iMUSIC: A Family of MUSIC-Like Algorithms for Integer Period Estimation

Srikanth Venkata Tenneti, *Student Member, IEEE*, and Palghat P. Vaidyanathan, *Fellow, IEEE*

***Abstract*—The MUSIC algorithm is one of the most popular techniques today for line spectral estimation. If the line spectrum is that of a periodic signal, can we adapt MUSIC to exploit the additional harmonicity in the spectrum? Important prior work in this direction includes the Harmonic MUSIC algorithm and its variations. For applications where the period of the discrete signal is an integer (or can be well approximated by an integer), this paper introduces a new and simpler class of alternatives to MUSIC. This new family, called iMUSIC, also includes techniques where simple integer valued vectors are used in place of complex exponentials for both representing the signal subspace, and for computing the pseudo-spectrum. It will be shown that the proposed methods not only make the computations much simpler than prior periodicity-adaptations of MUSIC, but also offer significantly better estimation accuracies for applications with integer periods. These advantages are demonstrated on examples that include repeats in protein and DNA sequences. The iMUSIC algorithms are based on the recently proposed Ramanujan subspaces and nested periodic subspaces. The resulting signal space bases are non-Vandermonde in structure. Consequently, many aspects of classical MUSIC that were based on the Vandermonde structure of complex-exponentials, such as guarantees for identifiability of the frequencies (periods in our case), are addressed in new ways in this paper.**

***Index Terms*—Period estimation, MUltiple sIgnal Classification (MUSIC), Ramanujan subspaces, nested periodic subspaces, protein repeats, iMUSIC.**

I. Introduction

The MUSIC algorithm (MUltiple SIgnal Classification) [51] is one of the most popular techniques for estimating line spectra in discrete time signals. It has widespread applications, including Direction of Arrival estimation [55], [56], [72], time delay estimation [45], neuro-imaging [27], [38], and many more. But when the signal of interest is periodic, its spectrum is not just arbitrary lines. There is a nice harmonic structure in the spectrum as shown in Fig. 1, which can be modeled mathematically as:

$$x(n) = \sum_{k=0}^{K-1} c_k e^{jk\omega_0 n} \tag{1}$$

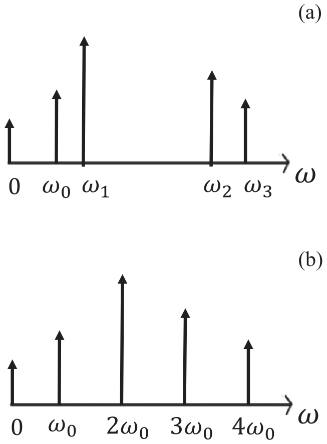

Fig. 1. (a) An arbitrary line spectrum. (b) The harmonic line spectrum of a periodic signal. Can we use the additional structure in the spectrum of a periodic signal to improve MUSIC?

where $2\pi/\omega_0$ (possibly not an integer) is usually considered as the 'period'. While MUSIC itself does not exploit this additional harmonic structure, it was shown in an important series of publications [11]–[13] that modifying MUSIC's search over complex exponentials so that we look for harmonically spaced peaks, improves the period estimates significantly. These methods were called Harmonic MUSIC (or HMUSIC). However, these methods are computationally much more complex than traditional MUSIC, especially when the input is a mixture of multiple periodic signals.

While (1) generically applies to several instances of periodicity such as speech, cardiology, EEG analysis and so forth, there is a second class of applications which have more structure than what is captured by (1). These are periodic signals whose periods are integers (or can be well approximated by integers) so that

$$x(n+P) = x(n) \qquad \forall\, n \in \mathbb{Z} \tag{2}$$

Here, the integer P is known as a *repetition index* of the signal, and the smallest positive repetition index is known as the *period*. Such applications include repeats in protein and DNA sequences. For instance, periodicity in the amino acid sequence of proteins often manifests as rich 3D repeating structures that play important roles in several diverse contexts (see Fig. 2) [1], [32]. Similarly, tandemly repeating nucleotide sequences in the DNA, known as micro-satellites, are widely used as bio-markers in forensics and kinship analysis, and are associated with several genetic disorders [4]. In these examples the period

Manuscript received March 21, 2018; revised August 14, 2018 and October 18, 2018; accepted October 26, 2018. Date of publication November 2, 2018; date of current version December 3, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Luca Venturino. This work was supported in part by the ONR Grants N00014-17-1-2732 and N00014-18-1-2390, in part by the NSF Grant CCF-1712633, and in part by the Amazon post doctoral fellowship. *(Corresponding author: Srikanth Venkata Tenneti.)*

The authors are with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: stenneti@caltech.edu; ppvnath@systems.caltech.edu).

Digital Object Identifier 10.1109/TSP.2018.2879039

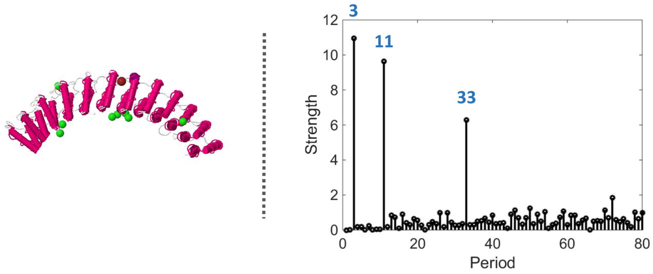

Fig. 2. Applications with Integer Periodicity: The protein AnkyrinR (PDB 1n11) that enables red blood cells to resist shear forces. Its period 33 structural repeats can be clearly identified in the plot on the right, produced by the proposed techniques.

is naturally an integer. In fact, many state of the art methods for conventional periodicity applications such as speech [10], [44], [71] are also based on integer period approximations. This paper shows that the simplicity of the integer period model (2) opens up the possibility for designing a more diverse class of MUSIC-like algorithms than prior works.

More specifically, for such signals with integer periods, this paper proposes a new formulation of MUSIC called *iMUSIC*, using the recently proposed Ramanujan Subspaces [68], [69] and Nested Periodic Subspaces (NPSs) [61]. The frequencies corresponding to signals with integer periods can be compactly represented by a non-uniform grid known as the Farey grid [70]. Based on the Farey grid, we propose an alternative to the classical MUSIC pseudospectrum: All the complex exponentials on the Farey grid belonging to a common Ramanujan subspace [68] are grouped together when computing this proposed pseudo-spectrum. It will be shown that the resulting algortihm yeilds much higher accuracies than classical MUSIC and its prior periodicity variants such as HMUSIC, while keeping the computational complexity very low.

Furthermore, the Ramanujan subspaces can alternatively be spanned by simple integer valued vectors instead of the Farey grid [70] (Fig. 3). In fact, using more general Nested Periodic Subspaces [61], one can construct many such examples of simple integer valued vectors that can be used to compute the proposed iMUSIC pseudospectrum instead of complex exponentials. Some of these new bases are very sparse, consisting of only 1's and 0's (Fig. 3(a)). These new representations give rise to a rich class of MUSIC-like algorithms that are well suited for integer period applications. Their advantages are demonstrated using examples that include Protein and DNA repeats. To the best of our knowledge, this is the first time MUSIC-like methods have been used on such bio-molecular repeats.

It should be mentioned here that there are other interesting algorithms such as the harmonic matching pursuit (HMP) [25], and expectation-maximization (EM) based algorithms [14] for taking advantage of the harmonicity in line spectra. While our focus in this work is only on MUSIC-like algorithms, we do include HMP and EM in our comparisons in Section V.

The mathematical formulation of classical MUSIC benefits greatly from the Vandermonde structure of complex exponentials. For instance, this is used in deriving the conditions for avoiding spurious peaks in the MUSIC psuedospectrum [51], [55]. The absence of a Vandermonde structure in NPSs introduces many new, but interesting challenges. For example, while we cannot guarantee the absence of additional (spurious) peaks in the iMUSIC pseudo-spectrum, *we can still prove that any such peak will not affect the estimated period*. These, and other such deviations from classical MUSIC will be rigorously addressed throughout the paper.

(a)

G_1	G_2	G_3		G_4	
1	1	1	0	1	0
1	0	0	1	0	1
1	1	0	0	0	0
1	0	1	0	0	0
1	1	0	1	1	0
1	0	0	0	0	1
1	1	1	0	0	0

(b)

S_1	S_2	S_3		S_4	
1	1	2	-1	2	0
1	-1	-1	2	0	2
1	1	-1	-1	-2	0
1	-1	2	-1	0	-2
1	1	-1	2	2	0
1	-1	-1	-1	0	2
1	1	2	-1	-2	0

Fig. 3. Simple Integer Alternatives to Complex-Exponentials: Bases of (a) The Natural Basis Subspaces, and (b) The Ramanujan Subspaces.

Before proceeding, we would like to make a small remark. The techniques we develop in this paper, and also MUSIC and its prior variants such as the HMUSIC algorithms [11]–[13], are based on the auto-correlation matrix of the signal. The reader may wonder if we can just estimate the period of a signal $x(n)$, $0 \leq n \leq L-1$, by looking at the peaks in the autocorrelation function instead:

$$r(k) = \frac{1}{L-k} \sum_{n=0}^{L-1-k} x(n)x^{*}(n+k) \tag{3}$$

If $x(n)$ has period P, we would expect $r(k)$ to have a peak whenever k is a multiple of P. Unfortunately, this does not work well often, such as for short datalengths, noisy inputs and for mixtures of periodic signals. For example, Fig 4(a) shows 100 samples of a signal that is a sum of randomly generated signals with periods 5, 7 and 13. Fig. 4 (b) plots the autocorrelation of this signal at 75 lags. As evident, there are no clear peaks at the true periods 5, 7 and 13. This happens even though there was no noise in the input signal to begin with. This is because, a mixture of periods 5, 7 and 13 has a resulting period $= \mathrm{LCM}(5, 7, 13) = 455$, which is where a peak in the autocorrelation is expected. From a theoretical perspective, there is no reason to expect peaks at each of the component periods in the mixture. Also, since the available datalength (100 samples) is much less than 455, we cannot even compute the autocorrelation at 455. All is not lost however, as these same 75 samples of the autocorrelation function can be used in a different way to obtain the plot shown in Fig. 4 (c), where there are clean peaks at periods 5, 7 and 13. This was done using one of the proposed iMUSIC techniques known as Farey MUSIC. Thus one way to look at the techniques

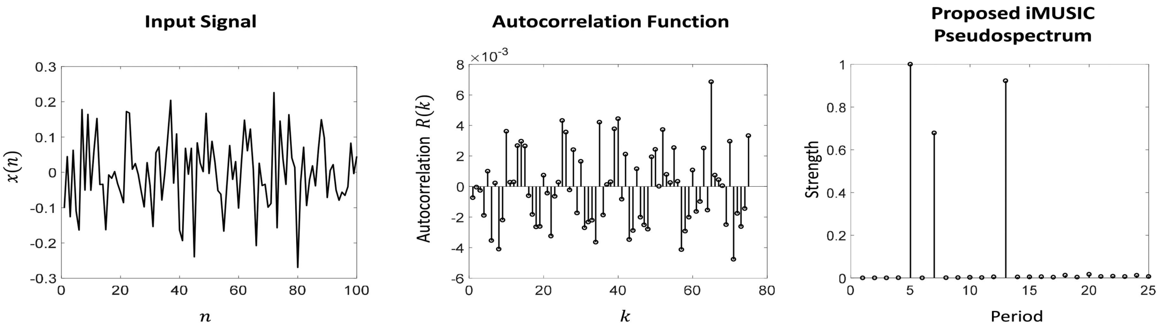

Fig. 4. MUSIC based techniques can be viewed as a more sophisticated way of estimating the periods using the auto-correlation function, than just looking for peaks in (3). Part (a) A sum of randomly generated signals with periods 5, 7 and 13. Part (b) The autocorrelation function at 75 lags. Part (c) The proposed Farey iMUSIC psudospectrum using the same autocorrelation values that were computed in Part (b). See Section I for details.

in the following sections is as more sophisticated ways of using the autocorrelation function than just looking for peaks in (3).

A. Outline

Section II summarizes MUSIC and its prior adaptations to periodic signals (the HMUSIC algorithms). Section III starts with a brief summary of Ramanujan and Nested Periodic Subspaces (NPSs). The proposed iMUSIC framework is then introduced, first using the Farey grid of complex exponentials. This is generalized to iMUSIC using other NPSs in Section IV, allowing the use of integer valued vectors for spanning the signal subspace. The conditions for identifiability of the true periods using such integer bases are rigorously derived here. Section V contains several simulations and comparisons with other techniques, including examples of protein and DNA repeats.

