EUSIPCO 2013 1569741109

ON THE INFLUENCE OF INHARMONICITIES IN MODEL-BASED SPEECH ENHANCEMENT

Sidsel Marie Nørholm, Jesper Rindom Jensen and Mads Græsbøll Christensen

Audio Analysis Lab, AD:MT, Aalborg University,
email: {smn, jrj, mgc}@create.aau.dk

ABSTRACT

In relation to speech enhancement, we study the influence of modifying the harmonic signal model for voiced speech to include small perturbations in the frequencies of the harmonics. A perturbed signal model is incorporated in the nonlinear least squares method, the Capon filter and the amplitude and phase estimation filter. Results show that it is possible to increase the performance, in terms of the signal reduction factor and the output signal-to-noise ratio, at the cost of increased complexity in the estimation of the model parameters. It is found that the perturbed signal model performs better than the harmonic signal model at input signal-to-noise ratios above approximately -10 dB, and that they are equally good below.

Index Terms— Single-channel speech enhancement, perturbed signal models, inharmonicity, parameter estimation

1. INTRODUCTION

In systems such as mobile phones, teleconferencing systems and hearing aids, noise interferes with the speech signal which has a detrimental effect on the quality of the resulting signal. Speech enhancement is therefore an important component in such systems. Speech enhancement can be performed using different approaches. A common one is filtering based on the noise statistics, e.g., using the Wiener filter. This method is very vulnerable to nonstationary noise because the problem of estimating noise statistics in the presence of speech is non-trivial [1, 2]. Another approach is to optimise filtering by assuming a model of the speech signal, as for example the harmonic signal model used in [2–6]. However, some problems arise when the harmonic signal model is used. The first is that only the voiced part of the speech signal can be modelled by a harmonic signal model. A second is due to the voiced speech being quasistationary, which means that the fundamental frequency changes over time. To minimise the effect of this, the processing is done on small segments, where the signal can be assumed periodic. A third problem is that voiced speech is not perfectly harmonic [7]. There are small perturbations in the frequencies of the harmonics and therefore they do not coincide completely with the harmonics of the assumed model. This causes unwanted distortion in the resulting speech signal when using a signal driven approach. The phenomenon of inharmonicity is well known from musical instruments, where the perturbations of the harmonics are very well defined and have to be taken into account, for example in the tuning of pianos [8]. Inharmonic models are also used in [6, 9] for fundamental frequency estimation in musical signals, but the research of the influence of inharmonicities in speech is very sparse. The inharmonicity in voiced speech is not as predictable as in musical instruments and a less restrictive model is therefore used in speech, (see e.g. [5, 7]). Inharmonicities are taken into account in the estimation of the amplitudes of the harmonics in [10], but the influence of using a perturbed signal model on the filter performance in speech enhancement has not been studied.

The purpose of this paper is, therefore, to investigate whether using a perturbed signal model will have an effect on filter performance, in terms of the signal reduction factor and the output signal-to-noise ratio (oSNR). The perturbations in synthetic signals and a set of voiced speech signals are estimated by incorporating the perturbed signal model in a nonlinear least squares (NLS) method [11] and the Capon and amplitude and phase estimation (APES) filters [12]. The estimated perturbations are then used in filtering of the signals with the APES filter [13] in order to find the gain in signal reduction factor and oSNR when compared to filtering based on the harmonic signal model.

In Section 2, the used signal model is presented along with the applied methods for estimation of the perturbations and filtering. In Section 3, the choices for the setup of experiments are explained followed by the results in Section 4, and Section 5 concludes the work.

2. METHODS

2.1. Signal model

A commonly used model of N samples of voiced speech or musical instrument recordings is given by a sum of complex

This work was funded by the Villum Foundation.

sinusoids, $s(n)$, corrupted by noise, $e(n)$, as

$$x(n) = \sum_{l=1}^{L} a_l e^{j\psi_l n} + e(n) = s(n) + e(n), \qquad (1)$$

where L is the model order. The l'th complex sinusoid has frequency ψ_l and complex amplitude $a_l = A_l e^{j\phi}$ with $A_l > 0$ and ϕ_l being the real amplitude and phase, respectively. The noise term, $e(n)$, is assumed to be zero mean and complex. Measurements of speech are real valued but can be converted to the complex representation by use of the Hilbert transform and be downsampled by a factor of two if N is sufficiently large [5].

