A Robust and Computationally Efficient Subspace-based Fundamental Frequency Estimator

Zhang, Johan Xi; Christensen, Mads Græsbøll; Jensen, Søren Holdt; Moonen, Marc

Published in:
IEEE Transactions on Audio Speech and Language Processing

DOI (link to publication from Publisher):
10.1109/TASL.2010.2040786

Publication date:
2010

Document Version
Accepted author manuscript, peer reviewed version

A Robust and Computationally Efficient Subspace-based Fundamental Frequency Estimator

Johan Xi Zhang, *Student Member, IEEE,* Mads Græsbøll Christensen, *Member, IEEE,* Søren Holdt Jensen, *Senior Member, IEEE,* and Marc Moonen, *Fellow, IEEE*

Abstract

In this paper, we present a method for high resolution fundamental frequency (F_0) estimation based on subspaces decomposed from a frequency-selective data model, effectively splitting the signal into a number of subbands. The resulting estimator is termed frequency-selective harmonic MUSIC (F-HMUSIC). Computational savings and robustness are expected due to the subband based approach. Additionally, a method for automatic subband signal activity detection is proposed which is based on information theoretic criterion where no subjective judgment is needed. The F-HMUSIC algorithm exhibits good statistical performance when evaluated with synthetic signals for both white and colored noise, while evaluations on real-life audio signal shows the algorithm to be competitive compared with other estimators. Finally, F-HMUSIC is concluded to be more computationally efficient and robust than other subspace based F_0 estimators while also showing robustness against recorded data with inharmonicities.

Part of this work will be presented at IEEE International Conference on Acoustics, Speech, and Signal Processing 2009.

The work of J. X. Zhang is supported by the Marie Curie EST-SIGNAL Fellowship, contract no. MEST-CT-2005-021175. The work of M. G. Christensen is supported by the Parametric Audio Processing project, Danish Research Council for Technology and Production Sciences, grant no. 274060521. J. X. Zhang, M. G. Christensen and S. H. Jensen are with Dept. of Electronic Systems (ES-MISP), Aalborg University, Aalborg, Denmark. Emails: {jxz, mgc, shj}@es.aau.dk

M. Moonen is with Dept. of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Leuven, Belgium. Email: marc.moonen@esat.kuleuven.be

A Robust and Computationally Efficient Subspace-based Fundamental Frequency Estimator

Index Terms

fundamental frequency estimation, pitch estimation, subband processing, subspace methods.

I. Introduction

The problem of estimating the fundamental frequency (F_0) or pitch in a recorded signal has been of interest to the signal processing community for many years. Many sophisticated algorithms have been proposed where the motivation for the intensive research in F_0 estimators is found in the wide usability, both inside and outside the field of engineering. The non-ideal characteristics of recorded data make the estimators especially challenging to design. For more details about F_0 properties of musical instruments we refer to, [1], [2]. In signal processing, the F_0 estimator is often a key component in speech and audio applications, such as linear prediction based speech coding, coding of speech and audio using a harmonic sinusoidal model, and musical information retrieval. Even in the field of linguistics, F_0 estimators can be applied when the analysis of tones (pitch) is an important part of understanding and classifying the language, such as for tonal languages [3], [4].

Most existing methods suffer from a degraded performance due to non-ideal characteristics of the recorded data such as low signal to noise ratio (SNR), missing partials, inharmonicity, signal transients and reverberations. Estimators are often time-domain techniques based on the auto-correlation function, cross-correlation function, averaged magnitude difference function, or average squared difference function. Other methods are mainly based on spectral extraction of the spectrogram. These methods are typically biased and primarily designed to solve the problem encountered in speech and audio applications. In most of the cases only a "rough" estimate of F_0 can be obtained. For a historical review of F_0 estimation methods, we refer to [5]–[8].

The harmonic structure of speech and audio signals can be modeled as follows: considering a set of

harmonic signals with frequencies ω_l for $l = l, ..., L$ in noise:

$$y(t) = \sum_{l=1}^{L} \beta_l e^{j\omega_l t} + e(t), \qquad \beta_l = \alpha_l e^{j\theta_l}, \tag{1}$$

where $t = 0, ..., N-1$ is the time index, L is the model order, α_l is the real-valued amplitude of l-th complex exponential, θ_l is its phase, and $e(t)$ is complex symmetric white Gaussian noise. For perfect harmonic signals, the frequencies of the harmonics are exact integer multiples of ω_0. This perfect harmonic model is not always valid. Depending on the instrument, different parametric models of the inharmonicity of the harmonics can be derived from physical models [1], [2]. A common model used for stiff-stringed instruments is $\omega_l = \omega_0 l\sqrt{1 + Bl^2}$ for $B \ll 1$ where B is normally referred to as the stiffness parameter, which is dependent on physical parameters of the string. The problem considered here is the estimation of ω_0 with or without estimation of the model order L in a time frame of N measured samples. The estimation problem associated with real valued signals can be cast as (1) by the use of analytic signals, which is valid when there is little or no spectral content of interest near 0 and π. In order to simplify our sinusoidal model as well as the algorithm, we only consider complex valued signals.

Recently, F_0 estimation algorithms based on subspace techniques have shown good estimation performance with a high accuracy in low SNR conditions, also providing flexibility for robust estimation on inharmonic signals [9]–[12], and to multi-pitch signals for known orders in [10] and for unknown orders in [12]. The main disadvantages of subspace based F_0 estimators are the high computational complexity of the subspace decomposition process, and the sensitivity to colored noise of the estimation of signal and noise subspaces.

In this paper we present an algorithm for high resolution fundamental frequency estimation based on subspaces decomposed from a frequency-selective (FS) data matrix model using inputs from a discrete Fourier transform (DFT). The resulting algorithm is termed Frequency-selective Harmonic MUSIC and represents a frequency domain extension of HMUSIC [9], [13]. F-HMUSIC adopts a subband based approach where the signal spectrum is divided into Q equally spaced subbands and where in each band an individual estimation problem is considered. This approach leads to a computationally more efficient algorithm compared to HMUSIC where the subspace decomposition is applied directly on the fullband covariance matrix, and furthermore the averaging of estimated F_0 from the different subbands is expected to lead to more robustness to colored noise. Moreover, the signal model order detection used in HMUSIC is limited to model orders $L \geq 1$, and therefore automatic signal presence detection in subbands is not possible [14]–[16]. Here, a new method for automatic signal activity detection in subbands is proposed, which is based on information theoretic criterion [17]. The main advantage of this subband detection

method is that no subjective judgment is required in the decision process. Based on this knowledge of the subband activity, additional computational savings can be achieved, e.g. by simplifications on the order estimation stage, and by estimating the noise subspace only in active subbands. For a more complete discussion on order detection, we refer the reader to [17]–[21]; for an overview of subspace based techniques, we refer to [22]–[25].

The performance of the automatic detection method is evaluated using Monte Carlo simulations where different parameters are examined. Furthermore, F-HMUSIC using this automatic detection in the subbands is evaluated on recorded musical signals [26] and its performance is compared with the performance of HMUSIC and YIN [7], [9]. Parameter selections and encountered problems during the evaluations are discussed. Additionally, the statistical properties of F-HMUSIC are evaluated using Monte Carlo simulations for synthetic signals with constant and Rayleigh distributed amplitudes in both white and colored noise, where the Rayleigh distribution is often used to model audio and speech signal amplitudes.

