Joint High-Resolution Fundamental Frequency and Order Estimation

Christensen, Mads Græsbøll; Jakobsson, Andreas; Jensen, Søren Holdt

Published in:
IEEE Transactions on Audio Speech and Language Processing

DOI (link to publication from Publisher):
10.1109/TASL.2007.899267

Publication date:
2007

Document Version
Peer reviewed version

Joint High-Resolution Fundamental Frequency and Order Estimation

Mads Græsbøll Christensen*, *Member, IEEE*, Andreas Jakobsson, *Senior Member, IEEE*, and Søren Holdt Jensen, *Senior Member, IEEE*

***Abstract*— In this paper, we present a novel method for joint estimation of the fundamental frequency and order of a set of harmonically related sinusoids based on the MUSIC estimation criterion. The presented method, termed HMUSIC, is shown to have an efficient implementation using FFTs. Furthermore, refined estimates can be obtained using a gradient-based method. Illustrative examples of the application of the algorithm to real-life speech and audio signals are given, and the statistical performance of the estimator is evaluated using synthetic signals, demonstrating its good statistical properties.**

I. Introduction

Fundamental frequency estimators form an integral part in many signal processing applications, and especially so in speech and audio processing. For example, long-term prediction in linear prediction-based speech coding requires that the fundamental frequency, or equivalently the pitch period, is estimated [1]. Similarly, parametric coding of speech and audio using a harmonic sinusoidal model is typically based on a fundamental frequency estimator [2]. Fundamental frequency estimators are also key components in music information retrieval applications such as automatic music transcription and in musical genre classification [3]. Many fundamental frequency estimators used in speech and audio processing are time-domain techniques based on the auto-correlation function, cross-correlation function, averaged magnitude difference function, or averaged squared difference function. These methods are typically biased methods primarily concerned with handling particular problems in speech and audio processing, such as the so-called first formant interaction problem, than obtaining high resolution frequency estimates. An illustrative example of such a recent approach can be found in [4]. For a historical review of fundamental frequency estimation methods, we refer to [5], [6], and for examples of more recent work, we refer to [4], [7]–[13].

The fundamental frequency estimation problem can be defined as follows. Consider a harmonic signal with the fundamental frequency ω_0 that is corrupted by an additive white complex circularly symmetric Gaussian noise, $w(n)$, for $n = 0, \ldots, N-1$,

$$x(n) = \sum_{l=1}^{L} A_l e^{j(\omega_0 l n + \phi_l)} + w(n), \tag{1}$$

where $A_l > 0$ and ϕ_l are the amplitude and the phase of the l'th harmonic, respectively. The frequency of the l'th harmonic is thus $\omega_l = \omega_0 l$, and the problem considered in this paper is to estimate the fundamental frequency ω_0, as well as the model order L, from a set of N measured samples, $x(n)$. We refer to the number of harmonics, L, as the model order. Herein, we focus on the most common case, for which the set of harmonics follows $\omega_l = \omega_0 l$, for $l = 1, \ldots, L$; however, other cases may be of interest where some of the harmonics, even the fundamental, may be missing [14]. It should be stressed that most estimators operate under the assumption of the model order being known. Typically, this requires an initial order estimation prior to the frequency estimation. For example, the nonlinear least-squares (NLS) method is well-known to be equivalent to the maximum likelihood estimator under the condition of white Gaussian noise, provided that the model order is known [15]. We refer the reader to [16], [17] for a more complete discussion of this difficult problem. We remark that real valued signals can be cast into the form of (1) via the down-sampled discrete-time analytic signal [18]. Here, we have used the complex formulation because of its notational simplicity and because it leads to computationally less complex algorithms.

It has recently been shown that the MUltiple SIgnal Classification (MUSIC) estimation criterion can be used for high-resolution estimation of the fundamental frequency [19]. The resulting estimator was shown to have good statistical performance, approaching the Cramér-Rao lower bound (CRLB), provided that the order L is known. In this paper, we further extend on this work. Specifically, we propose an algorithm that jointly estimates the fundamental frequency and the order, showing that this algorithm can be efficiently implemented using the fast Fourier transform (FFT). Also, refinements of the estimates can be obtained using a gradient-based method derived herein. We refer to the proposed estimator as the harmonic MUSIC (HMUSIC) estimator. Using simulated data, the proposed estimator is evaluated using Monte Carlo simulations. Furthermore, we compare to the asymptotic CRLB and the recent Markov-like weighted least squares (WLS) estimator published in [7]. Additionally, illustrative examples of the application of the proposed methods to speech and audio

This research was supported in part 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.

M. G. Christensen and S. H. Jensen are with the Dept. of Electronic Systems, Aalborg University, Niels Jernes Vej 12, DK-9220 Aalborg, Denmark (phone: +45 96 35 {86 20, 86 54}, fax: +45 96 15 15 83, email: {mgc,shj}@kom.aau.dk).

A. Jakobsson is with the Dept. of Electrical Engineering, Karlstad University, Universitetgatan 2, SE-651 88 Karlstad, Sweden (phone: +45 54 700 2330, fax: +46 54 700 2197, email andreas.jakobsson@ieee.org).

signal analysis are given.

The remaining part of the paper is organized as follows. In Section II, the proposed estimator is presented along with some implementation details. In Section III, numerical results and illustrative examples are presented, and Section IV concludes on the work.

II. The Proposed Estimator

A. Covariance Matrix Model

In this section, we present the fundamentals of the MUSIC algorithm [20], [21] (see also [22]) and introduce some useful vector and matrix definitions. We start out by defining $\tilde{\mathbf{x}}(n)$ as a signal vector containing M samples of the observed signal, i.e.,

$$\tilde{\mathbf{x}}(n) = \begin{bmatrix} x(n) & x(n-1) & \cdots & x(n-M+1) \end{bmatrix}^T, \quad (2)$$

with $(\cdot)^T$ denoting the transpose. Then, assuming that the phases of the harmonics are independent and uniformly distributed on the interval $(-\pi, \pi]$, the covariance matrix $\mathbf{R} \in \mathbb{C}^{M\times M}$ of the signal in (1) can be written as [16]

$$\begin{aligned} \mathbf{R} &= \mathrm{E}\left\{\tilde{\mathbf{x}}(n)\tilde{\mathbf{x}}^H(n)\right\} \\ &= \mathbf{A}\mathbf{P}\mathbf{A}^H + \sigma^2\mathbf{I}, \end{aligned} \quad (3)$$

where $\mathrm{E}\{\cdot\}$ and $(\cdot)^H$ denote the statistical expectation and the conjugate transpose, respectively. Note that for this decomposition to hold, the noise need not be Gaussian. Furthermore, $\mathbf{P}$ is a diagonal matrix containing the squared amplitudes, i.e.,

$$\mathbf{P} = \mathrm{diag}\left(\begin{bmatrix} A_1^2 & \cdots & A_L^2 \end{bmatrix}\right), \quad (4)$$

and $\mathbf{A} \in \mathbb{C}^{M\times L}$ a full rank Vandermonde matrix defined as

