Pitch Estimation and Tracking with Harmonic Emphasis On The Acoustic Spectrum

Karimian-Azari, Sam; Mohammadiha, Nasser; Jensen, Jesper Rindom; Christensen, Mads Græsbøll

Published in:
I E E E International Conference on Acoustics, Speech and Signal Processing. Proceedings

DOI (link to publication from Publisher):
10.1109/ICASSP.2015.7178788

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

PITCH ESTIMATION AND TRACKING WITH HARMONIC EMPHASIS ON THE ACOUSTIC SPECTRUM

Sam Karimian-Azari[⋆,1]*, Nasser Mohammadiha*[†,2]*, Jesper R. Jensen*[⋆,3]*, and Mads G. Christensen*[⋆]

[⋆] Audio Analysis Lab, AD:MT, Aalborg University, email: {ska, jrj, mgc}@create.aau.dk
[†] University of Oldenburg, email: n.mohammadiha@uni-oldenburg.de

ABSTRACT

In this paper, we use unconstrained frequency estimates (UFEs) from a noisy harmonic signal and propose two methods to estimate and track the pitch over time. We assume that the UFEs are multivariate-normally-distributed random variables, and derive a maximum likelihood (ML) pitch estimator by maximizing the likelihood of the UFEs over short time-intervals. As the main contribution of this paper, we propose two state-space representations to model the pitch continuity, and, accordingly, we propose two Bayesian methods, namely a hidden Markov model and a Kalman filter. These methods are designed to optimally use the correlations in the consecutive pitch values, where the past pitch estimates are used to recursively update the prior distribution for the pitch variable. We perform experiments using synthetic data as well as a noisy speech recording, and show that the Bayesian methods provide more accurate estimates than the corresponding ML methods.

Index Terms— Harmonic signal, frequency estimate, pitch estimation, Bayesian filter, Kalman filter

1. INTRODUCTION

Audio signals such as recordings of voiced speech and some music instruments can be modeled as a sum of harmonics with a fundamental frequency (or pitch). In practice, these signals are recorded in the presence of noise, and thus, the clean harmonic model will be less accurate. As a result, obtaining an accurate estimate of the pitch in noisy conditions is both challenging and very important for a wide range of applications such as enhancement, separation, and compression. Different pitch estimation methods have been investigated in [1, 2] which are based on a harmonic constraint. One common method to estimate the pitch is through the maximum likelihood (ML) framework [3]. In ML methods, consecutive pitch values are estimated independently, where obtaining a minimum-variance estimate is guaranteed [4, 5]. However, the pitch values in a sequence are usually highly correlated, which motivates the development of the Bayesian methods to optimally use the correlations. The Bayesian methods incorporate prior distributions, and can be used to derive the minimum mean square error (MMSE) estimator and the maximum a posteriori (MAP) estimator [6], e.g., [7].

State-of-the-art methods mostly track pitch estimates in a sequential process, e.g., [8–10]: first, pitch values are estimated in each time-frame, which is a sub-vector of the whole signal, and then they are smoothed, using a dynamic programming approach such as [11], without considering the noise statistics. For instance, the method in [8] uses a nonlinear smoothing method, which is a combination of median and low-pass filtering, and the method in [9] tracks pitch estimates based on a hidden Markov model (HMM). However, to obtain an optimal solution, the estimation and tracking have to be done jointly. One method that does this is proposed in [12], which operates in the time-domain and uses a HMM based system to utilize the temporal correlation. This estimator is optimal if the noise is stationary with known statistics, while it is suboptimal in the more practical scenario where the noise statistics are unknown. A simple method to improve the performance in this scenario is to update the signal and noise statistics over time using a low-pass filter with exponential forgetting factor [13].