The remaining part of the paper is organized as follows. In Section II, the development of F-HMUSIC is introduced where the frequency-selective data matrix model is reviewed, and an automatic subband detection method is proposed. The evaluation results from both recorded and synthetic signals are presented in Section III before the conclusions are drawn in Section IV.

II. Proposed Methods

A. Frequency-Selective Data Matrix Model

The given signal sequence (1) is first Fourier transformed using an N point DFT. Let us then assume that the components of interest lie in a prespecified subband is composed of the following Fourier frequencies:

$$\left\{ \begin{array}{cccc} \frac{2\pi}{N}k_1^m & \frac{2\pi}{N}k_2^m & \dots & \frac{2\pi}{N}k_M^m \end{array} \right\}, \tag{2}$$

where m denotes the subband index of $Q = N/M$ equally divided subbands, and $\{k_1^m ... k_M^m\}$ are M given consecutive integers. The number of components L_m lying in the m-th subband specified by (2) is assumed to be $L_m \leq L$.

In the derivation of the frequency-selective model the following definitions are used:

$$z_k = e^{j\frac{2\pi}{N}k}, \qquad k = 0, 1, \ldots, N-1 \tag{3}$$

$$\mathbf{u}_k = \left[\begin{array}{cccc} z_k & \ldots & z_k^s \end{array} \right]^T \tag{4}$$

$$\mathbf{v}_k = \left[\begin{array}{cccc} 1 & z_k & \ldots & z_k^{N-1} \end{array} \right]^T \tag{5}$$

$$\mathbf{y} = \left[\begin{array}{cccc} y(0) & \ldots & y(N-1) \end{array} \right]^T \tag{6}$$

$$Y_k = \mathbf{v}_k^* \mathbf{y}, \qquad k = 0, 1, \ldots, N-1 \tag{7}$$

$$\mathbf{e} = \left[\begin{array}{cccc} e(0) & \ldots & e(N-1) \end{array} \right]^T \tag{8}$$

$$E_k = \mathbf{v}_k^* \mathbf{e}, \qquad k = 0, 1, \ldots, N-1, \tag{9}$$

where $\mathbf{u}_k$ is the so-called phase shift vector and $\mathbf{v}_k$ is the Fourier vector for z_k, $\mathbf{y}$ is the signal vector, $\mathbf{e}$ is the noise vector, $*$ is the complex conjugate, T is the vector transpose, and s is a user parameter. The choice of s will be discussed later.

Let $\{\omega_l^m\}_{l=1}^{L_m}$ denote the components of interest lying in the m-th subband. The key equation of the FS data matrix model involving the DFT sequence Y_k is proved in [25], [27], and given as:

$$\mathbf{u}_k Y_k = \left[\mathbf{a}(\omega_1^m) \ldots \mathbf{a}(\omega_{L_m}^m) \right] \left[\begin{array}{c} \beta_1 \mathbf{v}_k^* \mathbf{b}(\omega_1^m) \\ \vdots \\ \beta_{L_m} \mathbf{v}_k^* \mathbf{b}(\omega_{L_m}^m) \end{array} \right] + \mathbf{\Gamma} \mathbf{u}_k + \mathbf{u}_k E_k, \tag{10}$$

vectors $\mathbf{a}(\omega_l^m)$ and $\mathbf{b}(\omega_l^m)$ is specified as:

$$\mathbf{a}(\omega_l^m) = \left[\begin{array}{cccc} e^{j\omega_l^m} & \ldots & e^{js\omega_l^m} \end{array} \right]^T \tag{11}$$

$$\mathbf{b}(\omega_l^m) = \left[\begin{array}{cccc} 1 & e^{j\omega_l^m} & \ldots & e^{j(N-1)\omega_l^m} \end{array} \right]^T, \tag{12}$$

which express the harmonic components of the signal. Matrix $\mathbf{\Gamma} \in \mathbb{C}^{s \times s}$ is a known matrix which was defined in [25], [27].

To separate the terms corresponding to these in-band components of interest from the out-of-band components in (10), we use the notation

$$\mathbf{A}_m = \left[\begin{array}{cccc} \mathbf{a}(\omega_1^m) & \ldots & \mathbf{a}(\omega_{L_m}^m) \end{array} \right] \tag{13}$$

$$\mathbf{x}_k = \left[\begin{array}{c} \beta_1 \mathbf{v}_k^* \mathbf{b}(\omega_1^m) \\ \vdots \\ \beta_{L_m} \mathbf{v}_k^* \mathbf{b}(\omega_{L_m}^m) \end{array} \right], \tag{14}$$

for the in-band components, and similarly $\tilde{\mathbf{A}}_m$ and $\tilde{\mathbf{x}}_k$ for the components that represent leakage signals in the subband. A compact matrix form of (10) for all DFT frequencies in the m-th subband is given as:

$$\mathbf{Y}_m = \mathbf{A}_m\mathbf{X}_m + \mathbf{\Gamma}\mathbf{U}_m + \tilde{\mathbf{A}}_m\tilde{\mathbf{X}}_m + \mathbf{E}_m, \tag{15}$$

where matrices in (15) are defined as:

$$\mathbf{Y}_m = \left[\begin{array}{ccc} \mathbf{u}_{k_1^m} Y_{k_1^m} & \ldots & \mathbf{u}_{k_M^m} Y_{k_M^m} \end{array} \right] \tag{16}$$

$$\mathbf{E}_m = \left[\begin{array}{ccc} \mathbf{u}_{k_1^m} E_{k_1^m} & \ldots & \mathbf{u}_{k_M^m} E_{k_M^m} \end{array} \right] \tag{17}$$

$$\mathbf{U}_m = \left[\begin{array}{ccc} \mathbf{u}_{k_1^m} & \ldots & \mathbf{u}_{k_M^m} \end{array} \right] \tag{18}$$

$$\mathbf{X}_m = \left[\begin{array}{ccc} \mathbf{x}_{k_1^m} & \ldots & \mathbf{x}_{k_M^m} \end{array} \right], \tag{19}$$

with $\mathbf{Y}_m \in \mathbb{C}^{s\times M}$. The third and fourth terms in (15) are, respectively, the out-of-band components and the noise term. Term $\mathbf{\Gamma}\mathbf{U}_m$ in (15) is eliminated by postmultiplying (15) with a projection matrix,

$$\mathbf{\Pi}_m^{\perp} = \mathbf{I} - \mathbf{U}_m^*(\mathbf{U}_m\mathbf{U}_m^*)^{-1}\mathbf{U}_m, \tag{20}$$

which is the orthogonal projection matrix onto the null space of $\mathbf{U}_m$ which is a $s \times M$ matrix, where s is chosen such that $M > s$. The out-of-band component $\tilde{\mathbf{X}}_m$ is assumed to be zero which is asymptotically the case. The resulting expression is written as:

$$\mathbf{Y}_m\mathbf{\Pi}_m^{\perp} = \mathbf{A}_m\mathbf{X}_m\mathbf{\Pi}_m^{\perp} + \mathbf{E}_m\mathbf{\Pi}_m^{\perp}, \tag{21}$$

