THE MULTI-PITCH ESTIMATION PROBLEM: SOME NEW SOLUTIONS

Mads Græsbøll Christensen[1]*, Petre Stoica*[2]*, Andreas Jakobsson*[3]*, and Søren Holdt Jensen*[1]

[1] Dept. of Electronic Systems
Aalborg University, Denmark
{mgc,shj}@es.aau.dk

[2] Dept. of Information Technology
Uppsala University, Sweden
peter.stoica@it.uu.se

[3] Dept. of Electrical Engineering
Karlstad University, Sweden
andreas.jakobsson@kau.se

ABSTRACT

In this paper, we formulate the multi-pitch estimation problem and propose a number of methods to estimate the set of fundamental frequencies. The methods, which are based on nonlinear least-squares, MUltiple SIgnal Classification (MUSIC) and the Capon principles, have in common the fact that the multiple fundamental frequencies are estimated by means of a one-dimensional search. The statistical properties of the methods are evaluated via Monte Carlo simulations.

***Index Terms*—** Acoustic signal analysis, spectral analysis, frequency estimation

1. INTRODUCTION

The problem of finding the fundamental frequency or the pitch of a periodic waveform occurs in many signal processing applications, for example in applications involving speech and audio signals. For instance, in audio processing the fundamental frequency plays a key role in automatic transcription and classification of music [1]. Given the importance of the problem, a wide variety of fundamental frequency estimation methods have been developed in the literature, e.g., [1–4]. In most cases, these methods are based on a model where only a single set of harmonically related sinusoids are present at the same time. Indeed, the multi-pitch estimation problem, i.e., the problem of estimating the fundamental frequencies of multiple sets of periodic waveforms, is a difficult one, and one that has received much less attention than the single-pitch case, though notable exceptions can be found in [1, 5, 6]. The multi-pitch scenarios occur regularly in music signals, perhaps even more frequently than the single-pitch case, and often also in speech processing. Typically, the situation occurs whenever multiple instruments or speakers are present at the same time or when multiple tones are being played on a musical instrument. The multi-pitch estimation problem can be defined as follows: consider a signal consisting of several, say K, sets of harmonics (hereafter referred to as sources) with fundamental frequencies ω_k, for $k = 1, \ldots, K$, that is corrupted by an additive white complex circularly symmetric Gaussian noise, $w(n)$, having variance σ^2, for $n = 0, \ldots, N-1$, i.e.,

$$x(n) = \sum_{k=1}^{K} \sum_{l=1}^{L} a_{k,l} e^{j\omega_k l n} + w(n), \tag{1}$$

where $a_{k,l} = A_{k,l} e^{j\phi_{k,l}}$, with $A_{k,l} > 0$ and $\phi_{k,l}$ being the amplitude and the phase of the l'th complex harmonic of the k'th source, respectively. The problem is then to estimate the fundamental frequencies $\{\omega_k\}$, or the pitches, from a set of N measured samples, $x(n)$. In the present work, we assume that the number of sources, K, is known and that the number of harmonics, L, of each source is also known and equal for all sources. This may seem like a restrictive assumption, but for many practical speech and audio applications, it is not always required that the order be known precisely. Provided that the order does not vary too much, it is often sufficient to simply know the average order. The role of an order estimate is mainly to avoid ambiguities in the cost functions that may cause spurious estimates at q/g times the true fundamental frequency (with $q, g \in \mathbb{N}$) such as the well-known problems of halvings and doublings.

In this paper, we propose and evaluate a number of estimators for finding the fundamental frequencies $\{\omega_k\}$ based on principles from statistical signal processing. In particular, we propose an approximate nonlinear least-squares (NLS) method, a MUltiple SIgnal Classification (MUSIC) based method as well as a Capon-based method. The proposed methods all have the following simple form:

$$\{\hat{\omega}_k\} = \arg\max_{\{\omega_k\}} \sum_{k=1}^{K} J(\omega_k), \tag{2}$$

meaning that an estimate of the set of fundamental frequencies can be obtained by evaluating a cost function $J(\omega_k)$ for various ω_k and then picking the K highest peaks, i.e., costly multi-dimensional searches are avoided.

