# New Results on Perceptual Distortion Minimization and Nonlinear Least-Squares Frequency Estimation

Christensen, Mads Græsbøll; Jensen, Søren Holdt

*Published in:*
I E E E Transactions on Audio, Speech and Language Processing

*DOI (link to publication from Publisher):*
10.1109/TASL.2011.2129507

*Publication date:*
2011

*Document Version*
Accepteret manuscript, peer-review version

# New Results on Perceptual Distortion Minimization and Nonlinear Least-Squares Frequency Estimation

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

***Abstract*— In a paper in this journal, a framework was presented wherein a number of practical methods for finding the perceptually most important sinusoids in audio signals could be related using a particular perceptually motivated distortion measure and it was argued that for Gaussian noise and a large number of samples, these methods should attain the Cramér-Rao lower bound. In this correspondence, we report some new results on this subject. Specifically, we analyze the finite-sample performance of these methods in experiments and we conclude that for a high number of samples, they perform close to the Cramér-Rao lower bound. However, for a low number of observations, we demonstrate that special care must be taken in designing the perceptually motivated distortion measure if high-resolution estimates are desired. In particular, the smoothness of the frequency response of the perceptual filter that implements the distortion measure is shown to be important.**

## I. Introduction

THE problem of finding the $L$ perceptually most important sinusoids for a given audio segment occurs in speech and audio applications involving, for example, modeling and coding. If the $L$ sinusoids are found such that a perceptual distortion measure is minimized, we can claim in audio coding applications that at a given bit-rate (a given number of sinusoids, assuming that, on average, a certain number of bits is used per sinusoid using, e.g., spherical or polar quantizers [1], [2]), the best possible performance is achieved. In audio modeling where only limited computational resources are available, it is likewise desirable the allowable number of sinusoids is spent the best way possible. In [3], a framework for perceptual distortion minimization and sinusoidal frequency estimation was presented based on the distortion measure presented in [4], which can be written as the 2-norm of an error signal filter by a perceptual filter. A number of well-known methods for perceptual frequency estimation, namely the weighted matching pursuit (WMP) [5], the pre-filtering method [6], and the perceptual matching pursuit (PMP) [7] (and thus also the cyclic matching pursuit (CMP) [8]) were related to the optimal solution within this framework. These method were then shown to be equivalent under certain conditions, and it was argued, based on results in estimation theory [9], that for a high number of samples, all these methods should achieve the best possible performance in a statistical sense in the presence of Gaussian noise, i.e. they attain the Cramér-Rao lower bound (CRLB). Other interesting examples of methods that aim at extracting components based on perceptual relevance as measured in various ways include those of [10]–[16]. The interested reader may also wish to consult [17] for a detailed discussion of the usage of perceptual weighting or pre-/post-processing in signal processing.

In this correspondence, we seek to investigate whether this is actually the case in experiments and we hypothesize, based on [9], that the perceptual distortion measure should be designed such that the corresponding filter response is smooth. The experiments are carried out in a simple and tractable way by applying the estimators to the problem of finding the frequency of a sinusoid in white Gaussian noise. For this problem, the original matching pursuit (MP) [18] based on the minimization of the 2-norm is identical to the optimal nonlinear least-squares (NLS) method and, hence, the maximum likelihood estimator. The question that we seek to answer with this experiment can be paraphrased as follows. If the finite sample performance of the estimators is indeed degraded by the use of the perceptual weighting matrix, for example by a bias, then perhaps the sinusoids should first be estimated in another way and perception be taken into account in a second step. Interestingly, it was observed in [7] that the perceptual weighting in WMP could cause a bias in the frequency estimates. It should be noted that the formulation used in this paper is more general than the one originally used in [3] in the sense that we here explicitly account for oversampling in the frequency domain and the use of a non-trivial window function.

The remaining part of this paper is organized as follows: First, we briefly review the framework of [3] in Section II. Then, we discuss the issue of the smoothness of the perceptual filters and show to obtain smooth filters in a computationally efficient way in Section III, before the experimental results are presented in Section IV. Finally, we conclude on the work in Section V.