In this paper, we use the relation between harmonics to estimate and track the pitch in a harmonic signal. Herein, we jointly estimate and track pitch incorporating both the harmonic constraints and noise characteristics. First, we analytically find an optimal ML pitch estimator in each time-frame using unconstrained frequency estimates (UFEs)[1], which are the perturbed frequencies of harmonics in Gaussian noise [20]. One of the key contributions of this work is to transfer the pitch estimation problem with the harmonic constraints into a state-space representation where the state equation is designed to model the pitch evolution. Consequently, we can use a state-of-the-art Bayesian method to estimate the pitch values. We propose a discrete state-space

This work was funded by the Villum Foundation[1], the Cluster of Excellence 1077 "Hearing4all" by the German Research Foundation (DFG)[2], and the Danish Council for Independent Research, grant ID: DFF 1337-00084[3].

[1] UFEs are multiple single-frequency tones, which are the location of peaks of spectral densities over frequency, assuming that the number of harmonics are known, e.g. using a method in [14, 15]. Different methods for estimation of the spectral density have been investigated in [16], e.g., using discrete Fourier transform (DFT), MUSIC [17], NLS [18], and Capon [19].

representation, an HMM, using which we develop a MAP estimator for the pitch. We also propose a continuous state-space, a Kalman filter (KF), which is used to obtain an MMSE estimate of the pitch. Both the HMM and KF based methods utilize the correlations and lead to recursive pitch estimates.

The rest of this paper is organized as follows: In Section 2, we present the signal model, and introduce the ML pitch estimator. For a sequence of observations, the Bayesian estimators are presented in Section 3. Then, in Section 4, some experimental results are presented. In closing, the work is concluded in Section 5.

2. PITCH ESTIMATION

2.1. Signal Model

We model a harmonic signal[2], e.g., voiced speech, as a sum of $L(n)$ sinusoids at the time instance n like

$$s(n) = \sum_{l=1}^{L(n)} \alpha_l \, e^{\,j\,(\omega_l(n)n+\varphi_l)}, \tag{1}$$

where $\omega_l(n) = l\omega_0(n)$, and α_l and φ_l are amplitude and initial phase of each sinusoid, respectively. In the signal sub-vector $\mathbf{s}(n) = [\, s(n), s(n-1), \cdots, s(n-M-1)\,]^T$, we assume that the signal parameters are approximately stationary, and collect the constrained frequencies like

$$\mathbf{\Omega}(n) = [\, \omega_1(n),\, \omega_2(n),\, \ldots, \omega_L(n)\,]^T = \mathbf{d}_L(n)\,\omega_0(n), \tag{2}$$

where the superscript ^{T} is the transpose operator, and $\mathbf{d}_L(n) = [\, 1, 2, \ldots, L(n)\,]^T$. We assume that the harmonic signal $s(n)$ is contaminated by additive Gaussian noise $v(n)$ with the variance of σ^2 and zero mean as

$$x(n) = s(n) + v(n), \tag{3}$$

i.e., $v(n) \sim \mathcal{N}(0, \sigma^2)$. If the narrowband signal-to-noise ratios (SNRs) of sinusoids are high enough, the observed signal of such harmonic model can be approximated by the angular noise $\Delta\omega_l(n)$ with a zero-mean normal distribution on each sinusoid [22] as

$$x(n) \approx \sum_{l=1}^{L(n)} \alpha_l \, e^{\,j\,(\omega_l(n)n+\Delta\omega_l(n)+\varphi_l)}. \tag{4}$$

Therefore, unconstrained frequency estimates (UFEs)—of the constrained frequencies—can be approximated as the summation of the true frequencies and an error term $\Delta\mathbf{\Omega}(n)$ that is defined as $\Delta\mathbf{\Omega}(n) = [\, \Delta\omega_1(n), \Delta\omega_2(n), \ldots, \Delta\omega_L(n)\,]^T$ [20], i.e.,

$$\begin{aligned}\hat{\mathbf{\Omega}}(n) &= [\, \hat{\omega}_1(n),\, \hat{\omega}_2(n),\, \ldots,\, \hat{\omega}_L(n)\,]^T \\ &= \mathbf{\Omega}(n) + \Delta\mathbf{\Omega}(n),\end{aligned} \tag{5}$$