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}(\omega_0) & \cdots & \mathbf{a}(\omega_0 L) \end{bmatrix}, \quad (5)$$

where $\mathbf{a}(\omega) = \begin{bmatrix} 1 & e^{-j\omega} & \cdots & e^{-j\omega(M-1)} \end{bmatrix}^T$. Also, σ^2 denotes the variance of the additive noise, $w(n)$, and $\mathbf{I}$ is the $M \times M$ identity matrix. We note that $\mathbf{A}\mathbf{P}\mathbf{A}^H$ has rank L. Let

$$\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H \quad (6)$$

be the eigenvalue decomposition (EVD) of the covariance matrix. Then, $\mathbf{U}$ contains the M orthonormal eigenvectors of $\mathbf{R}$, i.e., $\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_M \end{bmatrix}$ and $\mathbf{\Lambda}$ is a diagonal matrix containing the corresponding eigenvalues, λ_k, with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$. Let $\mathbf{G}$ be formed from the eigenvectors corresponding to the $M-L$ least significant eigenvalues, i.e.,

$$\mathbf{G} = \begin{bmatrix} \mathbf{u}_{L+1} & \cdots & \mathbf{u}_M \end{bmatrix}. \quad (7)$$

The noise subspace spanned by $\mathbf{G}$ will then be orthogonal to the Vandermonde matrix $\mathbf{A}$, i.e.,

$$\mathbf{A}^H\mathbf{G} = \mathbf{0}, \quad (8)$$

and $\mathcal{R}(\mathbf{A}) = \mathcal{R}(\mathbf{S})$ with $\mathcal{R}(\cdot)$ denoting the range and $\mathbf{S} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_L \end{bmatrix}$ being the eigenvectors that span the signal subspace. We here form a consistent estimate of the covariance matrix $\mathbf{R}$ as

$$\widehat{\mathbf{R}} = \frac{1}{N}\sum_{n=M}^{N} \tilde{\mathbf{x}}(n)\tilde{\mathbf{x}}^H(n). \quad (9)$$

The MUSIC estimate is found as the frequencies minimizing

$$J = \|\mathbf{A}^H\mathbf{G}\|_F^2 = \mathrm{Tr}\left\{\mathbf{A}^H\mathbf{G}\mathbf{G}^H\mathbf{A}\right\}, \quad (10)$$

with $\mathrm{Tr}\{\cdot\}$ and $\|\cdot\|_F$ denoting the trace and Frobenius norm, respectively. Note that for notational simplicity, we have omitted the dependency of $\mathbf{A}$ and $\mathbf{G}$ on the unknowns. For more on the performance of MUSIC see, e.g., [23], [24], and for more on subspace-based estimation techniques in general, see, e.g., [25], [26].

B. Harmonic MUSIC

Herein, we will extend the MUSIC estimation criterion in (10) for jointly estimating both the fundamental frequency and the model order. We note that the cost function in (10) varies with the order L and the size of the covariance matrix, M, and we must therefore first derive an appropriate scaling. By the Cauchy-Schwarz inequality, we have that

$$\|\mathbf{A}^H\mathbf{G}\|_F \leq \|\mathbf{A}^H\|_F\|\mathbf{G}\|_F. \quad (11)$$

As the $M-L$ columns of $\mathbf{G}$ are orthonormal, and all the L columns of $\mathbf{A}$ have norm $\sqrt{M}$, we get

$$\frac{\|\mathbf{A}^H\mathbf{G}\|_F}{\sqrt{LM(M-L)}} \leq 1. \quad (12)$$

The scale factor $\sqrt{LM(M-L)}$, which is due to the variable dimensions of $\mathbf{A}$ and $\mathbf{G}$, makes the noise floor of the cost function invariant to the matrix dimensions. When the order is unknown, which is generally the case, it was estimated in [19] as $L = \lfloor \frac{2\pi}{\omega_0} \rfloor - 1$, and the fundamental frequency was estimated as the value for which the Vandermonde matrix is closest to being orthogonal to the noise subspace, i.e.,

$$\hat{\omega}_0 = \arg\max_{\omega_0\in\Omega} \frac{LM(M-L)}{\|\mathbf{A}^H\mathbf{G}\|_F^2}, \quad (13)$$

where Ω is a set of candidate fundamental frequencies. However, this approach may lead to a wrong identification of the noise subspace in (7), and this, in turn, led to the problems of spurious estimates reported in [19]. The orthogonality in (8) will hold only when the estimated fundamental frequency equates the true frequency *and* the order L is chosen such that $\mathcal{R}(\mathbf{A}) = \mathcal{R}(\mathbf{S})$. Thus, we propose to jointly estimate the fundamental frequency and the order as follows. Define the two-dimensional cost function, depending on both the fundamental frequency and the order, as

$$P(\omega_0, L) = \frac{LM(M-L)}{\|\mathbf{A}^H\mathbf{G}\|_F^2}. \quad (14)$$

Then, exploiting (14), we proceed to estimate the fundamental frequency as

$$\hat{\omega}_0 = \arg\max_{\omega_0\in\Omega} \max_{L\in\mathcal{L}} P(\omega_0, L). \quad (15)$$

We term the resulting estimator the harmonic MUSIC (HMUSIC). Note that this order estimation principle holds in general for any set of linearly independent vectors. We stress that both $\mathbf{A}$ and $\mathbf{G}$ depend on L while only $\mathbf{A}$ depends on the fundamental frequency and that as a by-product of (15), we also get an estimate of the order L. In some cases,

the model order may not be of interest, but even then it is necessary to determine it to obtain a correct fundamental frequency estimate. It should be noted that the set of possible orders, $\mathcal{L}$, will depend on the fundamental frequency since the harmonics are bounded by 2π. Because of this dependency, the maximizations in (15) cannot be interchanged without modifying the sets Ω and $\mathcal{L}$ accordingly.

C. Efficient Implementation

The three major sources of computational complexity of the proposed method are the calculation of the EVD of the covariance matrix in (6), the inner product $\mathbf{A}^H\mathbf{G}$, and evaluating the Frobenius norm in (14). We will now show how the algorithm can be implemented efficiently. First, we define the Fourier matrix $\mathbf{F} \in \mathbb{C}^{F\times F}$, with $F \gg N$, as

$$\mathbf{F} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & z^1 & z^2 & \cdots & z^{F-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & z^{(F-1)} & z^{2(F-1)} & \cdots & z^{(F-1)(F-1)} \end{bmatrix}, \quad (16)$$

where $z = e^{-j2\pi\frac{1}{F}}$. Next, we define a matrix $\mathbf{D} \in \mathbb{R}^{F\times M}$ containing the squared absolute values of the inverse FFTs of the zero-padded eigenvectors in $\mathbf{U}$ as

$$[\mathbf{D}]_{lm} = \left| \left[\mathbf{F}^H \begin{bmatrix} \mathbf{U} \\ \mathbf{0} \end{bmatrix} \right]_{lm} \right|^2, \quad (17)$$

with $[\mathbf{D}]_{lm}$ being the (l,m)'th element of $\mathbf{D}$. For a candidate pair of a fundamental frequency of $2\pi\frac{f}{F}$ and an order L', the Frobenius norm in (14) can be calculated as

$$\|\mathbf{A}^H\mathbf{G}\|_F^2 = \sum_{m=L'+1}^{M} \sum_{l=1}^{L'} [\mathbf{D}]_{(fl+1)m}. \quad (18)$$

Thus, the complexity of calculating $\|\mathbf{A}^H\mathbf{G}\|_F^2$ for different ω_0 and L can be significantly reduced by calculating the inverse FFT of all the eigenvectors once for each given data set. We note that some of the eigenvectors, corresponding to the largest eigenvalues, can be excluded from definition of (17), since there is a lower bound on L'. Here, we have included all of them for notational simplicity.

