EURASIP Journal on Advances in Signal Processing

SpringerOpen

This Provisional PDF corresponds to the article as it appeared upon acceptance. Fully formatted PDF and full text (HTML) versions will be made available soon.

Joint DOA and multi-pitch estimation based on subspace techniques

EURASIP Journal on Advances in Signal Processing 2012, **2012**:1 doi:10.1186/1687-6180-2012-1

Johan XI Zhang (jxz@es.aau.dk)
Mads GRAESBOLL Christensen (mgc@imi.aau.dk)
Soren HOLDT Jensen (shj@es.aau.dk)
Marc Moonen (marc.moonen@esat.kuleuven.be)

ISSN 1687-6180

Article type Research

Submission date 26 March 2011

Acceptance date 2 January 2012

Publication date 2 January 2012

Article URL http://asp.eurasipjournals.com/content/2012/1/1

This peer-reviewed article was published immediately upon acceptance. It can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

For information about publishing your research in *EURASIP Journal on Advances in Signal Processing* go to

http://asp.eurasipjournals.com/authors/instructions/

For information about other SpringerOpen publications go to

http://www.springeropen.com

© 2012 Zhang *et al.* ; licensee Springer.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Joint DOA and multi-pitch estimation based on subspace techniques

Johan Xi Zhang*[1], Mads Græsbøll Christensen[2], Søren Holdt Jensen[1] and Marc Moonen[3]

[1]Department of Electronic Systems (ES-MISP), Aalborg University, Aalborg, Denmark

[2]Department of Architecture, Design and Media Technology, Aalborg University, Denmark

[3]Department of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Leuven, Belgium

*Corresponding author: jxz@es.aau.dk

Email addresses:

MGC: mgc@imi.aau.dk

SHJ: shj@es.aau.dk

MM: marc.moonen@esat.kuleuven.be

Abstract

In this article, we present a novel method for high-resolution joint direction-of-arrivals (DOA) and multi-pitch estimation based on subspaces decomposed from a spatio-temporal data model. The resulting estimator is termed multi-channel harmonic MUSIC (MC-HMUSIC). It is capable of resolving sources under adverse conditions, unlike traditional methods, for example when multiple sources are impinging on the array from approximately the same angle or similar pitches. The effectiveness of the method is demonstrated on a simulated an-echoic array recordings with source signals from real recorded speech and clarinet. Furthermore, statistical evaluation with synthetic signals shows the increased robustness in DOA and fundamental frequency estimation, as compared with to a state-of-the-art reference method.

Keywords: multi-pitch estimation; direction-of-arrival estimation; subspace orthogonality; array processing.

1. Introduction

The problem of estimating the fundamental frequency, or pitch, of a period waveform has been of interest to the signal processing community for many years. Fundamental frequency estimators are important for many practical applications such as automatic note transcription in music, audio and speech coding, classification of music, and speech analysis. Numerous algorithms have been proposed for both the single- and multi-pitch scenarios [1–5]. The problem for single-pitch scenarios is considered as well-posed. However, in real-world signals, the multi-pitch scenario occurs quite frequently [2,6]. The multi-pitch estimation algorithms are often based on, i.e., various modification of the auto-correlation function [1,7], maximum likelihood, optimal filtering, and subspace techniques [2,3,8]. In real-life recordings, problems such as frequency overlap of sources, reverberation, and colored noise will strongly limit the performance of multi-pitch estimator and estimator designed for single channel recordings often use simplified signal models. One widely used signal simplification in multi-pitch estimators, for example, is the sparseness of the signal, where the frequency spectrum of sources are assumed to not overlap [2]. This assumption may be appropriate when sources consist of mixture of several speech signals having different pitches [9]. However,

for audio signals it is less likely to be true. This is especially so in western music, where instruments are most often played in accord, something that causes the harmonics to overlap or even coincide. With only single-channel recording it is, therefore, hard, or perhaps even impossible, to estimate pitches with overlapping harmonics, unless additional information, such as a temporal or spectral model, is included.

Recently, multi-channel approaches have attracted considerable attention both in single- and multi-pitch scenarios. By exploring the spatial information of the sources, more robust pitch estimators have been proposed [10–14]. Most of those multi-channel methods are still mainly based on auto-correlation function-related approaches, however, although a few exceptions can be found in [15–18]. In direction-of-arrival (DOA) estimators, audio and speech signals are often modeled as broadband signal, and standard subspace methods such as MUSIC and ESPRIT are only defined for narrow-band signal model, which then fail to directly operate on broadband signals [19]. One often used concept is band-pass filtering of broadband signals into subbands, where narrow-band estimators can be applied to each subband [20]. In the narrow-band case, a delay in the signal is equivalent to a phase shifts according to the frequencies of complex exponentials. An alternative method is, however, as follows: since harmonic signals consist of sinusoidal components, we can model each source as multiple narrow-band signal with distinct frequencies arriving at the same DOA.

In this article, we propose a parametric method for solving the problem of joint fundamental frequency and DOA estimation based on subspace techniques where the quantities of interest are jointly estimated using a MUSIC-like approach. We term the proposed estimator Multi-channel multi-pitch Harmonic MUSIC (MC-HMUSIC). The spatio-temporal data model used in MC-HMUSIC is based on the JAFE data model [21,22]. Originally, the JAFE data model was used for estimating joint unconstrained frequencies and DOAs estimates of complex exponential using ESPRIT, which is referred as joint angle-frequency estimation (JAFE) algorithm. Other-related work with joint frequency-DOA methods includes [23–25]. In this article, we have parametrized the harmonic structure of periodic signals in the signal model to model the fundamental frequency and the DOA of individual sources. An estimator is constructed for jointly estimating the parameters of interest. Incorporating the DOA parameter in finding the fundamental frequency may give better robustness against a signal with overlapping harmonics. Similarly, it can be expected that the DOA can be found more accurately when the nature of the signal of interest is taken into account.

The remainder of this article is comprised four sections: Section 2, in which we will introduce some notation, the spatio-temporal signal model, for which we also derive the associated Cramér–Rao lower bound, along with the JAFE data mode; Section 3, where we then present the proposed method; Section 4, in which we present the experimental results obtained using the proposed method; and, finally, Section 5, where we conclude on our work.

2. Fundamentals

2.1. Spatio-temporal signal model

Next, the signal model employed throughout the article will be presented. Without multi-path propagation of sources, it is given as follows: the signal x_i received by microphone element i arranged in a uniform linear array (ULA) configuration, $i = 1, \ldots, M$, is given by

$$\begin{aligned} x_i(n) &= \sum_{k=1}^{K}\sum_{l=1}^{L_k} \beta_{l,k} \mathsf{e}^{j(\omega_k l n + \phi_k l (i-1))} + e_i(n), \\ \beta_{l,k} &= A_{l,k}\mathsf{e}^{j\gamma_{l,k}}, \end{aligned} \tag{1}$$

for sample index $n = 0, \ldots, N-1$, where subscript k denotes the kth source and l the lth harmonic. Moreover, $A_{l,k}$ is the real-valued positive amplitude of the complex exponential, L_k is the number of harmonics, K is number of sources, $\gamma_{l,k}$ is the phase of the individual harmonics, ϕ_k is the phase shift caused by the DOA, and $e_i(n)$ is complex symmetric white Gaussian noise. The phase shift between array elements is given as $\phi_k = \omega_k f_s \frac{d}{c}\sin(\theta_k)$, where d is the spacing between the elements measured in wavelengths, c is the speed of propagation in unit [m/s], θ_k is the DOA defined for $\theta_k \in [-90^\circ, 90^\circ]$, f_s is the signal sampling frequency. The problem of interest is to estimate ω_k and θ_k. We in the following assume that the number of sources K is known and the number of harmonics L_k of individual sources is known or found in some other, possibly joint, way. We note that a number of ways of doing this has been proposed in the past [26–28,2].

2.2. Cramér–Rao lower bound

We will now proceed to derive the exact Cramér–Rao lower bound (CRLB) for the problem of estimating the parameters of interest. First, we define the $M \times 1$ deterministic signal model vector $\mathbf{s}(n, \boldsymbol{\mu})$ with column element as

$$s_i(n, \boldsymbol{\mu}) = \sum_{k=1}^{K} \sum_{l=1}^{L_k} \beta_{l,k} \mathrm{e}^{j(\omega_k l n + \phi_k l(i-1))}, \quad \beta_{l,k} = A_{l,k} \mathrm{e}^{j\gamma_{l,k}}, \tag{2}$$

where $\mathbf{s}(n, \boldsymbol{\mu}) = \begin{bmatrix} s_1(n, \boldsymbol{\mu}) & \cdots & s_M(n, \boldsymbol{\mu}) \end{bmatrix}^T$. Furthermore, the parameter vector $\boldsymbol{\mu}$ is given by

$$\begin{array}{ccccccc} \boldsymbol{\mu} & = & [\omega_1 & \cdots & \omega_K & \theta_1 & \cdots \quad \theta_K \\ & & A_{1,1} & \gamma_{1,1} & \cdots & A_{L_K,K} & \gamma_{L_K,K}] \end{array}. \tag{3}$$

Recall that the observed signal vector with additive white noise is given by

$$\mathbf{x}(n) = \mathbf{s}(n, \boldsymbol{\mu}) + \mathbf{e}(n) = \begin{bmatrix} s_1(n, \boldsymbol{\mu}) \\ \vdots \\ s_M(n, \boldsymbol{\mu}) \end{bmatrix} + \mathbf{e}(n), \tag{4}$$

with $\mathbf{e}(n)$ being the noise column vector. The CRLB is defined as the variance of an unbiased estimate of the pth element of $\boldsymbol{\mu}$, which is lower bounded as

$$\mathrm{var}(\boldsymbol{\mu}_p) \geq \left[\mathbf{C}^{-1}\right]_{pp}, \tag{5}$$

where $\mathbf{C}$ is the so-called Fisher information matrix given by

$$\mathbf{C} = \frac{2}{\sigma^2} \mathrm{Re} \left(\sum_{n=0}^{N-1} \frac{\partial \mathbf{s}(n, \boldsymbol{\mu})^H}{\partial \boldsymbol{\mu}} \frac{\partial \mathbf{s}(n, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T} \right). \tag{6}$$