[2]Here, we utilize the discrete-time analytical signal, as in [21], to simplify the notation and reduce the resulting complexity.

where $\Delta\mathbf{\Omega}(n)$ is a zero-mean multivariate-normally-distributed variable with the covariance matrix defined as

$$\mathbf{R}_{\Delta\mathbf{\Omega}}(n) = \mathrm{E}\{\Delta\mathbf{\Omega}(n)\Delta\mathbf{\Omega}^T(n)\}, \tag{6}$$

where $\Delta\mathbf{\Omega}(n) = \hat{\mathbf{\Omega}}(n) - \mathrm{E}\{\hat{\mathbf{\Omega}}(n)\}$, and $\mathrm{E}\{\cdot\}$ denotes the mathematical expectation. In white Gaussian noise, the precision matrix (inverse of the covariance matrix) is given by [20]:

$$\mathbf{R}^{-1}_{\Delta\mathbf{\Omega}}(n) = \frac{2}{\sigma^2}\,\mathrm{diag}\Big\{\, \alpha_1^2,\, \alpha_2^2,\, \ldots,\, \alpha_L^2 \,\Big\}, \tag{7}$$

where $\mathrm{diag}\{\cdot\}$ denotes the diagonal matrix formed with the vector input along its diagonal. Consequently, for the time frame $\mathbf{x}(n) = [\, x(n), x(n-1), \cdots, x(n-M-1)\,]^T$, the probability density function (PDF) of the UFEs given the unknown pitch is approximately given by a multivariate normal distribution with the constrained and non-zero mean:

$$P(\hat{\mathbf{\Omega}}(n)|\omega_0(n)) \sim \mathcal{N}(\mathbf{d}_L(n)\,\omega_0(n), \mathbf{R}_{\Delta\mathbf{\Omega}}(n)). \tag{8}$$

2.2. ML pitch estimate

Assuming that pitch is a deterministic parameter, the maximum likelihood (ML) estimator can be used to obtain an estimate for the pitch, where the log-likelihood function of the UFEs is maximized:

$$\hat{\omega}_0(n) = \arg\max_{\omega_0(n)} \log P(\hat{\mathbf{\Omega}}(n)|\omega_0(n)). \tag{9}$$

The optimal ML pitch estimator can be obtained by taking the first derivative of the likelihood function with respect to $\omega_0(n)$ and setting it to zero, and is given by

$$\hat{\omega}_0(n) = \big[\mathbf{d}_L^T(n)\mathbf{R}^{-1}_{\Delta\mathbf{\Omega}}(n)\mathbf{d}_L(n)\big]^{-1}\mathbf{d}_L^T(n)\mathbf{R}^{-1}_{\Delta\mathbf{\Omega}}(n)\hat{\mathbf{\Omega}}(n).$$

In the particular case with white Gaussian noise, the ML pitch estimator is simplified to

$$\hat{\omega}_0(n) = \frac{1}{\sum_{l=1}^{L}(l\;\alpha_l)^2}\,[\, \alpha_1^2,\, 2\alpha_2^2,\, \ldots,\, L\alpha_L^2\,]\,\hat{\mathbf{\Omega}}(n), \tag{10}$$

which is the same result as the weighted least squared (WLS) pitch estimator in [5].

3. PITCH TRACKING

In general, the ML estimator is interesting because it is the minimum-variance unbiased estimator in Gaussian noise. Using M samples of a stationary signal, the minimum variance of the ML pitch estimator is inversely proportional to M^3 [1]. Speech signals generally are not stationary, but a voiced speech signal often has an stationary pitch during a short-time frame less than 30 ms that, consequently, limits the number of samples and the variance of the obtained pitch estimate. Moreover, pitch values are usually correlated in a sequence; this a priori information can be used to minimize the estimation error, which is the aim of this section.

