\begin{aligned} \hat{\mathbf{r}} &= \left(\mathbf{r} - \frac{1}{L}\beta\right)_+ \oslash \left(\mathbf{1}_F + 2\boldsymbol{\lambda}\right), \\ \mathbf{M} &= e^{\frac{L}{\zeta\epsilon}\boldsymbol{\lambda}^T} \left(e^{-\frac{1}{\epsilon}\mathbf{C}} \odot e^{\frac{L}{\zeta\epsilon}\boldsymbol{\Psi}}\right). \end{aligned}$$

The unique minimizer for $\mathbf{Q}$ follows from [20, Theorem 1], and is given by

$$\mathbf{Q} = 0, \quad \text{s.t.} \quad \|\boldsymbol{\Psi}\|_{1,\infty} \leq \frac{1}{L}\zeta\eta. \tag{15}$$

Plugging in these values into the Lagrangian yields the dual problem

$$\begin{aligned} \underset{\boldsymbol{\lambda}\in\mathbb{R}_+^F,\, \boldsymbol{\Psi}\in\mathbb{R}^{F\times G}}{\text{minimize}} \quad & \frac{1}{L}\zeta\epsilon\mathbf{v}^T(\mathbf{K}\odot\mathbf{W})\mathbf{1}_G \\ & + \frac{1}{2}\left\langle\mathbf{1}_F \oslash (\mathbf{1}_F + 2\boldsymbol{\lambda}), (\rho)_+^2\right\rangle \\ \text{s.t.} \quad & \|\boldsymbol{\Psi}\|_{1,\infty} \leq \frac{1}{L}\zeta\eta, \end{aligned}$$

where $\mathbf{v} = \exp\left(\frac{L}{\zeta\epsilon}\boldsymbol{\lambda}\right), \mathbf{K} = \exp\left(-\frac{1}{\epsilon}\mathbf{C}\right), \mathbf{W} = \exp\left(\frac{L}{\zeta\epsilon}\boldsymbol{\Psi}\right)$. □

A. Update of dual variables

We now want to show that in each iteration k, the updates for the dual variables are

$$\begin{aligned} \boldsymbol{\lambda}^{(k)} &= 2\theta\epsilon\left(\Omega\left(\boldsymbol{\xi}^{(k-1)}\right) - \frac{1}{4\theta\epsilon}\right)_+, \\ \boldsymbol{\Psi}^{(k)} &= \underset{\boldsymbol{\Psi}}{\arg\min}\ \theta\epsilon(\mathbf{v}^{(k)}\odot\mathbf{K})^T\mathbf{W}, \\ &\text{s.t.} \quad \|\boldsymbol{\Psi}\|_{1,\infty} \leq \theta\eta. \end{aligned}$$

Proof. We want to solve the minimization problem of (13) with respect to $\boldsymbol{\lambda}$ and $\mathbf{M}$. Fixating $\boldsymbol{\Psi}$ and collecting the terms related to $\boldsymbol{\lambda}$, we obtain the minimization problem

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}_+^N} \quad \theta\epsilon\mathbf{v}^T\left(\mathbf{K}\odot\mathbf{W}\right) + \frac{1}{2}\left\langle\mathbf{1}_F \oslash (\mathbf{1}_F + 2\boldsymbol{\lambda}), (\rho)_+^2\right\rangle.$$

Setting the gradient of this with respect to $\boldsymbol{\lambda}$ to 0, we obtain

$$\exp\left(\frac{L}{\theta\epsilon}\boldsymbol{\lambda}\right) \odot \left(\mathbf{K}\odot\mathbf{W}\right)\mathbf{1}_G - \frac{1}{2}(\rho)_+^2 \oslash \left(\mathbf{1}_F + 2\boldsymbol{\lambda}\right) = 0. \tag{16}$$

By taking the logarithm on both sides, collecting the $\boldsymbol{\lambda}$-terms, dividing by 2 and adding $\frac{1}{4\theta\epsilon} - \log(4\theta\epsilon)$ on both sides, we obtain

$$\frac{\mathbf{1}_F + 2\boldsymbol{\lambda}}{4\theta\epsilon} + \log\left(\frac{\mathbf{1}_F + 2\boldsymbol{\lambda}}{4\theta\epsilon}\right) \tag{17}$$

$$= \frac{1}{2}\log((\rho)_+^2) - \frac{1}{2}\log\left(\left(\mathbf{K}\odot\mathbf{W}\right)\mathbf{1}_G\right) + \frac{1}{4\theta\epsilon} - \log\left(4\theta\epsilon\right), \tag{18}$$

where we denote the right-hand side of the equation $\boldsymbol{\xi}$. We now use the Wright omega-function, i.e., the function $\Omega : \mathbb{R} \to \mathbb{R}_+$ mapping x to $\Omega(x)$ such that $\Omega(x) + \log\Omega(x) = x$, thus granting us the solution

$$\boldsymbol{\lambda} = 2\theta\epsilon\left(\Omega(\boldsymbol{\xi}) - \frac{1}{4\theta\epsilon}\right). \tag{19}$$

We lastly fixate $\boldsymbol{\lambda}$ and collect the terms related to $\boldsymbol{\Psi}$ and get the minimization problem

$$\boldsymbol{\Psi}^{(k)} = \min_{\boldsymbol{\Psi}}\ \theta\epsilon(\mathbf{v}^{(k)}\odot\mathbf{K})^T\mathbf{W}, \tag{20}$$

$$\text{s.t.} \quad \|\boldsymbol{\Psi}\|_{1,\infty} \leq \theta\eta, \tag{21}$$

the solution of which is found in accordance with Theorem 2 in [20]. □

References

[1] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Instantaneous Fundamental Frequency Estimation With Optimal Segmentation for Nonstationary Voiced Speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2354–2367, 2016.