B. Notations

1) $d|p$ denotes that d is a divisor of p.
2) (k,d) denotes the greatest common divisor (GCD) of k and d. The least common multiple of k and d is denoted by $\mathrm{LCM}(k,d)$.
3) $\phi(d)$ is the Euler-totient function of d. It is equal to the number of positive integers $\leq d$ and coprime to d.
4) Vectors are denoted by bold lower case font (e.g., $\mathbf{x}$), matrices by bold upper case font (e.g., $\mathbf{A}$) and sets by blackboard font (e.g., $\mathbb{B}$).
5) $\mathbf{x}^\dagger$ denotes the transpose conjugate of $\mathbf{x}$.

II. MUSIC and Periodicity: An Overview of Prior Works

In this section, we will briefly outline the MUSIC algorithm [51], [66], and its prior adaptations to periodic signals. Let us begin with the following signal model:

$$x(n) = \sum_{k=0}^{K-1} c_k e^{j\omega_k n} + e(n), \tag{4}$$

where ω_k are distinct frequencies in $[-\pi,\pi)$ and $e(n)$ is zero-mean white noise with variance σ_e^2. Most prior MUSIC-based works model $c_k \in \mathbb{C}$ as random variables [12], [34], [66]. But we will assume them to be constants here, since such is the case in most applications of periodicity. Note that (4) can also model real valued signals, in which case the frequencies occur in pairs $\{\omega_k, -\omega_k\}$, with the corresponding coefficients occurring as complex conjugates $\{c_k, c_k^*\}$. So the discussion in the following applies to both real and complex valued signals.

Assume that there are L samples of $x(n), 1 \leq n \leq L$, and define the ith block of data as

$$\mathbf{x}(i) = \left[x(i)\ x(i+1)\ \cdots\ x(i+N-1)\right]^T, \tag{5}$$

where $N < L$ is the blocksize. We can call $\mathbf{x}(i)$ the ith “snapshot” but note that successive blocks are not independent (they have an overlap of $N-1$ samples). There are

$$M = L - N + 1 \tag{6}$$

blocks. Note that we can write the ith block as

$$\mathbf{x}(i) = \left[\mathbf{a}_0(i)\ \mathbf{a}_1(i)\ \cdots\ \mathbf{a}_{K-1}(i)\right] \underbrace{\begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{K-1} \end{bmatrix}}_{\mathbf{c}} + \mathbf{e}(i) \tag{7}$$

where $\mathbf{a}_k(i)$ are Vandermonde vectors up to scale:

$$\begin{aligned}\mathbf{a}_k(i) &= \left[e^{j\omega_k i}\ e^{j\omega_k(i+1)}\ \cdots\ e^{j\omega_k(i+N-1)}\right]^T \\ &= e^{j\omega_k i} \underbrace{\overset{T}{\left[1\ e^{j\omega_k}\ \cdots\ e^{j\omega_k(N-1)}\right]}}_{\mathbf{w}_k}\end{aligned}$$

Thus

$$\mathbf{x}(i) = \mathbf{A}\mathbf{\Lambda}_\omega(i)\mathbf{c} + \mathbf{e}(i) \tag{8}$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ e^{j\omega_0} & e^{j\omega_1} & \cdots & e^{j\omega_{K-1}} \\ e^{j2\omega_0} & e^{j2\omega_1} & \cdots & e^{j2\omega_{K-1}} \\ \vdots & \vdots & \ddots & \vdots \\ e^{j(N-1)\omega_0} & e^{j(N-1)\omega_1} & \cdots & e^{j(N-1)\omega_{K-1}} \end{bmatrix} \tag{9}$$

is a Vandermonde matrix independent of i, and

$$\mathbf{\Lambda}_\omega(i) = \mathrm{diag}\,\{e^{j\omega_0 i}, e^{j\omega_1 i}, \ldots, e^{j\omega_{K-1} i}\} \tag{10}$$

Define the data matrix to be

$$\mathbf{X} = \left[\mathbf{x}(\mathbf{1})\ \mathbf{x}(2) \cdots\ \mathbf{x}(M)\right]. \tag{11}$$

Then the sample autocorrelation matrix is

$$\widehat{\mathbf{R}} = \frac{1}{M}\mathbf{X}\mathbf{X}^{\dagger} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{x}(i)\mathbf{x}^{\dagger}(i) \tag{12}$$

For large M this can be approximated as

$$\widehat{\mathbf{R}} \approx \mathbf{A}\mathbf{\Lambda}_c\mathbf{A}^{\dagger} + \sigma_e^2\mathbf{I}_N \tag{13}$$

where $\mathbf{\Lambda}_c = \text{diag}\ \{|c_0|^2, |c_1|^2, \ldots, |c_{K-1}|^2\}$. (Please see the Appendix for a proof of (13)).

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ be the eigenvalues of $\widehat{\mathbf{R}}$. Since $\text{Rank}(\mathbf{A}\mathbf{\Lambda}_c\mathbf{A}^{\dagger}) = K$, it can be shown that $\lambda_{K+1} = \lambda_{K+2} = \ldots = \lambda_N = \sigma_e^2$. These are commonly referred to as the noise eigenvalues, and their corresponding eigenvectors $\mathbf{U_e} = [\mathbf{u}_{K+1},\ \mathbf{u}_{K+2},\ \ldots\ , \mathbf{u}_N]$, as the noise eigenvectors. Using (13), we obtain

$$\mathbf{A}\mathbf{\Lambda}_c\mathbf{A}^{\dagger}\mathbf{U_e} = \mathbf{0} \tag{14}$$

As long as $N \geq K$, $\mathbf{\Lambda}_c$ will have a full rank and $\mathbf{A}$ will have a full column-rank, because $c_k \neq 0$ and ω_k are distinct in $[-\pi, \pi)$. So (14) is equivalent to:

$$\mathbf{A}^{\dagger}\mathbf{U_e} = \mathbf{0} \tag{15}$$

That is, the complex-exponentials in (4) turn out to be orthogonal to the noise eigenspace. So one can then use the following to estimate the ω_k:

$$\min_{\omega \in (-\pi, \pi]} \ \|\mathbf{a}^{\dagger}(\omega)\mathbf{U_e}\|_2^2 \tag{16}$$

where $\mathbf{a}(\omega) = [1,\ e^{j\omega}, e^{2j\omega},\ \ldots\ , e^{(N-1)j\omega}]^T$. It can be proved [51], [55] that as long as $N > K$, the only complex-exponentials that are orthogonal to the noise eigenspace are those in (4). Hence, there will be no spurious estimates when solving (16).

Now, for applications with periodicity, MUSIC by itself does not exploit the fact that the lines in the spectrum are harmonically spaced (Fig. 1). Taking this into account, Christensen et al. [12] proposed to modify (16) as

$$\min_{\omega \in (-\pi, \pi]} \ \min_{K} \ \frac{\|\mathbf{B}^{\dagger}(\omega)\mathbf{U_e}\|_F^2}{KN(N-K)} \tag{17}$$

where $\mathbf{B}(\omega) = [\mathbf{a}(0), \mathbf{a}(\omega),\ \mathbf{a}(2\omega),\ \ldots\ , \mathbf{a}((K-1)\omega)]$. The factor of $KN(N-K)$ is a normalization term. The resulting algorithm was called the Harmonic MUSIC (HMUSIC) algorithm. It was further generalized to the case of mixtures of periodic signals in [11] as follows:

$$\min_{\{K_l\}_{l=0}^{Q-1}} \ \min_{\{\omega_l\}_{l=0}^{Q-1}} \ \sum_{l=0}^{Q-1} \frac{\|\mathbf{B}_{K_l}^{\dagger}(\omega_l)\mathbf{U_e}\|_F^2}{KN(N-K)} \tag{18}$$

where Q is the number of component periodic signals in the mixture. The various ω_l, $0 \leq l \leq Q-1$, represent the Q fundamental frequencies. K_l is the number of spectral lines that corresponsd to harmonics of ω_l. $\mathbf{B}_{K_l}$ is a matrix similar to the $\mathbf{B}$ in (17), with columns being complex exponentials with frequencies ω_l and its harmonics. $K = \sum_l K_l$ is the total signal space dimension.

Both (17) and (18) were shown to offer better estimates than MUSIC in the context of pitch estimation [11], [12]. However, notice that both (17) and (18) involve computationally intensive integer optimizations, apart from the optimization over the ω's. The most commonly used approach for optimizing over the ω's in (16), (17) and (18) is to evaluate them over uniform frequency grids, as was done in [12]. The uniform grid allows to exploit FFT in the computations, as shown in [12]. We will be using the same in the simulations of Section V. An interesting analytical framework on how to select the grid resolution for such problems, considering the trade-off between estimation accuracy and computational time, is presented in [41]. For MUSIC, alternatives based on polynomial root finding (such as the Root-MUSIC algorithm in [3], [48]) have also been proposed in the literature.

For signals that can be approximated well by the integer period model of (2) (such as DNA and Protein repeats), we can develop much simpler techniques than the above methods, with a significantly higher accuracy as well. One of the proposed methods includes using a non-uniform frequency grid known as the Farey grid [70], which will be shown to be ideally suited for the integer period model. We shall present these next.

III. The Proposed Methods

We begin with a brief review of Ramanujan subspaces [68] and nested periodic subspaces (NPS) [61]. These were introduced recently for the representation of sequences with integer periods. Our proposed iMUSIC algorithms (Secs. III-B and IV) will be based on these.

A. Ramanujan and Nested Periodic Subspaces: An Overview

For any integer $q > 0$, the **Ramanujan subspace** $\mathcal{S}_q$ is the space of period-q signals spanned by

$$s_q^{(k)}(n) \triangleq e^{j2\pi kn/q} \tag{19}$$

where $1 \leq k \leq q$ and $(k, q) = 1$ (i.e., k is coprime to q). $\mathcal{S}_q$ has dimension $\phi(q)$ (Euler totient function, Section I-B). It can be shown [69] that any period-P signal ($P =$ integer) can be spanned by the signals from $\mathcal{S}_{q_i}$ where $q_i|P$ (i.e., q_i are divisors of P). So the number of basis functions involved is $\sum_{q_i|P} \phi(q_i)$, which turns out to be precisely P [28].

A dictionary for representing all sequences with integer periods $\leq P_{max}$ takes the form

$$\mathbf{D} = \begin{array}{cc} & \begin{array}{cccc} \phi(1) & \phi(2) & \cdots & \phi(P_{max}) \end{array} \\ N & \left(\begin{array}{c|c|c|c} \mathbf{D}_1 & \mathbf{D}_2 & \cdots & \mathbf{D}_{P_{max}} \end{array} \right) \end{array} \tag{20}$$

where $\mathbf{D}_q$ has $\phi(q)$ columns (basis for $\mathcal{S}_q$) as indicated. One choice for the vectors in $\mathbf{D}_q$ is the set of $\phi(q)$ complex Vandermode vectors

$$\left[1\ e^{j2\pi k/q}\ e^{j4\pi k/q}\ e^{j6\pi k/q}\ \ldots\right]^T \tag{21}$$

where $(k, q) = 1$ (with $1 \leq k \leq q$). Notice that these are nothing but a subset of $\phi(q)$ columns from the $q \times q$ DFT matrix (periodically extended). Another alternative set of $\phi(q)$ basis functions for $\mathcal{S}_q$ (i.e., columns of $\mathbf{D}_q$) is based on the Ramanujan sum

$$c_q(n) = \sum_{\substack{(k,q)=1 \\ 1 \leq k \leq q}} e^{j\frac{2\pi k}{q}n} \tag{22}$$

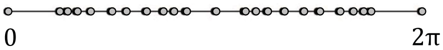

Fig. 5. The Non-Uniform Farey Grid: The Farey frequencies needed to span periods in the range $1 \leq P \leq 8$. See Section III for details.

defined in [47] and reviewed in [68]. Then the mth column of $\mathbf{D}_q$ has the form

$$\left[c_q(-m) \; c_q(1-m) \; c_q(2-m) \cdots \right]^T \tag{23}$$

for $0 \leq m \leq \phi(q) - 1$. Since $c_q(n)$ are known to be real integers, this forms an integer basis for $\mathcal{S}_q$. Examples of such Ramanujan basis are shown in Fig. 3(b). Now, given an integer P_{max}, the set of all columns of the form (21) represents frequencies of the form

$$2\pi \frac{k}{q}, \;\; 1 \leq k \leq q, (k, q) = 1 \quad \text{where} \;\; 1 \leq q \leq P_{max} \tag{24}$$

This is said to be the **Farey** frequency grid for P_{max} [70] (inspired by Farey sets [28]). It is the grid of frequencies represented by (20). Note that this is a nonuniform grid as demonstrated in Fig. 5, and has

$$\sum_{q=1}^{P} \phi(q) \approx \frac{3P_{max}^2}{\pi^2} \tag{25}$$

frequencies [28], [70]. We say that Eq. (20) is a *Farey dictionary* if Eq. (21) is used for the columns of the matrix $\mathbf{D}_q$, and a *Ramanujan dictionary* if Eq. (23) is used [70]. Even though the spaces $\mathcal{S}_q$ are orthogonal for different q [68], the columns from $\mathbf{D}_q$ are not exactly orthogonal for different q when $\mathbf{D}$ has only $N < \infty$ rows.