D. Refined Estimates

For many applications, only a coarse estimate of the fundamental frequency is needed, in which case the estimator in (15) may produce sufficiently accurate results at a reasonable complexity using the FFT-based method. If, however, very accurate estimates are desired, a refined estimate can be found as described next. For a given L, the gradient of the cost function (10) can be shown to be

$$\nabla J = \frac{\partial J}{\partial \omega_0} = 2\,\mathrm{Re}\left(\mathrm{Tr}\left\{ \mathbf{A}^H\mathbf{G}\mathbf{G}^H \frac{\partial}{\partial\omega_0}\mathbf{A} \right\} \right), \quad (19)$$

with $\mathrm{Re}(\cdot)$ denoting the real value, $\odot$ the Schur-Hadamard (element-wise) product, and

$$\frac{\partial}{\partial\omega_0}\mathbf{A} = -\mathbf{Y} \odot \mathbf{A} \quad (20)$$

with

$$\mathbf{Y} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ j & j2 & \cdots & jL \\ \vdots & \vdots & & \vdots \\ j(M-1) & j(M-1)2 & \cdots & j(M-1)L \end{bmatrix}. \quad (21)$$

This gradient can be used for finding refined estimates using standard methods. Here, we iteratively find a refined estimate of the fundamental frequency as

$$\hat{\omega}_0^{(i+1)} = \hat{\omega}_0^{(i)} - \delta\nabla J, \quad (22)$$

with i being the iteration index and δ a small, positive constant that is found adaptively using line search [27]. The method is initialized for $i = 0$ using the coarse estimate obtained from (15). As the order L is kept fixed, only the matrix $\mathbf{A}$ changes in each iteration. Note that instead of the gradient approach based on the HMUSIC cost function, the NLS cost function could be used. However, the NLS cost function (see, e.g., [7], [16]) is more complicated than the HMUSIC cost function, involving matrix inversion, and would hence be more computationally demanding. Also, the NLS cost function is known to be multimodal with an abundance of local minima.

III. Experimental Results

A. Signal Examples

We start out the experimental part of this paper by illustrating the application of the proposed method to analysis of an audio signal. Figure 1 shows a segment of a quasi-harmonic signal produced by a musical instrument, a violin. In Figure 2, the cost function (15) is shown for this signal for different values of the fundamental frequency and order. The combination of the fundamental frequency and the order estimates can be identified as large peaks in the landscape. It is interesting to note that a measure of the confidence of the estimate is how distinct the peak is. As can be seen from the figure, for the fairly stationary signal in Figure 1, the associated cost function has a very distinct peak. For non-stationary signals, or signals where the model does not hold, the cost function can be observed to be very noisy, with no distinct peak. Yet another example is shown in Figure 3, this time for a trumpet. The top panel shows the spectrogram (for low frequencies) of the signal while the bottom panel shows the estimated fundamental frequencies. The estimates can be seen to follow the fundamental frequency of the top panel. Note that there are some spurious estimates in the transition between notes at 4.25 s. For such non-stationary segments, or for segments containing multiple sets of harmonics, the signal model in (1) is invalid. The estimation of multiple fundamental frequencies can be incorporated into the proposed estimator at the cost of increased computational complexity. For these examples, the experimental setup was as follows: $N = 282$ samples were obtained at a sampling frequency of $f_s = 11025$ Hz. For each segment, the down-sampled discrete-time analytic signal was calculated using the FFT method [18]. Then, the sample covariance matrix of size $M = 110$ was calculated. The cost function in (15) was evaluated for fundamental frequencies

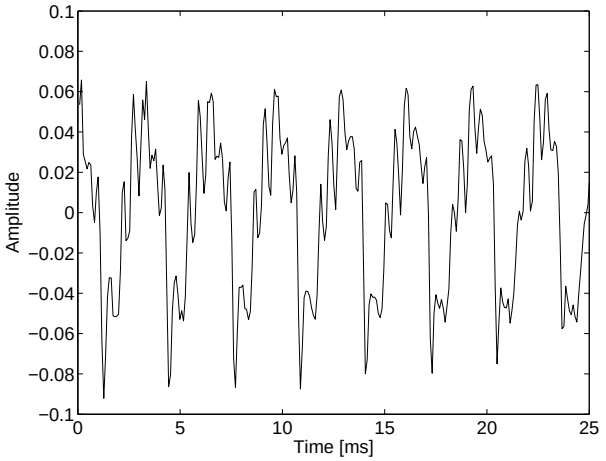

Fig. 1. Example of a quasi-harmonic signal, a segment of a violin signal.

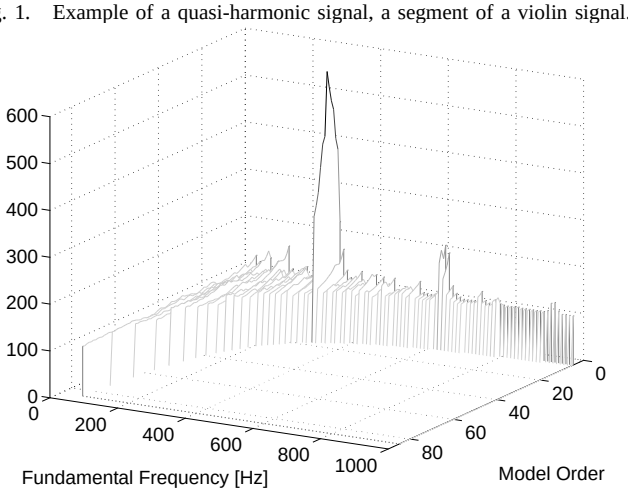

Fig. 2. HMUSIC cost function for different combinations of fundamental frequencies and orders for the signal in Figure 1.

corresponding to frequencies from approximately 60 to 1000 Hz in steps of 10 Hz using the FFT-based method with[1] $F = 1024$. For each possible fundamental frequency the model orders considered were $\mathcal{L} = \{5, \ldots, \lfloor \frac{2\pi}{\omega_0} \rfloor - 1\}$ with $\lfloor \cdot \rfloor$ denoting truncation.

Next, we demonstrate the application of the proposed method, along with the importance of the order estimate, to speech analysis. We have used speech sample dominated by voiced speech, namely a femaly speaker uttering "Why were you away a year Roy?". In Figure 4, the estimated fundamental frequencies are shown along with the spectrogram of the speech signal for various noise conditions, i.e., white Gaussian noise with signal-to-noise ratios (SNR) of 0, 10, 20, and 30 dB, respectively. The SNR is defined as $10\log_{10}(\bar{\sigma}^2/\sigma^2)$, with $\bar{\sigma}^2$ and σ^2 being the power of the speech and noise signals, respectively. The sampling frequency was $f_s = 8000$ Hz and segments with $N = 204$ (25.6 ms) were used to calculate the down-sampled discrete-time analytic signal which was then used to form a covariance matrix of size $M = 80$. The cost functions were evaluated on a 2 Hz grid from 60 to 400 Hz

[1]This value was chosen primarily for illustrative purposes. In practice, a higher value would most likely be desirable.