## II. Generalized Framework

Throughout the paper, we will make use of the complex notation and signals for two reasons. Firstly, because a simpler notation is obtained this way, and, secondly, because it leads to computationally more efficient algorithms. We therefore start out by calculating the $N$ so-called down-sampled discrete-time analytic signal samples $x(n)$ from $2N$ real input samples $y(n)$ (see, e.g., [19]). The analytic signal is defined as $\zeta(n) = y(n) + j\mathcal{H}\{y(n)\}$ where $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform. $x(n)$ is then obtained as $x(n) = \zeta(2n)$ for $n = 0, \ldots, N-1$. The problem of interest can now be stated as follows: Given an observed signal $\mathbf{x} = [\ x(0)\ \cdots\ x(N-1)\ ]^T$, consisting of the signal of interest and additive noise $\mathbf{e} = [\ e(0)\ \cdots\ e(N-1)\ ]^T$:

$$\mathbf{x} = \mathbf{Z}\mathbf{a} + \mathbf{e}, \tag{1}$$

with $\mathbf{a}$ being a vector containing the complex amplitudes, i.e., $\mathbf{a} = [\ a_1\ \cdots\ a_L\ ]^T$ and $\mathbf{Z}$ a matrix having complex sinusoids as columns, i.e.,

$$\mathbf{Z} = \begin{bmatrix} z_1^0 & \cdots & z_L^0 \\ z_1^1 & \cdots & z_L^1 \\ \vdots & & \vdots \\ z_1^{N-1} & \cdots & z_L^{N-1} \end{bmatrix}, \tag{2}$$

where $z_l = e^{j\omega_l}$, find the frequencies $\{\omega_l\}$. In this process, $\{a_l\}$ is also found implicitly such that a complete parametrization of the signal of interest $\mathbf{Z}\mathbf{a}$ is obtained.

The methods being studied here are those where the perceptually weighted distortion $D$ for a particular segment can be written as (see [4])

$$D = \sum_{k=0}^{K-1} P(k)|\hat{E}(k)|^2, \tag{3}$$



M. G. Christensen is with the Department of Architecture, Design and Media Technology, Aalborg University, Denmark (phone: +45 99 40 97 93, email: mgc@imi.aau.dk).

S. H. Jensen is with the Department of Electronic Systems, Aalborg University, Denmark (phone: +45 99 40 86 54, email: shj@es.aau.dk).

where $P(k)$ is a real, positive weighting function defined in the frequency domain for $k = 0, \ldots, K-1$ and $\hat{E}(k) = \sum_{n=0}^{N-1} w(n)\left[x(n) - \hat{x}(n)\right] e^{-j2\pi k/Kn}$ is the $K$ point Fourier transform of the weighted reconstruction error $\hat{e}(n) = x(n) - \hat{x}(n)$ with $\hat{x}(n)$ being an approximation of $x(n)$ and $w(n)$ a real, positive window function from which we define a diagonal matrix as $\mathbf{W} = \mathrm{diag}(w(0), \ldots, w(N-1))$. Here, $\hat{x}(n)$ is given by the sinusoidal components and will differ from $x(n)$ in two ways: firstly, due to the presence of stochastic signal components, and, secondly, due incomplete or imperfect parameter estimates. Hence, $\hat{e}(n)$ can be interpreted as an estimate of the additive noise $e(n)$ in (1).

The perceptual weighting function $P(k)$ is computed using the model proposed in [4]. Assuming that $K > N$, (3) can be written as

$$D = \left\| \mathbf{H} \begin{bmatrix} \mathbf{W}\hat{\mathbf{e}} \\ \mathbf{0} \end{bmatrix} \right\|_2^2 = \left\| \mathbf{H} \left( \begin{bmatrix} \mathbf{W}\mathbf{x} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{W}\hat{\mathbf{x}} \\ \mathbf{0} \end{bmatrix} \right) \right\|_2^2 \quad (4)$$

$$= \left\| \bar{\mathbf{H}}\mathbf{W}\left(\mathbf{x} - \hat{\mathbf{x}}\right) \right\|_2^2, \quad (5)$$

where model $\hat{\mathbf{x}} = \left[\, \hat{x}(0) \;\cdots\; \hat{x}(N-1) \,\right]^T$ and $\hat{\mathbf{e}} = \mathbf{x} - \hat{\mathbf{x}}$. The matrix $\mathbf{H}$ is a filtering matrix having a circulant structure, i.e.,

$$\mathbf{H} = \begin{bmatrix} h_0 & h_{K-1} & h_{K-2} & \cdots & h_1 \\ h_1 & h_0 & h_{K-1} & \cdots & h_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_{K-2} & h_{K-3} & \vdots & \ddots & h_{K-1} \\ h_{K-1} & h_{K-2} & h_{K-3} & \cdots & h_0 \end{bmatrix}, \quad (6)$$

with each entry being given by

$$h_n = \frac{1}{K} \sum_{k=0}^{K-1} \sqrt{P(k)} e^{j2\pi \frac{n}{K} k} \quad \text{for} \quad n = 0, \ldots, K-1. \quad (7)$$