It can be shown [68] that signals in $\mathcal{S}_q$ (hence the columns in $\mathbf{D}_q$) have period exactly q (they cannot be smaller, such as, a divisor of q). If a signal with period P is represented as a linear combination of elements in the subspaces $\mathcal{S}_{q_i}$, $1 \leq q_i \leq P_{max}$, then only those q_i that are divisors of P can have nonzero coefficients. In fact, once the representation $x(n) = \sum_i \alpha_i x_{q_i}(n)$ has been found where $x_{q_i}(n) \in \mathcal{S}_{q_i}$ and $\alpha_i \neq 0$, the period P can be shown to be

$$P = \text{LCM}\left\{q_{i_1}, q_{i_2}, \cdots\right\} \tag{26}$$

This is called the *LCM property* [69], [61]. That is, given an input signal, if we can find the exact set of Ramanujan subspaces that span the signal, then the LCM of the periods of those subspaces will be equal to the period of the input.

It was shown in [63], that the Ramanujan subspaces can alternatively be defined using the Exactly Periodic Subspaces of [39] and the Intrinsic Integer Periodic Functions of [46]. The *nested periodic subspaces* $\mathcal{N}_q$ introduced in [61] are a generalization of Ramanujan spaces $\mathcal{S}_q$. Examples include the so-called *natural basis*, and *random basis* for periodic signals defined in [61]. The natural basis is demonstrated in Fig. 3(a). Unlike $\mathcal{S}_q$, the spaces $\mathcal{N}_q$ are not necessarily orthogonal for different q. But just like $\mathcal{S}_q$, the space $\mathcal{N}_q$ has dimension $\phi(q)$. A dictionary similar to Eq. (20) can be defined based on such subspaces, and just like Ramanujan subspaces, any set of nested periodic subspaces can be used to represent periodic signals and enjoys the LCM property (26). The advantage of the generalized spaces $\mathcal{N}_q$ over $\mathcal{S}_q$ is that their basis functions can be very simple (consisting just 0's and 1's), as demonstrated in Fig. 3.

The LCM property easily extends to the case of mixtures of periodic signals. For example, if the input is a mixture of period 4 and period 6 signals, then the subspaces with periods 1, 2, 3, 4 and 6 are involved in spanning it (i.e., divisors of 4 and 6). More generally, if a mixture of periodic signals can be spanned by nested periodic subspaces of periods $\mathbb{P} = \{P_1, P_2, \ldots P_K\}$, then it can be shown that the component periods in the mixture are given by the following set[1] [61], [65]:

$$\mathbb{P}_H = \{P_i \in \mathbb{P} : MP_i \notin \mathbb{P} \;\forall M > 1\} \tag{27}$$

that is, those numbers in $\mathbb{P}$ that do not have any multiples also present in $\mathbb{P}$. Dictionaries based on NPSs have been shown to offer several new advantages over traditional period estimation techniques [16], [61], [63].

B. The Proposed iMUSIC Formulation

As explained in Section II, when the Vandermonde vectors (columns of $\mathbf{A}$) in Eq. (9) have a harmonic structure, it can be exploited to improve the MUSIC algorithm (e.g., HMUSIC [12]). Now, when $x(n)$ has integer period $\leq P_{max}$, the frequencies ω_i in (9) can only have the specific form (24). That is, the Vandermonde vectors are similar to the atoms in the Farey dictionary (20). In this case there is a different way to define the MUSIC spectrum which works much better than traditional MUSIC and HMUSIC. We refer to this as *Farey MUSIC*; as we shall see below, this is more than just restricting the computation of traditional MUSIC spectrum to the Farey grid. Since the Farey MUSIC algorithm is specifically designed to find integer periods, we also call it *iMUSIC* (where the i stands for integer period). Also, replacing the Farey atoms with other types of nested periodic bases leads to several generalizations of iMUSIC, as we shall see in Section IV.

Let us begin by assuming that $x(n)$ in (4) is a period-P signal. So the K columns of $\mathbf{A}$ in (9) are a subset of the atoms of the Farey dictionary, with periods being divisors of P. We can follow the derivation in Section II to obtain (15). That is, the atoms of the Farey dictionary that span the signal turn out to be orthogonal to the noise subspace. As long as N, the size of the snapshots in (5), is larger than K, no other Farey atoms will satisfy (15). At this point, we propose the following alternative to the MUSIC (Eq. (16)) and HMUSIC pseudo-spectra (Eqs. (17) and (18)): For every integer P, we compute

$$S_F(P) = \frac{1}{\phi(P)} \quad \sum_{m=1}^{\phi(P)} \frac{1}{\|\mathbf{U_e}^{\dagger}\mathbf{s}_P^{(m)}\|_2^2} \tag{28}$$

where $\{\mathbf{s}_P^{(m)}\}_{m=1}^{\phi(P)}$ are the $\phi(P)$ period-P atoms of the Farey dictionary. The $\phi(P)$ term in the denominator is a normalizing factor. *A plot of (28), with the integer P as the x-axis is the discrete iMUSIC pseudospectrum based on the Farey dictionary.*

[1] The question of the extent to which the component periods in a mixture can be uniquely identified is a fundamental problem, that was only recently addressed in [65]. Since those details are quite involved, we will skip them here and refer the reader to Section III in [65]. Eq. (27) in essence ensures that we do not declare a harmonic of a component period as another component period in the mixture.

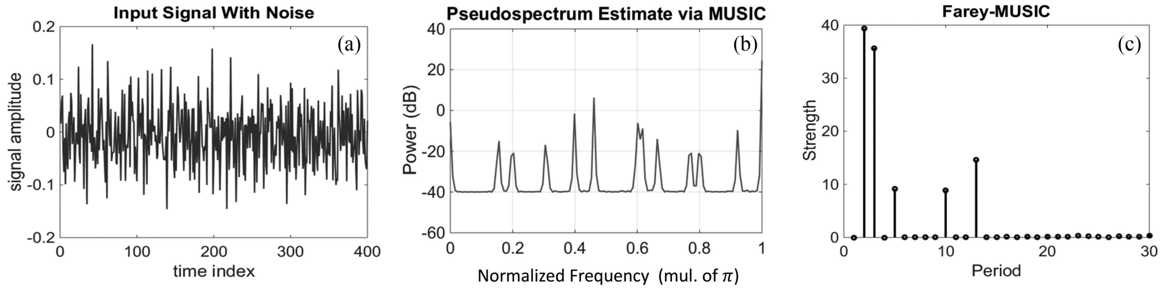

Fig. 6. Demonstrating the proposed iMUSIC method using a Farey dictionary on a mixture of periods 3, 10 and 13. (a) The noisy periodic signal, (b) conventional MUSIC, and (c) the new iMUSIC method in (28). See Section III-B for details.

C. Farey MUSIC Versus MUSIC and HMUSIC

Note that Eq. (28) is not just restricting ordinary MUSIC to a special non-uniform grid. It is differs from classical MUSIC and HMUSIC in the following ways:

1. Ramanujan Subspaces: Eq. (28) consolidates all the Farey atoms in each Ramanujan subspace into one sum. In this way each iMUSIC spectrum line is for one Ramanujan subspace. The LCM property of the Ramanujan subspaces applied to the peaks of (28), yields the period. The number of lines is therefore different from the number of lines in ordinary MUSIC.

Notice that HMUSIC in (17) and (18) also groups together harmonic multiples of a fundamental frequency. But while HMUSIC combines consecutive harmonics $\omega_l,\ 2\omega_l,\ 3\omega_l,\ \ldots,$ $K_l - 1\omega_l$ in (18) via the $\mathbf{B}_{K_l}$ matrix, in Farey-MUSIC, for every period P, we only combine those $e^{j2\pi kn/P}$ for which $(k,P)=1$. K_l itself in HMUSIC (18) is found by optimizing over all possible values. On the other hand, in Farey-MUSIC, the number of complex exponentials we associate with period P in (28) is fixed ($=\phi(P)$).

2. Mixtures of Periodic Signals: Unlike HMUSIC in (18), the complexity of iMUSIC does not increase with the number of component signals in a mixture, or with the number of harmonics for each component. The complexity of HMUSIC increases exponentially with the number of hidden periodic components Q in (18). This is because the number of ways in which K_l's in (18) can be chosen to add up to the total signal space dimension K increases exponentially with Q (see Appendix B). The proposed iMUSIC (28) does not need to compute the exact partition of the total signal space dimension into individual K_l's. We just need to compute (28) irrespective of the number of hidden periodic components.

3. The Period of a Complex-Exponential: There is a subtle distinction in how we interpret the period of a complex exponential. In prior works, the period of $e^{j2\pi kn/P}$ was interpreted as P/k. However, we follow the strict integer period definition as given in (2), so that the period is actually $P/\gcd(P,k)$.

All these differences when put together, result in significantly better accuracies and much simpler algorithmic complexity for integer period estimation, as will be seen in Section V. Before proceeding, we will show a simple demonstration of the iMUSIC equation (28). Fig. 6(a) shows a sum of randomly generated signals with periods 3, 10 and 13 and SNR 5dB. The total signal length (L in Section II) was 400. This signal was broken down into successive blocks of length 101 samples (N in (5)). K, the number of complex exponentials in (4), turns out to be 24 for this choice of periods. In practice, this true value of K is unknown a priori, so we estimate it here using a simple metric: all the eigenvalues of the auto-correlation matrix smaller than 5% of the maximum eigenvalue were considered as noise eigenvalues. Fig. 6(b) shows the conventional MUSIC pseudospectrum for reference. The peaks correspond to periods 12.79, 9.85, 6.56, 5.02, 4.34, 3.32, 3.24, 3.01, 2.59, 2.51 and 2.17. Notice that it is quite inconvenient to spot the true periods 3, 10 and 13 from this set. Fig. 6(c) shows the iMUSIC pseudospectrum computed using (28). It is easy to identify distinct peaks at periods 2, 3, 5, 10 and 13. Using the LCM property, we can deduce that these correspond to periods 3, 10 and 13.

IV. Generalizing iMUSIC From Farey Atoms to Other NPS Bases

Eq. (28) can alternatively be implemented using integer valued basis vectors instead of complex exponentials. This can be done using the Nested Periodic Subspaces (NPSs) [61], [63] described in Section III-A. The NPSs are generalizations of Ramanujan subspaces, and include several examples of integer bases for representing periodic sequences (Fig. 3). In fact, as explained in Section III-A, the Ramanujan subspaces can themselves be spanned by integer valued vectors instead of the Farey atoms.

Algorithmically, generalizing the iMUSIC spectrum using such NPSs is done as follows: We compute the following for every integer P instead of (28):

$$S_N(P) = \frac{1}{\phi(P)} \sum_{m=1}^{\phi(P)} \frac{1}{\|\mathbf{U}_{\mathrm{e}}{}^{\dagger}\mathbf{b}_P^{(m)}\|_2^2} \qquad (29)$$

where $\{\mathbf{b}_P^{(m)}\}_{m=1}^{\phi(P)}$ are the $\phi(P)$ period-P NPS basis vectors. Using the LCM property of NPSs, we can once again determine the period from the peaks of Eq. (29). For example, Fig. 7 shows plots of Eq. (29) vs. P for various integer valued NPS bases, for the signal shown in Fig. 6(a). When the atoms $\mathbf{b}_P^{(m)}$ come from a Ramanujan dictionary (Fig. 3(b)), we call (29) as Ramanujan MUSIC. Natural basis MUSIC and Random NPS MUSIC can be defined similarly using their respective dictionaries [61]. All these plots have clean peaks at periods 2, 3, 5, 10 and 13. Using the LCM property of NPSs, it is easy to see that they represent periods 3, 10 and 13.