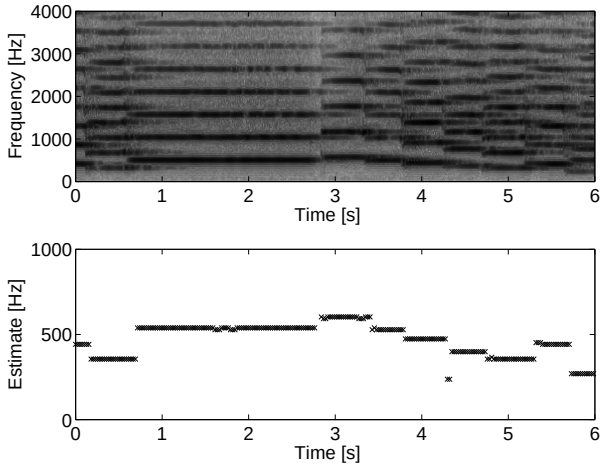

Fig. 3. Spectrogram of trumpet signal for low frequencies (top) and fundamental frequencies estimated using the proposed method (bottom).

using the FFT-based method. It can be seen from the figure that the signal is highly non-stationary, compared to the audio signals in Figures 1 and 3, with the fundamental frequency and order varying continuously throughout the duration of the signal. The crosses indicate the estimated fundamental frequencies as found using (15) where the order is estimated for each segment (denoted estimated order). The circles indicate the fundamental frequencies found by evaluating (13) for a fixed order of 5 (denoted fixed order). For SNRs above 0 dB, the proposed method can be seen to consistently estimate the correct fundamental frequency. At 0 dB, there are some erroneous estimates in low energy segments. For the fixed order case, however, spurious estimates can be observed for all SNRs, with the actual performance depending on how well the assumption of an order of 5 fits. This clearly shows that the importance of the order estimate in fundamental frequency estimation. Note that for SNRs below 0 dB, both methods do not return any meaningful estimates.

B. Statistical Evaluation

Next, we use an experimental evaluation similar to that of [7]. In assessing the statistical properties of the proposed estimator, we employ Monte Carlo simulations. In each trial, a signal is generated according to the model in (1), with the parameters and noise realizations being randomized. We will here use parameter values and constants that would be in the order of those used in speech and audio processing. Since the exact CRLB for the problem considered here varies with the parameters in a complicated way [7] and we here randomize the parameters, we instead compare the proposed method to the asymptotic CRLB. The asymptotic CRLB ($N \gg 1$), i.e. the lower bound on the variance of an unbiased estimator, for the fundamental frequency of the model in (1) can be shown to be (see Appendix I)

$$CRLB(\omega_0) = \frac{6\sigma^2}{N^3 \sum_{l=1}^{L} A_l^2 l^2}. \tag{23}$$

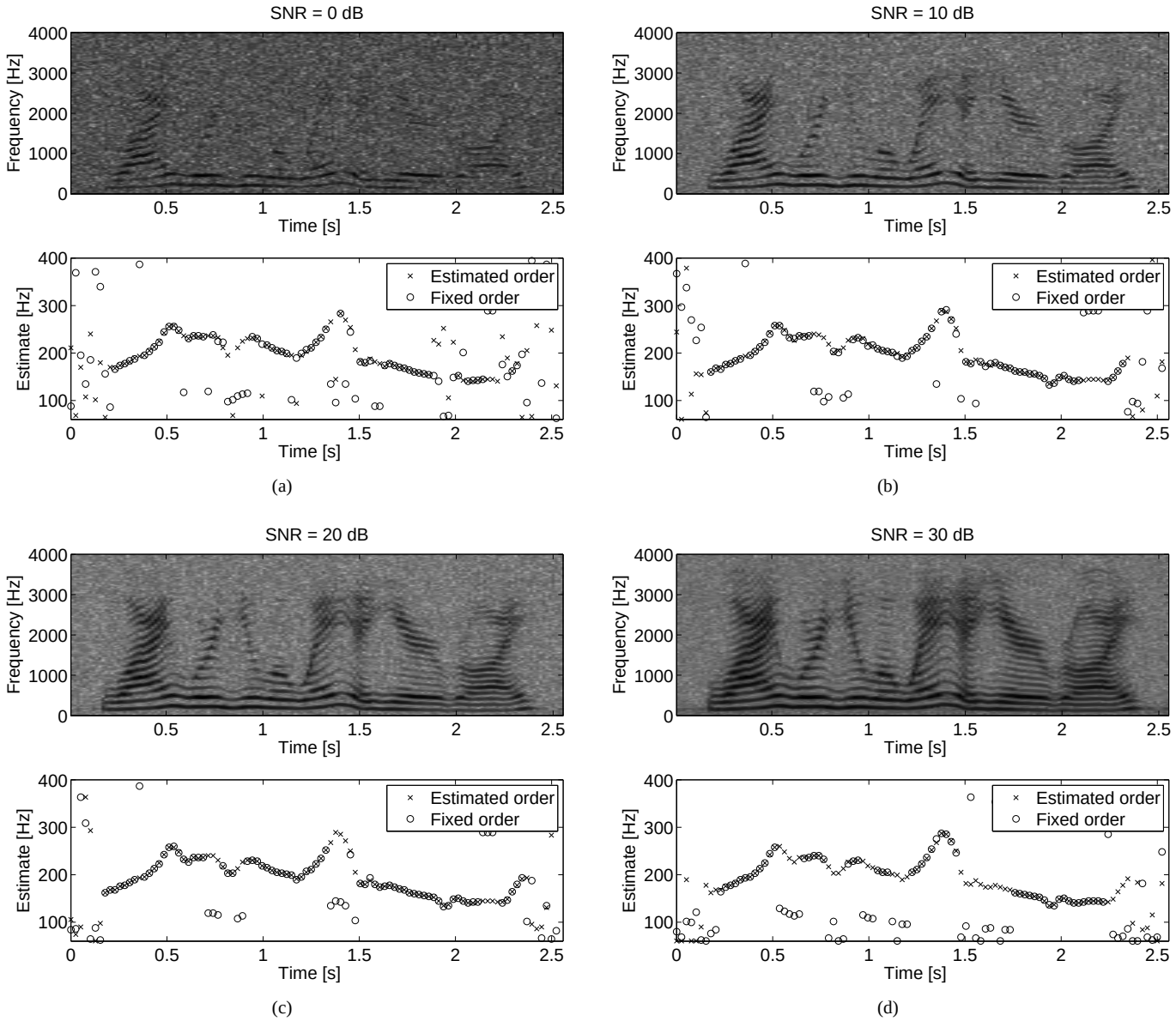

Fig. 4. Estimated fundamental frequencies (bottom panels) found using (13) with a fixed order (circles) and using (15) where also the order is estimated (crosses) for a speech signal in additive white Gaussian noise for various SNRs (top panels).