In the following subsections, we compute the likelihood of a given $\hat{\mathbf{\Omega}}(n)$ using (8), for which we need to compute the covariance matrix using (6). To evaluate (6), the expected value $\mathrm{E}\{\hat{\mathbf{\Omega}}(n)\}$ has to be computed first. Since the pitch is varying over time, we use an exponential moving average (EMA) method with a forgetting factor $0 < \lambda < 1$ to recursively update the time-varying mean value as:

$$\mathrm{E}\{\hat{\mathbf{\Omega}}(n)\} = \lambda\, \hat{\mathbf{\Omega}}(n) + (1-\lambda)\, \mathrm{E}\{\hat{\mathbf{\Omega}}(n-1)\}. \tag{11}$$

After computing $\mathrm{E}\{\hat{\mathbf{\Omega}}(n)\}$, we can compute $\mathbf{R}_{\Delta\mathbf{\Omega}}(n)$ using (6). For this purpose, we use an ML estimator for the covariance (from normally-distributed observations) among N estimates [6]:

$$\mathbf{R}_{\Delta\mathbf{\Omega}}(n) = \frac{1}{N} \sum_{i=n-N+1}^{n} \Delta\mathbf{\Omega}(i) \Delta\mathbf{\Omega}^T(i). \tag{12}$$

3.1. Discrete state-space: HMM

In this section, we assume that pitch is a discrete random variable and develop an HMM-based pitch estimation method to utilize the correlation between consecutive pitch values. For our problem, the hidden state corresponds to the pitch. HMM provides a simple and yet effective way to model the temporal correlations and has been widely used in speech processing [9, 23]. We discretize the interval that encloses the possible values of pitch into N_d centroids. In practice, since pitch is a continuous variable, the discretization may introduce a systematic bias in the estimation. However, this bias can be arbitrarily lowered by increasing N_d.

We use a first-order Markov model, where the state variable depends only on the one step past as:

$$P(\omega_0(n)|\omega_0(n-1), \cdots) = P(\omega_0(n)|\omega_0(n-1)), \tag{13}$$

where $P(\omega_0(n)|\omega_0(n-1))$ denotes the transition probability from $\omega_0(n-1)$ to $\omega_0(n)$, and $\sum_{\omega_0(n)} P(\omega_0(n)|\omega_0(n-1)) = 1$. By gathering all these probabilities, we obtain an $N_d \times N_d$ matrix which is usually referred to as the transition matrix. Since the neighboring pitch values are highly correlated, it is reasonable to assume that $\omega_0(n)$ is likely to be close to $\omega_0(n-1)$, and the probability of a pitch estimate far from $\omega_0(n-1)$ will be very small. In order to use this a priori information, we pre-define the transition matrix by sampling from a normal PDF. Hence, the diagonal elements of the transition matrix correspond to the maximum value of a normal PDF with the variance σ_t^2, and the neighboring values are sampled from the normal PDF in steps of one standard deviation.

In a hidden state-space model, we have a series of observations, i.e., UFEs, which indirectly relate to states, and each state has an emission distribution that is the same as the likelihood function in (8). We aim to estimate pitch (the hidden state) in a causal manner, i.e., given only the current and past observations $\{\hat{\mathbf{\Omega}}(n), \hat{\mathbf{\Omega}}(n-1), \ldots\}$. This yields a MAP estimate for pitch, and the common method to implement it is through the forward algorithm [23]:

$$\hat{\omega}_0(n) = \arg\max_{\omega_0(n)} \log P(\omega_0(n)|\hat{\mathbf{\Omega}}(n), \hat{\mathbf{\Omega}}(n-1), \cdots) \tag{14}$$

$$= \arg\max_{\omega_0(n)} \log P(\hat{\mathbf{\Omega}}(n)|\omega_0(n)) + \log P(\omega_0(n)|\hat{\mathbf{\Omega}}(n-1), \hat{\mathbf{\Omega}}(n-2), \cdots), \tag{15}$$

that maximizes the log-likelihood function plus the logarithm of the prior distribution, which appears as a regularization term. The prior distribution is recursively updated as

$$P(\omega_0(n)|\hat{\mathbf{\Omega}}(n-1), \hat{\mathbf{\Omega}}(n-2), \cdots) = \sum_{\omega_0(n-1)} P(\omega_0(n)|\omega_0(n-1)) P(\omega_0(n-1)|\hat{\mathbf{\Omega}}(n-1), \cdots). \tag{16}$$

Note that the maximization in (14) is simply choosing the maximum value in an N_d-dimensional vector.