The $K \times N$ matrix $\bar{\mathbf{H}}$ in (5) contains the $N$ first columns of $\mathbf{H}$, which results in matrix that is still circulant but not square. The perceptual distortion measure can be seen to implement a particular kind of linear transform, namely a linear filter. The eigenvectors of such a matrix are the Fourier basis vectors, and, asymptotically, sinusoids of arbitrary frequency are therefore eigenvectors of such a matrix [20]. The discrete Fourier transform (DFT) matrix is defined as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{q}_0 & \mathbf{q}_1 & \cdots & \mathbf{q}_{K-1} \end{bmatrix}, \quad (8)$$

with $\mathbf{q}_k = \left[\, q_k^0 \;\cdots\; q_k^{K-1} \,\right]^T$ and $q_k = 1/\sqrt{K} e^{-j2\pi k/K}$. Introducing the additional diagonal matrix $\mathbf{\Lambda} = \sqrt{K}\,\mathrm{diag}(\mathbf{Q}\mathbf{h})$ where $\mathbf{h} = \left[\, h_0 \;\cdots\; h_{K-1} \,\right]^T$ is the first column of $\mathbf{H}$, the eigenvalue decomposition (EVD) of this matrix can be expressed as

$$\mathbf{H} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^H. \quad (9)$$

As can be seen, the function of the perceptual weighting matrix can be interpreted as an unitary transformation followed by a weighting and an inverse unitary transformation. This is, as was argued in [3], what leads to the equivalence of a number of methods asymptotically but also in special cases for a finite number of samples.

We can now use the defined perceptual distortion measure to find the parameters of the signal of interest. More specifically, the perceptual nonlinear least-squares (PNLS) estimates of the frequencies $\{\omega_l\}_{l=1}^L$ are then the minimizers of the norm of the perceptually weighted error, i.e.,

$$\{\hat{\omega}_l\} = \arg\min_{\{\omega_l\}} \|\bar{\mathbf{H}}\mathbf{W}(\mathbf{x} - \mathbf{Z}\mathbf{a})\|_2^2. \quad (10)$$

Introducing $\mathbf{V} = \bar{\mathbf{H}}\mathbf{W}$, this can also be expressed as

$$\{\hat{\omega}_l\} = \arg\max_{\{\hat{\omega}_l\}} \mathbf{x}^H \bar{\mathbf{V}}\mathbf{Z} \left( \mathbf{Z}^H \mathbf{V}^H \mathbf{V}\mathbf{Z} \right)^{-1} \mathbf{Z}^H \mathbf{V}^H \mathbf{V}\mathbf{x}. \quad (11)$$

Additionally, the vector $\mathbf{a}$ can be estimated optimally given the frequencies $\{\omega_l\}$ as $\hat{\mathbf{a}} = \left( \mathbf{Z}^H \mathbf{V}^H \mathbf{V}\mathbf{Z} \right)^{-1} \mathbf{Z}^H \mathbf{V}^H \mathbf{V}\mathbf{x}$.

Solving (10) is, however, an intractable problem due to the nonlinear nature of the unknowns, the frequencies, and instead iterative, suboptimal estimation procedures are commonly used. We will now proceed to briefly recap a number of such methods that have been reported in the literature and operate on the distortion measure defined in (3). First, we will introduce some additional quantities, namely the residual vector at iteration $l$ as $\mathbf{r}_l = \left[\, r_l(0) \;\cdots\; r_l(N-1) \,\right]^T$ with $r_{l+1}(n) = r_l(n) - \hat{a}_l e^{j\hat{\omega}_l n}$ which is initialized as $r_1(n) = x(n)$. In the PMP [7] sinusoids are chosen iteratively one at the time as the minimizer of the perceptually weighted norm of this residual, i.e.,

$$\hat{\omega}_l = \arg\min_{\omega} \|\mathbf{V}(\mathbf{r}_l - \mathbf{z}a)\|_2^2. \quad (12)$$

with $\mathbf{z} = \left[\, e^{j\omega 0} \;\cdots\; e^{j\omega(N-1)} \,\right]^T$. This results in the following frequency estimation criterion[1]:

$$\hat{\omega}_l = \arg\max_{\omega} \frac{|\mathbf{z}^H \mathbf{V}^H \mathbf{V}\mathbf{r}_l|^2}{\|\mathbf{V}\mathbf{z}\|_2^2} \quad (13)$$

The estimates can be obtained using two FFTs per iteration. Consider now that we choose the signal model component $\mathbf{z}$ such that it is an eigenvector of the perceptual weighting matrix, or approximately so, i.e,

$$\mathbf{V}\mathbf{z} = \lambda \mathbf{z}. \quad (14)$$

For this relation to hold exactly, we require that $K = N$, i.e., $\bar{\mathbf{H}} = \mathbf{H}$ since $\bar{\mathbf{H}}$ cannot have an EVD otherwise, and that $\mathbf{W} = \mathbf{I}$ since the EVD of $\mathbf{V}$ is otherwise not given by (9) and (14) then cannot hold. One can of course use the approximation regardless; it will just not be as accurate.