The rest of the paper is organized as follows: first, in Section 2, we introduce some notation and definitions. In Section 3, we present the proposed multi-pitch estimators along with the assumptions they are based on. Then, in Section 4, we analyze the performance of the estimators using synthetic signals and Monte Carlo simulations and, finally, we conclude the work in Section 5.

2. PRELIMINARIES

We begin by introducing some useful notation, definitions and results. First, using $\mathbf{x} = [\, x(0) \; \cdots \; x(N-1) \,]^T$ and $\mathbf{w} = [\, w(0) \; \cdots \; w(N-1) \,]^T$, with $(\cdot)^T$ denoting the transpose, we note that the signal model in (1) can be written as

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{a}_k + \mathbf{w}, \tag{3}$$

where $\mathbf{Z}_k = [\, \mathbf{z}(\omega_k) \; \cdots \; \mathbf{z}(\omega_k L) \,]$, $\mathbf{z}(\omega) = [\, 1 \; e^{j\omega} \; \cdots \; e^{j\omega(B-1)} \,]^T$, with $B = N$ and $\mathbf{a}_k = [\, a_{k,1} \; \cdots \; a_{k,L} \,]^T$. Next, we define the covariance matrix as $\mathbf{R} = \mathrm{E}\left\{\tilde{\mathbf{x}}(n)\tilde{\mathbf{x}}^H(n)\right\}$ where $\tilde{\mathbf{x}}(n)$ is a signal vector formed from M consecutive samples of the observed signal,

The work of M. G. Christensen was supported by the Intelligent Sound project, Danish Technical Research Council grant no. 26–04–0092 and the Parametric Audio Processing project, Danish Research Council for Technology and Production Sciences grant no. 274–06–0521; the work of P. Stoica was supported by the Swedish Science Council (VR).

i.e., $\tilde{\mathbf{x}}(n) = [\, x(n) \ \cdots \ x(n+M-1) \,]^T$. Here, $\mathrm{E}\{\cdot\}$ and $(\cdot)^H$ denote the statistical expectation and the conjugate transpose, respectively. We note that $\tilde{\mathbf{x}}$ can be written similarly to $\mathbf{x}$ in (3) but with $B = M$. In practice, the covariance matrix is unknown and is replaced by the sample covariance matrix. For a single source and a high number of samples, i.e., $N \gg 1$, the (asymptotic) Cramér-Rao lower bound (CRLB), can be shown to be [4]

$$CRLB_k = \frac{6\sigma^2}{N^3 \sum_{l=1}^{L} A_{k,l}^2 l^2}. \tag{4}$$

The CRLB can be seen to depend on the pseudo signal-to-noise ratio (PSNR), defined as

$$PSNR_k = 10 \log_{10} \frac{\sum_{l=1}^{L} A_{k,l}^2 l^2}{\sigma^2} \ [\mathrm{dB}]. \tag{5}$$

Under the assumption that the sources are independent and that the harmonic frequencies are distinct, (4) can also be expected to hold approximately for the problem of estimating the fundamental frequencies in (1). However, for a low number of samples, the exact CRLB for a fundamental frequency will depend on the parameters of other sources as well.

3. SOME ESTIMATORS

3.1. Approximate NLS-based Method

The first estimator is based on the nonlinear least-squares method. Under the assumption of white Gaussian noise the NLS method is equivalent to the maximum likelihood method which is well-known to have excellent performance: it attains the CRLB provided that the number of samples is sufficiently high [7]. For the sinusoidal estimation problem, the NLS method has also been shown to achieve the asymptotic CRLB for large N in the colored Gaussian noise case [8], and, therefore, the NLS can be expected to be robust to the color of the noise. The NLS estimates are obtained as the set of fundamental frequencies that minimizes the 2-norm of the difference between the observed signal and the signal model, i.e.,

$$\{\hat{\omega}_k\} = \arg\min_{\{\omega_k\}} \left\| \mathbf{x} - \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{a}_k \right\|_2^2, \tag{6}$$

where $\|\cdot\|_2$ denotes the 2-norm. Assuming that all the frequencies in $\{\mathbf{Z}_k\}$ are distinct and well separated and that $N \gg 1$, (6) can be well-approximated by finding the fundamental frequency of the individual sources, i.e.,

$$\hat{\omega}_k = \arg\min_{\omega_k} \|\mathbf{x} - \mathbf{Z}_k \mathbf{a}_k\|_2^2. \tag{7}$$