3.2. Continuous state-space: Kalman filter (KF)

As it was discussed in Section 3.1, pitch is a continuous variable and, hence, it is theoretically preferred to model the variations of pitch using a continuous state-space representation, e.g., [24]. In this section, we develop such model, where the state-evolution equation is designed to take into account the correlation of the pitch values in the consecutive frames. For this purpose, we write the complete state-space representation as follows:

$$\omega_0(n) = \omega_0(n-1) + \delta(n),$$
$$\hat{\mathbf{\Omega}}(n) = \mathbf{d}_L(n)\, \omega_0(n) + \Delta\mathbf{\Omega}(n),$$

where $\delta(n) \sim \mathcal{N}(0, \sigma_t^2)$ and $\Delta\mathbf{\Omega}(n) \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_{\Delta\mathbf{\Omega}}(n))$ are the state and observation noise, respectively, which are assumed to be independent. Kalman filtering is a well-known method that computes the MMSE estimate of the hidden state variable in above [25], which is used here.

First, a pitch estimate is predicted using the past estimates as

$$\hat{\omega}_0(n|n-1) = \hat{\omega}_0(n-1|n-1) \tag{17}$$

where $\hat{\omega}_0(n|n-1)$ denotes the predicted estimate using the past observations until $\hat{\mathbf{\Omega}}(n-1)$, and $\hat{\omega}_0(n-1|n-1)$ denotes the updated estimate at time $n-1$ using all the past observations, including $\hat{\mathbf{\Omega}}(n-1)$. The variance of the prediction is also given by

$$\sigma_{\kappa}^2(n|n-1) = \sigma_{\kappa}^2(n-1|n-1) + \sigma_t^2, \tag{18}$$

where $\sigma_{\kappa}^2(n|n-1)$ and $\sigma_{\kappa}^2(n-1|n-1)$ denote the variance of the predicted estimate and updated estimate, respectively.

Second, the pitch estimate is updated. For this purpose, the error term (or innovation) is computed as

$$\mathbf{e}(n) = \hat{\mathbf{\Omega}}(n) - \mathbf{d}_L(n)\, \hat{\omega}_0(n|n-1). \tag{19}$$

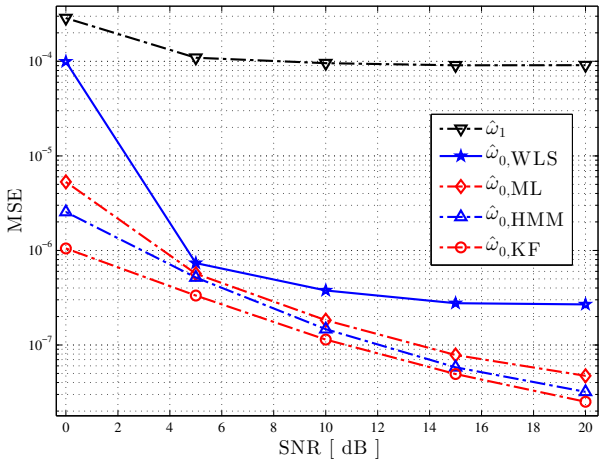

Fig. 1. Obtained MSE using the proposed methods as a function of SNR. See text for details.