Although the above idea is simple, the non-Vandermonde nature of the NPS bases introduces several challenges in the

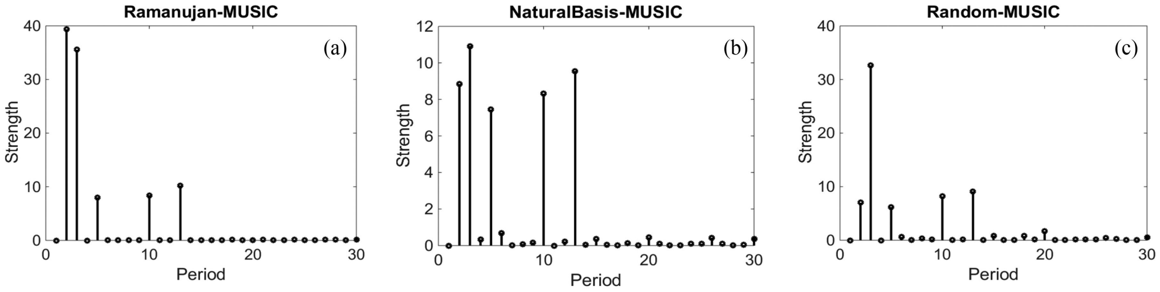

Fig. 7. Demonstrating the NPS based iMUSIC methods on the signal shown in Fig. 6 using (a) a Ramanujan dictionary, (b) a natural basis dictionary and (c) a randomly generated NPS dictionary. See Section IV for details.

mathematical formulation of (29) when compared to classical MUSIC. So in the remainder of this section, we will study (29) in a rigorous fashion. To start with, let us assume that $x(n)$ is a period P signal. Following the derivations of Section II, we can arrive at (14). Now, since $x(n)$ has integer period, the columns of $\mathbf{A}$ themselves have integer periods (atoms of the Farey dictionary), and hence can be spanned by any other set of NPSs (such as say natural basis). We can write this as:

$$\mathbf{A}_{N\times K} = \mathbf{B}_{N\times K'}\mathbf{T}_{K'\times K} \tag{30}$$

where the K' columns of $\mathbf{B}$ are the basis vectors of the other NPS. It can be shown [61] that the LCM property applied to the columns of either $\mathbf{A}$ or $\mathbf{B}$ yields the same answer, namely P. It is useful to consider two separate cases at this point, depending on whether K equals K'.

A. The Case When $K = K'$

In general, K and K' can be different. But to start with, we assume $K = K'$ since it is the most common situation in applications with integer periods. Conceptually, $K = K'$ means that one would require the same number of NPS basis vectors to span the snapshots, no matter which NPS is chosen. For instance, if $x(n)$ was randomly generated, say by repeating a $P \times 1$ Gaussian random vector $\mathbf{x}_P$, then it can be shown that $K = K'$ with probability 1 (Appendix C). This also applies to mixtures of periodic signals, when each component signal is randomly generated. In applications such as DNA and protein repeats, where the nucleotides or the amino acids are mapped to numbers using scales such as the molecular size, hydrophobicity etc. [1], it is quite natural to expect that $K = K'$. The case of $K \neq K'$ will be addressed later in Section IV-B.

We can re-write (14) using (30) as follows:

$$\mathbf{B}\mathbf{T}\mathbf{\Lambda}_c\mathbf{T}^{\dagger}\mathbf{B}^{\dagger}\mathbf{U_e} = \mathbf{0} \tag{31}$$

In (30), as long as $N > K$, $\mathbf{A}$ will have a full column rank (Vandermonde property). This implies that $\mathbf{B}$ and $\mathbf{T}$ will also have full column ranks K $(= K')$, and hence $\mathbf{B}\mathbf{T}\mathbf{\Lambda_c}\mathbf{T}^{\dagger}$ will have a full column rank in (31). So (31) is equivalent to:

$$\mathbf{B}^{\dagger}\mathbf{U_e} = \mathbf{0} \tag{32}$$

Notice that this is similar to (15), but involves the columns of an NPS dictionary rather than complex exponentials. So given such an NPS dictionary, plotting (29) for every period $\leq P_{max}$ will result in peaks at periods corresponding to the columns of $\mathbf{B}$. So we may think of using the LCM property on those peaks to estimate the period. But before we can do so, just like in classical MUSIC, we need to address the following question first: Can there be NPS basis vectors other than the columns of $\mathbf{B}$ that are also orthogonal to $\mathbf{U_e}$?

We have an interesting deviation from classical MUSIC in this aspect. While we cannot guarantee the absence of such additional (spurious) NPS basis vectors producing peaks in (29), we can nevertheless prove that *any such additional peaks will not affect the period estimate*. To see this, we first need the following result proved in [65]:

Theorem 1: Let $x(n)$ be a noiseless periodic signal whose period is known to lie in the integer set $\mathbb{P} = \{P_1, P_2, \ldots, P_K\}$. To be able to uniquely identify its period using L consecutive samples, it is both necessary and sufficient that:

$$L \geq L_{min} = \max_{P_i,P_j\in\mathbb{P}} P_i + P_j - (P_i, P_j) \tag{33}$$

◇

The above result is a fundamental identifiability result that is independent of which estimation technique is used [65]. We will use it to prove the following:

Theorem 2: Suppose the period of $x(n)$ in (4) is known a priori to lie in the integer set $\mathbb{P} = \{P_1, P_2, \ldots, P_K\}$. If N, the length of the snapshots in (5), satisfies:

$$N \geq L_{min} = \max_{P_i,P_j\in\mathbb{P}} P_i + P_j - (P_i, P_j) \tag{34}$$

then the LCM of the periods of all the NPS basis vectors that are orthogonal to $\mathbf{U}_e$, will be equal to the true period of the signal.

Proof: Let us assume that the input's period is P. As mentioned earlier, the LCM of the periods of the columns in $\mathbf{A}$ in (9) and $\mathbf{B}$ in (30) will be equal to P. Suppose $\mathbf{b}$ is an NPS basis vector that is not a column of $\mathbf{B}$, but still satisfies $\mathbf{b}^{\dagger}\mathbf{U}_e = \mathbf{0}$. There are two possibilities:

Case (i): Period of $\mathbf{b}$ divides P. In this case, even if $\mathbf{b}^{\dagger}\mathbf{U}_e = \mathbf{0}$, a peak in the psuedospectrum at period of $\mathbf{b}$ will not change the LCM estimate. So such a spurious peak will not lead to a false period estimate.

Case (ii): Period of $\mathbf{b}$ does not divide P. We will show using contradiction that such a $\mathbf{b}$ cannot exist. If there was such a $\mathbf{b}$, then $\mathbf{b}$, along with the columns of $\mathbf{B}$ will constitute $K + 1$ vectors in the K dimensional null-space of $\mathbf{U}_e^{\dagger}$. When N satisfies (34), it follows in particular that $N > P_{max} \geq K$. $N > K$ implies that $\mathbf{A}$ in (30) will have full column rank (Vandermonde property), and so $\mathbf{B}$ will also have full column rank K (recall that we assumed $K = K'$ to start with). This would mean that $\mathbf{B}$ is a basis for the null space of $\mathbf{U}_e^{\dagger}$, and so:

$$\mathbf{b} = \mathbf{B}\mathbf{v} \tag{35}$$

for some vector $\mathbf{v}$. Notice that the L.H.S. is a length N segment of a signal whose period does not divide P. The R.H.S. is a segment of a signal whose period necessarily divides P, since the columns in $\mathbf{B}$ are NPS basis vectors whose periods are divisors of P. As long as $N \geq L_{min}$ according to Theorem 1, such an ambiguity in identifying the period is not possible. Hence we arrive at a contradiction to the existence of such a $\mathbf{b}$. ■

Remark 1: When the set of possible periods in Theorem 2 is $\mathbb{P} = \{1,\ 2,\ 3,\ \ldots,\ P_{max}\}$, L_{min} turns out to be $2P_{max} - 2$. This is because, the pair of candidate periods P_{max} and $P_{max} - 1$ not only maximize $P_i + P_j$ in (34), but also have the minimum possible GCD.

Remark 2: Theorem 2 can be generalized to mixtures of periodic signals. If $x(n)$ were a mixture of M periodic signals with periods in $\{1,\ 2,\ 3,\ \ldots,\ P_{max}\}$, then the minimum N is approximately:

$$N \geq 2MP_{max} \tag{36}$$

For readers familiar with [65], (36) is in fact an approximation of the following precise lower bound:

$$N \ \geq \ N_{min} = \max_{\substack{\mathbb{P}_i,\mathbb{P}_j \subset \mathbb{P} \\ \mathbb{P}_i,\mathbb{P}_j \ \text{are} \\ \text{M}-\text{sets of size} = N}} \sum_{d \in \text{D.S.}(\{\mathbb{P}_i \cup \mathbb{P}_j\})} \phi(d) \tag{37}$$

The proof is based on the generalization of Theorem 1 to mixtures of periodic signals [65]. The details are quite involved for the scope of this paper, so we will skip the proof here, and refer the interested reader to [65] for directions.

Remark 3: Theorem 2 is tight in the following sense: It is possible to construct examples of NPSs for which spurious peaks will affect the period estimate when N doesn't satisfy (34). But for most NPSs, a smaller N may be sufficient. For instance, if we use the Farey atoms, it is easy to show using their Vandermonde structure [55] that we just need:

$$N > \max_{P_i \in \mathbb{P}} P_i \tag{38}$$

instead of (34) in Theorem 2. However, deriving the precise necessary and sufficient bounds for other NPSs that do not have a Vandermonde structure is difficult. Theorem 2 is useful in this regard.

So we have so far shown that as long as $K = K'$, and the snapshot length satisfies (34), the period of the signal can be estimated using (29). We will now discuss the case of $K \neq K'$.

B. The case of $K \neq K'$

Let us consider the following two cases separately:

Case A. $K > K'$: This will not happen as long as the snapshot length $N > K$, because then $\mathbf{A}$ in (30) will have full column rank K. So at least K linearly independent columns are needed in $\mathbf{B}$ in the R.H.S. of (30).

Case B. $K < K'$: This can occur in some cases. For instance, if $x(n) = e^{j2\pi n/P}$, then $K = 1$, as only one Farey column is required to span the snapshots of $x(n)$. But if we use natural basis subspaces (Fig. 1), then $K' = P$, as it can be shown that P basis vectors of the natural basis subspaces are needed to span each snapshot of this $x(n)$.

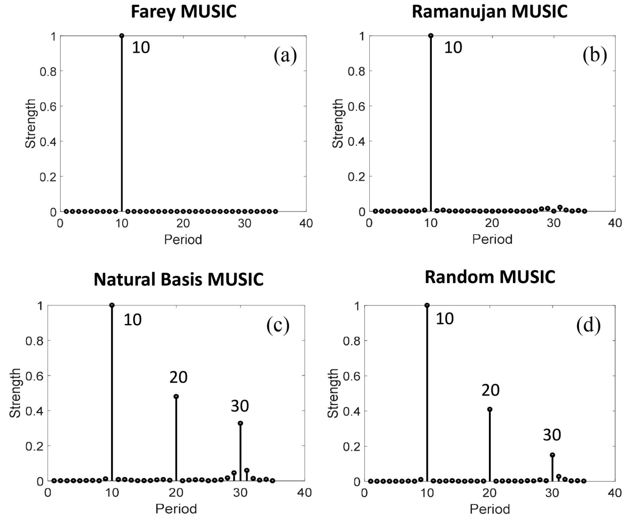

Fig. 8. The effect of $K < K'$ on the pseudo-spectrum of (a) Ramaujan Subspaces (Farey basis) (b) Ramanujan Subspaces (integer basis) (c) Natural Basis Subspaces and (d) Randomly generated NPSs. See Section V for details.

When $K < K'$, $\mathbf{T}$ in (30) will not have a full rank. Hence, (31) does not imply (32). So it is quite possible that some of the columns of $\mathbf{B}$ are not orthogonal to $\mathbf{U}_\mathrm{e}$. So using (29) and LCM property is not theoretically guaranteed to give the correct period estimate. This is a fundamental limitation of any non-Farey NPS basis. Nevertheless, it was experimentally observed that:

- For iMUSIC using Ramanujan subspaces, (28) gave the correct period estimates even for the non-Vandermonde integer basis vectors.
- For Natural Basis subspaces and randomly generated NPSs, the only spurious peaks observed were smaller peaks at multiples of the true period. So the period could still be estimated upto a multiple.

As an illustration, let us consider the following signal: $x(n) = e^{j2\pi n/10}$. For this signal, we would need only one Farey basis vector to span its snapshots, while we would need 10 basis vectors from Ramanujan integer basis, and similarly from the Natural Basis subspace. For randomly generated NPSs as well, we would need 10 basis vectors with probability 1. So for each of these other NPSs, we have $K' = 10$ in (30). Fig. 8 shows the pseudo-spectra obtained from each of these NPSs using (29). As is evident, $K < K'$ wasn't really an issue when using Ramanujan subspaces, even for the non-Vandermonde integer basis vectors. There were no spurious peaks or missing peaks. For the natural basis and the randomly generated NPSs, we can see spurious peaks at multiples of the true period, which is 10. So the period can only be estimated up to a multiple of the true period. Hence in practice, for signals such as pure sinusoids, it is recommended to use the Ramanujan subspaces instead of more general NPSs.