The CRLB can be seen to depend on the pseudo signal-to-noise ratio (PSNR)[2]:

$$PSNR = 10\log_{10}\frac{\sum_{l=1}^{L} A_l^2 l^2}{\sigma^2}\ [\text{dB}]. \tag{24}$$

In Appendix II, it is shown how this definition of the PSNR relates to the more commonly used definition of the SNR. The benefit of an adaptive order estimate is evident from the CRLB: the more harmonics that are present, the more accurate an estimate we can get and, due to the weighting by l^2, the higher harmonics are actually more important than the lower. It is interesting to note that the asymptotic CRLB in (23) does not depend on the fundamental frequency and that, as expected, the fundamental frequency can be estimated more accurately than any of the frequencies of the individual harmonics. For a low number of observations, the performance of estimators is expected to depend on the fundamental frequency, since the fundamental frequency determines how closely spaced the harmonics will be. Figure 5 shows the exact CRLB, given in [7], averaged over different realizations of the parameters and noise in (1), for a PSNR of 20 dB, as a function of the fundamental frequency, for a sampling frequency of 8000 Hz. The amplitudes in (1) were generated according to a Rayleigh probability density function (pdf) while the phases were distributed uniformly and $L = 10$. The exact CRLB can be seen to depend on the fundamental frequency with the bound increasing for low frequencies. For high frequencies and high N, the bound can be seen to approach the asymptotic CRLB in (23).

In the experiments to follow, we use the following setup: two cases are considered, namely for constant amplitudes, i.e.,

[2]The PSNR is defined similarly in [28] but differently in [7].

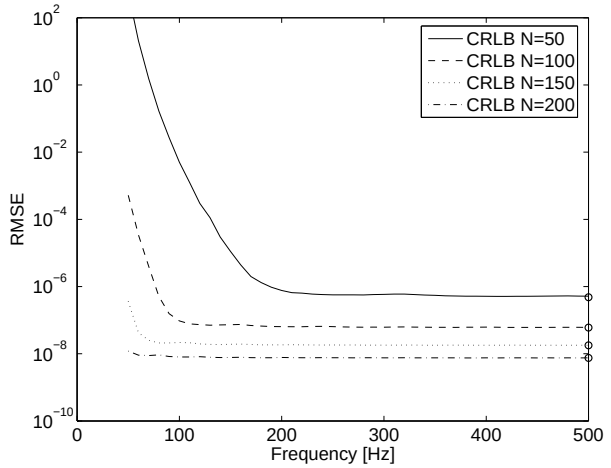

Fig. 5. The exact CRLB averaged over 1000 realizations as a function of the fundamental frequency (in Hz) for a sampling frequency of 8000 Hz. The circles at the right edge indicate the corresponding asymptotic CRLB.

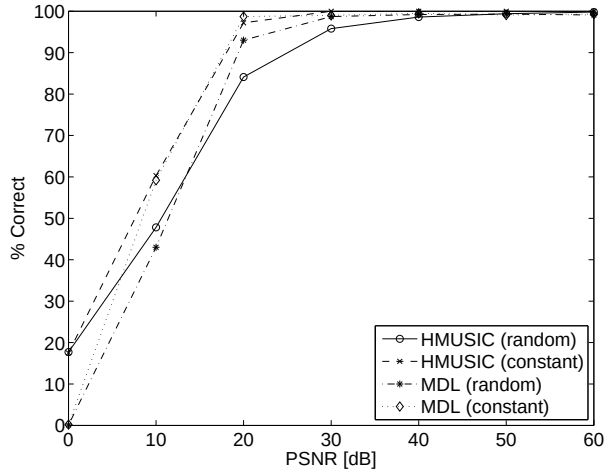

Fig. 6. Percentage of correctly estimated orders of the proposed method for varying PSNR for constant and randomized amplitudes, with $N = 100$.

$A_l = 1\ \forall l$, and for amplitudes that are generated according to a Rayleigh pdf. The motivation for testing the algorithm with amplitudes that are Rayleigh distributed is that, for speech and audio signals, the harmonics can typically not be assumed to have equal amplitudes. Hence, robustness towards such variations is desirable. In both cases, orders were generated from a uniform probability mass function (pmf) from 5 to 10 and with a fundamental frequency of $\omega_0 = 0.1963$. We see from Figure 5 that the asymptotic CRLB (indicated by circles) can be expected to hold for this value. The cost function in (15) was evaluated for fundamental frequencies in the interval $\Omega \in [0.04; 0.4]$ using the FFT-based method with $F \approx 8192M$. Note that this interval includes $2\omega_0$ and $\frac{1}{2}\omega_0$, so any potential problems with spurious estimates at these frequencies, as are often seen in fundamental frequency estimators, would show up in the statistical evaluation. Moreover, the MUSIC algorithm is generally sensitive to the choice of M relative to N. This is an inherent tradeoff between having many vectors in the averaging in (9) while retaining sufficient dimensions of the signal and noise subspaces. Here, we have used $M = \lfloor \frac{4}{5}N \rfloor$. For each possible fundamental frequency $\omega_0 \in \Omega$, the models orders considered were $\mathcal{L} = \{5, \ldots, \lfloor \frac{2\pi}{\omega_0} \rfloor - 1\}$. The noise was complex white Gaussian distributed while the phases were distributed uniformly on the interval $(-\pi; \pi]$.

First, we confirm that the proposed method results in an accurate order estimate. In Figure 6, the percentage of correctly estimated orders are shown for varying PSNR with $N = 100$. For each PSNR, 1000 trials were run. We here compare to the minimum description length (MDL) method [29]–[31] (see also [16], [32]). In finding the MDL estimates, the true fundamental frequency was used, the log-likelihoods were calculated using amplitudes that were estimated using least-squares (see [17], [33]), and the noise variance was estimated by subtracting the estimated sinusoids from the signal. As can be seen in the figure, the proposed method estimates the correct order for sufficiently high PSNRs. For randomized amplitudes, HMUSIC is slightly worse than MDL, if the latter is allowed to know the true fundamental frequency, for 20 and 30 dB, but better for 0 and 10 dB.

In the following, we evaluate the estimators in terms of the root mean squared estimation error (RMSE) defined as

$$RMSE = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\hat{\omega}_0^{(s)} - \omega_0\right)^2}, \qquad (25)$$

with ω_0 and $\hat{\omega}_0^{(s)}$ being the true fundamental frequency and the estimate, respectively, and with S being the number of Monte Carlo trials. This is done for various PSNRs, for a given N, as well as for different N for a given PSNR. The number of Monte Carlo trials was 200. As a reference method, we use the WLS estimator proposed in [7]. This is a computationally efficient method with good statistical performance. It operates in a two-step procedure where first the unconstrained frequencies of the individual harmonics are estimated and sorted according to their value. Then, a fundamental frequency estimate is formed from these frequencies in a weighted way. It should be noted that this method requires that the model order is known, and the weighting requires that the amplitudes are either known or well estimated (see also [33]). We stress that the proposed method requires neither the amplitudes nor the order to be known. Here, to allow for the most favorable implementation of WLS, we allow it to use the true model order, the actual amplitudes as well as estimate the frequencies using ESPRIT [34]. As a result, the estimators have comparable complexity, namely $\mathcal{O}(N^3)$, due to the EVD of the covariance matrix. In general, HMUSIC will have a higher complexity than WLS, as it requires a nonlinear grid search while the fitting procedure of WLS is in closed form.