Minimizing (7) with respect to the complex amplitudes $\mathbf{a}_k$ gives the estimates $\hat{\mathbf{a}}_k = \left(\mathbf{Z}_k^H \mathbf{Z}_k\right)^{-1} \mathbf{Z}_k^H \mathbf{x}$, which, when inserted in (7), yields

$$\hat{\omega}_k = \arg\max_{\omega_k} \mathbf{x}^H \mathbf{Z}_k \left(\mathbf{Z}_k^H \mathbf{Z}_k\right)^{-1} \mathbf{Z}_k^H \mathbf{x} \tag{8}$$

$$\approx \arg\max_{\omega_k} \mathbf{x}^H \mathbf{Z}_k \mathbf{Z}_k^H \mathbf{x} \tag{9}$$

where the last line follows from the assumption that $N \gg 1$. Cast in the framework of (2), the resulting cost function is

$$J(\omega_k) = \|\mathbf{Z}_k^H \mathbf{x}\|_2^2, \tag{10}$$

where the inner product $\mathbf{Z}_k^H \mathbf{x}$ can be implemented efficiently for a linear grid search over ω_k using a fast Fourier transform (FFT). The NLS method can be extended to deal with an unknown order for the single-pitch case and colored Gaussian noise in a computationally efficient manner [9]. An alternative interpretation of the approximate NLS estimator is that (10) can be written as $J(\omega_k) = \sum_{l=1}^{L} \|\mathbf{z}(\omega_k l)^H \mathbf{x}\|_2^2$ which is the periodogram power spectral density estimate of $\mathbf{x}$ evaluated at and summed over the harmonic frequencies $\omega_k l$.

3.2. MUSIC-based Method

We proceed to examine a subspace approach based on the orthogonality principle of MUSIC (see, e.g., [10]), i.e., that the signal and noise subspaces are orthogonal. In [4], it was shown that high resolution fundamental frequency estimates can be obtained using this principle, along with an accurate order estimate, and in [11] the approach was generalized to the multi-pitch estimation problem. We will here briefly review these ideas in the context of this paper, i.e., for the case of known order and number of sources. Assuming that the phases of the harmonics are independent and uniformly distributed on the interval $(-\pi, \pi]$, the covariance matrix and its eigenvalue decomposition (EVD) can be written as

$$\mathbf{R} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^H = \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{P}_k \mathbf{Z}_k^H + \sigma^2 \mathbf{I}, \tag{11}$$

where $\mathbf{U}$ is formed from the M orthonormal eigenvectors of $\mathbf{R}$, i.e., $\mathbf{U} = [\, \mathbf{u}_1 \ \cdots \ \mathbf{u}_M \,]$, $\mathbf{\Lambda}$ is a diagonal matrix with the eigenvalues on the diagonal and $\mathbf{P}_k = \mathrm{diag}\left(\left[\, A_{k,1}^2 \ \cdots \ A_{k,L}^2 \,\right]\right)$. First, note that

$$\mathrm{rank}\left(\sum_{k=1}^{K} \mathbf{Z}_k \mathbf{P}_k \mathbf{Z}_k^H\right) = KL, \tag{12}$$

and let $\mathbf{G}$ be the noise subspace formed from the eigenvectors corresponding to the $M - KL$ least significant eigenvalues. Then, it can be shown that the noise subspace spanned by $\mathbf{G}$ will be orthogonal to the Vandermonde matrices $\{\mathbf{Z}_k\}$ which span the signal subspace formed by the eigenvectors corresponding to the KL most significant eigenvalues. Therefore, the set of fundamental frequencies can be found as [11]

$$\{\hat{\omega}_k\} = \arg\min_{\{\omega_k\}} \sum_{k=1}^{K} \left\| \mathbf{Z}_k^H \mathbf{G} \right\|_F^2, \tag{13}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Finally, we define the cost function to be maximized for each individual source as

$$J(\omega_k) = \frac{1}{\|\mathbf{Z}_k^H \mathbf{G}\|_F^2}, \tag{14}$$

which can be evaluated efficiently using the FFT (see [4] for further details). Note that while the NLS methods is based on an asymptotic assumption that facilitates finding individual fundamental frequencies independently, there is no such approximation in the MUSIC approach. The covariance matrix decomposition in the MUSIC approach, however, is dependent on the distribution of the phases and the whiteness of the noise, while the NLS approach is still asymptotically efficient for colored noise. It should also be noted that the MUSIC approach is the only method, among those considered here, that requires a priori knowledge about the number of sources for the evaluation of the cost function.