[2] M. Muller, D. P. W. Ellis, A. Klapuri, and G. Richard, "Signal Processing for Music Analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, 2011.

[3] 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 Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1015–1022, 2009.

[4] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 04 2002.

[5] D. Talkin and W. B. Kleijn, "A Robust Algorithm for Pitch Tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[6] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient," *Signal Processing*, vol. 135, pp. 188–197, 2017.

[7] M. Christensen and A. Jakobsson, *Multi-Pitch Estimation*. Springer Nature, 2022.

[8] S. Gonzalez and M. Brookes, "PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.

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

[10] F. Elvander, T. Kronvall, S. Adalbjörnsson, and A. Jakobsson, "An adaptive penalty multi-pitch estimator with self-regularization," *Signal Processing*, vol. 127, pp. 56–70, 2016.

[11] L. Shi, J. R. Jensen, J. K. Nielsen, and M. G. Christensen, "Multipitch Estimation Using Block Sparse Bayesian Learning and Intra-Block Clustering," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 666–670.

[12] H. Fletcher, "Normal Vibration Frequencies of a Stiff Piano String," *The Journal of the Acoustical Society of America*, vol. 36, no. 1, pp. 203–209, 1964.

[13] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*. Springer Science & Business Media, 2012.

[14] R. A. Rasch and V. Heetvelt, "String Inharmonicity and Piano Tuning," *Music Perception*, vol. 3, no. 2, pp. 171–189, 1985.

[15] J. Fernandes, F. Teixeira, V. Guedes, A. Junior, and J. P. Teixeira, "Harmonic to Noise Ratio Measurement - Selection of Window and Length," *Procedia computer science*, vol. 138, pp. 280–285, 2018.

[16] C. Villani, *Optimal Transport: Old and New*. Springer, 2009, vol. 338.

[17] T. T. Georgiou, J. Karlsson, and M. S. Takyar, "Metrics for Power Spectra: An Axiomatic Approach," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 859–867, 2008.

[18] F. Elvander, A. Jakobsson, and J. Karlsson, "Interpolation and Extrapolation of Toeplitz Matrices via Optimal Mass Transport," *IEEE Transactions on Signal Processing*, vol. 66, no. 20, pp. 5285–5298, 2018.

[19] F. Elvander, I. Haasler, A. Jakobsson, and J. Karlsson, "Multi-marginal optimal transport using partial information with applications in robust localization and sensor fusion," *Signal Processing*, vol. 171, p. 107474, 2020.

[20] I. Haasler and F. Elvander, "Multi-Frequency Tracking via Group-Sparse Optimal Transport," *IEEE Control Systems Letters*, vol. 8, pp. 1048–1053, 2024.

[21] R. Flamary, C. Févotte, N. Courty, and V. Emiya, "Optimal spectral transportation with application to music transcription," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[22] F. Elvander, S. Adalbjörnsson, J. Karlsson, and A. Jakobsson, "Using optimal transport for estimating inharmonic pitch signals," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 331–335.

[23] F. Elvander, "Estimating Inharmonic Signals with Optimal Transport Priors," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[24] F. Elvander and A. Jakobsson, "Defining Fundamental Frequency for Almost Harmonic Signals," *IEEE Transactions on Signal Processing*, vol. 68, pp. 6453–6466, 2020.

[25] L. Marple, "Computing the discrete-time "analytic" signal via FFT," *IEEE Transactions on signal processing*, vol. 47, no. 9, pp. 2600–2603, 1999.

[26] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, "A Robust and Computationally Efficient Subspace-Based Fundamental Frequency Estimator," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 3, pp. 487–497, 2010.

[27] G. Peyré and M. Cuturi, "Computational Optimal Transport: With Applications to Data Science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[28] F. Elvander, J. Karlsson, and T. van Waterschoot, "Convex Clustering for Multistatic Active Sensing via Optimal Mass Transport," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 1730–1734.

[29] G. Flood and F. Elvander, "Multi-Source Localization and Data Association for Time-Difference of Arrival Measurements," *arXiv preprint arXiv:2403.10329*, 2024.

[30] R. Corless and D. Jeffrey, "The Wright ω Function." in *Artificial Intelligence, Automated Reasoning, and Symbolic Computation, Joint International Conferences*, Marseille, France, 2002, pp. 76–89.