The partial derivative matrix is denoted as

$$\frac{\partial \mathbf{s}(n, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = \begin{bmatrix} \frac{\partial s_1(n, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}} & \cdots & \frac{\partial s_M(n, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}} \end{bmatrix}, \tag{7}$$

where vector $\frac{\partial s_i(n, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}}$ is the partial derivatives with respect to the entries in the vector $\boldsymbol{\mu}$. The expression for the columns in $\frac{\partial \mathbf{s}(n, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}}$ is given as

$$\frac{\partial s_i(n, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = \begin{bmatrix} \sum_{l=1}^{L_1} jl \left(n + (i-1) f_s \frac{d}{c} \sin(\theta_1)\right) \beta_{l,1} \mathrm{e}^{j(\omega_1 l n + \phi_1 l(i-1))} \\ \vdots \\ \sum_{l=1}^{L_K} jl \left(n + (i-1) f_s \frac{d}{c} \sin(\theta_K)\right) \beta_{l,K} \mathrm{e}^{j(\omega_K l n + \phi_K l(i-1))} \\ \sum_{l=1}^{L_1} \left(jl (i-1) \omega_1 f_s \frac{d}{c} \cos(\theta_1)\right) \beta_{l,1} \mathrm{e}^{j(\omega_1 l n + \phi_1 l(i-1))} \\ \vdots \\ \sum_{l=1}^{L_K} \left(jl (i-1) \omega_K f_s \frac{d}{c} \cos(\theta_K)\right) \beta_{l,K} \mathrm{e}^{j(\omega_K l n + \phi_K l(i-1))} \\ \mathrm{e}^{j\gamma_{1,1}} \mathrm{e}^{j(\omega_1 n + \phi_1 (i-1))} \\ jA_{1,1} \mathrm{e}^{j\gamma_{1,1}} \mathrm{e}^{j(\omega_1 n + \phi_1 (i-1))} \\ \vdots \\ \mathrm{e}^{j\gamma_{L_K,K}} \mathrm{e}^{j(\omega_K L_K n + \phi_K L_K (i-1))} \\ jA_{L_K,K} \mathrm{e}^{j\gamma_{L_K,K}} \mathrm{e}^{j(\omega_K L_K n + \phi_K L_K (i-1))} \end{bmatrix}. \tag{8}$$

2.3. The JAFE data model

Next, we will introduce the specifics of the JAFE data model [22,29] that our method is based on. At a time instant n the received signal from the M array elements are $\mathbf{x}(n) = \begin{bmatrix} x_1(n) & x_2(n) & \cdots & x_M(n) \end{bmatrix}^T$, which can be written as

$$\mathbf{x}(n) = \mathbf{A}\mathbf{\Phi}\mathbf{b} + \mathbf{e}(n), \tag{9}$$

where $\mathbf{e}(n) \in \mathbb{C}^{M\times 1}$ is the noise vector, and $\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \cdots & \mathbf{A}_K \end{bmatrix}$ is a Vandermonde matrix containing parameters ω_k and θ_k for sources $k = 1, \ldots, K$, i.e.,

$$\mathbf{A}_k = \begin{bmatrix} \mathbf{a}(\theta_k, \omega_k 1) & \cdots & \mathbf{a}(\theta_k, \omega_k L_k) \end{bmatrix}, \tag{10}$$

with $\mathbf{a}(\theta, \omega)$ being the array steering vector given by

$$\mathbf{a}(\theta, \omega) = \begin{bmatrix} 1 & \cdots & \mathsf{e}^{j\omega f_s \frac{d}{c}(M-1)\sin(\theta)} \end{bmatrix}^T. \tag{11}$$

Here, $(\cdot)^T$ denotes the vector transpose. Unlike the steering vector defined in [22,21], where only the DOA is parametrized, here, a general definition of the vector (11) is used, in which it depends on both θ and ω [29]. The frequency components are expressed in $\mathbf{\Phi}^n = \operatorname{diag}\left(\begin{bmatrix} \mathbf{\Phi}_1^n & \cdots & \mathbf{\Phi}_K^n \end{bmatrix}\right)$ where the matrix for each source is given by

$$\mathbf{\Phi}_k = \operatorname{diag}\left(\begin{bmatrix} \mathsf{e}^{j\omega_k} & \cdots & \mathsf{e}^{j\omega_k L_k} \end{bmatrix}\right). \tag{12}$$

The complex amplitudes for involving components are represented by the following vector:

$$\mathbf{b} = \begin{bmatrix} \beta_{1,1} & \cdots & \beta_{L_1,1} & \cdots & \beta_{1,K} & \cdots \beta_{L_K,K} \end{bmatrix}^T. \tag{13}$$

To capture the temporal behavior, N time-domain data samples of the array output $\mathbf{x}(n)$ are collected to form the $M \times N$ data matrix $\mathbf{X}$, which is defined as

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}(n) & \cdots & \mathbf{x}(N) \end{bmatrix}. \tag{14}$$

Due to the structure of the harmonic components, the data matrix is given by

$$\mathbf{X} = \mathbf{A}\begin{bmatrix} \mathbf{b} & \mathbf{\Phi}\mathbf{b} & \cdots & \mathbf{\Phi}^{N-1}\mathbf{b} \end{bmatrix} + \mathbf{E}, \tag{15}$$

where $\mathbf{E} \in \mathbb{C}^{M\times N}$ is a matrix containing N samples of the noise vector **e**(n).

In speech and audio signal processing, it is common to model each source as a set of multiple harmonics with model order $L_k > 1$. Due to the narrow-band approximation of the steering vector, the multiple complex components with distinct frequencies impinge on the array with identical DOA will result in a non-unique spatial frequencies which cause a harmonic structure in the spatial frequencies $\phi_k l \ \forall l$ as well. The multiple sources impinge on the array with different DOAs consisting of various frequency components may, for certain frequency combinations, give the same array steering vector, which cause the matrix $\mathbf{A}$ to be rank deficient. Normally, this ambiguous mapping of the steering vector is mitigated by band-pass filtering the signal into its subbands, where the DOA of the signal is uniquely modeled by the narrow-band steering vector [20, Chap. 9].

Here, the ambiguities and the rank-deficiency are avoided by introducing temporal smoothness in order to restore the rank of $\mathbf{A}$. The temporally smoothed data matrix is obtained by stacking t times temporally shifted versions of the original data matrix [22,21,29], given as

$$\mathbf{X}_t = \begin{bmatrix} \mathbf{A}\begin{bmatrix} \mathbf{b} & \mathbf{\Phi}\mathbf{b} & \cdots & \mathbf{\Phi}^{N-t}\mathbf{b} \end{bmatrix} \\ \mathbf{A}\mathbf{\Phi}\begin{bmatrix} \mathbf{b} & \mathbf{\Phi}\mathbf{b} & \cdots & \mathbf{\Phi}^{N-t}\mathbf{b} \end{bmatrix} \\ \vdots \\ \mathbf{A}\mathbf{\Phi}^{t-1}\begin{bmatrix} \mathbf{b} & \mathbf{\Phi}\mathbf{b} & \cdots & \mathbf{\Phi}^{N-t}\mathbf{b} \end{bmatrix} \end{bmatrix} + \mathbf{E}_t, \tag{16}$$

where $\mathbf{X}_t \in \mathbb{C}^{tM\times N-t+1}$ is the temporally smoothed data matrix, and $\mathbf{E}_t$ is the noise term constructed from $\mathbf{E}$ in a similar way as $\mathbf{X}_t$. In using the signal model where the amplitudes are assumed stationary for $n = 0, \ldots, N-1$, $\mathbf{X}_t$ can be factorized as

$$\mathbf{X}_t = \begin{bmatrix} \mathbf{A} \\ \mathbf{A}\mathbf{\Phi} \\ \vdots \\ \mathbf{A}\mathbf{\Phi}^{t-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} & \mathbf{\Phi}\mathbf{b} & \cdots & \mathbf{\Phi}^{N-t}\mathbf{b} \end{bmatrix} + \mathbf{E}_t. \tag{17}$$

With some additional definitions, we can also write this expression more compactly as

$$\mathbf{X}_t = \overline{\mathbf{A}}_t\mathbf{B}_t + \mathbf{E}_t, \tag{18}$$

where $\overline{\mathbf{A}}_t = \left[\begin{array}{cccc} \mathbf{A} & \mathbf{A}\mathbf{\Phi} & \cdots & \mathbf{A}\mathbf{\Phi}^{t-1} \end{array} \right]^T$ and $\mathbf{B}_t = \left[\begin{array}{cccc} \mathbf{b} & \mathbf{\Phi}\mathbf{b} & \cdots & \mathbf{\Phi}^{N-t}\mathbf{b} \end{array} \right]$. The temporally smoothed data matrix $\mathbf{X}_t$ can maximally resolve up to $tM \geq \sum_{k=1}^{K} L_k$ complex exponentials, where $\overline{\mathbf{A}}_t$ is linearly independent for any distinct θ and ω [30].

When multiple sources with distinct DOA with the same fundamental frequency impinge on the array, it will result in correlation between the underlying signals, which will make it harder to separate the corresponding components into its eigenvectors [22,31]. To mitigate this problem, spatial smoothing is introduced, which works as follows. An array of M sensors is subdivided into S subarrays. In this article, the subarrays are spatially shifted with one element in each subarrays, the number of elements in each subarray being $M_S = M - S + 1$. For $s = 1, \ldots, S$, let $\mathbf{J}_s \in \mathbb{C}^{tM_s \times tM}$ be the selection matrix corresponding to the sth subarray for the data matrix $\mathbf{X}_t$. Then, the spatio-temporally smoothed data matrix $\mathbf{X}_{t,S} \in \mathbb{C}^{tM_s \times S(N-t+1)}$ is given by

$$\mathbf{X}_{t,s} = \left[\begin{array}{ccc} \mathbf{J}_1\mathbf{X}_t & \cdots & \mathbf{J}_S\mathbf{X}_t \end{array} \right]. \tag{19}$$

Furthermore, $\mathbf{X}_{t,s}$ can be factorized as

$$\mathbf{X}_{t,s} = \left[\begin{array}{ccc} \mathbf{J}_1\overline{\mathbf{A}}_t & \cdots & \mathbf{J}_S\overline{\mathbf{A}}_t \end{array} \right] \left[\begin{array}{ccc} \mathbf{B}_t & & \\ & \ddots & \\ & & \mathbf{B}_t \end{array} \right] + \mathbf{E}_{t,s}, \tag{20}$$

where $\mathbf{E}_{t,s}$ is the noise term constructed from $\mathbf{E}$ in a similar way as $\mathbf{X}_{t,s}$. Using the shift invariance structure in $\mathbf{A}_m$, the term $\mathbf{J}_s\mathbf{A}_m$ for $s = 1, \ldots, S$ is given by

$$\mathbf{J}_s\overline{\mathbf{A}}_t = \mathbf{J}_1\overline{\mathbf{A}}_t\mathbf{\Theta}^{s-1}, \tag{21}$$

where

$$\mathbf{\Theta} = \operatorname{diag}\left\{ \left[\begin{array}{ccc} \mathrm{e}^{j\phi_1 1} \cdots \mathrm{e}^{j\phi_1 L_1} & \cdots & \mathrm{e}^{j\phi_K 1} \cdots \mathrm{e}^{j\phi_K L_K} \end{array} \right] \right\}, \tag{22}$$

which is simply the phase difference between array elements. With (21), the matrix $\mathbf{X}_{t,s}$ can be written in a compact form as

$$\mathbf{X}_{t,s} = \mathbf{J}_1\overline{\mathbf{A}}_t \left[\begin{array}{cccc} \mathbf{B}_t & \mathbf{\Theta}\mathbf{B}_t & \cdots & \mathbf{\Theta}^{S-1}\mathbf{B}_t \end{array} \right] + \mathbf{E}_{t,s}, \tag{23}$$

with selection matrix expressed as

$$\mathbf{J}_1 = \mathbf{I}_t \otimes \left[\begin{array}{cc} \mathbf{I}_{M_s} & \mathbf{0} \end{array} \right], \tag{24}$$

where $\mathbf{I}_t \in \mathbb{R}^{t \times t}$ and $\mathbf{I}_{M_s} \in \mathbb{R}^{M_s \times M_s}$ are the identity matrices, $\otimes$ is the Kroneker product as defined in [22].