In Figure 7, the RMSEs are shown for different cases. In Figures 7(a) and 7(c), the experiments of [7] are repeated with $A_l = 1\ \forall l$. As can be seen, both WLS and HMUSIC have very good statistical performance for PSNRs above 20 dB. In Figures 7(b) and 7(d), the amplitudes are randomized according to a Rayleigh pdf. This can be seen to have an impact on the performance of both estimators. It is interesting to note that, as before, HMUSIC breaks down for PSNRs

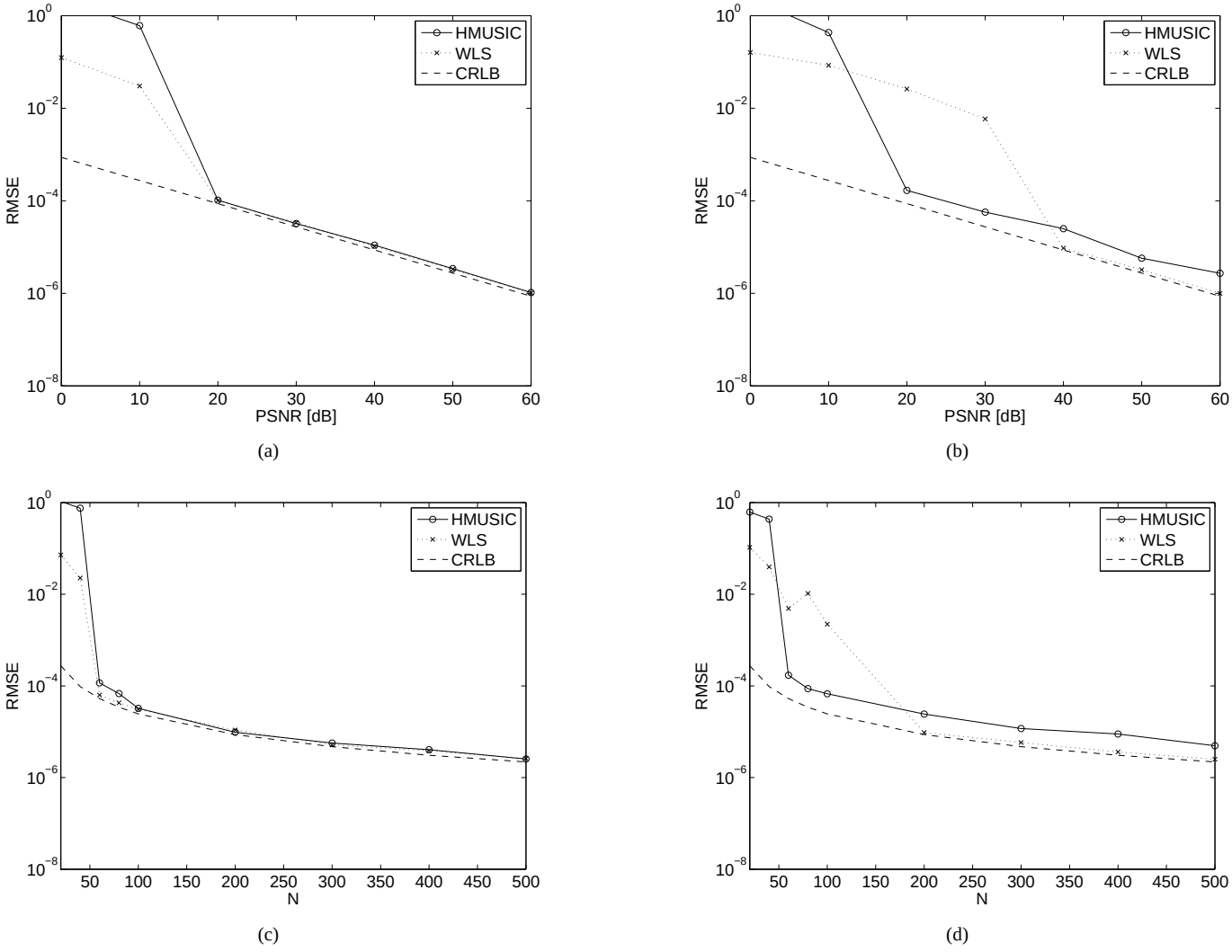

Fig. 7. (a) RMSE as a function of PSNR for $N = 200$ with constant amplitudes, (b) RMSE as a function of PSNR for $N = 200$ with randomized amplitudes, (c) RMSE as a function of N for $PSNR = 40$ dB with constant amplitudes, and (d) RMSE as a function of N for $PSNR = 40$ dB with randomized amplitudes.

below 20 dB[3]. However, WLS breaks down at PSNRs below 40 dB. Therefore, we conclude that HMUSIC is more robust than WLS. Above 40 dB, however, the performance of WLS exceeds that of HMUSIC with WLS being closer to the CRLB. A separate experiment was carried out to determine whether the gap between the CRLB and the RMSE of HMUSIC was due to a erroneous order estimates. However, simulations showed no increase in performance in RMSE for a known order. An explanation of the gap is that for low amplitudes, eigenvectors belonging to the signal subspace are likely to be interchanged with noise subspace eigenvectors although it appears from Figure 6 that the order is still estimated correctly. Also, it should be noted that the distance between the RMSE and the CRLB is due to an increased variance rather than an increased bias. The bad performance of WLS below 40 dB in Figure 7(b) can largely be attributed to erroneously estimated frequencies in the unconstrained estimator, in this case ESPRIT, due to the high probability of small amplitudes in the Rayleigh pdf. We note that the WLS fitting procedure may easily result in erroneous estimates for large errors since these may cause a wrong ordering of the frequencies. In Figure 7(c), both estimators can be seen to follow the CRLB closely as a function of the number of observations N, with WLS having slightly better performance than HMUSIC. In Figure 7(d), the trend of 7(b) is continued for different N, and HMUSIC can again be seen to be more robust.

[3]By breakdown, we mean that the RMSE of the estimator deviates from the CRLB by an order of magnitude. This kind of behavior is often seen in practical estimators and is also known as a thresholding effect. This effect is predicted by the Barankin and other bounds [35], [36], but not by the CRLB.

IV. Conclusion

In this paper, we have presented a method for high-resolution estimation of the fundamental frequency of a set of harmonically related sinusoids, assuming an unknown model order. The method, which is based on the MUSIC estimation criterion, jointly estimates the fundamental frequency and the number of harmonics. Since many estimators, such as the nonlinear least-squares method, require that the model order is

known, this is a significant advantage of the proposed method. It has been shown how the method can be implemented efficiently using FFTs and how refined estimates can be obtained by a gradient-based method. The application of the proposed method to analysis of audio signals has been illustrated with signal examples. The statistical performance of the method, in terms of the mean squared error, has been evaluated and compared to the asymptotic Cramér-Rao lower bound and the Markov-like weighted least squares (WLS) method. The simulations show that the proposed method has good statistical performance and that it is more robust to noise than the WLS method, even in the case when the latter is allowed to know the true model order and the true signal amplitudes.