The matrix $\mathbf{Y}_m\mathbf{\Pi}_m^{\perp}$ obtained for m-th subband can be decomposed using either a singular value decomposition (SVD) or an eigen value decomposition (EVD), i.e. [9], [27]:

$$\mathbf{Y}_m\mathbf{\Pi}_m^{\perp} = \mathbf{H}_m\mathbf{\Lambda}_m\mathbf{V}_m. \tag{22}$$

The matrix $\mathbf{H}_m$ in (22) is written as

$$\mathbf{H}_m = \left[\begin{array}{cccc} \mathbf{h}_1 & \mathbf{h}_2 & \ldots & \mathbf{h}_s \end{array} \right], \tag{23}$$

where the columns of $\mathbf{H}_m$ contain the singular vectors defining the signal and noise subspace, and $\mathbf{\Lambda}_m$ is a diagonal matrix containing the corresponding singular values. Furthermore, let $\mathbf{S}_m$ and $\mathbf{G}_m$ be the orthonormal subspaces denoted as follows:

$$\mathbf{S}_m = \left[\begin{array}{cccc} \mathbf{h}_1 & \mathbf{h}_2 & \ldots & \mathbf{h}_{L_m} \end{array} \right] \tag{24}$$

$$\mathbf{G}_m = \left[\begin{array}{cccc} \mathbf{h}_{L_m+1} & \mathbf{h}_{L_m+2} & \ldots & \mathbf{h}_s \end{array} \right], \tag{25}$$

with $\mathbf{S}_m$ connected to the signal subspace associated with the L_m principal singular values, and $\mathbf{G}_m$ connected to the orthonormal noise subspace associated with $s - L_m$ singular values. The noise subspace spanned by $\mathbf{G}_m$ is then orthogonal to the Vandermonde matrix $\mathbf{A}_m$ defined in (13), i.e.,

$$\mathbf{A}_m^H \mathbf{G}_m = \mathbf{0}. \tag{26}$$

for frequencies ω_l^m where $l = 1, ..., L_m$.

B. F-HMUSIC

In this part, F-HMUSIC is formulated with a subband based approach for jointly estimating both F_0 and the model order L for harmonics with frequencies $\omega_l = \omega_0 l\sqrt{1 + Bl^2}$, $l = 1, ..., L$. The spectrum from 0 to π is divided into Q equally spaced subbands where the number of active subbands containing harmonics have index Q'. For simplicity, Q' is assumed to be known. In the next subsection, the proposed subband activity detection method will be described.

The harmonic model order L of (1) is given as:

$$L = \sum_{m=1}^{Q} L_m, \tag{27}$$

with L_m denoting the number of harmonics in subband m. The number of harmonics in each subband is further derived from the laws of inharmonicity written as:

$$L_m = \left\lfloor L'_m - \sum_{i=1}^{m-1} \left\lfloor L'_i \right\rfloor \right\rfloor, \tag{28}$$

where,

$$L'_m = \sqrt{-\frac{1}{2B} + \left(\left(\frac{1}{2B} \right)^2 + \left(\frac{\frac{2\pi}{N} k_M^m}{\omega_0} \right)^2 \right)^{1/2}} \tag{29}$$

is derived from $\frac{2\pi}{N} k_M^m > \omega_0 L_m \sqrt{1 + BL_m^2}$. Note, expression in (29) is valid for $B > 0$. When $B = 0$, the number of harmonics is $L'_m = \left\lfloor \frac{\frac{2\pi}{N} k_M^m}{\omega_0} \right\rfloor$. In this paper B is assumed to be known. In F_0 estimations on recorded piano notes averaged B measured from various pianos can often be used, as an example in [28] good estimation results have been shown using average measured B in estimations on recorded piano notes. If B is unknown, it can be estimated as a parameter of interests in the extended cost function [11].

The Vandermonde matrix $\mathbf{A}_m$ in (13) has been derived without consideration of the harmonic structure of the signal. If the harmonic structure is taken account, then $\mathbf{A}_m$ can be written as:

$$\mathbf{A}_m = \left[\begin{array}{ccc} \mathbf{a}(\omega_1^m) & \dots & \mathbf{a}(\omega_{L_m}^m) \end{array} \right]. \tag{30}$$

The two dimensional cost function for the joint order and fundamental frequency estimation is given as:

$$J(\omega_0, L) = \frac{1}{Q'} \sum_{m=1}^{Q'} \frac{\left\| \mathbf{A}_m^H \mathbf{G}_m \right\|_F^2}{s \min(L_m, s - L_m)}, \tag{31}$$

where the denominator is a scaling factor that makes the noise floor of the cost function invariant to the changing matrix dimensions of $\mathbf{A}_m$ and $\mathbf{G}_m$ based on the angle between subspaces [14], [15]. More specifically the measure is the average over cosine to all the non-trivial angles between the subspaces spanned by the column of $\mathbf{A}_m$ and $\mathbf{G}_m$. The estimates for the order L and the fundamental frequency ω_0 are obtained by minimizing (31),

$$\hat{\omega}_0 = \arg \min_{\omega_0 \in \Omega} \min_{L \in \mathcal{L}} J(\omega_0, L), \tag{32}$$

where Ω is the searching space for the fundamental frequency, and $\mathcal{L}$ is the search space for the order estimation.

The performance of the proposed method depends on a number parameters such as the data length N, the number of subband Q, and the user parameter s. In general, the resolution is mainly dependent on parameters s, N and Q. Increasing N leads to a resolution improvement, while increasing Q reduces the resolution. Previous experience with similar approaches show that user parameter s may be selected as large as possible to increase the number of linearly independent vectors in the noise subspace, but still less than M in order to achieve a correct estimate of the FS data matrix model [25], [27], [29].

The cost function in (31) can be computed using either an FFT based method or a gradient based method. Both methods are described in [2]. A coarser estimate is achieved when the efficient FFT based method is used. However, for applications that require accurate estimates for a given model order, a gradient search algorithm with minor modifications compared to the method described in [2] can be used.

C. Subband detection

The proposed subband activity detection method is formulated using the information theoretic criterion for model selections described in [17].It is known from [30] that for a given Toeplitz matrix $\mathbf{R}$, an asymptotically equivalent circulant $M \times M$ matrix $\mathbf{C}$ can be constructed, under the condition of $\lim_{M\to\infty} \frac{1}{\sqrt{M}} \left\| \mathbf{C} - \mathbf{R} \right\|_F = 0$, where $\left\| \cdot \right\|_F$ is the Forbenius norm and the limit is taken over the dimensions of $\mathbf{C}$ and $\mathbf{R}$. Circulant matrix $\mathbf{C}$ can then be written as $\mathbf{C} = \mathbf{Q\Gamma Q}^H$ where $\mathbf{Q}$ is the Fourier matrix. Therefore, the absolute square magnitude of DFT elements are asymptotically equal to the eigenvalues

of **R**. The DFT elements can be written as $X_k^2 = |l_k|^2 e^{j\angle X_k}$, where $|l_k|^2$ is the squared magnitudes and $e^{j\angle X_k}$ is the phase.