It is interesting to note that the noise term $\mathbf{E}_{t,s}$ is no longer white due to the spatio-temporal smoothing procedure, as correlation between the different rows of (23) is obtained. A pre-whitening step can be implemented in (23) to mitigate this. We note, however, that according to results reported in [22], pre-whitening step is only interesting for signals with low SNR where minor estimation improvement can be achieved. In this article, the main interest is to propose a multi-channel joint DOA and multi-pitch estimator, for which reason the whitening process is left without further description, but we refer the interested reader to [22]. We also note that aside from spatial smoothing, forward–backward averaging could also be implemented to reduce the influence of the correlated sources [22,31,19].

3. The proposed method

3.1. Coarse estimates

From the final spatio-temporally smoothed data matrix, a basis for the signal and noise subspaces can be obtained as follows. The singular value decomposition (SVD) of the data matrix (23) is given by

$$\mathbf{X}_{t,s} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H, \tag{25}$$

where the columns of $\mathbf{U}$ are the singular vectors, i.e.,

$$\mathbf{U} = \left[\begin{array}{ccc} \mathbf{u}_1 & \cdots & \mathbf{u}_{tM_S} \end{array} \right]. \tag{26}$$

A basis of the orthogonal complement of the signal subspace, also called the noise subspace, is formed from singular vector associated with the $mM_S - Q$ least significant singular values, i.e.,

$$\mathbf{G} = \left[\begin{array}{ccc} \mathbf{u}_{Q+1} & \cdots & \mathbf{u}_{mM_S} \end{array} \right], \tag{27}$$

with $Q = \sum_{k=1}^{K} L_k$ being the total number of complex exponentials in the signal. Similarly, the signal subspace is spanned by the Q largest singular values, i.e.,

$$\mathbf{S} = \left[\begin{array}{ccc} \mathbf{u}_1 & \cdots & \mathbf{u}_Q \end{array} \right]. \tag{28}$$

The defined signal subspace and noise subspace have similar property as traditional subspaces where estimators such as joint DOA and frequency, or fundamental frequency estimators can be constructed using the principle used in MUSIC [19,32,27, 26,4]. According to the signal noise subspace orthogonality principle, the following relationship holds:

$$\mathbf{J}_1\overline{\mathbf{A}}_t\mathbf{G} = \mathbf{0}, \tag{29}$$

where we, for notational simplicity, have introduced $\mathbf{J}_1\overline{\mathbf{A}}_t = \mathbf{A}_{ts}$. The matrix $\mathbf{A}_{ts}$ is comprised Vandermonde matrices for sources $k = 1, \ldots, K$. The matrix for each individual source is given by

$$\mathbf{A}_{ts,k} = \begin{bmatrix} 1 & \cdots & 1 \\ \mathsf{e}^{j\phi_k} & & \mathsf{e}^{j\phi_k L_k} \\ \vdots & & \vdots \\ \mathsf{e}^{j\phi_k S} & \cdots & \mathsf{e}^{j\phi_k L_k S} \\ \vdots & & \vdots \\ \mathsf{e}^{j\omega_k(t-1)} & \cdots & \mathsf{e}^{j\omega_k L_k(t-1)} \\ \mathsf{e}^{j\phi_k}\mathsf{e}^{j\omega_k(t-1)} & & \mathsf{e}^{j\phi_k L_k}\mathsf{e}^{j\omega_k L_k(t-1)} \\ \vdots & & \vdots \\ \mathsf{e}^{j\phi_k S}\mathsf{e}^{j\omega_k(t-1)} & \cdots & \mathsf{e}^{j\phi_k L_k S}\mathsf{e}^{j\omega_k L_k(t-1)} \end{bmatrix}. \tag{30}$$

The cost function of the proposed joint DOA and multi-pitch estimator is then

$$J(\omega_k, \theta_k) = \left\| \mathbf{A}_{ts,k}^H \mathbf{G} \right\|_F^2, \tag{31}$$

where $\|\cdot\|_F$ is the Frobenius Norm. Note that this measure is closely related to the angles between the subspaces as explained in [33] and can hence be used as a measure of the extent to which (29) holds for a candidate fundamental frequency and DOA. The pair of fundamental frequency and DOA can, therefore, be found as the combination that is the closest to being orthogonal to $\mathbf{G}$, i.e.,

$$\{\omega_k, \theta_k\}_{k=1}^K = \arg \min_{\{\theta_k\}_1^K, \{\omega_k\}_1^K} \left\| \mathbf{A}_{ts,k}^H \mathbf{G} \right\|_F^2. \tag{32}$$

The multi-channel estimators will have a cost function which is more well-behaved compared to those of single channel multi-pitch estimators (see, e.g., [26,32,28] for some examples of such).

3.2. Refined estimates

For many applications, only a coarse estimate of involved fundamental frequencies and DOAs are needed, in which case the cost function in (32) is evaluated on pre-defined search region with some specified granularity. If, however, very accurate estimates are desired, a refined estimate can be found as described next. For a rough estimate of the parameter of interests, refined estimates are obtained by minimizing the cost function in (32) using a cyclic minimization approach. The gradient of the cost function (32) for fundamental frequency and DOA are given as

$$\frac{\partial}{\partial \omega_k} J(\omega_k, \theta_k) = 2\mathrm{Re}\left(\mathrm{Tr}\left\{ \mathbf{A}_{ts,k}^H \mathbf{G}\mathbf{G}^H \frac{\partial}{\partial \omega_k} \mathbf{A}_{ts,k} \right\} \right), \tag{33}$$

$$\frac{\partial}{\partial \theta_k} J(\omega_k, \theta_k) = 2\mathrm{Re}\left(\mathrm{Tr}\left\{ \mathbf{A}_{ts,k}^H \mathbf{G}\mathbf{G}^H \frac{\partial}{\partial \theta_k} \mathbf{A}_{ts,k} \right\} \right), \tag{34}$$

with $\mathrm{Re}\,(\cdot)$ denoting the real value. The gradient can be used for finding refined estimate using standard methods.

Here, we iteratively find a refined estimate using a cyclic approach. During an iteration, ω_k is first estimated with

$$\hat{\omega}_k^{i+1} = \hat{\omega}_k^i - \delta \frac{\partial}{\partial \omega_k} J(\hat{\omega}_k^i, \hat{\theta}_k^i), \tag{35}$$

where i is the iteration index and δ is a small positive constant that is found using line search. The estimated $\hat{\omega}_k^{i+1}$ is then used to initialize the minimization function for DOA, which is then found as

$$\hat{\theta}_k^{i+1} = \hat{\theta}_k^i - \delta \frac{\partial}{\partial \theta_k} J(\hat{\omega}_k^{i+1}, \hat{\theta}_k^i). \tag{36}$$

The method is initialized for $i = 0$ using the coarse estimates obtained from (32).

4. Experimental results

4.1. Signal examples

We start the experimental part of this article by illustrating the application of the proposed method to analyzing a mixed signal consisting of speech and clarinet signals, sampled at $f_s = 8000\,\text{Hz}$. The single-channel signals are converted into a multi-channel signal by introducing different delays according to two pre-determined DOA to simulate a microphone array with $M = 8$ channels. The simulated DOAs of the speech and the clarinet signals are, respectively, $\theta_1 = -45^\circ$ and $\theta_2 = 45^\circ$. The spectrogram of the mixed signal of the first channel is illustrated in Fig. 1. To avoid spatial ambiguities, the distance between two sensor is half the wavelength of the highest frequency in the observed signal, here $d = 0.0425\,\text{m}$. The mixed signal is segmented into 50% overlapped signal segments with $N = 128$. The user parameter selected in this experiment is $t = \lfloor \frac{2N}{3} \rfloor$ and $s = \lfloor \frac{M}{2} \rfloor$. The cost function is evaluated with a Vandermonde matrix with $L = 5$ complex exponentials, and the noise subspace is formed from an overestimated signal subspace with assumption of signal subspace containing $N/2 = 64$ complex exponentials. The signal subspace overestimation technique is usually used when the true order of the signal subspace is unknown, the signal subspace is assumed to be larger than the true one which can minimize the signal subspace components in the noise subspace. An added benefit of posing the problem as a joint estimation problem is that the multi-pitch estimation problem can be seen as several single-pitch problems for a distinct set of DOAs, one per source. Therefore, it is less important to select an exact signal model order than single-channel multi-pitch estimators would need [28]. The cost function is evaluated for frequencies from 100 to 500 with granularity of $0.52\,\text{Hz}$. The evaluated results are illustrated in Fig. 2 where the upper panel contains the fundamental frequency estimates and lower panel the DOA estimates. It can be seen that the proposed algorithm can track the fundamental frequency and the DOA of the speech signal well, with only a few observed errors on regions with low signal energy. The clarinet signal's DOA and fundamental frequencies have also been estimated well for all segments.

For the purpose of further comparison, the same signal will be analyzed using a standard time delay-and-sum beamformer [34] for DOA estimates and a single-channel maximum-likelihood based pitch estimator applied on the beamformed output signals [2]. The results are shown in Fig. 3. The figure clearly shows that the delay-sum beamformer cannot satisfactory resolve the DOAs with $M = 8$ array elements which will further affect the performance of the single-channel pitch estimator, as shown in the upper panel. In this example, the proposed algorithm shown in Fig. 2 is superior compared to reference method shown in Fig. 3. The low resolution performance of the reference method will make the statistical evaluation of this method uninteresting, and we, therefore, will not be using it any further in the experiments to follow.

4.2. Statistical evaluation

Next, we use Monte Carlo simulations evaluated on synthetic signals embedded in noise in assessing the statistical properties of the proposed method and compare it with the exact CRLB. As a reference method for pitch and DOA estimation, we use the JAFE algorithm proposed in [22] for jointly estimating unconstrained frequencies and DOAs. Next, the unconstrained frequencies are grouped according to their corresponding DOAs where closely related directions are grouped together. A fundamental frequency is formed from these grouped frequencies in a weighted way as proposed in [35]. We refer this as the WLS estimator. In order to remove the errors due to the erroneous estimate of amplitudes, we assume WLS having the exact signal amplitude given. The WLS estimator is a computationally efficient pitch estimation method with good statistical properties. The reference DOA estimate is easily obtained in a similar way from the mean value of these grouped DOAs according to [22].