APPENDIX I
ASYMPTOTIC CRAMÉR-RAO LOWER BOUND

In this appendix, we derive the asymptotic CRLB for the estimation problem considered in this paper. First, we define the model of signal signal of interest as

$$\hat{x}(n,\boldsymbol{\theta}) = \sum_{l=1}^{L} A_l e^{j(\omega_0 l n + \phi_l)} \tag{26}$$

being a function of the parameter vector

$$\boldsymbol{\theta} = [\ \omega_0\ A_1\ \phi_1\ \ \ldots A_L\ \phi_L]\,. \tag{27}$$

The variance of an unbiased estimate of the i'th parameter of $\boldsymbol{\theta}$ is then lower bounded as

$$\mathrm{var}\,(\boldsymbol{\theta}_i) \geq [\mathbf{B}(\boldsymbol{\theta})]_{ii}\,, \tag{28}$$

where $\mathbf{B}(\boldsymbol{\theta})$ is referred to as the CRLB matrix. Defining the vector

$$\hat{\mathbf{x}}(\boldsymbol{\theta}) = \left[\ \hat{x}(0,\boldsymbol{\theta}) \quad \hat{x}(1,\boldsymbol{\theta}) \quad \hat{x}(N-1,\boldsymbol{\theta})\ \right]^T, \tag{29}$$

and assuming that noise in (1) does not depend on any of the parameters in $\boldsymbol{\theta}$ as well as being Gaussian distributed with covariance matrix $\mathbf{Q}$, the exact CRLB is given by the so-called Slepian-Bangs formula (see, e.g., [37])

$$\mathbf{B}^{-1}(\boldsymbol{\theta}) = 2\,\mathrm{Re}\left\{\frac{\partial \hat{\mathbf{x}}^H(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\mathbf{Q}^{-1}\frac{\partial \hat{\mathbf{x}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right\}. \tag{30}$$

The bound can be seen to depend on the matrix

$$\frac{\partial \hat{\mathbf{x}}^H(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left[\ \tfrac{\partial \hat{x}(0,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad \ldots \quad \tfrac{\partial \hat{x}(N-1,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\ \right], \tag{31}$$

where the partial derivatives are given by

$$\frac{\partial \hat{x}(n,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \sum_{l=1}^{L} jln A_l e^{j(\omega_0 l n + \phi_l)} \\ e^{j(\omega_0 1 n + \phi_1)} \\ jA_1 e^{j(\omega_0 1 n + \phi_1)} \\ \vdots \\ e^{j(\omega_0 L n + \phi_L)} \\ jA_L e^{j(\omega_0 L n + \phi_L)} \end{bmatrix}. \tag{32}$$

In this paper, we make the assumption that the noise is also white, i.e., $\mathbf{Q}^{-1} = \frac{1}{\sigma^2}\mathbf{I}$. Inserting this into (30), the CRLB matrix can be seen to depend on the matrix $\mathbf{C}$, defined as

$$\mathbf{C} = \mathrm{Re}\left\{\frac{\partial \hat{\mathbf{x}}^H(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial \hat{\mathbf{x}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right\} \tag{33}$$

$$= \mathrm{Re}\left\{\begin{bmatrix} \boldsymbol{\chi}^H\boldsymbol{\chi} & \boldsymbol{\chi}^H\boldsymbol{\Psi}_1 & \ldots & \boldsymbol{\chi}^H\boldsymbol{\Psi}_L \\ \boldsymbol{\Psi}_1^H\boldsymbol{\chi} & \boldsymbol{\Psi}_1^H\boldsymbol{\Psi}_1 & \ldots & \boldsymbol{\Psi}_1^H\boldsymbol{\Psi}_L^H \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Psi}_L^H\boldsymbol{\chi} & \boldsymbol{\Psi}_L^H\boldsymbol{\Psi}_1 & \ldots & \boldsymbol{\Psi}_L^H\boldsymbol{\Psi}_L^H \end{bmatrix}\right\}, \tag{34}$$

where

$$\boldsymbol{\chi} = \left[\ \tfrac{\partial \hat{x}(0,\boldsymbol{\theta})}{\partial \omega_0} \quad \ldots \quad \tfrac{\partial \hat{x}(N-1,\boldsymbol{\theta})}{\partial \omega_0}\ \right]^T, \tag{35}$$

and

$$\boldsymbol{\Psi}_l = \begin{bmatrix} \frac{\partial \hat{x}(0,\boldsymbol{\theta})}{\partial A_l} & \ldots & \frac{\partial \hat{x}(N-1,\boldsymbol{\theta})}{\partial A_l} \\ \frac{\partial \hat{x}(0,\boldsymbol{\theta})}{\partial \phi_l} & \ldots & \frac{\partial \hat{x}(N-1,\boldsymbol{\theta})}{\partial \phi_l} \end{bmatrix}^T. \tag{36}$$

Then it can easily be seen that

$$\mathrm{Re}\left\{\boldsymbol{\Psi}_l^H\boldsymbol{\Psi}_l\right\} = \begin{bmatrix} N & 0 \\ 0 & A_l^2 N \end{bmatrix}. \tag{37}$$

Furthermore, assuming that ω_0 is not close to zero and that N is large, we can make the following approximations:

$$\mathrm{Re}\left\{\boldsymbol{\Psi}_l^H\boldsymbol{\Psi}_m\right\} \approx \mathbf{0} \quad \text{for} \quad l \neq m \tag{38}$$

$$\mathrm{Re}\left\{\boldsymbol{\chi}^H\boldsymbol{\chi}\right\} \approx \sum_{l=1}^{L} A_l^2 l^2 \frac{N(N-1)(2N-1)}{6} \tag{39}$$

$$\mathrm{Re}\left\{\boldsymbol{\Psi}_l^H\boldsymbol{\chi}\right\} \approx \begin{bmatrix} 0 \\ A_l^2 l \frac{N(N-1)}{2} \end{bmatrix}. \tag{40}$$

Inserting these expressions into (34), we get the following structured, sparse matrix

$$\mathbf{C} = \mathrm{Re}\left\{\begin{bmatrix} \boldsymbol{\chi}^H\boldsymbol{\chi} & \boldsymbol{\chi}^H\boldsymbol{\Psi}_1 & \boldsymbol{\chi}^H\boldsymbol{\Psi}_1 & \ldots & \boldsymbol{\chi}^H\boldsymbol{\Psi}_L \\ \boldsymbol{\Psi}_1^H\boldsymbol{\chi} & \boldsymbol{\Psi}_1^H\boldsymbol{\Psi}_1 & 0 & \ldots & 0 \\ \boldsymbol{\Psi}_2^H\boldsymbol{\chi} & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \boldsymbol{\Psi}_L^H\boldsymbol{\chi} & 0 & \ldots & 0 & \boldsymbol{\Psi}_L^H\boldsymbol{\Psi}_L^H \end{bmatrix}\right\}.$$

Next, partition the matrix $\mathbf{C}$ as follows, with $\mathbf{d}$ being a vector containing the first column, $\mathbf{E}$ a diagonal matrix, and c a scalar,

$$\mathbf{C} = \begin{bmatrix} c & \mathbf{d}^H \\ \mathbf{d} & \mathbf{E} \end{bmatrix} \tag{41}$$

then, from the matrix inversion lemma (see, e.g., [16]), we have that

$$\left[\mathbf{C}^{-1}\right]_{11} = \left(c - \mathbf{d}^H\mathbf{E}^{-1}\mathbf{d}\right)^{-1}, \tag{42}$$

yielding the asymptotic CRLB for the fundamental frequency estimation problem, i.e.,

$$[\mathbf{B}(\boldsymbol{\theta})]_{11} = \frac{\sigma^2}{2}\left[\mathbf{C}^{-1}\right]_{11} = \frac{6\sigma^2}{\sum_{l=1}^{L} A_l^2 l^2 N(N^2-1)} \tag{43}$$

$$\approx \frac{6\sigma^2}{N^3 \sum_{l=1}^{L} A_l^2 l^2}. \tag{44}$$

APPENDIX II
RELATION BETWEEN PSNR AND SNR

In this appendix, we relate the PSNR, defined in (24), to the more commonly used SNR, which for the signal model in (1), for a particular segment (with $N \gg 1$), is

$$SNR = 10\log_{10}\frac{\sum_{l=1}^{L} A_l^2}{\sigma^2}\ [\text{dB}]. \tag{45}$$