DFT elements in a subband are sorted by descending magnitudes, with the new sorted index denoted as k'. The sorting operation used here is similar to the sorting procedure applied on eigenvalues in EVD. The sorted magnitudes of DFT is then inserted into the cost function derived in [17], given as:

$$\begin{aligned} MDL(k') &= -\log\left(\frac{\prod_{n=k'+1}^{M}|l_n|^{2/(M-k')}}{\frac{1}{M-k'}\sum_{n=k'+1}^{M}|l_n|^2}\right)^{(M-k')N} \\ &+ \frac{1}{2}k'(2M-k')\log N, \end{aligned} \tag{33}$$

The first term in (33) is in fact the log-likelihood of the maximum likelihood estimator of the model parameters and the second term is a penalty term [17], [19]–[21]. In the proposed method only the activity of the band is of interest therefore when the minimum of (33) is $k' > 0$ the subband is decided as active.

Algorithm outline:

1) Extract the Fourier transformed segment of the specified subband m on Y_k, with index defined in (2)
2) Sort $|Y_k|$ in a descending order with the biggest magnitude first. The new sorted index is denoted as k'.
3) Insert sorted magnitudes into (33). Find the argument which gives minimum value of (33).
4) Detect subband activities using the following rules:
   $k' > 0$, subband is active.
   $k' = 0$, subband is not active.

In this paper, when subband signal detection method is used, the active subband is assumed to have full model order L. The search range for $\mathcal{L}$ is bounded by (28). Therefore the simplified cost function will be denoted as:

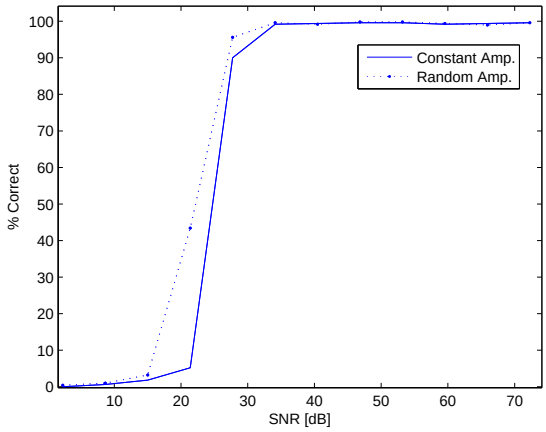
$$\hat{\omega}_0 = \arg\min_{\omega_0 \in \Omega} J(\omega_0), \tag{34}$$

where L is fixed and founded using (27).

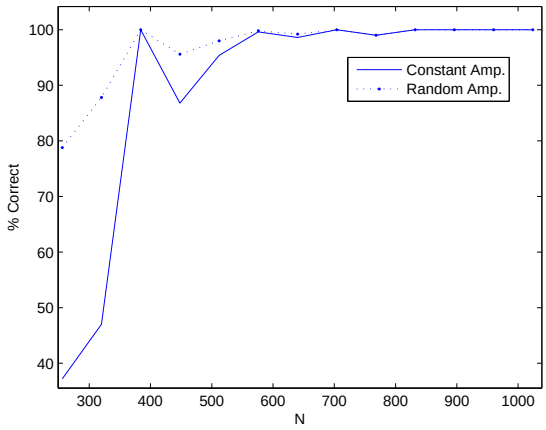
III. Experimental Results

A. Statistical Evaluation of subband detection algorithm

Before evaluating F-HMUSIC on real recorded signals the proposed subband signal detection method is evaluated with Monte Carlo simulations where errors are measured as correctness in detection. The test signal is generated according to (1) where the signal is perfectly harmonic $B = 0$ with corresponding



(a)



(b)

Fig. 1: *Percentages of correctly estimated subband activity on the proposed method evaluated on: a) different SNR under the white noise conditions b) a varying frame length N with fixed SNR=25dB.*

model number set to $L = \left\lfloor \frac{\pi}{\omega_0} \right\rfloor$. Two types of signal amplitudes are evaluated one with constant amplitudes and second type is random amplitudes generated according to Rayleigh distribution. The active subband detection errors are measured on subband with band index $m = 1$ where three different scenarios will be evaluated.

First, we start with an experiment of detection performance versus signal-to-noise ratio (SNR) where the sample length is fixed at $N = 512$. For each SNR, 500 Monte Carlo simulations are evaluated. The signals have the frequency $\omega_0 = 0.23$ and the performance is shown in Fig. 1a. From the simulations, it

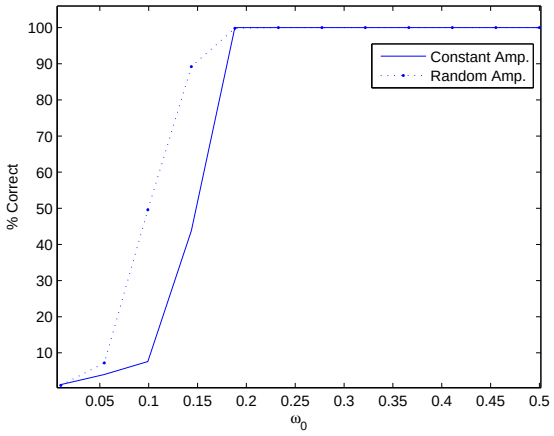


Fig. 2: *Percentages of correctly estimated subband activity on the proposed method evaluated on varying* F_0 *with fixed SNR*=30*dB and* $N = 1024$

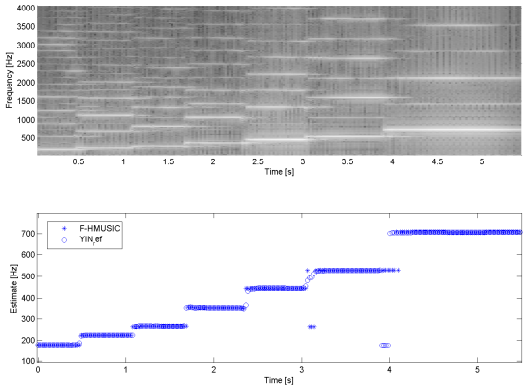


Fig. 3: *Top) Spectrogram of the clarinet signal. Bottom) Fundamental frequencies estimated using F-HMUSIC and YIN.*

can be seen that almost 100% accuracy can be achieved when the SNR is above 30dB. Next, the same signal setup is used with SNR fixed at 30dB and applied on different sample lengths evaluated from $N = 256$ to $N = 1024$. The results are shown in Fig. 1b, and the detection algorithm can be seen to be very accurate for sample length above $N = 512$. With an increased sample length, a better approximation of DFT magnitudes to the eigenvalues is achieved. Last test is to evaluate the performance when F_0 is varying from 0.01 to 0.5 with frame length fixed at $N = 1024$ and SNR at 30dB. The simulation results are shown in Fig. 2. The difficulty in this test is mainly on the lower F_0 where the harmonics are more

closely spaced than at higher F_0. This will be clarified later on.