This completes the formulation of the NPS based iMUSIC algorithms.

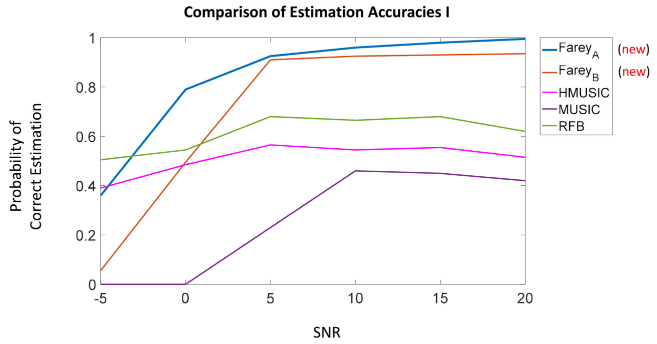

Fig. 9. Probability of Estimating both the component periods exactly. Comparison of the proposed Farey-MUSIC with other techniques. See Section V for details.

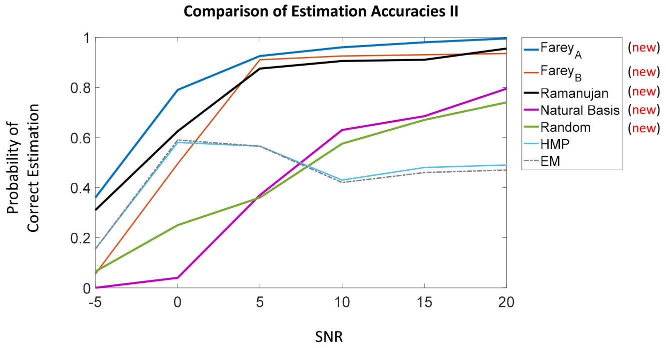

Fig. 10. Probability of Estimating both the component periods exactly. Comparison of the various NPSs for iMUISC, and also HMP and EM methods. The Farey-MUSIC plots from Fig. 9 have been repeated for reference. See Section V for details.

V. Experiments

In this section, we present several examples and comparisons to highlight various aspects of the proposed methods.

A. Comparison of Estimation Accuracies

Figs. 9 and 10 compare the accuracy of period estimation for several techniques as a function of SNR. For each SNR shown, 200 Monte Carlo trials were carried out with randomly generated signals, each signal being an additive mixture two periods randomly chosen from the interval[2] $[1, 25]$. The total datalength for each signal was $L = 500$ samples. The snapshot length (N in (5)) was chosen as 301. For simplicity, we assumed that the value of the total signal space dimension K is known to all the methods here, including MUSIC and HMUSIC. In subsequent simulations, we will show that the iMUSIC techniques are quite robust to using very simple estimators for K. P_{max}, the maximum period that is searched for in the signal, was chosen to be 35. The probability of correct estimation is plotted for each method,[3] which is the fraction of trials in which the detected periods were exactly equal to the input periods. For visual clarity, we have split the different methods into two figures, Figs. 9 and 10. Our observations are as follows:

1. Traditional MUSIC and HMUSIC: The MUSIC pseudospectrum (i.e., a plot of the inverse of $\|\mathbf{a}^{\dagger}(\omega)\mathbf{U_e}\|_2^2$ in (16) vs ω) is expected to contain peaks at all the harmonics of the component periods in the input. To identify the periods using MUSIC, we first computed a list of $2\pi/\omega$ values for each peak in the pseudospectrum. This gives a list of candidate periods, some of which might just be harmonics of larger periods. So we eliminated those numbers in this list that have a larger multiple also present in the same list. This way, if one of the peaks was a harmonic of a larger period, it will not be declared as an independent period in the mixture.

While MUSIC and the proposed iMUSIC algortihms just require an estimate of the total signal space dimension K, the HMUSIC algorithm in (18) requires to compute the exact partition of K into the K_l's. Recall that K_l is the number of lines in the spectrum that correpond to the l^{th} fundamental frequency. In Fig. 9, HMUSIC was given the benefit of knowing a priori the true values of K_l's, which is usually not available in practice. For a fair comparison, the final period estimates of both MUSIC and HMUSIC [11]–[13] were rounded to the nearest integers. Both these methods, evaluated on a frequency grid that has the same size as the Farey grid, yield probability of correct estimation close to 0 even at high SNRs. They required at least five times denser grids than the Farey grid to reach the performances shown in Figs. 9 and 10. Further increase in the grid size did not improve their accuracy significantly. As is seen, HMUSIC and MUSIC do not perform as well as the proposed iMUSIC methods, especially when using Farey and integer Ramanujan based methods. At lower SNR's however, HMUSIC does seem to offer good accuracies when compared with the Natural Basis and Random NPS based iMUSIC.[4]

2. Prior Ramanujan-Subspace Based Methods: An alternate way of using NPSs for period estimation is using compressed-sensing based dictionary methods [61], [70]. While they work very well compared to other methods for very short datalengths [61], [63], for the parameters considered here, the dictionaries turn out to be tall. Least squares based approach using such

[2] We ensured that the two numbers are not multiples of each other, as in that case, there is fundamentally no way to determine whether the input is a mixture of two periods, or just one periodic signal. Please see Section III of [65] for details on to what extent the component periods in a mixture can be uniquely identified in general.

[3] Mean Squared Error (MSE), a popular metric in general, is not appropriate here due to two reasons: (a) Different methods could detect different number of component periods. It is not straightforward to compare vectors of different lengths using MSE. (b) Estimating the period upto a multiple may be acceptable in many applications. For example, in Fig. 2, proteins with Ankyrin repeats are known to have periods in the range 30–40. So it might be more acceptable to estimate 66 as the period, instead of say 40, since we can readily deduce that 66 might actually indicate period 33 repeats. MSE on the other hand penalizes 66 more than 40. In any case, probability of correct estimation is in fact a stricter metric than MSE.

[4] An important practical aspect when using MUSIC-based techniques is how we pick the peaks from the psedo-spectra. In the experiments of this section, for MUSIC, we first applied a minor thresold (7.5% of the max. peak) to zero out very small peaks in the spectum. We then selected those frequencies as peaks where the value of the pseudo-spectrum is larger than at the immediately neighboring points. This latter step is also used in the implementations of HMUSIC provided by the authors of [13].[5] For the iMUSIC techniques, we selected all those peaks that are larger than a certain $x\%$ of the largest peak in the pseudo-spectrum. We noticed empirically that different NPSs seemed to require different values of x. To be specific, we used: Ramanujan: 5%, Farey: 10%, Random: 20% and Natural Basis: 30%.

[5] https://www.morganclaypool.com/page/multi-pitch

dictionaries can be efficiently implemented as a filter bank called the Ramanujan Filter Bank (RFB) [64], [67]. As seen in Fig. 9, the proposed iMUSIC methods based on farey and integer Ramanujan bases easily outperform the RFB. At lower SNRs though, the RFB performs better than the Natural Basis and Random NPSs. It is useful to note however that the most appropriate applications for the RFB are signals exhibiting localized or time varying periodicity such as chirps [64], [67].

3. Harmonic Matching Pursuit and Expectation Maximization: Fig. 9 also shows the performance of two other multi-pitch methods: Expectation Maximization [13], [14] and the Harmonic Matching Pursuit [13], [25] algorithms. Once again, both these methods give close to 0 probability of correct estimation when their frequency grid sizes are comparable to those of the Farey dictionary. Both of them required at least 10 times denser grids than the Farey grid to achieve the accuracies shown in Fig. 9, which made them significantly more computationally expensive.

4. The Proposed iMUSIC Methods: As seen in Figs. 9 and 10, iMUSIC using the Farey dictionary (denoted as $\mathrm{Farey}_{\mathrm{A}}$ in the plots) clearly outperforms all other methods considered here. An interesting observation in Fig. 10 is that, although Farey and Ramanujan dictionaries both span the same (Ramanujan) subspaces, Farey based iMUSIC performs slightly better. This might be because the Farey columns tend towards orthogonality for large enough N, while the Ramanujan integer basis vectors do not. Nevertheless, at almost all SNR levels shown, both these methods outperform all the other techniques considered in this simulation. The Natural Basis and Random NPSs based iMUSIC algorithms have good accuracies compared to other non-iMUSIC techniques at SNR's larger than about 7.5 dB. It is worth noting however, that at -5 dB, HMUSIC (with 5 times larger grid size, and the additional knowledge of K_l's) and the Ramanujan Filter Bank do perform well compared to the proposed methods.

For simplicity, we assumed that the iMUSIC techniques here have an a priori knowledge of the true total signal space dimension K. $\mathrm{Farey}_{\mathrm{B}}$ however uses a simple approach to estimate K: namely, all the eigenvalues of the auto-correlation matrix smaller than 5% of the maximum eigenvalue were considered as noise eigenvalues. As seen, the change in accuracy is not too severe compared to $\mathrm{Farey}_{\mathrm{A}}$, especially at higher SNRs. This will also be seen in the experiments in the following subsections, including the examples of proteins and DNA repeats. When using such a method to estimate K, we noticed empirically that the exact percentage to use depends on the SNR level. Using lower threshold levels for high SNRs and vice versa gave the best results.

Remark: Apart from the methods considered above, there are several other techniques that are popular in the literature such as [9], [10], [18], [19], [31], [52], [57]. While being the state-of-the-art for applications such as pitch estimation, these methods cannot be directly used in the above example since they are not easily extended to the case of mixtures of periodic signals. It is important to note that, although iMUSIC outperforms the other techniques in Figs. 9 and 10, these other methods, including the aforementioned papers on single pitch estimation, can handle the more general case of non-integer periods. Whether the iMUSIC algorithms can be adapted to such applications requires a detailed analysis in itself, and will be a part of our future research. In a following subsection, we will compare the iMUSIC methods with the state-of-the-art for an application with inherently integer periods: namely, protein repeats.

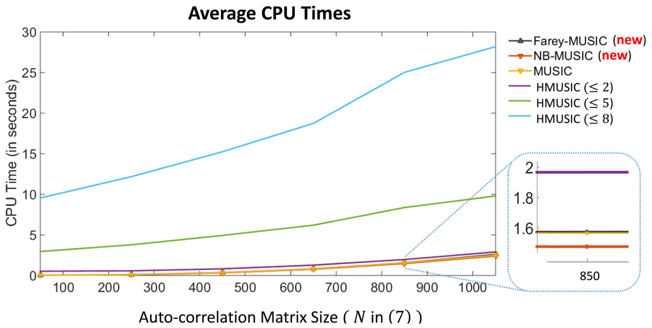

Fig. 11. A comparison of the CPU Times. See Section V for details.

B. Comparison of CPU Times for Eigenspace Methods

To show the computational savings that iMUSIC algorithms achieve over prior variants of MUSIC, Fig. 11 compares the average CPU times (MATLAB 2014b on a 2.4 GHz CPU with 8 GB RAM) as a function of the size of the autocorrelation matrix (which is also the size of the snapshots N in (5)). The total datalength of the signal, L was chosen as $3N$, and the dimension of the signal subspace K was fixed at 25 for simplicity. MUSIC and HMUSIC were implemented with a uniform frequency grid of the same size as the Farey grid. Recall however that both these methods typically require much more denser grids than Farey MUSIC (Section V-A). Unlike in Section V-A, here HMUSIC was not given a priori knowledge of the true values of K_l's in (18). As derived in Appendix B, significant computational complexity is incurred by HMUSIC while finding the correct partition of the signal space dimension K into the K_l's. Notice that our natural basis (NB) MUSIC is the fastest in Fig. 11. Farey-MUSIC and MUSIC are similar to each other in terms of CPU time due to identical grid sizes. In Fig. 11, HMUSIC$(\leq T)$ denotes using (18) with the prior knowledge that the number of hidden periodic components in the signal $Q \leq T$. It is shown in Appendix B that the complexity of HMUSIC increases exponentially with the number of hidden periodic components. In contrast, since we check the NPS basis vectors one-by-one in (28), the complexity of our proposed techniques does not depend on T. From Figs. 10 and 11, it is evident that our methods offer much better accuracy for integer period estimation than prior variants of MUSIC, while keeping the computational complexity low at the same time.