Defining a subvector of samples $\mathbf{x}(n) = [\, x(n)\ x(n-1)\ \ldots\ x(n-M+1)\,]^T$, where $M \leq N$ and $(\cdot)^T$ denotes the transpose, the signal model can be written as

$$\mathbf{x}(n) = \mathbf{Z} \begin{bmatrix} e^{-j\psi_1 n} & & 0 \\ & \ddots & \\ 0 & & e^{-j\psi_L n} \end{bmatrix} \mathbf{a} + \mathbf{e}(n), \qquad (2)$$

where $L < M$ and $\mathbf{Z} \in \mathbb{C}^{M \times L}$ is a matrix with Vandermonde structure given by

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}(\psi_1) & \mathbf{z}(\psi_2) & \ldots & \mathbf{z}(\psi_L) \end{bmatrix}, \qquad (3)$$

$$\mathbf{z}(\psi_l) = \begin{bmatrix} 1 & e^{-j\psi_l} & \ldots & e^{-j\psi_l(M-1)} \end{bmatrix}^T, \qquad (4)$$

$\mathbf{a} = [a_1\ \ldots\ a_L]^T$ is a vector containing the complex amplitudes of the signal and $\mathbf{e}(n)$ is defined like $\mathbf{x}(n)$, but containing the noise terms $e(n)$.

Often, voiced speech is characterised using a harmonic signal model obtained by setting $\psi_l = \omega_0 l$. The harmonics are then exact multiples of the fundamental frequency, ω_0. In many musical instruments, the frequencies of the harmonics deviate slightly in a very predictable manner, leading to $\psi_l = \omega_0 l\sqrt{1 + Bl^2}$, where $B \ll 1$ is an instrument dependent stiffness parameter [5]. In speech, perturbations of the harmonics are also present, however, they are not as predictable as in music, leading to a less restrictive model for speech with [5].

$$\psi_l = \omega_0 l + \Delta_l. \qquad (5)$$

Here, the perturbations, Δ_l, are assumed to be small and evenly distributed in the interval $P_l = [-\delta_l, +\delta_l]$, where δ_l is a small and positive number. Further, it is assumed that $\psi_l < \psi_k \,\forall\, l < k$.

The considered problem can either be solved by estimating ψ_l and from this find estimates of ω_0 and Δ_l [14], or the fundamental frequency can be estimated first and thereafter Δ_l. The second approach is taken in this paper and the fundamental frequency is therefore assumed known. Further, the model order is assumed to be known as well. Both the fundamental frequency and the model order can be found, e.g., using one of the methods in [13].

2.2. Nonlinear least squares method

The maximum a posteriori estimatior, which is asymptotically optimal, will, under the assumption of white Gaussian noise and a uniform distribution of Δ_l in P_l, reduce to the NLS method [5]. NLS minimises the error between the recorded data and the signal model from (2) with $M = N$ [5]

$$\{\hat{\Delta}_l\} = \arg \min_{\mathbf{a}, \{\Delta_l \in P_l\}} \|\mathbf{x}(n) - \mathbf{Z}\mathbf{a}\|_2^2, \qquad (6)$$

with $\|\cdot\|_2$ denoting the ℓ_2-norm. Minimisation of (6) with respect to $\mathbf{a}$ followed by insertion of the result in (6) will lead to the concentratred NLS estimator of the perturbations given by [5]

$$\{\hat{\Delta}_l\} = \arg \max_{\{\Delta_l \in P_l\}} \mathbf{x}^H \mathbf{Z}(\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{x}, \qquad (7)$$

where $(\cdot)^H$ denotes the Hermitian transpose.

When the noise is colored or when several speakers are present, the NLS estimator might not be the optimal choice and therefore it is instructive to look at other estimation methods as well.