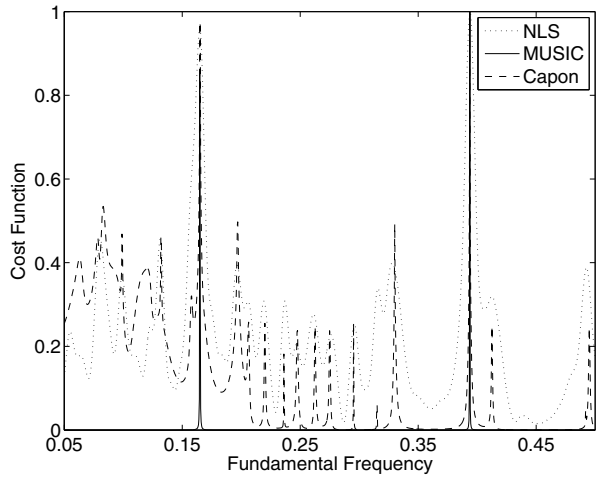

Fig. 1. Example of the cost functions for two sources having five harmonics each and true fundamental frequencies of 0.1650 and 0.3937 for $N = 160$ and $PSNR = 40$ dB.

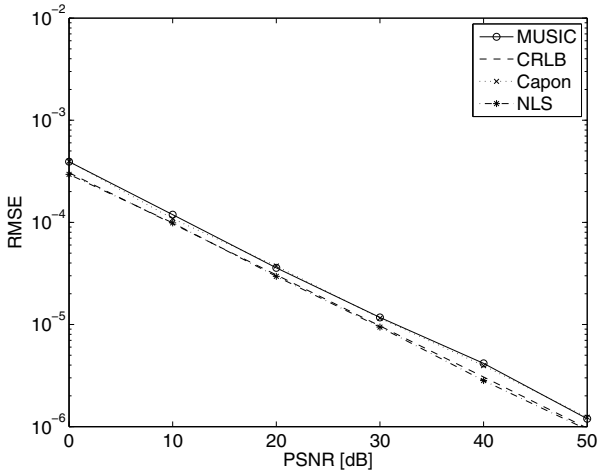

Fig. 2. RMSE as a function of PSNR for $N = 400$ for one source.

3.3. Capon-based Method

Finally, we introduce an estimator based on the Capon approach (see, e.g., [10]), which relies on the design of a set of filters that pass power undistorted at specific frequencies, here the harmonic frequencies, while minimizing the power at all other frequencies. Defining the filter bank matrix $\mathbf{H}^H$, consisting of L filters of length M, the filter design problem can be stated as the optimization problem:

$$\min_{\mathbf{H}} \mathrm{Tr}\left[\mathbf{H}^H\mathbf{R}\mathbf{H}\right] \quad \text{subject to} \quad \mathbf{H}^H\mathbf{Z}_k = \mathbf{I}, \tag{15}$$

where $\mathbf{I}$ is the $L \times L$ identity matrix. The set of filters $\mathbf{H}$ solving (15) are given by (see, e.g., [10])

$$\mathbf{H} = \mathbf{R}^{-1}\mathbf{Z}_k\left(\mathbf{Z}_k^H\mathbf{R}^{-1}\mathbf{Z}_k\right)^{-1}. \tag{16}$$