C. Effect of Increasing the Number of Hidden Periods

In Section V-A, we considered input signals that were mixtures of two hidden periods. How do the proposed iMUSIC algorithms perform as the number of hidden periods increases? Fig. 12 plots the estimation accuracy *vs.* the number of hidden periods for various NPS based iMUSIC methods. The total datalength was fixed at 750 samples, the autocorrelation window length at 301 samples, and the SNR at 10 dB. The number of hidden periods was increased from 1 to 8, with the hidden

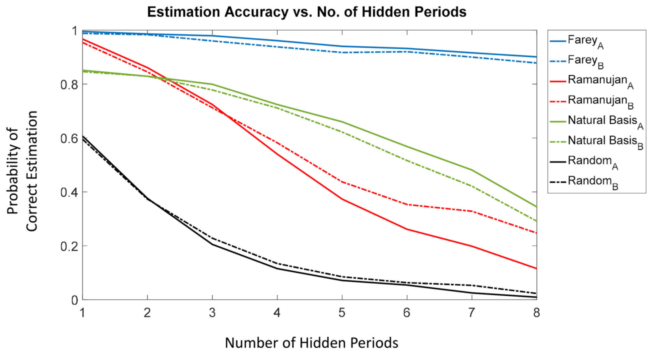

Fig. 12. Effect of increasing the number of hidden periods on the accuracy of the proposed iMUSIC techniques. See Section V-C for details.

TABLE I
PROTEIN REPEATS COMPARISON

Protein	1n11	1dfj	3du1
True Period	**33**	**57**	**5**
HMUSIC (5%)	66	82	40
HMUSIC (10%)	66	59	5
MUSIC (5%)	33	28	39
MUSIC (10%)	33	28	5
RADAR	27	26	23
TRUST	33	57	10
HHrepID	33	57	10
iMUSIC Farey	33	57	5
iMUSIC Ramanujan	33	57	5
iMUSIC Natural Basis	33	57 (s)	5 (s)
iMUSIC Random	33	57	5

periods themselves being chosen uniformly at random from the set $[1, 20]$. For each value of number of hidden periods, 1000 realizations of randomly generated signals were used to compute the probabilities of correct estimation. For each NPS, the estimation accuracies for both the following cases are shown: (a) using the iMUSIC algorithms with an a priori knowledge of the true value of K (denoted by the subscript A in Fig. 12), and (b) estimating K using the simple percentage metric described in the previous subsection (denoted by subscript B). For the chosen SNR level, we noticed that using a threshold value of 1% of the maximum eigenvalue gave the best results.

Our observations are as follows: First, it is natural to expect that the probability of correctly estimating all the component periods in a mixture of N signals will decrease as N increases. This is indeed seen in Fig. 12. The accuracy of the Farey iMUSIC method however does not decrease as quickly as the other NPSs, indicating that it is the best choice when the number of hidden periods is large. At this point, we do not have a theoretical justification for why this happens, but it will be interesting to investigate the same. Another interesting observation in these plots is that using an estimated K using the simple percentage rule described above, seems to give better performance than using the true value of K for certain NPSs like the Ramanujan (integer) case. Once again, it will be interesting to study why this happens from a fundamental perspective.

D. Examples of Protein Repeats and DNA Microsatellites

We will now demonstrate the proposed iMUSIC algorithms on repeats in proteins. A protein is essentially a chain of amino acids taken from an alphabet of 20 possible amino acids. Protein repeats are segments within the amino acid sequence that exhibit periodicity. For example, the sequence $\ldots MWACFACFACSY \ldots$ has 2 complete, and a partial cycle of the repeat ACF. Such repeats manifest as characteristic 3D periodic structures (for e.g., see Fig. 2), and play important roles in several diverse contexts [1]. For instance, they are known to admit a much higher mutation rate (substitution and insertion-deletion error) than usual. Such mutations have been associated with several diseases, including addictive behaviors to alcohol, nicotine and so on [42].

Detecting these repeats is not easy in practice, due to the high mutation rate. Several techniques have been proposed in this regard [6], [26], [30], [40], [58], [59]. In the following examples, we will demonstrate that the proposed iMUSIC algorithms can also be used for identifying such repeats. To the best of our knowledge, this is the first time any MUSIC based approach is being used for this application. In the following examples, we used the Kyte-Doolittle hydrophobicity scale [36] to map amino acids to numbers. This scale quantifies the hydropathy of amino acids, and has been widely used in studying protein structures and domains. To estimate the signal space dimension, we noticed empirically that using the simple percentage metric described above with thresholds between 5% and 10% of the maximum eigenvalue gave the best results.

In our first example, we consider the protein AnkyrinR (PDB 1n11) that enables red blood cells to resist shear forces during circulation. The period 33 repeats in AnkyrinR can easily be identified in the pseudo-spectra shown in Fig. 13. These plots show the results of applying the proposed methods using Ramanujan (integer basis), Natural Basis and Random Integer NPSs (The Farey basis can also be used; it was shown earlier in Fig. 1). All four plots have clear peaks at 33 and its divisors. Notice that the Ramanujan (integer basis) plot in Fig. 13(c) has a weak peak at 33. This is not a problem, since there are significant peaks at periods 11 and 3 in this case. The iMUSIC algorithm takes an LCM of all the strong peaks in the pseudospectrum, which in this case turns out to be 33 again.

Fig. 14 shows a second example, the protein Ribonuclease Inhibitor, which contains 15 luciene-rich repeats, alternately 28 and 29 residues long [35]. So according to our definition of periodicity in (2), the period of these repeats is 57. This can once again be easily identified in the pseudo-spectrum plots shown in Fig. 14. In all our experiments, we observed that the Ramanujan Subspaces (both Farey and the integer basis) gave the cleanest plots, followed by randomly generated NPSs. The Natual Basis Subspaces, although showing the tallest peaks at the correct periods, often had smaller spurious peaks. Recall that the natural basis subspaces performed well in Section V-A at higher SNRs. But the noise there was additive, whereas here we have substitution and insertion-deletion errors. It seems that the natural basis subspaces are more sensitive than the other NPSs for such errors.

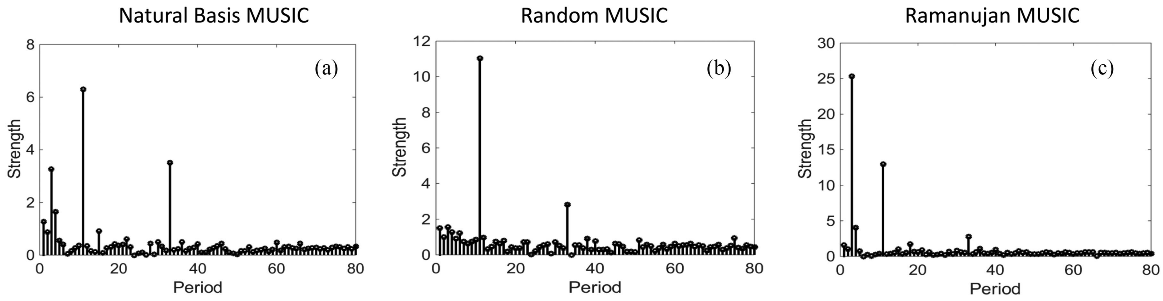

Fig. 13. Pseudospectra of the proposed NPS based techniques for the Ankyrin protein repeats shown in Fig. 2. See Section V for details.

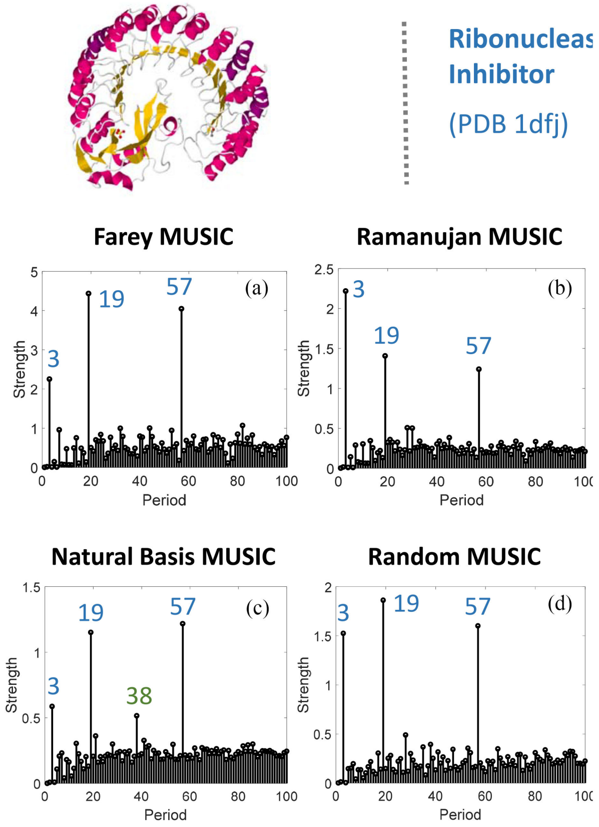

Fig. 14. Top: The protein Ribonuclease Inhibitor (PDB: 1dfj) exhibiting luciene-rich repeats. The psuedo-spectra obtained from (a) Ramanujan Subspaces (Farey basis) (b) Ramanujan Subspaces (Integer Basis) (c) Natural Basis Subspaces and (d) a randomly generated NPS, are shown. See Section V for details.

For comparison, Table I shows the period estimates of various techniques for three examples of protein repeats.[6] Apart from the examples shown in Figs. 13 and 14, we also consider the protein HetL (PDB:3du1), which is 237 amino acids long and contains tandem pentapeptide (period 5) repeats. This protein is known to play an important role in the nitrogen fixation process in cyanobacteria [43]. In Table I, notice that the estimates of HMUSIC and MUSIC are not as accurate as those of iMUSIC.

TABLE II
REPEATS IN HUMAN GENOME SEQUENCE AC010136

Location	Period	Copy No.	% match	% in-del	iMUSIC	TRF
11792-11805	3	6.7	95	0	√	
62794-62809	3	5	87	6.7	√	
75567-75586	4	5	95	10	√	
67043-67076	4	7	100	0	√	√
30684-30707	6	4	88	0	√	
58508-58536	6	4	77	0	√	
67043-67076	4	7	100	0	√	
42034-42706	11	59	69	11	√	√
11432-11492	13	5.2	72	22		√
42755-42827	33	2.2	90	0		√
83591-83676	39	2.3	80	7		√

Moreover, it can be seen that HMUSIC and MUSIC were very sensitive to errors in the estimation of signal space dimension. On the other hand, iMUSIC was very robust in this regard. While Figs. 13 and 14 used a cut-off of 5% for identifying the noise eigenvalues, the plots were very similar at 10% as well. The '(s)' next to natural basis iMUSIC's estimates in Table I indicates the presence of smaller spurious peaks in its pseudospectrum (such as the one at period 38 in Fig. 14).

Table I also compares three state of the art techniques used for protein repeats. RADAR [30] and TRUST [58] algorithms are based on self-alignment techniques (trace matrices and dynamic programming), while HHrepID [6] uses Hidden Markov Models. These methods were desinged speicifically for finding repeats from symbolic sequences (in this case, proteins). Once again, iMUSIC performs well in comparison to these methods.

As the last set of examples, we consider tandem repeats in the human DNA. Such repeats are of significance in a number of contexts, and several interesting techniques have been proposed in the past for identifying them [2], [4], [7], [23], [33], [54], [60]. For example, repeats in the DNA are the primary bio-markers used today in DNA fingerprinting, kinship analysis etc. [4], [20], [50]. They are also associated with several genetic disorders such as the fragile X syndrome, myotonic dystrophy, Huntington's disease and Friedreich's ataxia [4]. Fig. 15 shows an example of repeats (GenBank G08921) that are used in DNA fingerprinting. We mapped the nucleotides to numbers that were randomly generated using a Gaussian distribution. The iMUSIC psuedospectra in Fig. 15(a) to (d) clealry identify the period 4 repeats.

The iMUSIC methods can be used with other one-to-one mappings of nucleotides to numbers as well. For example, Table II shows examples of repeats from the GenBank sequence AC010136, found on the second chromosome of the human DNA. Here we used the simple mapping of assigning $\{A, C, T, G\}$ to the integers $\{1, 2, 3, 4\}$ respectively instead of a random mapping. For reference, one of the most popular

[6] HMUSIC, as described in [12], [13], uses a computationally intensive discrete optimization to find the signal space dimension. However, the implementation of HMUSIC provided by its author Christensen (as a part of [13]) requires the user to specify the signal space dimension as an input. For simplicity, we used the same 5% (and 10%) rule that we used with iMUSIC, for estimating the signal space dimension for HMUSIC and traditional MUSIC as well.