Here, we consider a $M = 8$ element ULA with sensor distance $d = 0.0425$ with a sampling frequency of $f_s = 8000$. The estimators are evaluated for two signal setups, first with two sources having $\omega_1 = 252.123$ and $\omega_2 = 300.321$ with $L_{1,2} = 3$, and second with one harmonic source of $\omega_1 = 252.123$ and $L_1 = 3$. All amplitudes on individual harmonics are set to unity $A_{k,l} = 1$ for tractability. Both sources are assumed to be far-field sources impinging on the array with DOAs at $\theta_1 = -43.23^\circ$ and $\theta_2 = 70^\circ$, respectively, and for one source having a DOA of $\theta_1 = -43.23^\circ$. All simulation results are based on 100 Monte Carlo runs. The performance is measured using the root mean squared estimation error (RMSE) as defined in [28, 32, 26, 27]. The user parameter for JAFE data model is selected to the optimal values as proposed in [22] with temporal and spatial smoothness parameters, $t = \lfloor \frac{2N}{3} \rfloor$ and $s = \lfloor \frac{M}{2} \rfloor$, respectively. We note that in practical applications, the computational complexity has to also be considered in selecting the appropriate parameters t and s. An example of the 2-dimensional (2D) cost function of our proposed method evaluated on two mixed signal is illustrated in Fig. 4, where a coarser estimate of the DOA and fundamental estimates can be identified from the two peaks in the 2D cost function.

In the first simulation, we evaluate the proposed method's statistical properties in a single source scenario for varying sample lengths and SNRs. The RMSEs on signal with varying N are shown in Fig. 5, and with varying SNR in Fig. 6. It can be seen

from these figures that both estimators perform well for all SNR above $0\,\text{dB}$ with WLS being slightly better for fundamental frequency estimation while the proposed estimator is better in DOA estimation. Both methods are also able to follow CRLB closely for around sample length $N > 60$. The better DOA estimation capabilities of the proposed method can be explained by the joint estimation of the fundamental frequency and DOA, which leads to increased robustness under adverse conditions. Both estimators can be considered as consistent in the single-pitch scenario.

Next, we evaluate our method for the multi-pitch scenario. The so-obtained RMSEs for varying N and SNR are depicted in Figs. 7 and 8. In Fig. 7, it clearly shows that the proposed method is better than the WLS estimator for short sample lengths. The WLS estimator is not following CRLB until $N > 80$ samples while the proposed estimator is for $N > 64$. The remaining gap between CRLB and both evaluated estimators for $N > 80$ are due to the mutual interference between the harmonic sources. The slowly converging performance of WLS is mainly due to the bad estimate of the unconstrained frequency estimate using the JAFE method. With our selected simulation setup, the JAFE estimator is not giving consistent estimates for all harmonic components, which, in turn, results in poor performance in the WLS estimates. In general, the WLS estimator is sensitive to spurious estimate of the unconstrained frequencies. Moreover, the proposed estimator, which is jointly estimating both the DOA and the fundamental frequency, yields better estimates for smaller sample length N. The results in terms of RMSEs for varying SNRs are shown in Fig. 8. This figure shows that the proposed estimator is again more robust than the WLS estimator for both DOA and fundamental frequency estimation.

In next two experiments, we will study the performance as a function of the difference in fundamental frequencies and DOAs for multiple sources. We start with studying the RMSE as a function of the difference between the fundamental frequencies of two harmonic sources, i.e., $\Delta\omega = |\omega_1 - \omega_2|$, with $\theta_1 = -43.321^\circ$ and $\theta_2 = 70^\circ$. Here, we use an SNR set to $40\,\text{dB}$, and a sample length $N = 64$ with $M = 8$ array elements. The obtained RMSEs are shown in Fig. 9. The figure clearly shows that both methods can successfully estimate the fundamental frequencies and DOAs. Once again the proposed estimator gives more robust estimates, close to the CRLB. Additionally, it should be noted that both methods are correctly estimating the DOA even when the both fundamental frequencies are identical $\omega_1 = \omega_2$, something that would not be possible with only a single channel. MC-HMUSIC has the ability to estimate the fundamental frequencies when both harmonics are identical provided that the DOAs are distinct and vice versa. Estimation of the parameters of signals with overlapping harmonics is a crucial limitation in multi-pitch estimation using only single-channel recordings. In the final experiment, the RMSE as a function of the difference between the DOAs of two harmonic sources $\Delta\theta = |\theta_1 - \theta_2|$ is analyzed for an SNR set to $40\,\text{dB}$ and a sample length of $N = 64$ with $M = 8$ array elements. The fundamental frequencies are $\omega_1 = 252.123$ and $\omega_2 = 300.321$, respectively. The observations and conclusions are basically the same as before, with the proposed method outperforming the reference method so far.

5. Conclusion

In this article, we have generalized the single-channel multi-pitch problem into a multi-channel multi-pitch estimation problem. To solve this new problem, we propose an estimator for joint estimation of fundamental frequencies and DOAs of multiple sources. The proposed estimator is based on subspace analysis using a time–space data model. The method is shown to have potential in applications to real signals with simulated anechoic array recording, and a statistical evaluation demonstrates its robustness in DOA and fundamental frequency estimation as compared to a state-of-the-art reference method. Furthermore, the proposed method is shown to have good statistical performance under adverse conditions, for example for sources with similar DOA or fundamental frequency.

Competing interests

The authors declare that they have no competing interests.

Acknowledgment

The study of Zhang was supported by the Marie Curie EST-SIGNAL Fellowship, Contract No. MEST-CT-2005-021175.

References

[1] A Klapuri, Automatic music transcription as we know it today. J. New Music Res. **33**, 269–282 (2004)

[2] MG Christensen, A Jakobsson, *Multi-Pitch Estimation*, Synthesis Lectures on Speech and Audio Processing, 2009

[3] L Rabiner, On the use of autocorrelation analysis for pitch detection. IEEE Trans. Signal Process. **44**, 2229–2244 (1996)

[4] JX Zhang, MG Christensen, SH Jensen, M Moonen, A robust and computationally efficient subspace-based fundamental frequency estimator. IEEE Trans. Acoust. Speech Language Process. **18**(3), 487–497 (2010)

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

[6] DL Wang, GJ Brown, *Computational Auditory Scene Analysis: Principle, Algorithm, and Applications* (Wiley, IEEE Press, New York, 2006)

[7] A Klapuri, Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. IEEE Trans. Speech Audio Process. **11**, 804–816 (2003)

[8] V Emiya, D Bertrand, R Badeau, A parametric method for pitch estimation of piano tones, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2007, pp. 249–252

[9] S Rickard, O Yilmaz, Blind separation of speech mixtures via time–frequency masking. IEEE Trans. Signal Process. **52**, 1830–1847 (2004)

[10] M Wohmayr, M Kepsi, Joint position-pitch extraction from multichannel audio, in *Proceedings of the Interspeech*, 2007

[11] X Qian, R Kumaresan, Joint estimation of time delay and pitch of voiced speech signals, in *Record of the Asilomar Conference on Signals, Systems, and Computers*, vol. 2, 1996

[12] SN Wrigley, GJ Brown, Recurrent timing neural networks for joint F0-localisation based speech separation, in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007

[13] F Flego, M Omologo, Robust F0 estimation based on a multi-microphone periodicity function for distant-talking speech, in *EUSIPCO*, 2006

[14] L Armani, M Omologo, Weighted auto-correlation-based F0 estimation for distant-talking interaction with a distributed microphone network, in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2004, pp. 113–116

[15] D Chazan, Y Stettiner, D Malah, Optimal multi-pitch estimation using the em algorithm for co-channel speech separation, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1993

[16] G Liao, HC So, PC Ching, Joint time delay and frequency estimation of multiple sinusoids, in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, 2001, pp. 3121–3124

[17] Y Wu, HC So, Y Tan, Joint time-delay and frequency estimation using parallel factor analysis. Elsevier Signal Process. **89**, 1667–1670 (2009)

[18] LY Ngan, Y Wu, HC So, PC Ching, SW Lee, Joint time delay and pitch estimation for speaker localization, in *Proceedings of the IEEE International Symposium on Circuits and Systems*, 2003, pp. 722–725

[19] P Stoica, R Moses, *Spectral Analysis of Signals* (Prentice-Hall, Upper Saddle River, 2005)

[20] M Brandstein, D Ward, *Microphone Arrays* (Springer, Berlin, 2001)

[21] AJ van der Veen, M Vanderveen, A Paulraj, Joint angle and delay estimation using shift invariance techniques. IEEE Trans. Signal Process. **46**, 405–418 (1998)

[22] AN Lemma, AJ van der Veen, EF Deprettere, Analysis of joint angle-frequency estimation using ESPRIT. IEEE Trans. Signal Process. **51**, 1264–1283 (2003)

[23] M Viberg, P Stoica, A computationally efficient method for joint direction finding and frequency estimation in colored noise, in *Record of the Asilomar Conference on Signals, Systems, and Computers*, vol. 2, 1998, pp. 1547–1551

[24] JD Lin, WH Fang, YY Wang, JT Chen, FSF MUSIC for joint DOA and frequency estimation and its performance analysis. IEEE Trans. Signal Process. **54**, 4529–4542 (2006)

[25] S Wang, J Caffery, X Zhou, Analysis of a joint space–time doa/foa estimator using MUSIC, in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, 2001, pp. B138–B142

[26] MG Christensen, P Stoica, A Jakobsson, SH Jensen, Multi-pitch estimation. Elsevier Signal Process. **88**(4), 972–983 (2008)

[27] MG Christensen, A Jakobsson, SH Jensen, Joint high-resolution fundamental frequency and order estimation. IEEE Trans. Acoust. Speech Signal Process. **15**(5), 1635–1644 (2007)

[28] JX Zhang, MG Christensen, SH Jensen, M Moonen, An iterative subspace-based multi-pitch estimation algorithm. Elsevier Signal Process. **91**, 150–154 (2011)

[29] AN Lemma, ESPRIT based joint angle-frequency estimation algorithms and simulations. PhD Thesis, Delft University, 1999

[30] T Shu, XZ Liu, Robust and computationally efficient signal-dependent method for joint DOA and frequency estimation. EURASIP J. Adv. Signal Process., vol. **2008**, Article ID 10.1155/2008/134853 (2008)

[31] H Krim, M Viberg, Two decades of array processing research-the parametric approach. IEEE SP. Mag. (1996)

[32] MG Christensen, A Jakobsson, SH Jensen, Multi-pitch estimation using Harmonic MUSIC, in *Record of the Asilomar Conference on Signals, Systems, and Computers*, 2006, pp. 521–525

[33] MG Christensen, A Jakobsson, SH Jensen, Sinusoidal order estimation using angles between subspaces. EURASIP J. Adv. Signal Process., Article ID 948756, 1–11 (2009)

[34] BDV Veen, KM Buckley, Beamforming: a versatile approach to spatial filtering. IEEE ASSP Mag. **5**, 4–24 (1988)

[35] H Li, P Stoica, J Li, Computationally efficient parameter estimation for harmonic sinusoidal signals. Elsevier Signal Process. 1937–1944 (2000)

Fig. 1: **The mixed spectrogram of the real recorded speech and clarinet signal.**

Fig. 2: **The estimation results using the proposed methods**: **(a)** fundamental frequency, **(b)** the DOA with the horizontal axis denoting time axis.