This data and frequency dependent filter bank can then be used to estimate the fundamental frequencies by maximizing the power of the filter's output, i.e., $\mathrm{Tr}\left[\mathbf{H}^H\mathbf{R}\mathbf{H}\right]$. Inserting (16) into this expression yields

$$\hat{\omega}_k = \arg\max_{\omega_k} \mathrm{Tr}\left[\left(\mathbf{Z}_k^H\mathbf{R}^{-1}\mathbf{Z}_k\right)^{-1}\right], \tag{17}$$

which can be seen to depend only on the Vandermonde matrix $\mathbf{Z}_k$ and the inverse covariance matrix $\mathbf{R}^{-1}$. Defining $\mathbf{Y} = \mathbf{Z}_k^H\mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}}$, the cost function can be evaluated for different ω_k as

$$J(\omega_k) = \mathrm{Tr}\left[\left(\mathbf{Y}\mathbf{Y}^H\right)^{-1}\right], \tag{18}$$

where $\mathbf{Y}$ can be computed using the FFT once the EVD of $\mathbf{R}$ has been evaluated. The form (18) is preferred over, e.g., a Cholesky-based implementation because numerical issues can easily be resolved. Alternatively, the filter bank design in (15) can be formulated as the design of a single filter which is subject to L constraints, one for each harmonic. Interestingly, such an approach has some conceptual similarities with the comb-filtering approach of [12].

4. NUMERICAL RESULTS

In this section, we evaluate the performance of the introduced estimators. Figure 1 shows the cost functions of the proposed estimators for a signal of length $N = 160$ consisting of $K = 2$ sources having five unit amplitude harmonics each with $PSNR = 40$ dB. The two sources have fundamental frequencies 0.1650 and 0.3937, respectively. As can be seen, the cost functions have distinct peaks at those frequencies with the MUSIC- and Capon-based method having narrower peaks than the approximate NLS. Also worth noting is the multi-modal nature of the cost functions with a number of fairly sharp false peaks. Indeed, this shows why the fundamental frequency estimation problem is a difficult one. At first sight, this appears to be less of an issue for the MUSIC-based approach, but upon closer inspection, it can be observed that MUSIC generally suffers from this problem too. We proceed to evaluate the proposed estimators using Monte Carlo simulations by generating signals according to the model (1) with the phases and the noise being randomized over realizations. In the first experiment, the estimators are evaluated for two sources with well-separated fundamental frequencies, again 0.1650 and 0.3937, and with $L = 3$. We compare the root mean square estimation error (RMSE) of the estimators with the asymptotic CRLB given in (4). The RMSEs are calculated jointly over both sources. In order to have the same CRLB for both fundamental frequencies, we set all amplitudes to unity, i.e., $A_{k,l} = 1,\ \forall k, l$. The estimates are obtained as follows: First, the cost functions (10), (14), and (18) are evaluated on a coarse grid. Then, these coarse estimates are used to initialize gradient-based methods that are used to obtain refined estimates. For the MUSIC- and Capon-based methods, the gradients of (14), and (18) are used whereas for NLS, the gradient for the approximate cost function (10) was found not to produce high resolution estimates. Instead, the gradient was derived for this case based on (8). For the MUSIC-based method, we choose $M = \lfloor N/2 \rfloor$ while for the Capon-based method we used $M = \lfloor 2N/5 \rfloor$[1]. We note that the cost function (8) is approximate, being based on the negligence of the inner products between the sources. The experiments are run for a fixed number of observations, $N = 400$, and varying PSNRs for one harmonic source and two harmonic sources, respectively. The signals are generated as described in the previous experiment and for all combinations of parameters, 100 Monte Carlo trials are run. The RMSE are shown in Figures

[1]These values were found empirically to result in good performance. The reason for having different M for the two methods is that they exhibit different sensitivity to the choice of M.

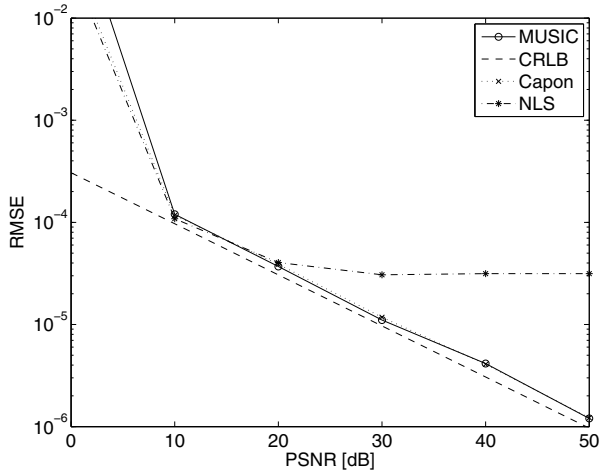

Fig. 3. RMSE versus the PSNR for $N = 400$ for two sources.