Then, the predicted estimate is updated:

$$\hat{\omega}_0(n|n) = \hat{\omega}_0(n|n-1) + \mathbf{h}_{\mathrm{K}}(n)\mathbf{e}(n), \qquad (20)$$

where $\mathbf{h}_{\mathrm{K}}(n)$ denotes the Kalman gain and is given by

$$\mathbf{h}_{\mathrm{K}}(n) = \sigma^2_{\mathrm{K}}(n|n-1)\mathbf{d}^T_L(n)\left[\mathbf{\Pi}_L(n)\sigma^2_{\mathrm{K}}(n|n-1)+\mathbf{R}_{\Delta\mathbf{\Omega}}(n)\right]^{-1},$$

where $\mathbf{\Pi}_L(n) = \mathbf{d}_L(n)\mathbf{d}^T_L(n)$. The variance of the updated estimate is also recursively updated using

$$\sigma^2_{\mathrm{K}}(n|n) = \left[1 - \mathbf{h}^T_{\mathrm{K}}(n)\mathbf{d}_L(n)\right]\sigma^2_{\mathrm{K}}(n|n-1). \qquad (21)$$

4. EXPERIMENT RESULTS

We perform simulations to estimate and track the pitch in synthetic and real speech signals using the proposed methods. In the first experiment, we estimate the frequency of a sinusoid signal with the sampling frequency $f_s = 8.0$ kHz. A 65536-point discrete Fourier transform (DFT) was applied on data samples during 10 ms, i.e., $M = 80$. The forgetting factor λ in (11) was set to 0.6, and $N = 50$ observations were used to estimate the noise covariance matrix in (12). The sinusoid signal in this experiment was a linear chirp signal with $L = 5$ harmonics with random phases and identical amplitudes during 0.1 s, which was then perturbed by additive white Gaussian noise at various signal-to-noise ratios (SNRs). The starting pitch of the chirp signal was $400\pi/f_s$ and it increases with a rate of $r = 100$ Hz/s. For the HMM-based pitch estimator, the frequency range $\omega \in [150, 280] \times (2\pi/f_s)$ was discretized into $N_d = 1000$ samples. The variance related to the state transition for both HMM- and KF-based methods was set to be proportional to the linear chirp rate, i.e., $\sigma_t = \sqrt{2}\pi r/f_s^2$. Fig. 1 shows the obtained Mean Square Error (MSE), using 100 Monte-Carlo simulations for each SNR. As can be seen, the HMM- and KF-based pitch estimates have lower MSE than the corresponding ML pitch estimate, $\hat{\omega}_{0,\mathrm{ML}}$, and a state-of-the-art pitch estimator from [5], which is denoted by $\hat{\omega}_{0,\mathrm{WLS}}$. Moreover, the figure shows that the first harmonic of the UFEs (denoted by $\hat{\omega}_1$) results in significantly larger errors than all the other methods.

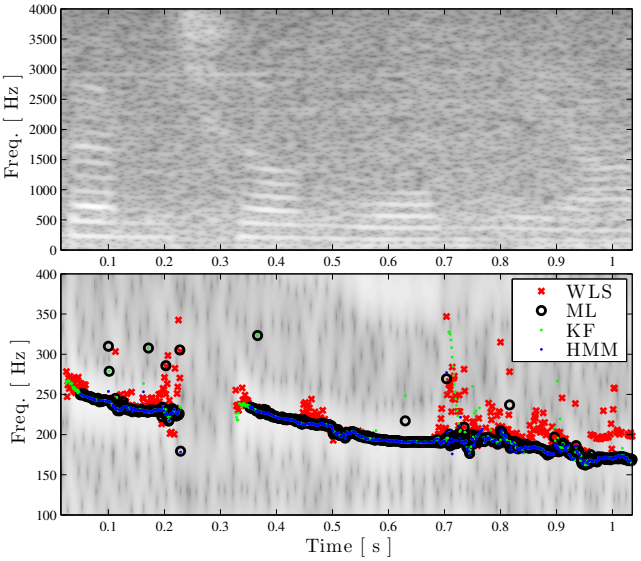

Fig. 2. Spectrogram of a speech signal in the presence of car noise at SNR = 5 dB (top), and estimated pitch values, superimposed on the spectrogram (bottom).