Fig. 3: The estimates of **(a)** the fundamental frequency using maximum-likelihood estimator at the output of the beamformer, **(b)** the DOA using a delay-sum beamformer.

Fig. 4: **Example of cost functions for two synthetic sources having three harmonic each,** $N = 64$ **and** $M = 8$**.** The true fundamental frequency of $\omega_1 = 252.123$ and $\omega_2 = 300.321$ having DOA $\theta_1 = -43.23^\circ$ and $\theta_2 = 70^\circ$, respectively.

Fig. 5: **RMSE as a function of N for SNR $= 40$ dB evaluated on single-pitch signal with unit amplitude**: **(a)** fundamental frequency estimates; **(b)** DOA estimates.

Fig. 6: **RMSE as a function of SNR for $N = 64$ evaluated on single-pitch signal with unit amplitudes**: **(a)** fundamental frequency estimates; **(b)** DOA estimates.

Fig. 7: **RMSE as a function of N for SNR $= 40$ dB evaluated on multi-pitch signal with unit amplitudes**: **(a)** joint fundamental frequency estimates; **(b)** joint DOA estimates.

Fig. 8: **RMSE as a function of SNR for $N = 64$ evaluated on multi-pitch signal with unit amplitudes**: **(a)** joint fundamental frequency estimates; **(b)** joint DOA estimates.

Fig. 9: **RMSE as a function of $\Delta\omega$**: **(a)** joint fundamental frequency estimates; **(b)** joint DOA estimates.

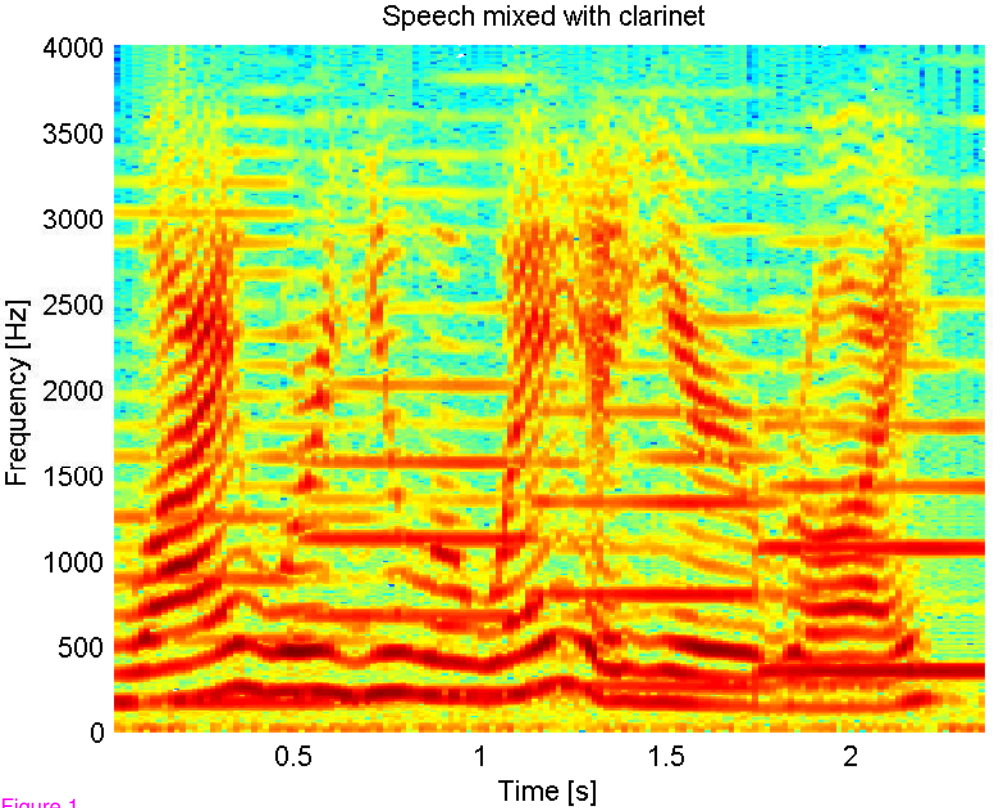

Speech mixed with clarinet
4000
3500
3000
2500
2000
1500
1000
500
0
Frequency [Hz]
0.5
1
1.5
2
Time [s]

Figure 1

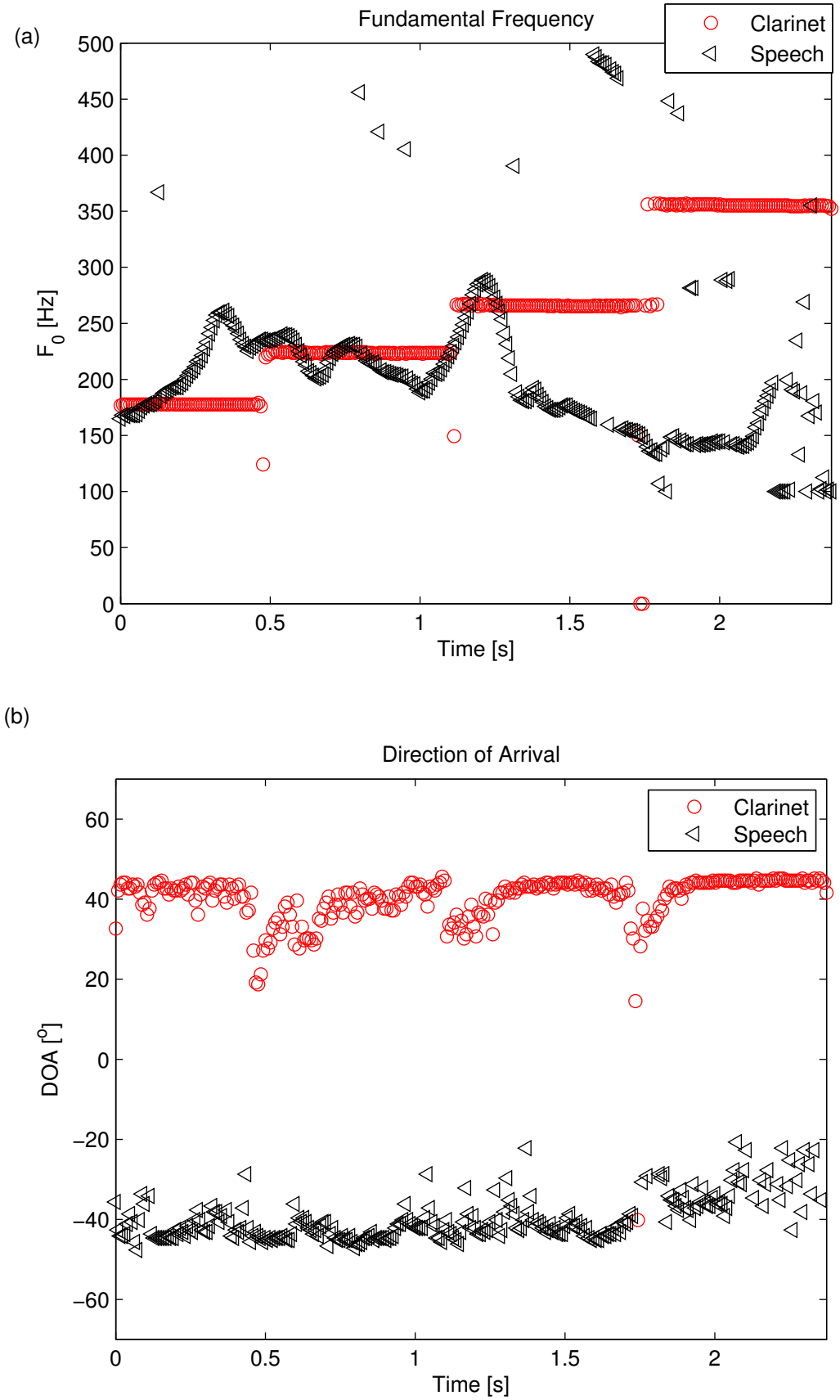
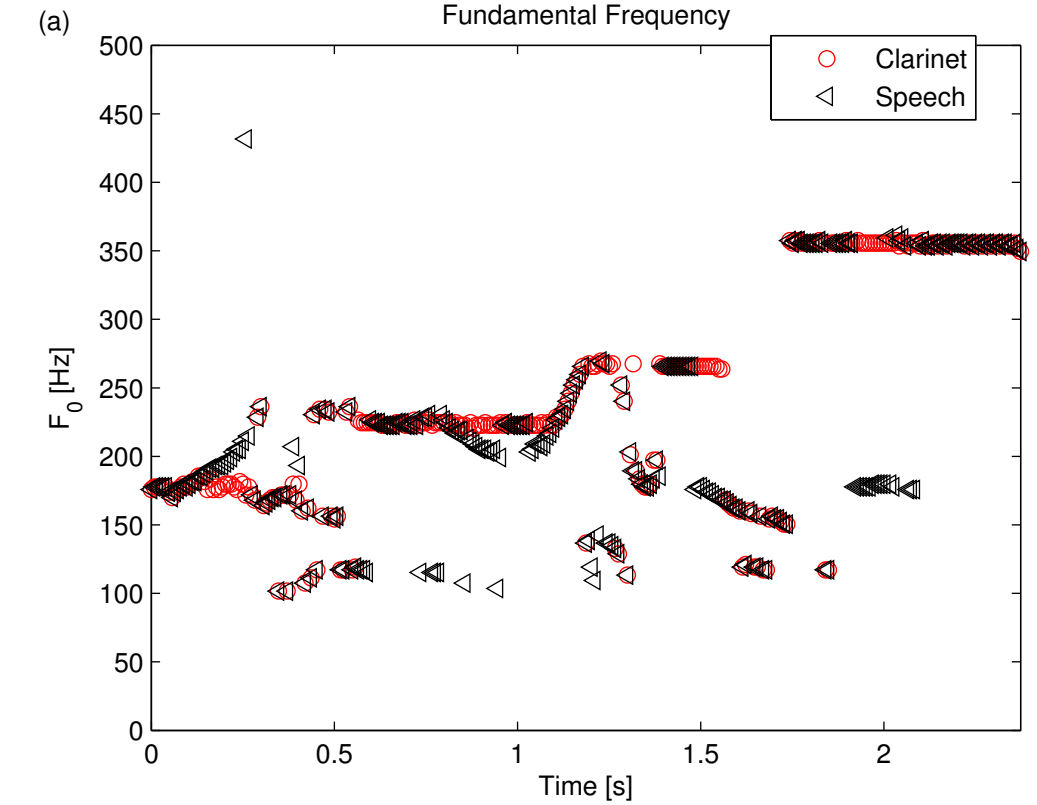