2.3. Capon filter

The Capon filter is designed to minimise the output of the filter while having unit gain at the harmonic frequencies. This minimisation problem can be expressed as [5]

$$\min_{\mathbf{h}} \mathbf{h}^H \mathbf{R}_x \mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^H \mathbf{Z} = \mathbf{1}, \qquad (8)$$

where $\mathbf{h} = [h(0)\, h(1)\ \ldots\ h(M-1)]^H$ is the filter response, $\mathbf{1} = [1\ \ldots\ 1]^T$ and $\mathbf{R}_x$ is the covariance matrix of $\mathbf{x}$ defined as

$$\mathbf{R}_x = \mathrm{E}\{\mathbf{x}(n)\mathbf{x}^H(n)\}, \qquad (9)$$

with $\mathrm{E}\{\cdot\}$ denoting statistical expectation. When $s(n)$ and $e(n)$ are uncorrelated, the covariance matrix of $\mathbf{x}$ is given by the sum of the covariance matrices of the signal, $\mathbf{R}_s$, and the noise, $\mathbf{R}_e$, i.e., $\mathbf{R}_x = \mathbf{R}_s + \mathbf{R}_e$. However, none of these are known and $\mathbf{R}_x$ has to be estimated as, e.g.,

$$\widehat{\mathbf{R}}_x = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{x}(n)\mathbf{x}^H(n). \qquad (10)$$

The filter that minimises (8) is given by [5]

$$\mathbf{h} = \mathbf{R}_x^{-1}\mathbf{Z}(\mathbf{Z}^H\mathbf{R}_x^{-1}\mathbf{Z})^{-1}\mathbf{1}. \qquad (11)$$

By maximising the output power of this filter, the perturbations can be estimated as

$$\{\hat{\Delta}_l\} = \arg \max_{\{\Delta_l \in P_l\}} \mathbf{1}^H(\mathbf{Z}^H\mathbf{R}_x^{-1}\mathbf{Z})^{-1}\mathbf{1}, \qquad (12)$$

2.4. Amplitude and phase estimation filter

The APES filter uses the same principle as the Capon filter. The only difference is that another covariance matrix is used in (8) which is estimated by subtracting from $\mathbf{R}_x$ the covariance corresponding to the part of $\mathbf{x}$ that resembles the signal model [13]

$$\widehat{\mathbf{R}}_e = \widehat{\mathbf{R}}_x - \mathbf{G}^H \mathbf{W}^{-1} \mathbf{G}, \tag{13}$$

with

$$\mathbf{G} = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{w}(n)\mathbf{x}^H(n), \tag{14}$$

$$\mathbf{W} = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{w}(n)\mathbf{w}^H(n), \tag{15}$$

where $\mathbf{w}(n) = [\, e^{j\psi_1 n} \; \ldots \; e^{j\psi_L n} \,]^T$.

The optimisation problem for the APES filter is then given by (8) with $\mathbf{R}_x$ replaced by $\widehat{\mathbf{R}}_e$ and the solutions for the optimal filter and the perturbations are given by (11) and (12) also with $\mathbf{R}_x$ replaced by $\widehat{\mathbf{R}}_e$.

2.5. Numerical optimisation

The estimation of the perturbations by means of (7) or (12) is a multidimensional, nonlinear and nontrivial problem. Direct estimation is therefore not feasible [11] and approximate solutions have been found as explained in what follows.

The perturbations are found one at a time by a grid search in the intervals P_l. An approximate position of the maximum is found at first, followed by a Fibonacci search [15] to give an increased resolution. If the cost functions in (7) and (12) for a given harmonic have no peak inside P_l, the perturbation is set to zero.

The NLS algorithm needs information about the perturbations of all harmonics in order to find the minimum distance between $\mathbf{x}(n)$ and the signal model $\mathbf{Za}$ in (6). In the first approach, denoted NLS-I, the perturbations are initialised with zeros and continuously updated with the estimated values of the perturbations. In the second approach, denoted NLS-II, the perturbations are initialised with the correct values of the perturbations and only the value of the perturbation under investigation is changed. With this second approach, the estimation of the perturbations is not influenced by errors in the frequencies of the other harmonics. Estimates based on NLS-II are therefore expected to reach the Cramér-Rao bound (CRB) and can in that case be used to bound the performance of other methods. It will of course only be possible to use NLS-II on synthetic signals where the perturbations are known. Using the Capon and APES filters for estimation, it is found that the best results are obtained using a single order filter fitted to the harmonic under investigation, compared to using a filter of order L. Therefore, first order filters have been used.

3. EXPERIMENTAL SETUP

The different ways to estimate Δ_l were evaluated through Monte Carlo simulations (MCS). A signal of the form (1) with $\{\psi_l\}$ given by (5) was generated and the performance of the different methods was evaluated by means of the mean squared error (MSE), $\frac{1}{LK}\sum_{l=1}^{L}\sum_{k=1}^{K}(\Delta_{l,k} - \hat{\Delta}_{l,k})^2$, where K is the number of MCS. The MSE was evaluated as a function of the input signal-to-noise ratio (iSNR) and the number of samples, N, and compared to the CRB for unconstrained frequency estimation [11].