As we saw earlier, the perceptual weighting matrix may be seen as a unitary transformation followed by a perceptual weighting of the individual directions. From this perspective, (14) can be interpreted as the special property of the chosen model that it is invariant to the unitary transformation of the perceptual weighting matrix, meaning that only the length of the vector changes during the transformation. Based on (14), the estimation criterion of the PMP can be simplified into the so-called pre-filtering method used, e.g., in [6]. Specifically, the signal is filtered before estimation (or, as is the case in [6], quantization), i.e.,

$$\hat{\omega}_l = \arg\min_{\omega} \|\mathbf{V}\mathbf{r}_l - \lambda \mathbf{z}a\|_2^2 = \arg\max_{\omega} \frac{\left|\mathbf{z}^H \mathbf{V}\mathbf{r}_l\right|^2}{N}. \quad (15)$$

This estimator is very simple and can be implemented efficiently using the FFT of the pre-filtered residual. Noting that the inner product can be written as $\mathbf{z}^H \mathbf{V}\mathbf{z}_l = \lambda^* \mathbf{v}^H \mathbf{r}_l$ by using the eigenvector approximation once again, we obtain

$$\hat{\omega}_l = \arg\max_{\omega} |\lambda|^2 \frac{\left|\mathbf{z}^H \mathbf{r}_l\right|^2}{N}. \quad (16)$$

The eigenvalue $\lambda$, which depends on $\omega$, can now be seen to implement a simple, frequency-dependent weighting, and the estimation criterion in (16) is therefore identical to that of WMP [5].

In conclusion, we can now make some interesting observations by comparing the optimal estimator (10) with the iterative, suboptimal approximations in (13), (15), and (16):

- For a distinct set of frequencies and a large number of samples, the estimators can be expected to yield similar results since the interactions between the individual components will become smaller as $N$ grows.

[1] We here ignore the amplitude estimates since these can be found from the same inner products that are used in the frequency estimates and are not the subject of the present study.

- Therefore, one would expect that the estimates will differ when $N$ is small or the sinusoids are not well-separated in frequencies. This happens, for example, for complicated mixtures of signals with many harmonics, as is commonly the case in music signals.
- Similarly, it can be expected that the difference also will depend on the number of sinusoids that are to be extracted.

## III. Perceptual Filter Smoothing

It is well-known that for the problem of finding the frequency of a single sinusoid in white Gaussian noise, the maximum likelihood estimator and the nonlinear least-squares method are equivalent. However, the question is whether the use of a perceptual distortion measure will affect the estimators in a statistical sense. In [9], it was shown that the nonlinear least-squares estimator is still asymptotically efficient when the noise is colored provided that the noise power spectral density is smooth around the frequencies of the sinusoids, i.e., that the corresponding pre-whitening filter is smooth. Thus, it was argued in [3], for a high number of samples, the performance of perceptual frequency estimators should be unaffected by the coloring effect of the perceptual weighting matrix if the perceptual weighting filter is smooth. For more details and results on these matters, we refer the interested reader to the paper [3].

Masking curves, and thus the corresponding perceptual filters considered here, are typically calculated in the frequency domain for a discrete set of frequency points (see, for example, the ISO 11172-3 (MPEG-1) Psycho-acoustic Model 1 described in [21]). This is also the way that the model proposed in [4] is implemented. From filter design theory, it is well-known that direct manipulation of the coefficients of a discrete Fourier transform is a problematic, yet tempting, way of designing digital filters. It is sometimes referred to as the frequency sampling design method. The resulting filters may have some poor characteristics such as excessive ripples in between frequency points. Another related issue is that masking curves may exhibit rather extreme dynamics, i.e., there may be huge differences between valleys and peaks. The threshold in quiet, for example, exhibits a difference of about 80 dB. Since we are here concerned with high-resolution estimation of model parameters, it is quite possible that such filters may ruin the performance, and we here hypothesize that this is actually the case. Intuitively it also makes sense. A sharp dip in the perceptual filter near the true frequency of a sinusoid can be suspected to cause a bias in the estimates, at least one can easily see that this would be the case for the weighted matching pursuit and the pre-filtering method.