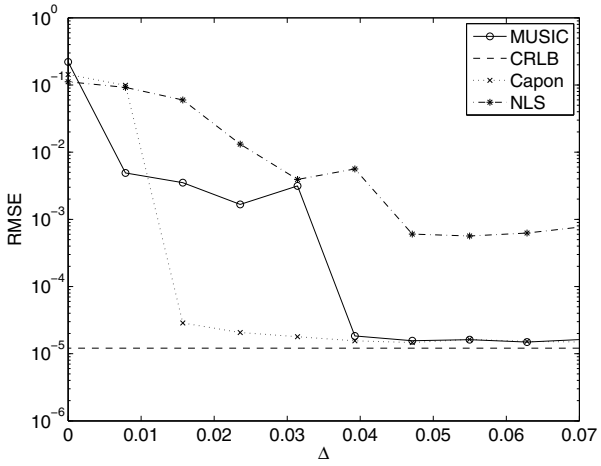

Fig. 4. RMSE versus the difference between the fundamental frequencies of two sources for $N = 160$ and $PSNR = 40$ dB.

2 and 3. It can be seen that for the case of one harmonic source, all estimators perform well for all tested PSNRs with NLS having the best performance. For one harmonic source, the NLS method is exact, meaning that there is no approximation in the estimate (8). For two sources, however, the NLS method can be seen to saturate at PSNRs above 20 dB while both the MUSIC- and Capon-based methods follow the CRLB closely. It can also be observed that all methods exhibit thresholding effects below 10 dB. Similar observations can be made about the behavior of the estimators as a function of the number of observations, N. In a final experiment, the RMSE is studied as a function of the difference between the fundamental frequencies of two harmonic sources, i.e., $\Delta = |\omega_1 - \omega_2|$, for a PSNR of 40 dB and $N = 160$. The results are shown in Figure 4. It can be seen that the Capon-based approach performs the best for closely spaced harmonics and that the approximate NLS performs the worst.

5. CONCLUSIONS

We have considered the problem of estimating the fundamental frequencies of superpositions of periodic waveforms, also known as the multi-pitch estimation problem. We have proposed a number of estimators that are based on one-dimensional evaluations of cost functions, namely the approximate nonlinear least-squares, MUSIC- and Capon-based techniques. The basic assumptions for these methods to work for the multi-pitch estimation problem have been outlined and their finite sample performance has been evaluated using Monte Carlo simulations. It has been found that the MUSIC- and Capon-based methods have good statistical performance for both the multi- and single-pitch cases, following the Cramér-Rao lower bound closely. The approximate NLS, however, has excellent performance for the single-pitch case but does not perform well for the multi-pitch case. For closely spaced fundamental frequencies the Capon-based approach has been found to have a performance superior to that of the MUSIC-based method.

6. REFERENCES

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

[2] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Signal Processing*, vol. 80, pp. 1937–1944, 2000.

[3] K. W. Chan and H. C. So, "Accurate frequency estimation for real harmonic sinusoids," *IEEE Signal Processing Lett.*, vol. 11(7), pp. 609–612, July 2004.

[4] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. on Audio, Speech and Language Processing*, Apr. 2006, submitted.

[5] R. Gribonval and E. Bacry, "Harmonic Decomposition of Audio Signals with Matching Pursuit," *IEEE Trans. Signal Processing*, vol. 51(1), pp. 101–111, Jan. 2003.

[6] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech and Audio Processing*, vol. 11(6), pp. 804–816, 2003.

[7] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, 1993.

[8] P. Stoica, A. Jakobsson, and J. Li, "Cisiod parameter estimation in the coloured noise case: Asymptotic Cramér-Rao bound, maximum likelihood, and nonlinear least-squares," in *IEEE Trans. Signal Processing*, Aug. 1997, vol. 45(8), pp. 2048–2059.

[9] M. G. Christensen and S. H. Jensen, "Variable order harmonic sinusoidal parameter estimation for speech and audio signals," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2006.

[10] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Pearson Prentice Hall, 2005.

[11] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation using harmonic MUSIC," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2006.

[12] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34(5), pp. 1124–1138, Oct. 1986.