The signal was generated with $L = 5$, $A_l = 1 \; \forall \; l$, random phase, fundamental frequency and perturbations in the intervals $\phi_l \in [0, 2\pi]$, $f_0 \in [150, 250]$ Hz, $\Delta_l \in [-15, 15]$ Hz, and δ_l was chosen to be 30 Hz. The Fibonacci search was performed with 14 iterations. The noise was white Gaussian with a standard deviation calculated from the desired iSNR. When N was varied, the iSNR was set to 10 dB, whereas when the iSNR was varied, N was fixed at 200. In the Capon and APES filters, the filter length was set to $\lfloor N/4 \rfloor$, with $\lfloor \cdot \rfloor$ denoting the floor operator. According to [4], this should be a good choice of filter length for both filter types. The number of MCS was $K = 500$. The importance of including perturbations in the filter design was tested by making APES filters with the estimated perturbations included and comparing them to a filter based on the harmonic assumption, $\Delta_l = 0 \, \forall \, l$. The APES filter was chosen since it was found to perform better than the Capon filter, when filtering based on already estimated frequency components is considered, which is consistent with frequency and amplitude estimation results in [12]. The performance of the filters with a perturbed and a harmonic signal model was evaluated by calculation of the signal reduction factor, $\xi_{\text{sr}}(\mathbf{h})$, and the oSNR$(\mathbf{h})$ given by [2]

$$\xi_{\text{sr}}(\mathbf{h}) = \frac{\sigma_s^2}{\sigma_{s,\text{nr}}^2} = \frac{\sigma_s^2}{\mathbf{h}^H \mathbf{R}_s \mathbf{h}}, \tag{16}$$

$$\text{oSNR}(\mathbf{h}) = \frac{\sigma_{s,\text{nr}}^2}{\sigma_{e,\text{nr}}^2} = \frac{\mathbf{h}^H \mathbf{R}_s \mathbf{h}}{\mathbf{h}^H \mathbf{R}_e \mathbf{h}}, \tag{17}$$

where σ_s and $\sigma_{s,\text{nr}}$ are the variances of the signal before and after filtering and $\sigma_{e,\text{nr}}$ is the variance of the noise after filtering. Without signal distortion, the variance of the desired signal before and after filtering is the same, and, therefore, $\xi_{\text{sr}}(\mathbf{h})$ should preferably be one. However, even though $\xi_{\text{sr}}(\mathbf{h}) = 1$, the signal can still be distorted in subbands [2]. Further, better performance after filtering requires oSNR$(\mathbf{h}) >$ iSNR.

In order to test the perturbed signal model on voiced speech, recordings from the Keele database [16] were used. Four different speakers were used, two men and two women. The speech signal was downsampled to have a sample frequency of 8 kHz and divided into four non-overlapping segments, one for each speaker. Voiced sections and uncertain voiced sections with periodicity in the laryngograph were treated as voiced speech and extracted from the speech signal. Hereafter, voiced speech segments with a length shorter

than $3N$ were discarded. In total, the performance measures were calculated for 49013 samples of voiced speech and averaged. Random white noise was added to give the desired iSNR and the performance was evaluated for the harmonic signal model and for perturbations estimated with NLS-I and Capon. Since the lowest fundamental frequency in the speech signal was 57 Hz, δ_l was set to 25 Hz.

4. EXPERIMENTAL RESULTS

The MSEs of the estimated perturbations were averaged over all harmonics and are shown in Fig. 1 as a function of N and the iSNR. NLS-II reaches the CRB for all N, whereas NLS-I and Capon follow the same course from 100 samples and up with a small but constant gap to the CRB. The APES filter does not perform well for estimation of the perturbations, as was also found in [12] in the case of fundamental frequency estimation. No method reaches the CRB at low iSNRs, but above 0 dB the tendency is the same as when N was varied. It should be kept in mind, that when no peak was found in the search interval, the perturbation was set to zero, which is seen to have an influence on the result at low iSNRs as well as for the APES filter at $N = 50$.