In the next experiment, we estimate the pitch in a speech signal degraded by car noise at SNR = 5 dB. We select voiced speech segments using the normalized low frequency energy ratio [26], and estimate the number of harmonics using the MAP order estimation [15]. A fixed $\sigma_t = 0.0318\pi/f_s$ was used for both HMM- and KF-based methods. The other parameters were set: $M = 240$, $\lambda = 0.9$, and $N = 150$, as the best choise for this experiment. Fig. 2 depicts the estimated pitch values on the spectrogram of the noisy signal. As can be observed, the HMM-based method tracks the pitch values smoothly and more accurately compared to the other methods.

5. CONCLUSION

The work presented in this paper has focused on pitch estimation. We have formulated an ML estimator for the pitch, which was then extended to utilize the correlations between consecutive pitch values to achieve higher accuracy and continuity for sequential pitch estimates. We have proposed HMM- and KF-based pitch estimation methods from the unconstrained frequency estimates, where noise characteristics were updated recursively. These characteristics make a contour over the frequency and time evolution, which were considered in the joint pitch estimation and tracking. Experimental results showed that both HMM- and KF-based methods outperform the corresponding optimal ML pitch estimator and another state-of-the-art method, based on the weighted least squares. Moreover, results using a real speech signal showed that the HMM-based method tracks the pitch more accurately and smoothly than the KF-based method.

6. REFERENCES

[1] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Process.*, vol. 5, no. 1, pp. 1–160, 2009.

[2] W. J. Hess, "Pitch and voicing determination of speech with an extension toward music signals," *Springer Handbook of Speech Processing*, pp. 181–212, 2008.

[3] D. Rife and R. Boorstyn, "Single tone parameter estimation from discrete-time observations," *IEEE Trans. Inf. Theory*, vol. 20, pp. 591–598, Sep 1974.

[4] P. Stoica, A. Jakobsson, and J. Li, "Cisoid parameter estimation in the colored noise case: asymptotic Cramér-Rao bound, maximum likelihood, and nonlinear least-squares," *IEEE Trans. Signal Process.*, vol. 45, pp. 2048–2059, Aug. 1997.

[5] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Elsevier Signal Process.*, vol. 80(9), pp. 1937–1944, 2000.

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

[7] L. Parra and U. Jain, "Approximate kalman filtering for the harmonic plus noise model," in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pp. 75–78, IEEE, 2001.

[8] L. Rabiner, M. Sambur, and C. Schmidt, "Applications of a nonlinear smoothing algorithm to speech processing," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 23, no. 6, pp. 552–557, 1975.

[9] M. Wohlmayr, M. Stark, and F. Pernkopf, "A probabilistic interaction model for multipitch tracking with factorial hidden markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 19, no. 4, pp. 799–810, 2011.

[10] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[11] H. Ney, "A dynamic programming algorithm for nonlinear smoothing," *Signal Processing*, vol. 5, pp. 163–173, March 1983.

[12] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Speech Audio Process.*, vol. 12, pp. 76 – 87, Jan. 2004.

[13] M. G. Christensen, "A method for low-delay pitch tracking and smoothing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 345–348, IEEE, 2012.

[14] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, pp. 36–47, Jul. 2004.

[15] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, pp. 2726 –2735, Oct. 1998.

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

[17] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, pp. 276–280, Mar. 1986.

[18] P. Stoica and A. Nehorai, "Statistical analysis of two non-linear least-squares estimators of sine waves parameters in the colored noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 2408–2411 vol.4, 1988.

[19] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, pp. 1408–1418, Aug. 1969.

[20] S. Karimian-Azari, J. R. Jensen, and M. G. Christensen, "Robust pitch estimation using an optimal filter on frequency estimates," in *Proc. European Signal Processing Conf.*, Sept. 2014.

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

[22] S. Tretter, "Estimating the frequency of a noisy sinusoid by linear regression (corresp.)," *IEEE Trans. Inf. Theory*, vol. 31, no. 6, pp. 832–835, 1985.

[23] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[24] C. Dubois and M. Davy, "Joint detection and tracking of time-varying harmonic components: a flexible bayesian approach," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 4, pp. 1283–1295, 2007.

[25] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Prentice Hall, 2000.

[26] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4559–4571, 2008.