In obtaining smooth perceptual filters, we use the following method. We seek to design an FIR filter having coefficient vector $\mathbf{h}$, the first column of the matrix $\mathbf{H}$ in (6), of length $G$ and we here require that the filter length is less than the length $K$ of the original filter. The frequency response of the filter should be as close as possible to some desired response, in our case the square root of the weighting function $P(k)$, evaluated at some frequency points, here $2\pi\frac{k}{K}$ for $k = 0, \ldots, K-1$. This leads to a vector containing the desired response, defined as

$$\mathbf{p} = \left[ \begin{array}{ccc} \sqrt{P(0)} & \cdots & \sqrt{P(K-1)} \end{array} \right]^T. \tag{17}$$

First, we define the following $K \times G$ matrix implementing the Fourier transform:

$$\mathbf{F} = \left[ \begin{array}{ccc} e^{-j2\pi\frac{0}{K}0} & \cdots & e^{-j2\pi\frac{0}{K}(G-1)} \\ \vdots & & \vdots \\ e^{-j2\pi\frac{K-1}{K}0} & \cdots & e^{-j2\pi\frac{K-1}{K}(G-1)} \end{array} \right]. \tag{18}$$

Now we can write the frequency response of the filter $\mathbf{h}$ as $\mathbf{Fh}$, i.e., as the Fourier transform of $h_n$ and we may state our filter design problem as the following approximation problem

$$\mathbf{Fh} \approx \mathbf{p}. \tag{19}$$

The next question is in what sense the frequency response of the designed filter should match the desired response. Here, we use the 2-norm since it leads to a computationally simple algorithm that has a closed-form solution, namely

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|\mathbf{Fh} - \mathbf{p}\|_2^2 \tag{20}$$

$$= \left(\mathbf{F}^H\mathbf{F}\right)^{-1}\mathbf{F}^H\mathbf{p}. \tag{21}$$

We note that the pseudo-inverse $\mathbf{F}^\dagger = \left(\mathbf{F}^H\mathbf{F}\right)^{-1}\mathbf{F}^H$ can be pre-calculated and that the smoothing method therefore reduces to the calculation of the matrix-vector product $\mathbf{F}^\dagger\mathbf{p}$ for each segment of audio. It can be seen that (21) reduces to the frequency sampling method for $G = K$.

Note that we do not claim this approximation to be the best possible. It is, however, a practical method that can easily be applied in complexity sensitive applications. We here use it mainly to show that the smoothness of the filters matters.

## IV. Experimental Results

In this section, we will examine the finite sample performance of the various estimators for a single sinusoid in the presence of white Gaussian noise and compare to the CRLB. For the estimation problem considered here, the CRLB, i.e., the lower bound for the variance of an unbiased estimator, can for large $N$ and a distinct set of frequencies $\{\omega_l\}$ be shown to be (see, e.g., [22]),

$$CRLB = \frac{6\sigma^2}{A_l^2 N^3}, \tag{22}$$

where $\sigma^2$ is the variance of the Gaussian noise $e(n)$. The CRLB can be seen to depend on the signal-to-noise ratio (SNR) $A_l^2/\sigma^2$. If, as argued in [3], the performance of the estimators is unaffected by the use of the perceptual distortion measure, the performance is expected to asymptotically, i.e., for large $N$, be the same both with and without the perceptual distortion measure. It may very well be, however, that for a low number of samples, the various estimator exhibit different performance. At this point, it should also be noted that the underlying CRLB does not change by a linear transformation, such as, in our case, a filter, although the matter is here actually somewhat complicated as the filter depends on the signal parameters.

The performance of the respective estimators is measured in terms of the root mean square estimation error (RMSE) which is calculated in Monte Carlo simulations. Here, 200 runs is used for each combination of $N$ and SNR, where the noise and the phases are randomized in each run. Special care must be taken in selecting the frequency and amplitude, however, since the masking curve and thus also the perceptual filter depend on these parameters. For example, a sinusoid having a frequency in between two frequency points of the masking curve may not estimated as accurately as one having a frequency near the frequency points and the steepness of the filter depends on the amplitude. The frequency is chosen to be in the middle of the audible range and is chosen such that it does not fall on the sampling points of the FFT. In these experiments, we assume a sampling frequency of 44.1 kHz and use a rectangular window. With respect to the amplitude of the sinusoids, we investigate two cases. One where the amplitude is small such that the masking curve is relatively flat, i.e., dominated by the threshold in quiet, and one where the amplitude is high such that the masking curve is steep near the frequency. The difference between the two amplitudes is 30 dB. Note that the absolute values of the amplitudes matter in the use

and calibration of the perceptual model. For each realization of the signal, the perceptual filter is calculated using the model in [23] with $K = 4096$, and the various estimators are then applied to the problem of finding the frequency. For all the estimators, initial estimates are found using FFTs of size 4096 and then a gradient-based method, operating on their respective cost functions, is used for obtaining high-resolution estimates. We will conduct the experiments with and without the filter smoothing with $G = K/4$. Note that the WMP and the pre-filtering (denoted PRE) method are equivalent in this particular experimental setup.