The performance measures according to the perturbations found in Fig. 1 are shown in Fig. 2 along with the performance of a filter based on the harmonic signal model, i.e., $\Delta_l = 0 \,\forall\, l$. NLS-I, NLS-II and Capon perform equally well and better than both APES and the harmonic signal model when the sample length is larger than 50 and the iSNR is larger than -10 dB. The similarity between the performance using NLS-I, NLS-II and Capon means that it is not crucial to use an estimation method for the perturbations that reaches the CRB. The signal distortion is clearly decreased when taking perturbations into account. When the perturbations are estimated with NLS-I, NLS-II and Capon, $\xi_{\text{sr}}(\mathbf{h})$ is very close to 0 dB independently of N and iSNR, whereas it is increasing as a function of both N and iSNR when a harmonic signal model is used. The oSNR$(\mathbf{h})$ is also increased using the perturbed signal model. When using NLS-I instead of the harmonic signal model, the gains in oSNR$(\mathbf{h})$ are 3.1 dB and 10.5 dB at iSNRs of 0 dB and 10 dB, respectively. The performance on real speech is shown in Fig. 3 as a function of the iSNR. The tendency here is the same as in the case of synthetic signals, and the perturbed signal model leads to improvements in both $\xi_{\text{sr}}(\mathbf{h})$ and oSNR$(\mathbf{h})$. The speech signal is more distorted than the synthetic signal in Fig. 2, but, nevertheless, when using NLS-I, $\xi_{\text{sr}}(\mathbf{h})$ is lowered by 2.1 dB and 3.4 dB compared to the harmonic signal model at 0 dB and 10 dB, respectively. The gain in oSNR$(\mathbf{h})$ is 2.2 dB and 3.8 dB at the same iSNRs.

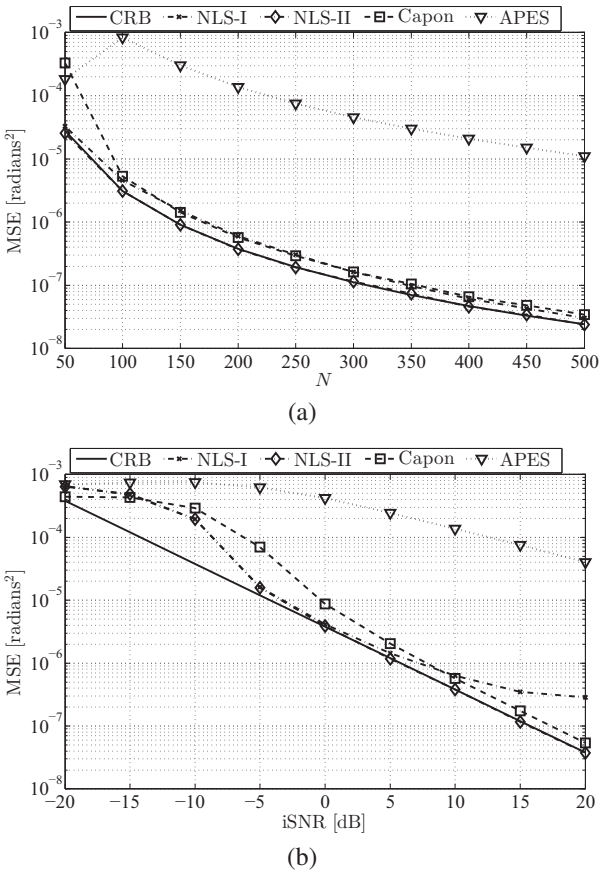


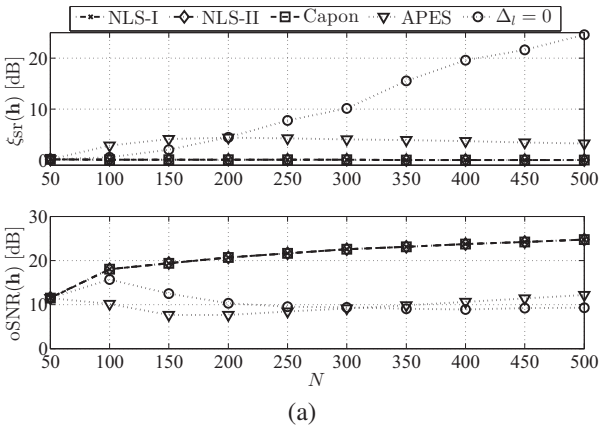
Fig. 1. Mean squared error (MSE) of the estimated perturbations as a function of (a) N and (b) iSNR.