From the simulation results, we can clearly confirm that the proposed subband detection method can sufficiently detect subband activity under different circumstances. In all cases, the performances with random generated signal amplitudes are better than constant amplitudes which can be explained by the limited sample length where DFT magnitudes are far away from equality to the asymptotically equal eigenvalues. In the case of asymptotically equality between Fourier power magnitudes and eigenvalues, every harmonic should only have one element representation in the Fourier spectrum. This is usually never the case when sample length is limited, and frequency smearing of the harmonics in the frequency domain will be obtained. The smearing effect is not crucial on white noise because perfect white noise has a flat spectrum distribution. Therefore, the proposed methods performs better on random generated amplitudes since interfering elements might be treated as noise elements when the amplitudes power is close to the noise variance.

B. Signal Examples

We start this subsection by demonstrating the proposed method on a recorded sequence of clarinet playing an up going arpeggio. The clarinet signals are assumed to be perfectly harmonic with $B = 0$. Spectrogram of the signal is shown in Fig. 3 top panel while estimates of F_0 using F-HMUSIC and YIN are shown in the bottom panel. According to Fig. 3, our algorithm can successfully estimate the fundamental frequency except on some boarder region where the signal is not well defined due to non-ideal circumstances such as reverberation in the room, that may cause a multi-pitch scenario where our model in (1) is invalid. The setup used in F-HMUSIC in this example was on a signal with sampling frequency $f_s = 11025$Hz processed with a frame length of $N = 512$, and 50% overlaps. The model order of F-HMUSIC is $s = \lfloor 0.9M \rfloor$. The method is generally sensitive to the choice of s. For short frames, large number of s is preferred. Two subbands are selected where the active subbands are automatically detected using the proposed detection algorithm. The cost function was evaluated from 100Hz to 1000Hz.

In this part, we evaluated F-HMUSIC on recordings from a database consisting of transcribed notes played by pianos [26]. The database is recorded under different reverberation environment, with three loudness levels (piano, mezzo-forte and forte). For each note, we selected a test set consisting of recordings played with six different pianos. In average, 1000 frames of data is processed for each note. The onset and offset time of the note is provided by the database, which provides a challenging test data where both metallic thumps of hammers against strings, and constantly degradation of SNR during the release of the note are involved. One example of a note spectrogram on test data at 466Hz is shown in Fig. 4,

which clearly shows the non-ideal signal conditions during onset and release of the note. This forces us to estimate under different circumstances with a general fixed parameter setup during the entire evaluation. Furthermore, the subband signal detection performance on low F_0 is more sensitive for frames with low SNR, with the statistical performance for low F_0 being shown in Fig. 2. The main intention of this evaluation is to evaluate the robustness of subspace based methods on real-recorded data. Our proposed methods will be compared with both HMUSIC[1] and YIN[2]. Previous studies of YIN have often referred it as a very robust single pitch estimator while HMUSIC has primarily shown good performance on synthetic signals and on small set of recorded signals.

During the evaluations, each estimated F_0 is quantized to the nearest note in the musical scale with A4 tuned to 440 Hz. Errors are then measured as incorrect MIDI note estimates. The evaluated signals are analyzed with a window length of $N = 1024$ and sliding forward in time with 50% overlap. For computational simplicities, signal is downsampled to $F_s = 11025$Hz. Parameters used in F-HMUSIC and HMUSIC are selected as follows: $\Omega \in [103.83, 4310]$ Hz, $Q = 2$, and $s = \lfloor 0.85M \rfloor$, $s' = \lfloor 0.60N \rfloor$ where s' is the user parameter for HMUSIC. Piano notes will be evaluated on MIDI notes 45 to 108. In this evaluation, the lowest possible F_0 is selected to be a bit higher than the lowest note that can be produced by a piano. This is because at low F_0, closely space sinusoids will give a rank deficient Vandermond matrix (30) which gives inaccurate estimations and cost function evaluated on rank deficient region will degrade the overall performance of F_0 [29]. Therefore it is important to select Ω which does not include rank deficient points in order to stabilize the overall performance. Note that F-HMUSIC use less data samples which make it more sensitive to rank deficiency problem than HMUSIC. Due to the limited data length inserted into the FS data model, s needs to be selected close to M in order to reduce the noise influence of the data. The stiffness parameters B for different F_0 is averaged from the results presented in [2, page 365]. The estimation errors evaluated on MIDI notes are reported in Fig. 5, it shows clearly that both subspace based methods suffer from degradation in estimation robustness on recorded signals. This can be explained by model mismatches where subspace based F_0 estimators do not make additional adaptation to the model changes. Model mismatch situations are most probably to be observed during onset and release of the notes.

The significantly reduced performance on higher MIDI notes can be explained with that model order L is decreasing for increased F_0. This reduce the estimation performance according to the asymptotic

[1]The HMUSIC used in evaluation on recorded data is based on fixed order where the order $L = \left\lfloor \frac{\pi}{\omega_0} \right\rfloor$

[2]Parameters used in YIN is the standard parameter found on authors webpage

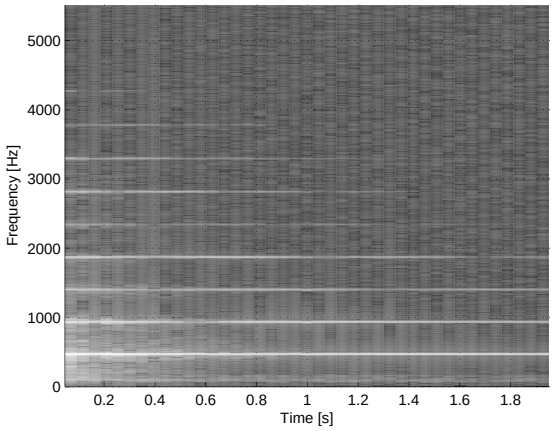


Fig. 4: *Spectrogram example of one note on MIDI note* 70 *with fundamental frequency* 466.2*Hz.*

TABLE I: *Summarized errors of MIDI notes* $[45, 95]$

MIDI Notes 45-95	F-HMUSIC	HMUSIC	YIN
$\%$ Mean Errors	6.4%	7.4%	6.2%
1/2 Octave Errors	6.3%	18.7%	9.83%
2 Octave Errors	17.6%	45.6%	27.2%

CRLB described in [9]. Regarding temporal aspects of the test signal, the reduced detection performance is related with physical properties of the piano sound where amplitudes of harmonics decay[3] rapidly for frequency above 2800Hz [2, page 384] which has the effect to significantly increase the number of frames with low SNR.

Too make the comparison fair between involved methods, errors will be discussed for MIDI notes $[45, 95]$ which in our point of view is the operating region for the involved algorithms. The errors are summarized in Table I where it shows that the performance of subspace based F_0 estimators are comparable with YIN. It also shows that HMUSIC is more sensitive to octave errors than F-HMUSIC but no significant differences between the performance have been observed. Even thought F-HMUSIC make less 2-octave and 1/2-octave errors, those errors are hard to avoid. Nevertheless, our proposed estimator does not significantly improve the robustness of F_0 estimations but high resolution of estimates can be obtained, something that is not possible using YIN or similar time-domain methods. Another

[3]The speed for amplitude decreases is often referred as decay time [2, page 384]

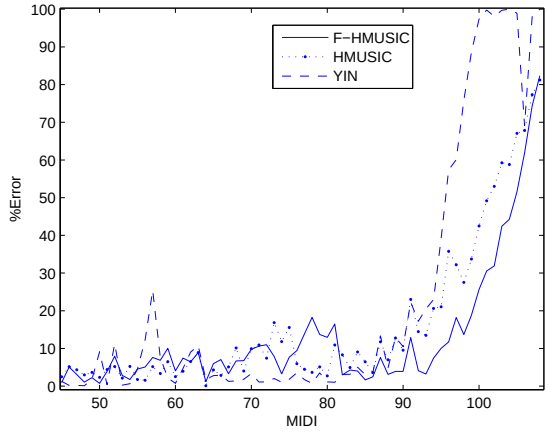


Fig. 5: *Percentage errors of the quantized MIDI notes evaluated from* $[45, 108]$.