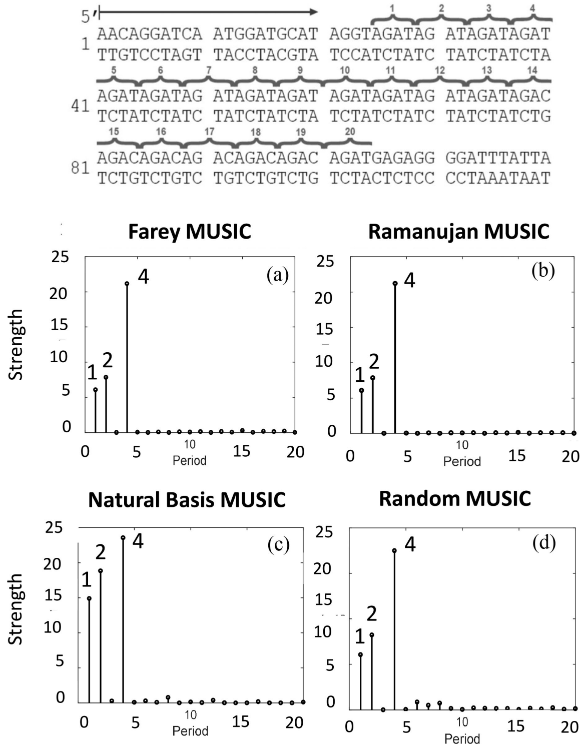

Fig. 15. Top: An example of DNA microsatellites that are used in DNA fingerprinting. The psuedo-spectra obtained from (a) Ramanujan Subspaces (Farey basis) (b) Ramanujan Subspaces (Integer Basis) (c) Natural Basis Subspaces and (d) a randomly generated NPS, are shown. See Section V for details.

techniques for identifying DNA repeats, the Tandem Repeat Finding algorithm [4], is compared with the Farey based iMUISC method. The %-match and the %- indel metrics are a meaure of substitution and insertion-deletion errors respectively [60]. In Table II, a $\surd$ indicates that a particular repeat was correctly identified by the corresponding method. Table II shows an interesting observation that we consistently observed in all our DNA experiments. The iMUSIC methods seem very good at idenitfying repeats with short periods (known in the literature as DNA micro-satellites). The TRF method on the other hand was better at detecting larger periods with very small number of repeating copies. Here, we used a cut-off of 2.5% to identify the noise eigenspace for iMUSIC. One may be able to improve the performance of the iMUSIC methods for larger periods by changing this threshold. This will be investigated as a part of our future work. In a recent work [60], this sequence AC010136 is analyzed in greater detail using a related Ramanujan subspaces based method known as the Ramanujan Filter Bank. We refer the reader to [60] in this regard.

A note on numerical mappings: An important practical aspect when using iMUSIC for protein and DNA repeats is the mapping used to convert them to numerical sequences. Notice that such mappings introduce geometric distances between the amino acids/nucleotides. Are such induced geometric distances appropriate for finding repeats in these symbolic sequences? The answer to this could be slightly different for protein and DNA repeats.

For protein repeats, we are generally trying to infer periodicity in the 3D structure of the protein from its amino acid sequence [1]. If there were no errors in the sequence, one may use any one-to-one mapping to identify periodicity in the amino acid sequence. However, when there are substitution errors, it is known that certain substitutions of amino acids do not change the 3D structure as much as other substitutions. In the presence of such substitution errors, even though there may not be exact periodicity in the amino acid sequence, there will still be periodicity in the 3D structure of the protein. Numerical mappings such as the Kyte Doolittle (KD) scale that we have used here are very useful in capturing this phenomena. The KD scale quantifies the similarity between amino acids based on their hydrophobic nature. Since the hydropathy of the constituent amino acids plays an important role in how a protein folds in solutions, the KD scale is very popular in studies that identify protein structure from sequence. In particular, it has been used in popular works on protein repeat detection, such as [26], [40]. The rASA (relative accessible surface area) is another such mapping used in [40]. Similarity score matrices (or substitution matrices) have also been used in works such as REPPER [26] to quantify this notion of 'closeness' between amino acids in the context of substitutions.

In contrast, in applications of DNA repeats such as forensics, kinship analysis etc. [4], [20], [50], we are primarily interested in repeats in the nucleotide sequence itself (and not concerned as much about the 2D or 3D structure). So as long as each nucleotide is mapped to a unique number, periodicity in the nucleotide sequence can be detected using periodicity in the numerical sequence. We showed two different kinds of mappings in the examples above. However, in practice, we noticed that the iMUSIC pseudo-spectrum obtained from one mapping may reveal the periods in a much more clearer fashion than a different mapping. The best mapping to use with iMUSIC methods is something the authors would like to explore in the future. A starting point in this direction could be [23], which proposes a way to define a spectrum of a DNA sequence based on a minimum entropy condition. It is also worth noting that some techniques such as [2], [4], by default, do not rely on an explicit numerical mapping.

Regarding modeling of noise: The iMUSIC formulation in Section III-B models noise as additive and white. Is this appropriate for DNA and proteins? To investigate this, let us start by noting that there are two kinds of noise in such sequences: substitution and insertion-deletion errors. When a sequence is corrupted by substitution errors, in terms of the corresponding numeric sequences, the difference between the original sequence $x(n)$ and the corrupted sequence $x_e(n)$, can surely be modeled as additive:

$$x_e(n) = x(n) + e(n) \tag{39}$$

The whiteness assumption on the noise essentially means that we assume these substitution errors to occur independent of each other. The aforementioned mappings based on substitution probabilities and similarity measures ensure that $e(n)$ is usually small. *An additive noise model unfortunately cannot take into account insertion deletion errors.* We noticed in our simulations however (for e.g., see Table II), that for reasonable amount of insertion deletion errors, the iMUSIC methods are still able to identify the repeats. It is also worth noting that many popularly used repeat finding works such as [7], [26], [40], [54] do not model insertion deletion errors explicitly in any manner when formulating their algorithms.

While the above examples do demonstrate the proposed methods as good candidates for these applications, a more thorough experimental evaluation of their performance in comparison with prior works in these application domains is still necessary. Such an analysis merits a much broader discussion than the scope of this paper. Our focus here has been to introduce and establish these methods on a sound theoretical footing. Optimizing them for specific applications such as DNA and protein repeats will be a part of our future work.

VI. Conclusion and Future Directions

This paper presents a new family of MUSIC-like algorithms for integer period estimation, based on Ramanujan subspaces [69] and nested periodic subspaces [61]. These new algorithms offer very simple integer valued basis vectors for spanning the signal space, and result in significantly better accuracy and computational simplicity than existing techniques for integer periods. The non-Vandermonde nature of the basis vectors introduces a number of subtle differences from the traditional MUSIC formulation. These were carefully addressed in the paper. A number of simulation experiments were presented demonstrating these algorithms, including examples from protein and DNA repeats.

While the model in (2) is especially relevant to applications with inherent integer periodicity, many state of the art methods for conventional periodicity applications such as in speech [10], [44], [71] also use integer period approximations. Adapting our techniques for such applications will be of interest to us in our future work. Even for integer period applications such as proteins and DNA repeats, we are interested in specifically tailoring our algorithms for each of these applications in a more thorough fashion by optimizing over larger databases, and comparing their performance with the existing state of the art methods in those domains. We used simple metrics (the 5% and 10 % rules) to estimate the total signal space dimensions in our simulations here. There have been more carefully designed criteria proposed in the literature [15], [17], [22], especially for sinusoidal model order estimation, and it will be useful to extend them to the case of general nested periodic subspaces. We also used a simplistic model of the noise being additive while formulating the iMUSIC algorithms. For repeats in proteins and DNA, the noise in practice takes the form of substitution and insertion-deletion errors. While substitution errors can still be modeled as additive noise, the exact effect of insertion-deletion noise on the the signal model (especially the autocorrelation matrix) will be interesting to study in the future. Finally, apart from MUSIC, techniques such as ESPRIT [49] and the recent atomic norm based methods [5], [8] are also popularly used for various line spectral applications. While we specifically focused on MUSIC in this paper, it will be very interesting to see if we can similarly adapt these other techniques for harmonic spectra in the future.

Appendix A

Proof of (13): Substituting from (8), the signal component of (12) is

$$\mathbf{V}\left(\frac{1}{M}\sum_{i=1}^{M}\mathbf{\Lambda}_{\omega}(i)\mathbf{c}\mathbf{c}^{\dagger}\mathbf{\Lambda}_{\omega}^{\dagger}(i)\right)\mathbf{V}^{\dagger}. \tag{40}$$

The noise component is

$$\frac{1}{M}\sum_{i=1}^{M}\mathbf{e}(i)\mathbf{e}^{\dagger}(i), \tag{41}$$

and the cross terms are

$$\mathbf{V}\left(\frac{1}{M}\sum_{i=1}^{M}\mathbf{\Lambda}_{\omega}(i)\mathbf{c}\mathbf{e}^{\dagger}(i)\right), \tag{42}$$

and its transpose conjugate. The matrix inside bracketts in Eq. (40) has ml-th element

$$c_m c_l^* \frac{1}{M}\sum_{i=1}^{M} e^{j(\omega_m-\omega_l)i} \tag{43}$$

For $m \neq l$ we have $\omega_m - \omega_l \neq 0 \bmod 2\pi$, so (43) approaches zero for large M. So, (40) has the form $\mathbf{V}\mathbf{\Lambda}_c\mathbf{V}^{\dagger}$ where $\mathbf{\Lambda}_c$ is a diagonal matrix with diagonal elements $|c_k|^2$. Secondly, for large M Eq. (41) approaches $\sigma_e^2\mathbf{I}$. Thirdly the matrix inside bracketts in Eq. (42) has ml-th element $c_m \sum_{i=1}^{M} e^{j\omega_m i} e_l^*(i)/M$. This is a zero-mean random variable with variance $|c_m|^2\sigma_e^2/M \rightarrow 0$ for large M. These three observations justify (13). ▽ ▽ ▽

Appendix B

Complexity of HMUSIC: Solving (18) incurs an exponential complexity in Q. Here is the justification for the same. Let us assume that the total signal space dimension $K = \sum_{l=0}^{Q-1} K_l$ has been estimated using the eigenvalue distribution of the autocorrelation matrix. Notice that, to solve (18), for each choice of the integer parameters $\{K_l\}_{l=0}^{Q-1}$, we must solve a continuous variables optimization in $\{\omega_l\}_{l=0}^{Q-1}$. So the complexity is proportional to the number of positive integer solutions for $\{K_l\}_{l=0}^{Q-1}$ such that:

$$K_0 + K_1 + K_2 + \cdots + K_{Q-1} = K \tag{44}$$

This is a standard partitions problem, whose number of solution can be shown to be:

$$\binom{K-1}{Q-1} \tag{45}$$

We would like to study the behaviour of (45) as Q increases. It is reasonable to expect K, the total number of lines in the spectrum, to increase with the number of periodic signals in the mixture, i.e., Q. For simplicity in analysis, let us assume a linear depenedence in the following expressions: $K \approx MQ$. Notice that each periodic component in the mixture will add at least one line to the net spectrum, so it is reasonable to consider $M > 1$ in this model. So (45) can be written as:

$$\binom{MQ-1}{Q-1} \tag{46}$$

We can now use Stirling's approximation to factorials [21], which claims:

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^{n} \tag{47}$$

where e is the Euler's number. Using this, it can be shown that (46) is approximately equal to the following for large Q:

$$\frac{1}{\sqrt{2\pi QM(M-1)}}\left[\frac{M^M}{(M-1)^{M-1}}\right]^Q \tag{48}$$

For large enough Q, the exponential term dominates the denominator, which is a sub-polynomial in Q. Hence, the complexity is exponential in Q.

Appendix C

Explanation of why $K = K'$ occurs with probability 1 for random periodic signals: Suppose $x(n)$ was a period P signal generated by repeating a $P \times 1$ Gaussian random vector $\mathbf{x}_P$. Let $\mathbf{A} \in \mathbb{C}^{P \times P}$ be any full rank matrix. To span $\mathbf{x}_P$ using the columns of $\mathbf{A}$, with probability 1 we will need to use all P columns of $\mathbf{A}$. Now, the P basis vectors of any Nested Periodic Subspaces (NPS), with periods P and its divisors, are essentially periodically extended verisons of $P \times P$ full rank matrices called Nested Periodic Matrices [61]. In other words, spanning $x(n)$ using NPSs is equivalent to spanning $\mathbf{x}_P$ using the corresponding $P \times P$ Nested Periodic Matrix. Hence, to span such an $x(n)$, one would need all P NPS basis vectors with probability 1, irrespective of which family of NPSs is chosen. In particular, this implies that $K = K' = P$ in the context of Section IV-A, with probability 1. This same idea applies to the case where $x(n)$ is a mixture of multiple periodic signals, where each component has been generated using Gaussian random vectors.

References

[1] M. A. Andrade, C. P. Iratxeta, and C. P. Ponting, "Protein repeats: Structures, functions, and evolution," *J. Struct. Biol.*, vol. 134, pp. 117–131, 2001.

[2] Arora, Sethares, and Bucklew,, "Latent Periodicities in Genome Sequences," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 3, pp. 332–342, Jun. 2008.

[3] A. Barabell, "Improving the resolution performance of eigenstructure-based direction-finding algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Boston, USA, 1983, pp. 336–339.