(a)
Fundamental Frequency
Clarinet
Speech
500
450
400
350
300
250
200
150
100
50
0
F0 [Hz]
0
0.5
1
1.5
2
Time [s]
(b)
Direction of Arrival
Clarinet
Speech
60
40
20
0
−20
−40
−60
DOA [°]
0
0.5
1
1.5
2
Time [s]

Figure 2


(a)
Fundamental Frequency
Clarinet
Speech
500
450
400
350
300
250
200
150
100
50
0
F0 [Hz]
0
0.5
1
1.5
2
Time [s]

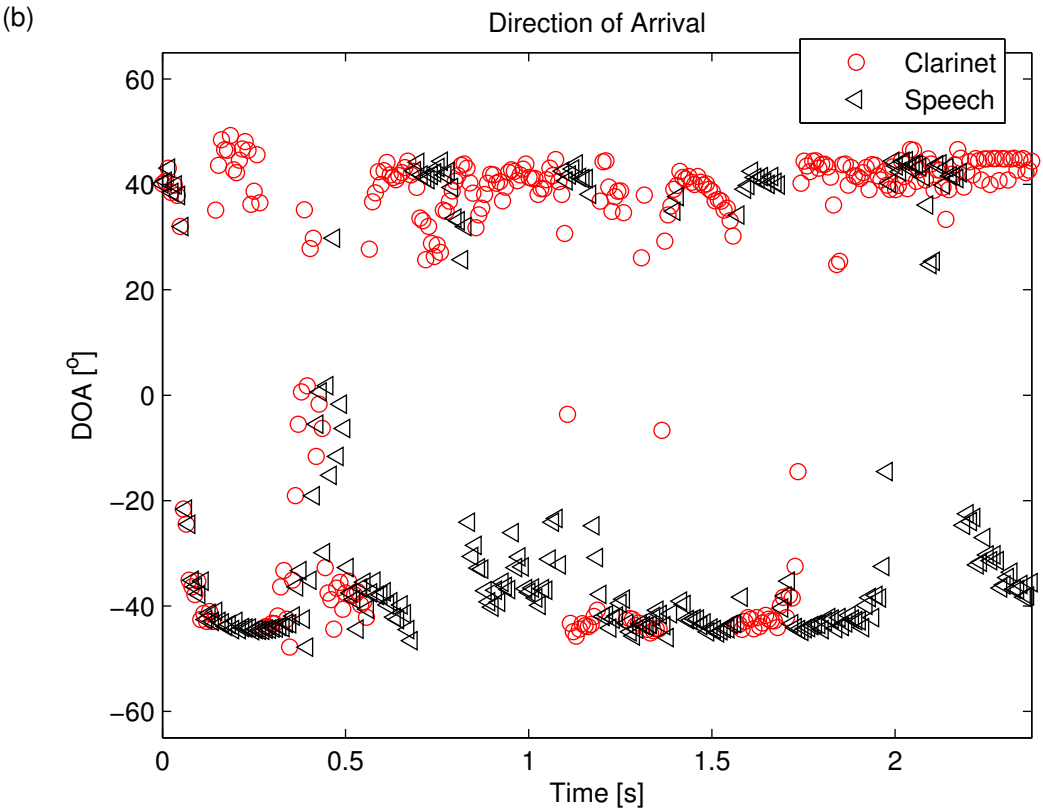

(b)
Direction of Arrival
Clarinet
Speech
60
40
20
0
−20
−40
−60
DOA [°]
0
0.5
1
1.5
2
Time [s]

Figure 3

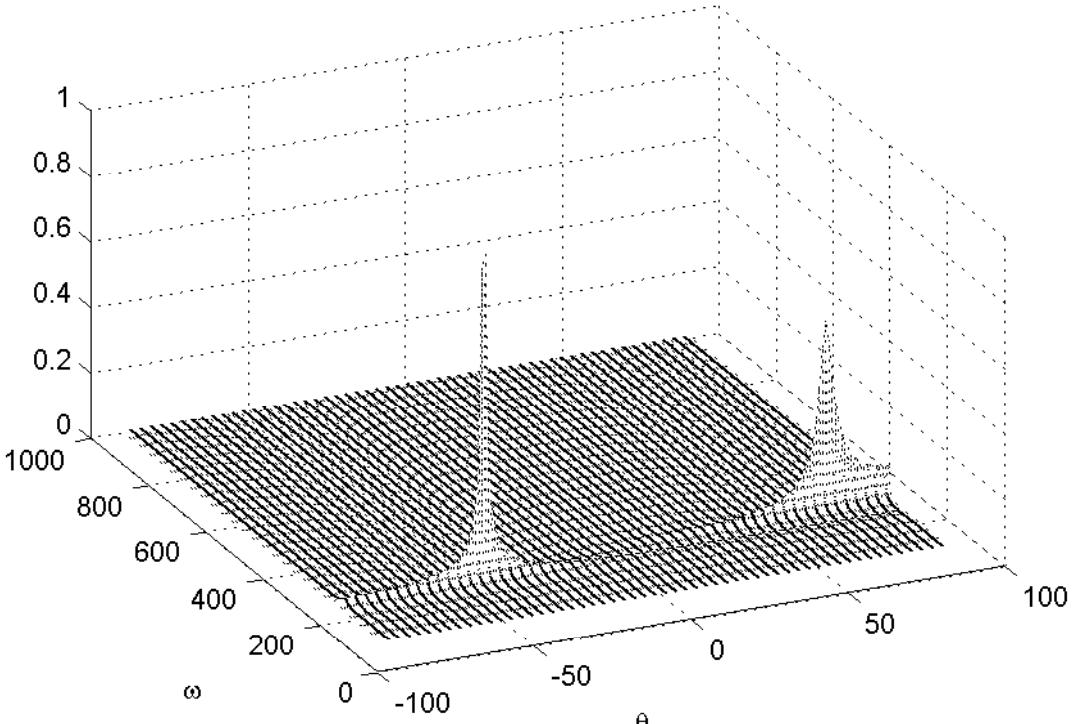
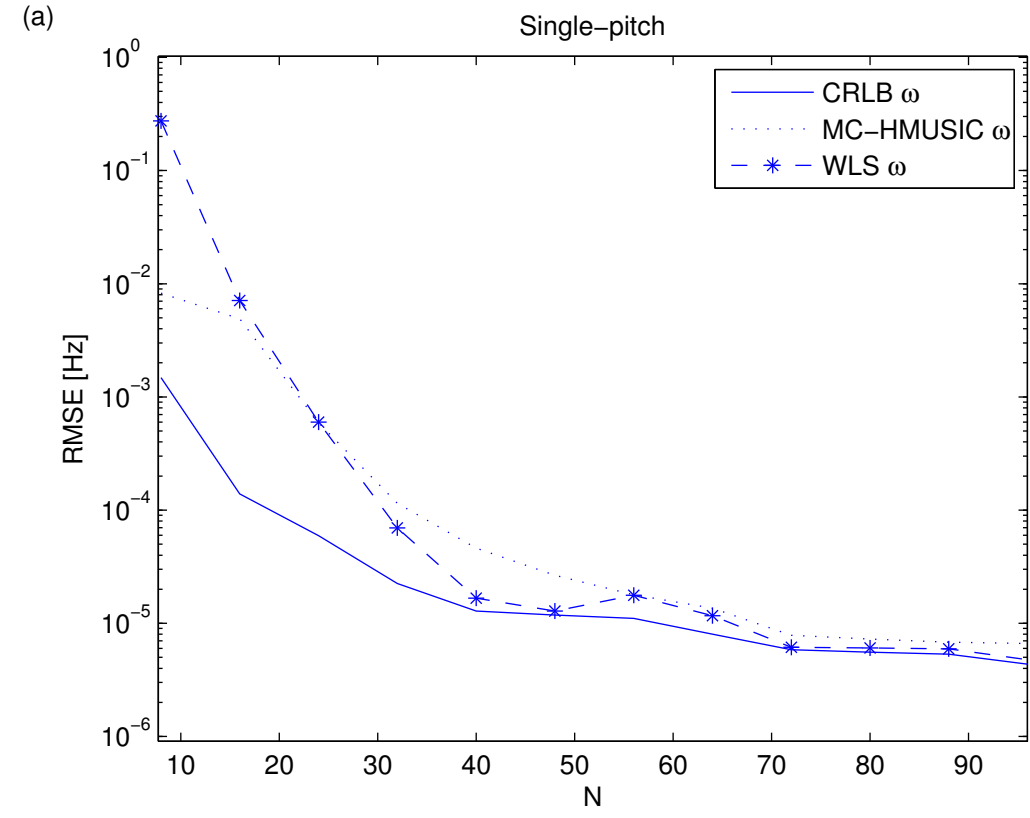