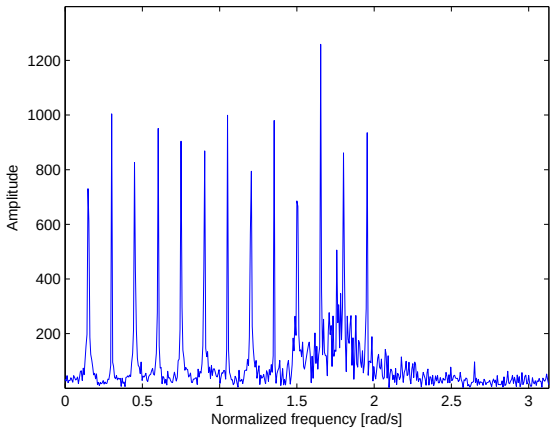
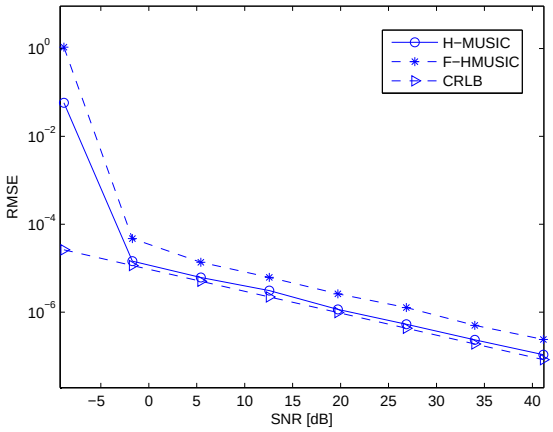


Fig. 6: *Frequency domain representation of one realization on the harmonic signal embedded in colored noise. The SNR is at* $11dB$ *with the constant amplitudes* $\alpha_l = 1 \forall l$.

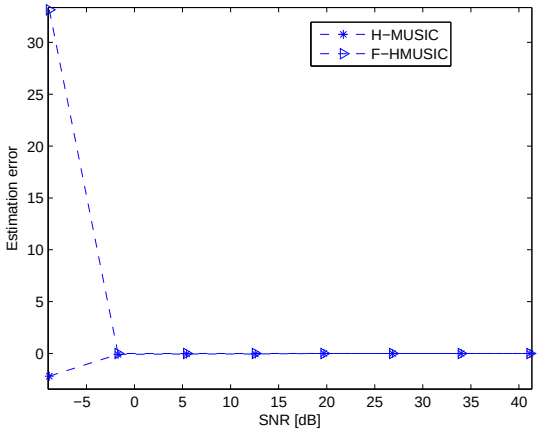
advantage of F-HMUSIC is the computational complexity is relatively low compared with other subspace based F_0 estimators. As an example, in HMUSIC the computational complexity using EVD on fullband covariance matrix is of order $\mathcal{O}\left((N)^3\right)$, and by splitting up the estimation problem into subproblems the computational load will be reduced by $\frac{1}{(2Q)^3}$ when frequency samples from region 0 to π are used.

C. Statistical Evaluation of F-HMUSIC

Next, the statistical properties of the proposed method is evaluated using the Monte Carlo simulations under both white and colored noise conditions. In this part of the evaluation only statistical properties



(a)



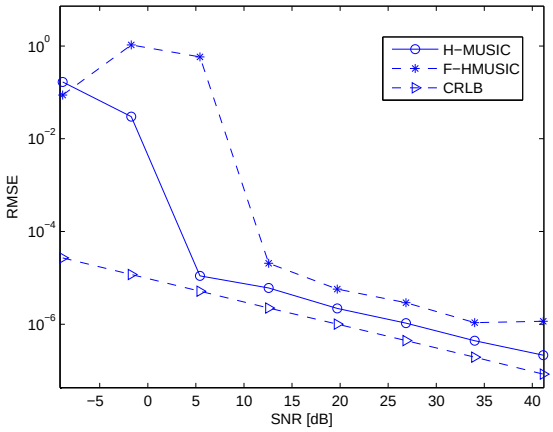
(b)

Fig. 7: *a) The RMSE with a varying SNR in the case of constant amplitudes embedded in white noise. b) Corresponding model order estimation errors*

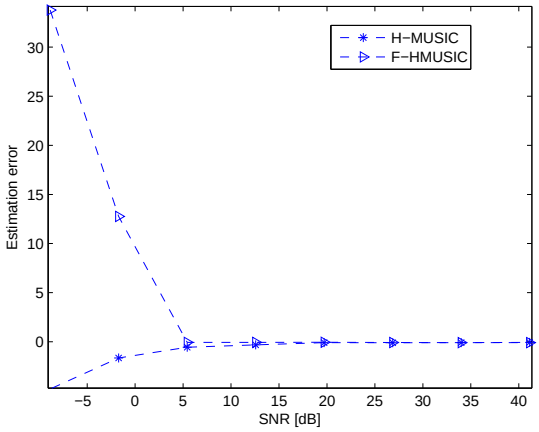
of the algorithm is of interest and errors generated due to the automatic subband signal detection are not preferred therefore the subband containing the signal is assumed to be known. The signal is perfect harmonic with $B = 0$.

In each trial, a signal segment is generated according to the model in (1), with the noise being randomized. The estimators are evaluated in terms of the root mean square error (RMSE) defined as:

$$RMSE = \sqrt{\frac{1}{D}\sum_{d=1}^{D}(\hat{\omega}_0 - \omega_0)^2}, \tag{35}$$



(a)



(b)

Fig. 8: *a) The RMSE with a varying SNR in the case of random distributed amplitudes embedded in white noise. b) Corresponding model order estimation errors.*

with ω_0 and $\hat{\omega}_0$ being the true fundamental frequency and the estimate, respectively, and with D being the number of Monte Carlo simulations. In this paper $D = 200$. This is done for various SNR defined as:

$$SNR = 10 \log_{10} \sum_{l=1}^{L} \frac{\alpha_l^2}{\phi(\omega_l)}, \tag{36}$$

where the function $\phi(\omega_l)$ is the power spectrum of the noise at frequency $\omega_l = \omega_0 l$. In the case of white noise, power spectrum is equal to the variance of noise, and in colored noise the power spectrum is white noise filtered with a AR process. The SNR is calculated with (36) where the function $\phi(\omega_l)$ is