The results, the RMSE as a function of the number of observations (the segment length) $N$, are shown in Figures 1 and 2 for a low and high amplitude, respectively, for an SNR of 0 dB, and in Figure 3, the results are shown with smoothing of the perceptual filters for the high amplitude case. A number of observations can be made from these figures. First of all, we see from Figure 1 that, as expected, the MP attains the CRLB meaning that it achieves the best possible performance. We also see, however, that there is a considerable gap for low $N$ between the CRLB and both the PMP and the WMP/PRE methods for the high amplitude case whereas for the low amplitude case shown in Figure 2, this gap is much smaller. The source of the of difference between the RMSEs observed in Figure 1 and 2 is most likely the perceptual filters, since there is no reason to believe that the amplitude of a sinusoid in itself should affect the performance as the SNR is fixed. It can also be seen that the use of the perceptual weighting appears to cause a threshold effect for a low number of samples. Such threshold effects are by no means uncommon phenomena in estimators, but it is stressed here since the MP does not appear to suffer from this at this particular SNR and the range of $N$ tested here. However, we also see that as the segment length is increased, the performance of the WMP and the PMP approach the CRLB, and the the general conclusion is, therefore, that for large $N$, the estimators appear to be efficient, despite the use of the perceptual distortion measure, and these findings support the arguments put forth in [3]. From Figure 3, we see that when the perceptual filters are replaced by smooth approximations, the RMSE of the estimators is improved, i.e. closer to the CRLB. It is likely that the poor performance observed for low $N$ in Figure 1 can be explained by the properties of the perceptual filters. With this in mind, the behavior that was observed in Figures 1 can likely be explained. Since the sinusoid becomes more localized as $N$ grows, the frequency region, for which the smoothness of the perceptual filter matters, will become smaller and the perceptual filter will thus appear more smooth.

We have observed, in additional experiments not reported in detail here, that the gap for low $N$ like that in Figure 1 depends on a number of experimental conditions, namely the amplitude of the sinusoid, and thereby the steepness of the perceptual filter around the frequency of the sinusoid, and also the SNR. If the contribution of estimation errors that are due to the observation noise exceeds the error that is presumably caused by the perceptual filter, the RMSEs approach the CRLB. Also, it should be noted that for the smooth perceptual filters, the PMP and WMP exhibit similar performance for the range of $N$ and the SNR tested here.

Next, we will illustrate the properties of the perceptual filter. In Figure 4, the desired smooth perceptual weighting function (dashed) is depicted along with the actual frequency response of the perceptual filter (solid) for a sinusoid having a frequency of 6 kHz. For visual clarity, a filter length of 256 was used here. As can be seen, the perceptual filter suffers from severe ripples in between the frequency points, in fact the ripples exceed 20 dB near the frequency of the sinusoid. We have also observed in our experiments, that the higher the amplitude of the sinusoid, and thus the steeper the filter, the more severe these ripples also appear to be. This indicates that the question of what exactly causes the degraded performance of the estimators is difficult to answer. For a particular set of realizations, and thus a

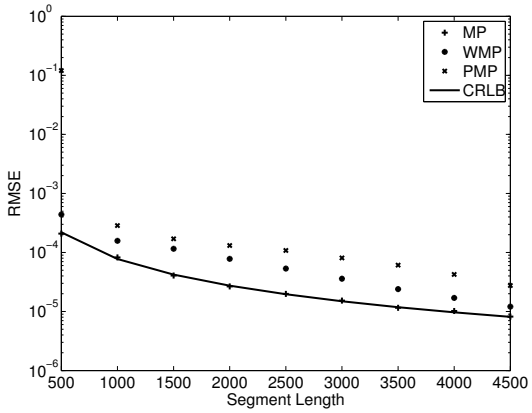


Fig. 1. RMSE of the various estimators as a function of the number of observations for a sinusoid having a high amplitude without smoothing of the perceptual filter.

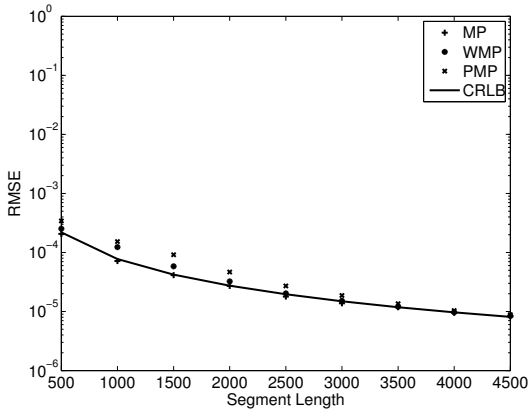


Fig. 2. RMSE of the various estimators as a function of the number of observations for a sinusoid having a low amplitude without smoothing of the perceptual filter.