[4] G. Benson, "Tandem repeats finder: A program to analyze DNA sequences," *Nucl. Acids Res.*, vol. 27, no. 2, pp. 573–580, 1999.

[5] B. N. Bhaskar, G. Tang, and B. Recht, "Atomic norm denoising with applications to line spectral estimation," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 5987–5999, Dec. 1, 2013.

[6] A. Biegert and J. Soding, "De novo identification of highly diverged protein repeats by probabilistic consistency," *Bioinformatics*, vol. 24, pp. 807–814, 2008.

[7] M. Buchner and S. Janjarasjitt, "Detection and visualization of tandem repeats in DNA sequences," *IEEE Trans. Signal Process.*, vol. 51, no. 9, pp. 2280–2287, Sep. 2003.

[8] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[9] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 124, no. 3, pp. 1638–1652, 2008.

[10] A. de Cheveigna and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, pp. 1917–1930, 2002.

[11] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Multi-Pitch estimation using harmonic music," in *Proc. 40th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, 2006, pp. 521–524.

[12] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1635–1644, Jul. 2007.

[13] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures Speech Audio Process.*, vol. 5, pp. 1–160, 2009.

[14] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Elsevier Signal Process.*, vol. 88, pp. 972–983, Apr. 2008.

[15] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Sinusoidal order estimation using angles between subspaces," *EURASIP J. Adv. Signal Process.*, vol. 2009, pp. 1–11, 2009.

[16] Deng and Han, "Ramanujan subspace pursuit for signal periodic decomposition," *Mech. Syst. Signal Process.*, vol. 90, pp. 79–96, Jun. 2017.

[17] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726–2735, Oct. 1998.

[18] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. Interspeech*, 2011, pp. 1973–1976.

[19] H. Duifhuis, L. F. Willems, and R. J. Sluyter, "Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception," *J. Acoust. Soc. Amer.*, vol. 71, no. 6, pp. 1568–1580, 1982.

[20] A. Edwards, A. Civitello, H. A. Hammond, and C. T. Caskey, "DNA typing and genetic mapping with trimeric and tetrameric tandem repeats," *Amer. J. Human Genetics*, vol. 49, pp 746–756, 1991.

[21] W. Feller, *An Introduction to Probability Theory and Its Applications*. vol. 1, 3rd ed. New York, NY, USA: Wiley, 1968, pp. 50–53.

[22] J.-J. Fuchs, "Estimating the number of sinusoids in additive white noise," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 36, no. 12, pp. 1846–1853, Dec. 1988.

[23] L. Galleani and R. Garello, "The minimum entropy mapping spectrum of a DNA sequence," *IEEE Trans. Inf. Theory*, vol. 56, no. 2, pp. 771–783, Feb. 2010.

[24] O. Gotoh, "An improved algorithm for matching biological sequences," *J. Mol. Biol.*, vol. 162, no. 3, pp. 705–708, Dec. 1982.

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

[26] M. Gruber, J. Soding, and A. N. Lupas, "REPPER–repeats and their periodicities in fibrous proteins," *Nucleic Acids Res.*, vol. 33, no. 2, pp. w239–w243, Jul. 2005.

[27] M. Hamalainen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa, "Magnetoencephalographytheory, instrumentation, and applications to noninvasive studies of the working human brain," *Rev. Modern Phys.*, vol. 65, no. 2, pp. 413–497, Apr. 1993.

[28] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*. London, U.K.: Oxford Univ. Press, 2008.

[29] G. H. Hardy and S. Ramanujan, "Asymptotic formulae in combinatory analysis," in *Proc. London Math. Soc.*, 1918, vol. 17, pp. 75–115.

[30] A. Heger and L. Holm, "Rapid automatic detection and alignment of repeats in protein sequences," *Proteins*, vol. 41, no. 2, pp. 224–237, Nov. 2000.

[31] D. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Amer.*, vol. 83, no. 1, pp. 257–264, 1988.

[32] A. V. Kajava, "Tandem repeats in proteins: From sequence to structure," *J. Struct. Biol.*, vol. 179, pp. 279–288, Sep. 2012.

[33] R. Kolpakov, G. Bana, and G. Kucherov, "mreps: Efficient and flexible detection of tandem repeats in DNA," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3672–3678, 2003.

[34] S. M. Kay, *Modern Spectral Estimation: Theory and Application*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.

[35] B. Kobe and J. Deisenhofer, "A structural basis of the interactions between leucine-rich repeats and protein ligands," *Nature*, vol. 374, pp 183–186, 1995.

[36] J. Kyte and R. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol.*, vol. 157, pp. 105–132, 1982.

[37] H. Luo and H. Nijveen, "Understanding and identifying amino acid repeats," *Briefings Bioinf.*, vol. 15, no. 4, pp. 582–591, Jul. 2014.

[38] J. C. Mosher, P. S. Lewis, and R. M. Leahy, "Multiple dipole modeling and localization from spatio-temporal MEG data," *IEEE Trans. Biomed. Eng.*, vol. 39, no. 6, pp. 541–557, Jun. 1992.

[39] D. D. Muresan and T. W. Parks, "Orthogonal, exactly periodic subspace decomposition," *IEEE Trans. Signal Process.*, vol. 51, no. 9, pp. 2270–2279, Sep. 2003.

[40] K. B. Murray, D. Gorse, and J. M. Thornton, "Wavelet transforms for the characterization and detection of repeating motifs," *J. Mol. Biol.*, vol. 316, pp. 341–363, Feb. 2002.

[41] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Grid size selection for nonlinear least-squares optimisation in spectral estimation and array processing," in *Proc. Eur. Signal Process. Conf.*, Budapest, 2016, pp. 1653–1657.

[42] M. J. Neville, E. C. Johnstone, and R. T. Walton, "Identification and characterization of ANKK1: A novel kinase gene closely linked to DRD2 on chromosome band 11q23.1," *Hum Mutation*, vol 23, pp 540–545, Jun. 2004.

[43] S. Ni, G. M. Sheldrick, M. M. Benning, and M. A. Kennedy, "The 2 A resolution crystal structure of HetL, a pentapeptide repeat protein involved in regulation of heterocyst differentiation in the cyanobacterium Nostoc sp. strain PCC 7120," *J. Struct. Biol.*, vol. 165, pp. 47–52, Jan. 2009.

[44] A. M. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum and a maximum likelihood estimate," in *Proc. Symp. Comput. Process. Commun.*, 1970, vol. 19, pp. 779–797.

[45] M. Oziewicz, "On application of MUSIC algorithm to time delay estimation in OFDM channels," *IEEE Trans. Broadcast.*, vol. 51, no. 2, pp. 249–255, Jun. 2005.

[46] S.-C. Pei and K.-S. Lu, "Intrinsic integer-periodic functions for discrete periodicity detection," *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1108–1112, Aug. 2015.

[47] S. Ramanujan, "On certain trigonometrical sums and their applications in the theory of numbers," *Trans. Cambridge Philos. Soc.*, vol. 22, no. 13, pp. 259–276, 1918.

[48] B. D. Rao and K. V. S. Hari, "Performance analysis of Root-Music," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 1939–1949, Dec. 1989.

[49] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[50] C. M. Ruitberg, D. J. Reeder, and J. M. Butler, "STRBase: A short tandem repeat DNA database for the human identity testing community," *Nucleic Acids Res.*, vol. 29, no. 1, pp 1320–1322, 2001.

[51] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[52] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental frequency measurement," *J. Acoust. Soc. Amer.*, vol. 43, no. 4, pp. 829–834, 1968.

[53] W. A. Sethares and T. W. Staley, "Periodicity transforms," *IEEE Trans. Signal Process.*, vol. 47, no. 11, pp. 2953–2964, Nov. 1999.

[54] D. Sharma *et al.*, "Spectral repeat finder (SRF): Identification of repetitive sequences using Fourier transformation," *Bioinformatics*, vol. 20, pp. 1405–1412, 2004.

[55] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2005.

[56] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramer–Rao bound," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 5, pp. 720–741, May 1989.

[57] X. Sun, "Pitch determination and voice quality analysis using Subharmonic-to-Harmonic Ratio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, pp. 333–336.

[58] R. Szklarczyk and J. Heringa, "Tracking repeats using significance and transitivity," *Bioinformatics*, vol. 20 (Suppl. 1), pp. 311–317, Aug. 2004.

[59] S. V. Tenneti and P. P. Vaidyanathan, "Detection of protein repeats using Ramanujan filter bank," in *Proc. 50th Asilomar Conf. Signals, Syst. Comput.*, Monterey, CA, USA, 2016, pp. 343–348.

[60] S. V. Tenneti and P. P. Vaidyanathan, "Detecting tandem repeats in DNA using Ramanujan Filter Bank," in *Proc. Int. Symp. Circuits Syst.*, Montreal, Canada, May 2016, pp. 21–24.

[61] S. V. Tenneti and P. P. Vaidyanathan, "Nested periodic matrices and dictionaries: New signal representations for period estimation," *IEEE Trans. Signal Process.*, vol. 63, no. 14, pp. 3736–3750, Jul. 2015.

[62] S. V. Tenneti and P. P. Vaidyanathan, "MUSIC and Ramanujan: MUSIC-like algorithms for integer periods using nested-periodic-subspaces," in *Proc. 51st Asilomar Conf. Signals, Syst. Comput.*, Oct. 2017, pp. 1997–2001.

[63] S. V. Tenneti and P. P. Vaidyanathan, "A unified theory of union of subspaces representations for period estimation," *IEEE Trans. Signal Process.*, vol. 64, no. 20, pp. 5217–5231, Oct. 2016.

[64] S. V. Tenneti and P. P. Vaidyanathan, "Ramanujan filter banks for estimation and tracking of periodicities," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, Australia, 2015, pp. 3851–3855.

[65] S. V. Tenneti and P. P. Vaidyanathan, "Minimum data length for integer period estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2733–2745, May, 2018.

[66] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 1992.

[67] P. P. Vaidyanathan and S. V. Tenneti, "Properties of Ramanujan filter banks," in *Proc. 23rd Eur. Signal Process. Conf.*, France, 2015, pp. 2816–2820.

[68] P. P. Vaidyanathan, "Ramanujan sums in the context of signal processing: Part I: Fundamentals" *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4145–4157, Aug. 2014.

[69] P. P. Vaidyanathan, "Ramanujan sums in the context of signal processing: Part II: FIR representations and applications" *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4158–4172, Aug. 2014.

[70] P. P. Vaidyanathan and P. Pal, "The Farey dictionary for sparse representation of periodic signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 360–364.

[71] J. D. Wise, J. Caprio, and T. W. Parks, "Maximum likelihood pitch estimation" *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 5, pp. 418–423, Oct. 1976.

[72] X. Zhang, L. Xu, L. Xu, and D. Xu, "Direction of departure (DOD) and direction of arrival (DOA) estimation in MIMO radar with reduced-dimension MUSIC," *IEEE Commun. Lett.*, vol. 14, no. 12, pp. 1161–1163, Dec. 2010.

Srikanth Venkata Tenneti (S'14) received the Ph.D. degree in electrical engineering from the California Institute of Technology (Caltech), Pasadena, CA, USA, in 2014, and the B.Tech degree in electrical engineering from the Indian Institute of Technology—Bombay, Mumbai, India, in 2012. He is currently a Postdoctoral Scholar with Caltech, pursuing research in the intersection of Deep Learning and DSP. His current research interests include developing a new framework for periodicity analysis using a combination of ideas from classical number theory (Ramanujan Sums), compressed sensing, filter design, and convolutional neural networks.

Palghat P. Vaidyanathan (S'80–M'83–SM'88–F'91) was born in Calcutta, India, on October 16, 1954. He received the B.Sc. (Hons.) degree in physics and the B.Tech. and M.Tech. degrees in radiophysics and electronics, all from the University of Calcutta, Kolkata, India, in 1974, 1977, and 1979, respectively, and the Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara, CA, USA, in 1982. He is the Kiyo and Eiko Tomiyasu Professor of Electrical Engineering at the California Institute of Technology, Pasadena, CA, USA, where he has been on the faculty since 1983. He has authored more than 500 papers and four books in the signal processing area. He is a past Distinguished Lecturer of the IEEE Signal Processing Society and recipient of the IEEE CAS Golden Jubilee Medal and the F. E. Terman Award of the ASEE. He has received multiple awards for excellence in teaching at the California Institute of Technology, including the Northrop Grumman prize for excellence in teaching. In 2016, he was recipient of the IEEE Gustav Robert Kirchhoff Award for "fundamental contributions to digital signal processing." He is a recipient of the IEEE Signal Processing Society's Technical Achievement Award (2002), Education Award (2012), and Society Award (2016).