Figure 4


(a)
Single−pitch
CRLB ω
MC−HMUSIC ω
WLS ω
RMSE [Hz]
N

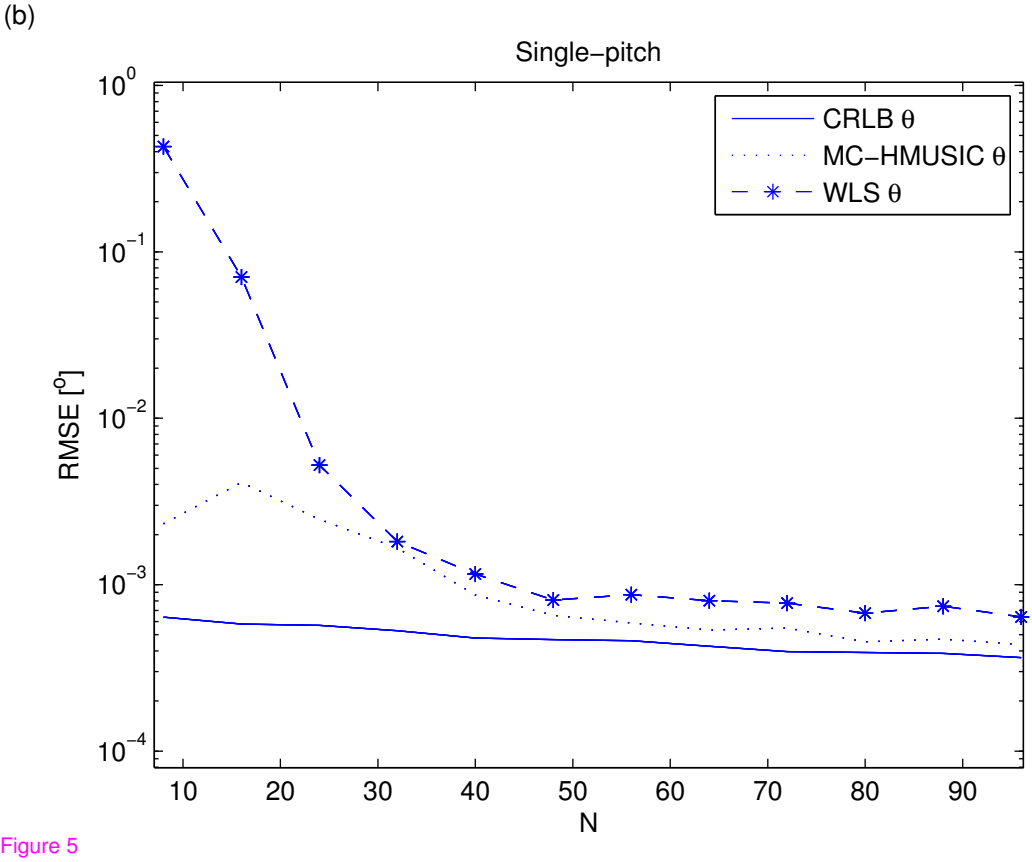
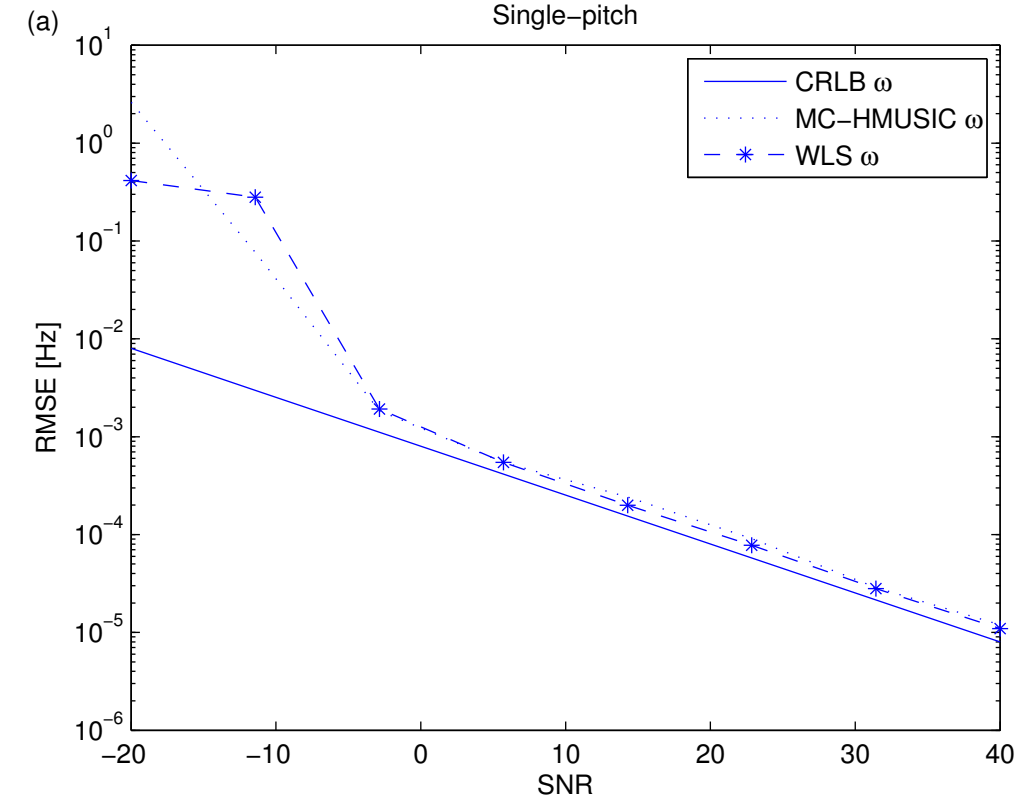

(b)
Single−pitch
CRLB θ
MC−HMUSIC θ
WLS θ
RMSE [°]
N

Figure 5


(a)
Single−pitch
CRLB ω
MC−HMUSIC ω
WLS ω
RMSE [Hz]
SNR
10^1
10^0
10^−1
10^−2
10^−3
10^−4
10^−5
10^−6
−20
−10
0
10
20
30
40

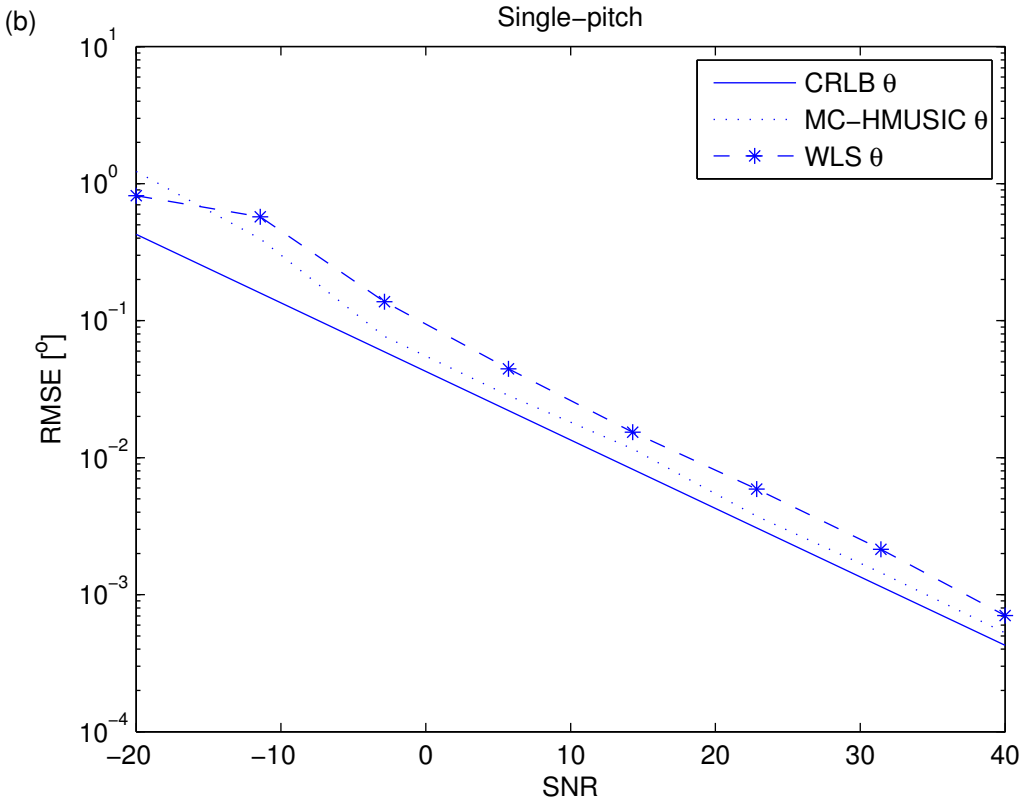
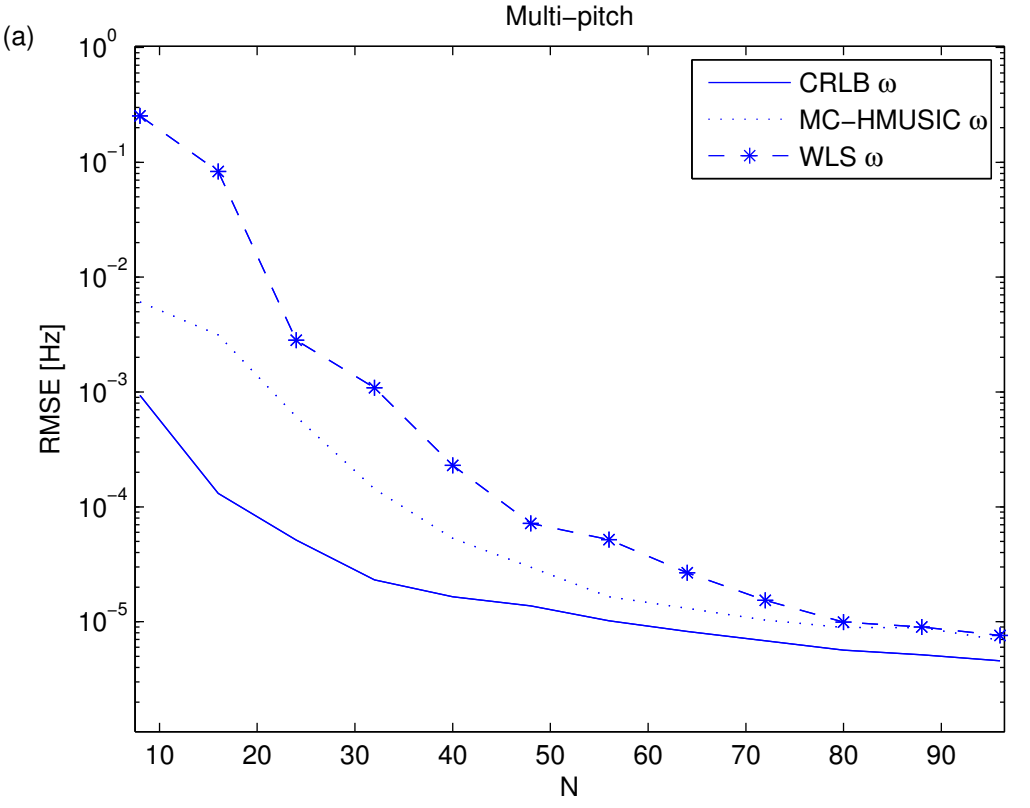

(b)
Single−pitch
CRLB θ
MC−HMUSIC θ
WLS θ
RMSE [°]
SNR
10^1
10^0
10^−1
10^−2
10^−3
10^−4
−20
−10
0
10
20
30
40