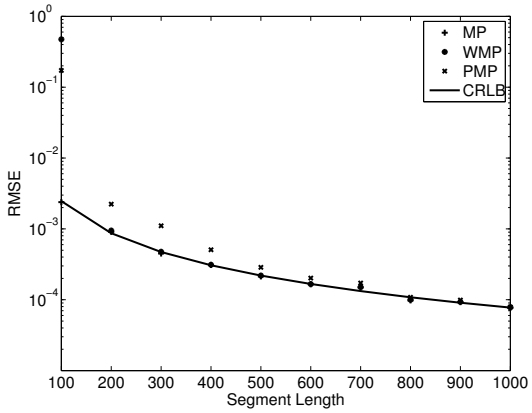


Fig. 3. RMSE of the various estimators as a function of the number of observations for a sinusoid having a high amplitude with smoothing of the perceptual filter.

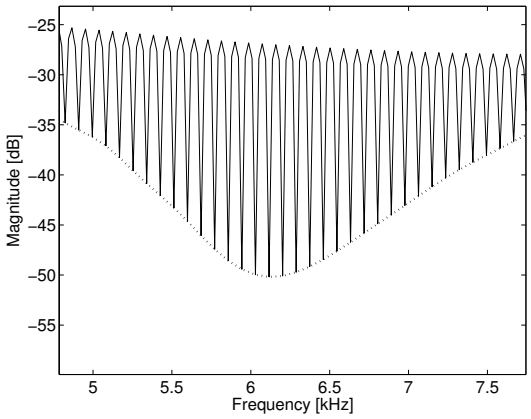


Fig. 4. Excerpt of an actual perceptual filter response (solid) and the desired smooth perceptual weighting function (dashed).

particular realization of the perceptual filter, the error that is caused by the perceptual filter is essentially a bias.

Interestingly, the properties of the perceptual filters may explain the observation made in [24] that further optimization of the parameters that were found using the FFT-based implementation of the PMP did not lead to any improvements in the perceived quality of the synthesized signal. Another possible explanation for this is of course that those differences cannot be detected by the human auditory system. However, we cannot dismiss that the estimation errors that are caused by the perceptual filter may indeed be audible, depending on the number of observations, SNR and signal parameters. In closing, it is also worth noting that the PMP was was reported in [7] to outperform WMP in terms of perceived quality evaluated in listening tests.

## V. Conclusion

In this correspondence, we have investigated the finite sample performance of a number of well-known methods for finding the perceptually most important sinusoids based on a perceptual distortion measure, and in the process, we have generalized the previously presented framework. The performance of the various methods has been investigated using Monte Carlo simulations, wherein the root mean square estimation errors of said estimators have been estimated and compared to a theoretical lower bound, the Cramér-Rao lower bound. For a high number of observations, the methods have been shown to perform very close to the Cramér-Rao lower bound, confirming the theoretical arguments made in a previous paper. Additionally, it has been shown that for a low number of observations the estimators perform far from this, meaning that they are suboptimal from an estimation point of view and that the performance can be improved by a smoothing of the perceptual filters that implement the distortion measure. These findings are important in that they show that the use of a perceptual distortion measure is not without its problems and that two step methods wherein sinusoids are first estimated where after perceptual analysis is performed to determine the perceptually most important sinusoids may in fact be preferable in some applications, for example when only a limited number of observations are available.

## References

[1] P. Korten, J. Jensen, and R. Heusdens, "High-resolution spherical quantization of sinusoidal parameters," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15(3), pp. 966–981, 2007.

[2] R. Vafin and W. B. Kleijn, "Entropy-constrained polar quantization: Theory and an application to audio coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2002, pp. 1837–1840.

[3] M. G. Christensen and S. H. Jensen, "On perceptual distortion minimization and nonlinear least-squares frequency estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14(1), pp. 99–109, Jan. 2006.

[4] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. on Applied Signal Processing*, vol. 9, pp. 1292–1304, 2005.

[5] T. S. Verma and T. H. Y. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, Mar. 1999, pp. 981–984.

[6] G. D. T. Schuller, B. Yu, D. Huang, and B. Edler, "Perceptual audio coding using adaptive pre- and post-filters and lossless compression," in *IEEE Trans. Speech and Audio Processing*, vol. 10(6), Sept. 2002, pp. 379–390.

[7] R. Heusdens, R. Vafin, and W. B. Kleijn, "Sinusoidal modeling using psychoacoustic-adaptive matching pursuits," *IEEE Signal Processing Lett.*, vol. 9(8), pp. 262–265, Aug. 2002.

[8] M. G. Christensen and S. H. Jensen, "The cyclic matching pursuit and its application to audio modeling and coding," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2007, pp. 550–554.

[9] 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*, vol. 45(8), Aug. 1997, pp. 2048–2059.

[10] H. Purnhagen, N. Meine, and B. Edler, "Sinusoidal coding using loudness-based component selection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, May 2002, pp. 1817–1820.