When this quantity is averaged over a number of segments, we get the so-called segmental SNR (see, e.g., [38]). Although we see from the CRLB in (23) that the estimation problem does not depend on the SNR but rather on the PSNR in a straightforward way, the PSNR and SNR can be related in some special cases. Assuming unit amplitudes, i.e., $A_l = 1\ \forall l$, we get

$$SNR = 10\log_{10}\frac{L}{\sigma^2} \tag{46}$$

$$= 10\left[\log_{10} L - 2\log_{10}\sigma\right] \tag{47}$$

whereas for the PSNR in (24), we get

$$\begin{aligned} PSNR &= 10\log_{10}\frac{\sum_{l=1}^{L} A_l^2 l^2}{\sigma^2} = 10\log_{10}\frac{\sum_{l=1}^{L} l^2}{\sigma^2} \\ &= 10\log_{10}\frac{L(L+1)(L+2)}{6\sigma^2} \\ &= 10\big[\log_{10} L + \log_{10}(L+1) \\ &\quad + \log_{10}(L+2) - \log_{10} 6 - 2\log_{10}\sigma\big]. \end{aligned} \tag{48}$$

The difference between the two definitions of the SNR can be seen to depend only on the number of harmonics L, i.e.,

$$\Delta = PSNR - SNR \tag{49}$$

$$= 10\big[\log_{10}(L+1) + \log_{10}(L+2) - \log_{10} 6\big]. \tag{50}$$

In Figure 8 this difference is shown, in dB, as a function of the number of harmonics. As can be seen, the PSNR is higher than the SNR for $L > 1$ and the difference grows larger for more harmonics. In practice, this means that for speech and single instrument audio signals with a fairly typical number of harmonics, in the range of 20–40, a PSNR of 20 dB would correspond to an SNR below 0 dB.

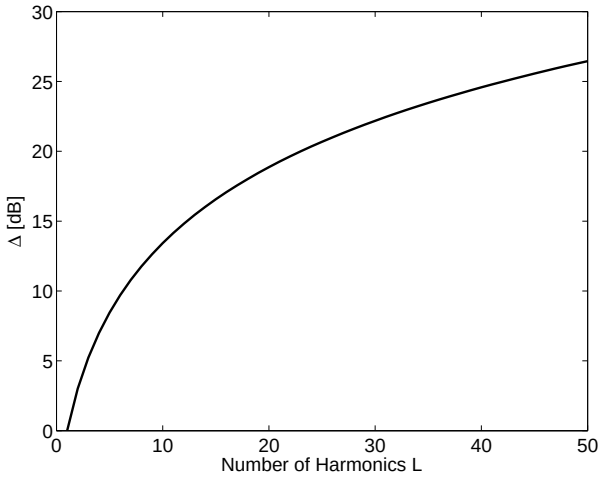

Fig. 8. The difference Δ between the two definitions of SNR in dB as a function of the number of harmonics.

REFERENCES

[1] P. Kroon and W. B. Kleijn, "Linear-prediction based analysis-by-synthesis coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., chapter 3, pp. 79–119. Elsevier Science B.V., 1995.

[2] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., chapter 4, pp. 121–174. Elsevier Science B.V., 1995.

[3] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech and Audio Processing*, vol. 10(5), pp. 293–302, July 2002.

[4] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111(4), pp. 1917–1930, Apr. 2002.

[5] W. Hess, *Pitch Determination of Speech Signals*, Springer-Verlag, Berlin, 1983.

[6] W. Hess, "Pitch and voicing determination," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sohndi, Eds., pp. 3–48. Marcel Dekker, New York, 1992.

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

[8] 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.

[9] 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.

[10] M. Davy, S. Godsill, and J. Idier, "Bayesian analysis of western tonal music," *J. Acoust. Soc. Am.*, vol. 119(4), pp. 2498–2517, Apr. 2006.

[11] S. Godsill and M. Davy, "Bayesian computational models for inharmonicity in musical instruments," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 2005, pp. 283–286.

[12] S. Godsill and M. Davy, "Bayesian harmonic models for musical pitch estimation and analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, vol. 2, pp. 1769–1772.

[13] Anssi Klapuri and Manuel Davy, Eds., *Signal Processing Methods for Music Transcription*, Springer, New York, 2006.

[14] T. D. Rossing, *The Science of Sound*, Addison-Wesley Publishing Company, 2nd edition, 1990.

[15] S. M. Kay, *Modern Spectral Estimation: Theory and Application*, Prentice-Hall, 1988.

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

[17] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE SP Mag.*, vol. 21(4), pp. 36–47, July 2004.

[18] S. L. Marple, "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Signal Processing*, vol. 47, pp. 2600–2603, Sept. 1999.

[19] M. G. Christensen, S. H. Jensen, S. V. Andersen, and A. Jakobsson, "Subspace-based fundamental frequency estimation," in *Proc. European Signal Processing Conf.*, 2004, pp. 637–640.

[20] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. 34(3), pp. 276–280, Mar. 1986.

[21] G. Bienvenu, "Influence of the spatial coherence of the background noise on high resolution passive methods," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1979, pp. 306–309.

[22] G. H. Händel, "On the history of music," *IEEE Signal Processing Magazine*, p. 13, Mar. 1999.

[23] A. Eriksson, P. Stoica, and T. Töderström, "Asymptotical analysis of MUSIC and ESPRIT frequency estimates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1993, vol. 4, pp. 556–559.

[24] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramer-Rao bound," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37(5), pp. 720–741, May 1989.

[25] A.-J. van der Veen, E. F. Deprettere, and A. L. Swindlehurst, "Subspace-based signal analysis using singular value decompostion," *Proc. IEEE*, vol. 81(9), pp. 1277–1308, Sept. 1993.

[26] H. Krim and M. Viberg, "Two decades of array signal processing research–the parametric approach," *IEEE SP Mag.*, July 1996.

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

[28] 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.
[29] J. Rissanen, "Modeling by shortest data description," *Autmatica*, vol. 14, pp. 468–478, 1978.
[30] J. Rissanen, "Stochastic complexity in statistical inquiry," *Singapore: World Scientific*, 1989.
[31] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, pp. 461–464, 1978.
[32] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Processing*, vol. 46, pp. 2726–2735, Oct. 1998.
[33] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: Survey, new results and an application," *IEEE Trans. Signal Processing*, vol. 48(2), pp. 338–352, Feb. 2000.
[34] R. Roy and T. Kailath, "ESPRIT – estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37(7), July 1989.
[35] Robert J. McAulay and Edward M. Hofstetter, "Barankin bounds on parameter estimation," *IEEE Trans. Information Theory*, vol. 17, no. 6, pp. 669–676, Nov. 1971.
[36] David C. Rife and Robert R. Boorstyn, "Single tone parameter estimation from discrete-time observations," *IEEE Trans. Information Theory*, vol. 20, no. 5, pp. 591–598, Sept. 1974.
[37] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, 1993.
[38] J. H. L. Hansen, J. G. Proakis, and J. R. Deller, Jr., *Discrete-Time Processing of Speech Signals*, Prentice-Hall, 1987.