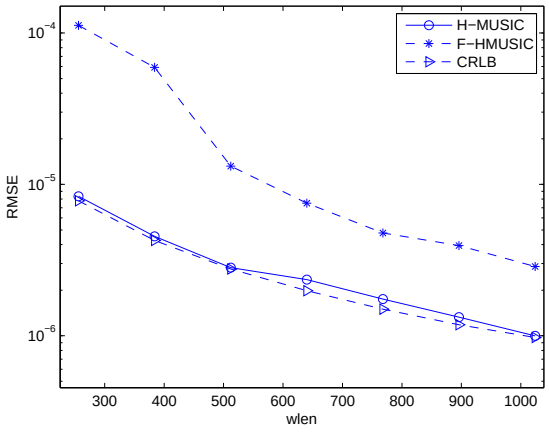


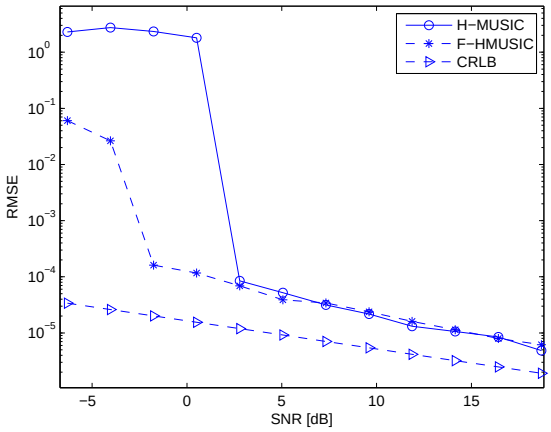
Fig. 9: *The RMSE performance with a varying frame length N where amplitudes is constant distributed and SNR fixed at $25dB$*

power spectrum at frequency ω_l. Furthermore, model order errors on the harmonic signals is defined as the difference between the estimated order subtracted on the true order. The results are compared with the exact CRLB for both white and colored noise cases using equations stated in [31], [32].

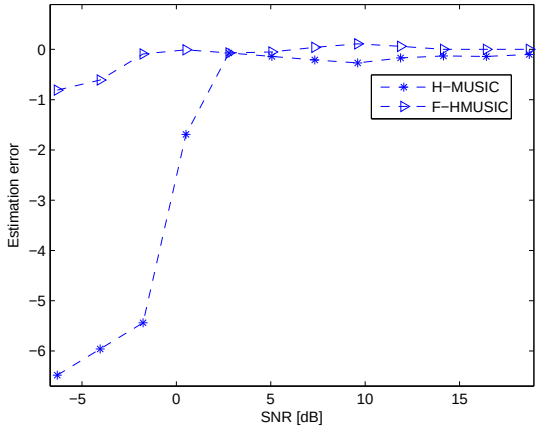
In the experiments to follow, we use the following signal and noise setup. The signal will consist of $L = 13$ complex exponentials embedded in noise with a fundamental frequency of $\omega_0 = 0.15$. Both white and colored noise are evaluated. Two cases of amplitudes are considered one with constant amplitudes of $\alpha_l = 1 \ \forall \ l$ and the other with a randomized amplitude generated from a Rayleigh probability density function. The Rayleigh distribution provides a good model for amplitudes from speech and musical instruments. For both F-HMUSIC and HMUSIC the parameters are set in common where the searching candidates of ω_0 is set to $\Omega \in [0.06, 0.4]$, the model order considered were $\mathcal{L} \in [5, \lfloor \pi/\omega_0 \rfloor - 1]$. Note that the interval for ω_0 includes both $2\omega_0$ and $\frac{1}{2}\omega_0$ which is normally referred as octave errors. The user parameter for F-HMUSIC is selected to be $s = \lfloor 0.5M \rfloor$, and for HMUSIC $\lfloor 0.5N \rfloor$.

In the first example, F-HMUSIC is evaluated in white noise scenario where the amplitudes of harmonics are constant. The corresponding results are shown in Fig. 7a where in estimated RMSE versus a varying SNR is shown and in Fig.7b the associated model order estimation errors is plotted. The performance curve of F-HMUSIC is closely following CRLB for region above the break down[4] region of the algorithm.

[4]The algorithm performance break down problem is a common problem at low SNR region which is also referred as subspace swapping problem where the high noise level cause part of the signal subspace erroneously determined as signal subspace.



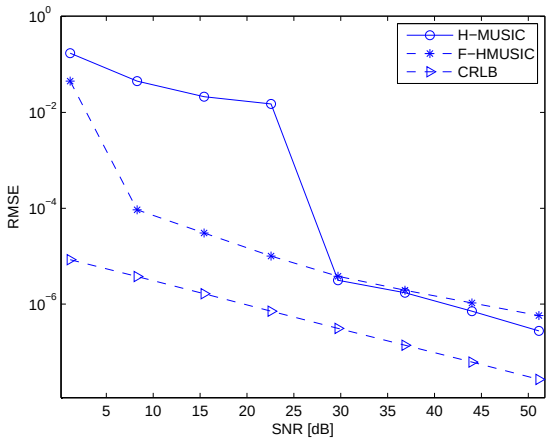
(a)



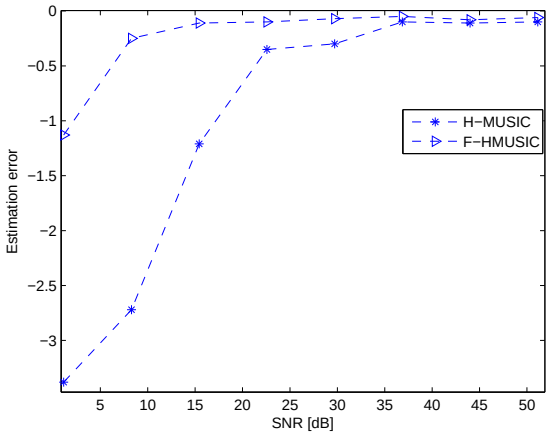
(b)

Fig. 10: *a) The RMSE with a varying SNR in the case of constant amplitudes embedded in colored noise. b) Corresponding model order estimation error*

With the consideration of computational savings the performance of F-HMUSIC is still comparable with HMUSIC as shown in Fig. 7a where both algorithms provides estimate close to CRLB. Next, the performance is measured on a set of harmonic signals whose signal amplitudes are generated according to a Rayleigh distribution. The performance curve of RMSE with varying SNR is shown in Fig. 8a, the subspace methods are suffering some performance degradation where the breakdown point has been raised compared to constant amplitude case. Overall, F-HMUSIC has shown good statistical performance for harmonic signals embedded in white noise. In the operation region above the breakdown point the



(a)



(b)

Fig. 11: *a) The RMSE with a varying SNR in the case of random amplitudes embedded in colored noise. b) Corresponding model order estimation error.*

order estimates have also shown good accuracies, and this is also the case in the remaining experiments. From Fig. 7b and 8b shows that the estimates of L is close to true value when estimate of F_0 is closing to CRLB.

In the next example, the RMSE performance is evaluated on a varying window length with SNR fixed at 25dB and amplitudes kept fixed at $\alpha_l = 1, \forall l$. For various frame length the user parameters are selected to be $\lfloor 0.85M \rfloor$ and $\lfloor 0.85N \rfloor$, respectively for F-HMUSIC and HMUSIC. The performance is reported in Fig. 9 where it shows that both algorithms can be operating on any frame length between 256 to 1024

with wisely selected user parameters.