[11] T. Painter and A. S. Spanias, "Perceptual segmentation and component selection for sinusoidal representations of audio," *IEEE Trans. Speech and Audio Processing*, vol. 13(2), pp. 149–162, Mar. 2005.

[12] K. Vos, R. Vafin, R. Heusdens, and W. B. Kleijn, "High-quality consistent analysis-synthesis in sinusoidal coding," in *Proc. Audio Eng. Soc. 17th Conf: High Quality Audio Coding*, Sept. 1999, pp. 244–250.

[13] R. Vafin, S. V. Andersen, and W. B. Kleijn, "Exploiting time and frequency masking in consistent sinusoidal analysis-synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, June 2000, pp. 901–904.

[14] P. Vera-Candeas, N. Ruiz-Reyes, J. Cuevas-Martinez, M. Rosa-Zurera, and F. Lopez-Ferreras, "Sinusoidal modelling using perceptual matching pursuits in the bark scale for parametric audio coding," *Vision, Image and Signal Processing, IEE Proceedings*, vol. 153(4), pp. 431–435, 2006.

[15] H. Najaf-Zadeh, R. Pichevar, L. Thibault, and H. Lahdili, "Perceptual matching pursuit," in *Audio Eng. Soc. Convention*, 2008.

[16] N. R. Reyes and P. Vera Candeas, "Adaptive signal modeling based on sparse approximations for scalable parametric audio coding," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18(3), pp. 447–460, Mar. 2010.

[17] M. G. Christensen, "On perceptual distortion measures and parametric modeling," in *In Proc. Acoustics'08 Paris*, 2008, invited.

[18] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41(12), pp. 3397–3415, Dec. 1993.

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

[20] R. M. Gray, *Toeplitz and circulant matrices: A review*. Foundations and Trends in Communications and Information Theory, 2006.

[21] T. Painter and A. S. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88(4), pp. 451–515, Apr. 2000.

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

[23] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new psychoacoustical masking model for audio coding applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, May 2002, pp. 1805 – 1808.

[24] D. Kloosterman, R. Heusdens, and J. Jensen, "Estimation of sinusoidal model parameters using newton optimization and a perceptual distortion measure," in *Proc. IEEE Benelux Signal Processing Symposium*, 2004, pp. 199–202.

**Mads Græsbøll Christensen** (S'00–M'05)) was born in Copenhagen, Denmark, in March 1977. He received the M.Sc. and Ph.D. degrees from Aalborg University, Denmark, in 2002 and 2005, respectively.

He was formerly with the Department of Electronic Systems, Aalborg University, and is currently an Associate Professor in the Department of Architecture, Design and Media Technology. He has been a Visiting Researcher at Philips Research Labs, Ecole Nationale Supérieure des Télécommunications (ENST), University of California, Santa Barbara (UCSB), and Columbia University. He has published more than 75 papers in peer-reviewed conference proceedings and journals and is coauthor (with A. Jakobsson) of the book *Multi-Pitch Estimation* (Morgan & Claypool Publishers, 2009). His research interests include digital signal processing theory and methods with application to speech and audio, in particular parametric analysis, modeling, and coding.

Dr. Christensen has received several awards, namely an IEEE International Conference on Acoustics, Speech and Signal Processing Student Paper Contest Award, the Spar Nord Foundation's Research Prize for his Ph.D. dissertation, and a Danish Independent Research Councils Young Researcher's Award. He is an Associate Editor for the IEEE Signal Processing Letters.

**Søren Holdt Jensen** (S'87–M'88–SM'00) received the M.Sc. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 1988, and the Ph.D. degree in signal processing from the Technical University of Denmark, Lyngby, in 1995.

Before joining the Department of Electronic Systems, Aalborg University, he was with the Telecommunications Laboratory of Telecom Denmark, Ltd, Copenhagen, Denmark, the Electronics Institute of the Technical University of Denmark, the Scientific Computing Group of Danish Computing Center for Research and Education (UNI·C), Lyngby, the Electrical Engineering Department of Katholieke Universiteit Leuven, Leuven, Belgium, and the Center for PersonKommunikation (CPK), Aalborg University. He is Full Professor and is currently heading a research team working in the area of numerical algorithms and signal processing for speech and audio processing, image and video processing, multimedia technologies, and digital communications.

Prof. Jensen was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and Elsevier Signal Processing, and is currently Member of the Editorial Board of the EURASIP Journal on Advances in Signal Processing. He is a recipient of an European Community Marie Curie Fellowship, former Chairman of the IEEE Denmark Section, and Founder and Chairman of the IEEE Denmark Section's Signal Processing Chapter. In January 2011, he was appointed as member of the Danish Council for Independent Research—Technology and Production Sciences by the Danish Minister for Science, Technology and Innovation.