5. CONCLUSION

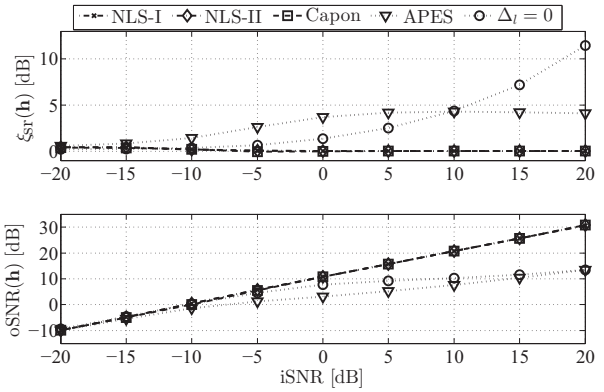
The influence of using the perturbed signal model as a basis for filtering of voiced speech signals was investigated and evaluated by means of the signal reduction factor and output signal-to-noise ratio. It was found that the performance was increased for input signal-to-noise ratios above approximately -10 dB when compared to the harmonic signal model. The perturbed and the harmonic signal models perform equally well for input signal-to-noise ratios below -10 dB. The perturbed signal model definitely has a potential of increasing the quality of the filtered speech signal, but with the perturbations found by grid searches, it comes with the cost of increased complexity in the estimation process.

6. REFERENCES

[1] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.



(a)



(b)

Fig. 2. Performance measures of synthetic signal as a function of (a) N and (b) iSNR.

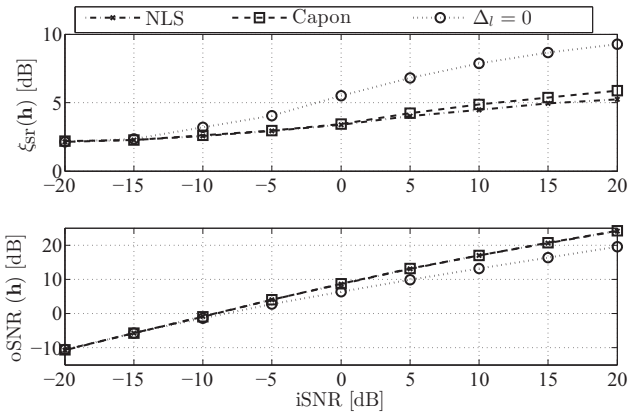


Fig. 3. Performance measures of speech signal as a function of iSNR.

[2] J. R. Jensen, *Enhancement of Periodic Signals: with Applications to Speech Signals*, Ph.D. thesis, Aalborg University, Jul. 2012.

[3] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 5, pp. 1124–1138, Oct. 1986.

[4] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: survey, new results, and an application," *IEEE Trans. on Signal Process.*, vol. 48, no. 2, pp. 338–352, Feb. 2000.

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

[6] V. Emiya, B. David, and R. Badeau, "A parametric method for pitch estimation of piano tones," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2007, vol. 1, pp. 249–252.

[7] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. on Speech and Audio Process.*, vol. 5, no. 5, pp. 389 –406, Sep. 1997.

[8] R. A. Rasch and V. Heetvelt, "String inharmonicity and piano tuning," *Music Perception: An Interdisciplinary Journal*, vol. 3, no. 2, pp. 171–189, Winter 1985.

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

[10] Y. Pantazis, O. Rosec, and Y. Stylianou, "Iterative estimation of sinusoidal signal parameters," *Signal Processing Letters, IEEE*, vol. 17, no. 5, pp. 461–464, 2010.

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

[12] A. Jakobsson and P. Stoica, "Combining Capon and APES for estimation of spectral lines," *Circuits, Systems and Signal Processing*, vol. 19, pp. 159–169, Mar. 2000.

[13] M. G. Christensen and A. Jakobsson, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Trans. on Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

[14] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Signal Process.*, vol. 80, no. 9, pp. 1937 – 1944, Sep. 2000.

[15] A. Antoniou and W. S. Lu, *Practical Optimization - Algorithms and Engineering Applications*, Springer Science+Business Media, 2007.

[16] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. Eurospeech*, Sep. 1995, pp. 837–840.