A final example will demonstrate F-HMUSIC performance evaluated in colored noise scenario. Signal setup is the same as previous examples except that, here the embedded noise is filtered with a second order AR process ($\frac{1}{1+0.3z^{-1}+0.8z^{-2}}$), where the main power of the noise is mainly concentrated on subband $m = 2$, and one realization of the signal embedded in colored noise is shown in Fig. 6. To enhance the performance of both methods a slightly different setup have been used for both F-HMUSIC and HMUSIC where the searching space for ω_0 is $\Omega \in [0.1, 0.8]$. Remaining parameters are same as in earlier examples. In the colored noise case the evaluation results for the constant distributed amplitudes is shown in Fig. 10 where the algorithm breakdown region for F-HMUSIC is lower than HMUSIC. Rayleigh distributed amplitudes are also evaluated with Monte Carlo Simulations, significantly better performances of F-HMUSIC has been shown in Fig. 11. Due to the noise properties, subband with index $m = 1$ contains white noise characteristic which provides good estimates with reduced subspace swapping properties than estimates in subband $m = 2$. By averaging the estimates from both subbands a more robust estimation is then achieved. From simulations shown in Fig 11, it can be clearly seen that F-HMUSIC is more robust against the colored noise than HMUSIC both in fixed and Rayleigh distributed amplitudes.

IV. Conclusion

In this paper, a high resolution fundamental frequency estimator termed F-HMUSIC with automatic subband signal detection has been proposed. This algorithm is a frequency domain based estimator using subspaces decomposed from FS data matrix model to efficiently estimate the fundamental frequency, where a subband based approach is adopted to reduce the sensitivity to the colored noise and increase the computational efficiency. Additionally, an automatic subband signal detection method has been proposed which is based on information theoretic criterion where no subjective judgment is needed. The performance of F-HMUSIC has been evaluated on both synthetic and recorded signals. From simulations on synthetic data shows that F-HMUSIC is more robust against colored noise than HMUSIC. Furthermore, robustness of the method has been demonstrated by evaluation on recorded signals where F-HMUSIC has shown performance close to YIN for MIDI notes between $[45, 95]$, and for MIDI note above 95 our algorithm performs better than YIN. Overall the performance of F-HMUSIC is considered as accurate and robust for the operating region. In the operation region the price we paid for computational complexity and robustness to colored noise is with estimation accuracy.

References

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

[2] N. H. Fletcher and T. D. Rossing, *The Physics of MUSICAL instruments*, Springer, 1998.

[3] P. J. Rose, "On the non-equivalence of fundamental frequency and pitch in tonal description," *In Prosodic Analysis and Asian Linguistics*, 1988.

[4] R. Herman, M. Beckman, and K. Honda, "Linguistic models of F0 use, physiological models of F0 control, and the issue of "mean response time"," *Language and Speech*, vol. 42, pp. 373–399, 1999.

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

[6] W. Hess, "Pitch and voicing determination," *Advances in Speech Signal Processing*, pp. 3–48, 1992.

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

[8] R. Badeau, B. David, and G. Richard, "A new perturbation analysis for signal enumeration in rotational invariance techniques," *IEEE Trans. Signal Processing*, vol. 54, no. 2, pp. 450–458, 2006.

[9] M.G. Christensen, A. Jakobsson, and S.H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 15, no. 5, pp. 1635–1644, 2007.

[10] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Elsevier Signal Processing*, vol. 88(4), pp. 972–983, 2008.

[11] M.G. Christensen, P. Vera-Candeas, S.D. Somasundaram, and A. Jakobsson, "Robust subspace-based fundamental frequency estimation," *IEEE Int. Conf. Acoust., Speech and Signal Processing*, pp. 101–104, 2008.

[12] M.G. Christensen, A. Jakobsson, and S.H. Jensen, "Multi-pitch estimation using Harmonic MUSIC," *in Rec. Asilomar Conf. Signals, Systems, and Computers*, pp. 521–525, 2006.

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

[14] M.G. Christensen, A. Jakobsson, and S.H. Jensen, "Sinusoidal order estimation using angles between subspaces," *Submitted to: IEEE Transaction in Signal Processing*, 2008.

[15] M.G. Christensen, A. Jakobsson, and S.H. Jensen, "Sinusoidal order estimation using the subspace orthogonality and shift-invariance properties," *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2007.

[16] A. Jakobsson, M. G. Christensen, and S. H. Jensen, "Frequency selective sinusoidal order estimation," *IEEE Electronic Letters*, vol. 43(21), pp. 1164–1165, 2007.

[17] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, pp. 387– 392, 1985.

[18] M. S. Bartlett, "A note on the multiplying factors for various x^2 approximations," *J. Roy. Stat. Soc,*, vol. Ser. B,16, pp. 296–298, 1954.

[19] H. Akaike, "Information theory and an extenstion of the maximum likelihood principle," in *in Proc. 2nd Int. Symp. Inform. Theory*, 1973.

[20] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 6, pp. 461–464, 1978.

[21] G. Schwartz, "Estimating the dimention of a model," *Ann. Stat.*, vol. 6, pp. 461–464, 1978.

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

[23] J. Gunnarsson and T. McKelvey, "High SNR performance analysis of F-ESPRIT," in *Rec. of 38th Asilomar Conference on Signals, Systems and Computers*, 2004.

[24] T. McKelvey and M. Viberg, "A robust frequency domain subspace algorithm for multi-component harmonic retrieval," in *Rec. of 34th Asilomar Conference on Signals, Systems and Computers*, 2001.

[25] P. Stoica, N. Sandgren, Y. Seln, L. Vanhamme, and S. van Huffel, "Frequency-domain method based on the singular value decomposition for frequency-selective NMR spectroscopy," *Journal of Magnetic Resonance*, vol. 165 (1), pp. 80–88, 2003.

[26] V. Emiya, *Transcription automatique de la musique de piano*, Ph.D. thesis, École Nationale Supérieure des Télécommunications, 2008.

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

[28] V. Emiya, B. David, and R. Badeau, "A parametric method for pitch estimation of piano tones," *IEEE Int. Conf. Acoust., Speech and Signal Processing.*, 2007.

[29] P. Stoica and T. Soderstrom, "Statistical analysis of music and subspace rotation estimates of sinusoidal frequencies," *IEEE Trans. on Signal Processing,*, vol. 39, pp. 1836–1847, 1991.

[30] R. M. Gray, "Toeplitz and circulant matrices: A review,"," *Foundations and Trends in Communications and Information Theory*, vol. 2(3), pp. 155–239, 2006.

[31] J. M. Francos and B. Friedlander, "Bounds for estimation of complex exponentials in unknown colored noise," *IEEE Trans. Signal Process.*, vol. 43, pp. 2176–2185, 1995.

[32] P. Stoica, A. Jakobsson, and J. Li, "Cisoid parameter estimation in the colored noise case: asymptotic cramer-rao bound, maximum likelihood, and nonlinear least-squares," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 45, no. 8, pp. 2048–2059, 1997.