Figure 6


(a)
Multi−pitch
CRLB ω
MC−HMUSIC ω
WLS ω
RMSE [Hz]
N

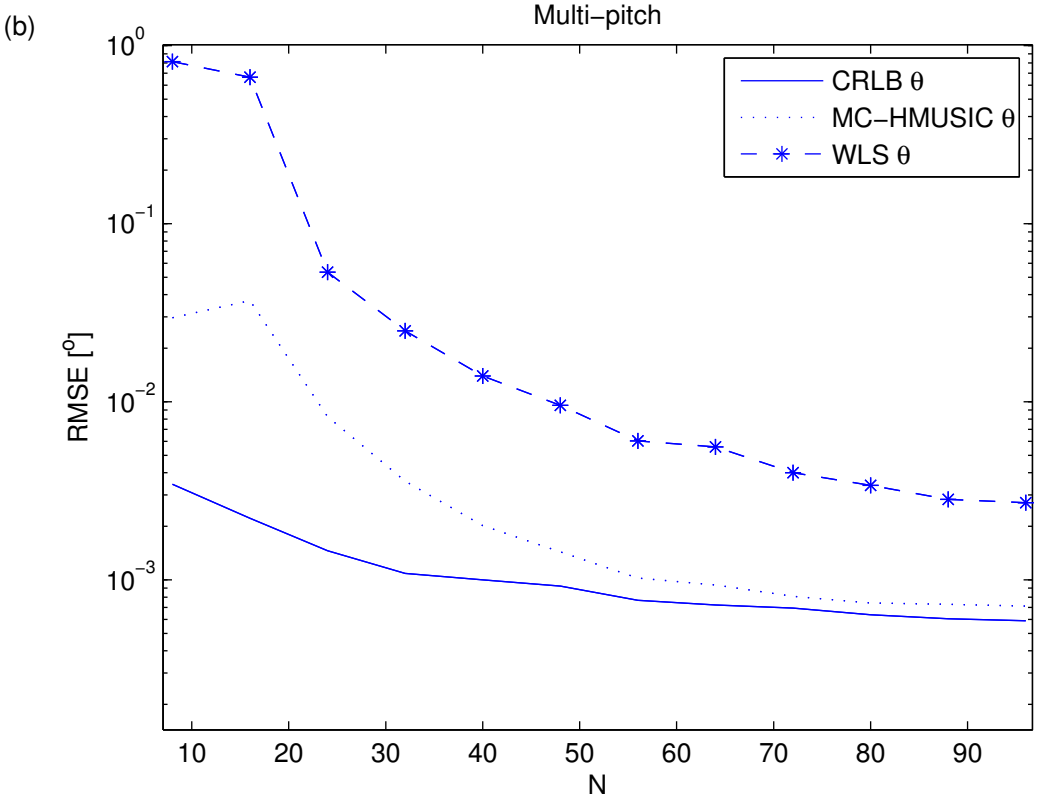
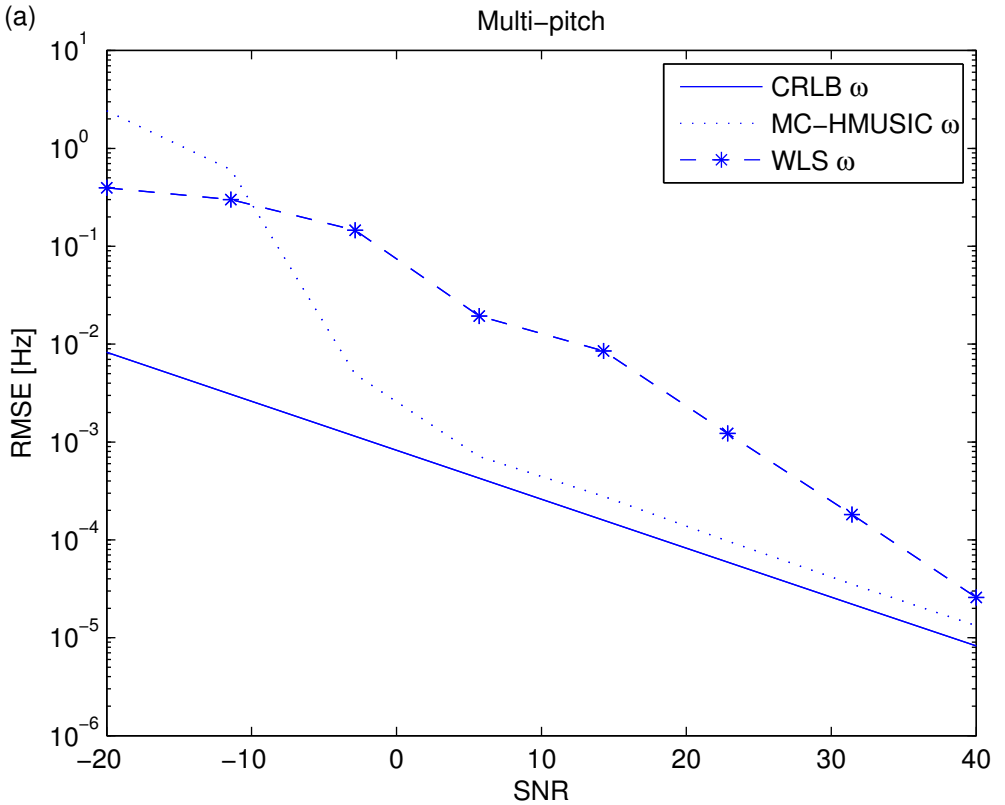

(b)
Multi−pitch
CRLB θ
MC−HMUSIC θ
WLS θ
RMSE [°]
N

Figure 7


(a)
Multi−pitch
CRLB ω
MC−HMUSIC ω
WLS ω
RMSE [Hz]
SNR

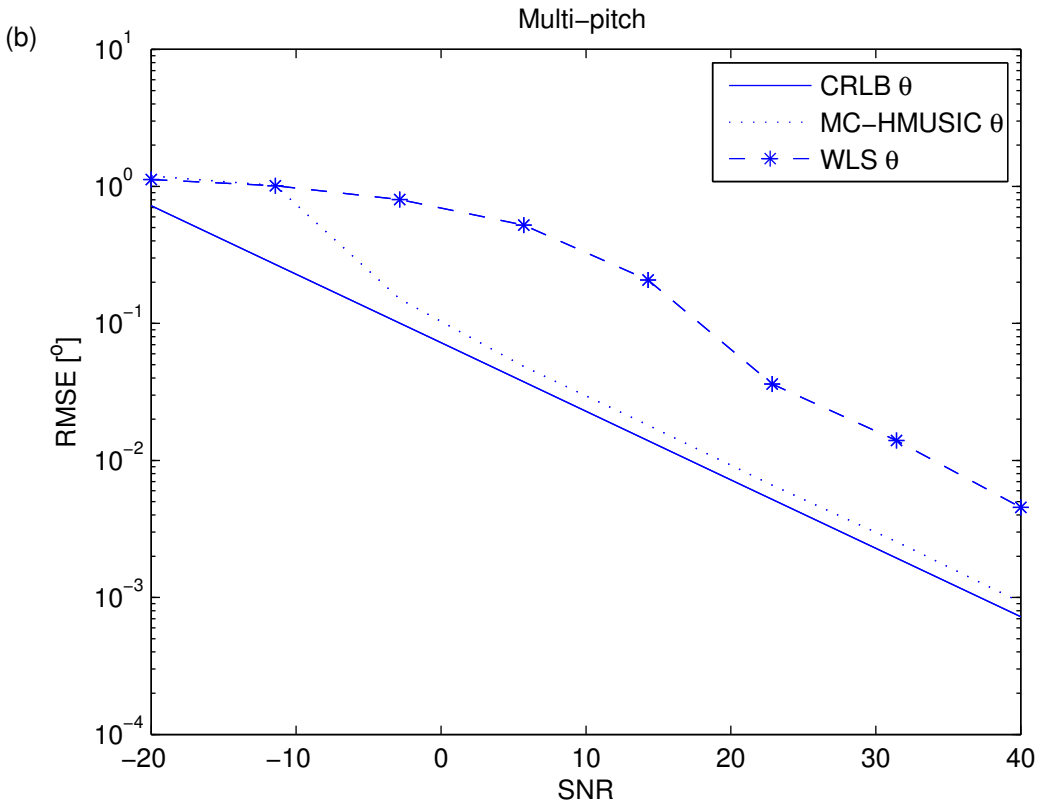

(b)
Multi−pitch
CRLB θ
MC−HMUSIC θ
WLS θ
RMSE [°]
SNR

Figure 8

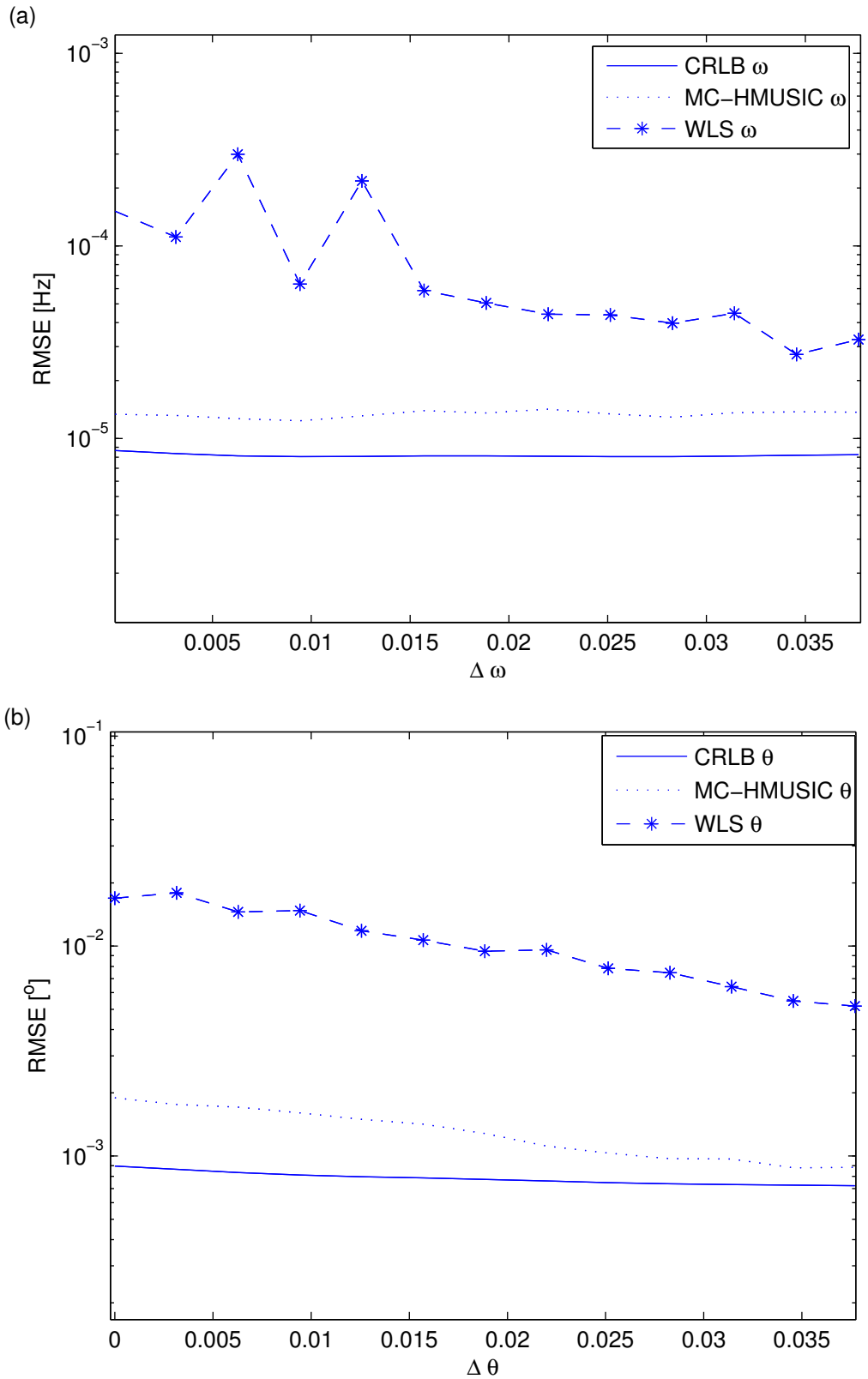

(a)
CRLB ω
MC−HMUSIC ω
WLS ω
RMSE [Hz]
Δ ω
(b)
CRLB θ
MC−HMUSIC θ
WLS θ
RMSE [°]
Δ θ

Figure 9