Maximum $A$-Posteriori Probability Multiple Pitch Tracking Using the Harmonic Model

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE M.Sc. DEGREE

By Amitai Koretz

Supervised by: Dr. Joseph Tabrikian

December 2009
Abstract

In this thesis, a new method for multiple fundamental frequency estimation for speech and music signals is proposed. Applications of audio and speech processing include many well-reviewed algorithms for estimating the fundamental frequency of monophonic speech and music signals. In the case of polyphonic signals, it is more difficult to successfully estimate each of the fundamental frequencies, as reflected by the dearth of existing methods addressing this problem. In this work, a new method based on the combination of the maximum likelihood (ML) and maximum \textit{a-posteriori} probability (MAP) criteria is derived for fundamental frequencies tracking where each one of the fundamental frequencies is modeled by a first-order Markov process. The dominant signal is modeled as a harmonic source with unknown deterministic amplitudes, while the remaining signals, including other harmonic signals, are modeled as Gaussian interference sources with an unknown covariance matrix. After estimation of the dominant source, it is removed from the signal by projection of the signal into the null subspace spanned by the estimated signal. This procedure is iterated for all the harmonic sources in the data. The algorithm is tested with speech, music, and synthetic signals where in each case, two harmonic sources of the same kind were mixed. The performance of the proposed algorithm is evaluated and compared to an existing reference method in terms of root-mean-square error (RMSE) and gross-error-rate (GER) as a function of signal-to-interference ratio (SIR).
Acknowledgment

I am deeply indebted to my advisor, Dr. Joseph Tabrikian, for his guidance, dedication, assistance, and especially his patience and constant support.
# Contents

1 Introduction .......................................................... 1
   1.1 Introduction ................................................................ 1

2 Background ..................................................................... 3
   2.1 Harmonic Model - Review ....................................... 3
   2.2 Problem Definition Using the Harmonic Model .......... 4
   2.3 Multidimensional Maximum-Likelihood .................... 4
   2.4 MUSIC ........................................................................ 5
      2.4.1 MUSIC Criterion ............................................. 6
      2.4.2 Single-Pitch Estimation ................................ 8
      2.4.3 Multi-Pitch Estimation ................................. 10
      2.4.4 Root-MUSIC .................................................. 12
      2.4.5 MUSIC - Discussion ...................................... 15
      2.4.6 Equalized MUSIC ........................................ 19

3 The Proposed ML-MAP Method ..................................... 22
3.1 MVDR / ML Estimator .................................................. 23
   3.1.1 MVDR Criterion .................................................. 23
   3.1.2 ML DOA Estimation in the Presence of Unknown Gaussian Inter-
        ference ............................................................. 24
3.2 Single Source Pitch Tracking ......................................... 25
   3.2.1 Single Harmonic Source + Interference Model ................. 25
   3.2.2 Log-likelihood for Single Source + Interference Model ....... 26
   3.2.3 Forward-Backward Tracking .................................... 30
3.3 Multiple Pitch Tracking With Projection .......................... 36
   3.3.1 Algorithm Outline .............................................. 39

4 Simulation Results .......................................................... 41
   4.1 Synthetic Signals ................................................... 41
   4.2 Real Signals ....................................................... 42
      4.2.1 Example 1 - Log-likelihood Illustration ...................... 44
      4.2.2 Example 2 - Speaker Combinations .......................... 46
      4.2.3 Example 3 - Music Combination ............................. 48

5 Discussion and Conclusions ............................................. 54
   5.1 Discussion and Conclusions ....................................... 54

A Narrowband model used in array processing .......................... 61
Abbreviations

AWGN - Additive white Gaussian Noise
ML - Maximum likelihood
MAP - Maximum \textit{a-posteriori} Probability
MSE - Mean-square error
MUSIC - MUltiple SIgnal Classification
PDF - Probability density function
RMSE - Root mean-square error
SIR - Signal-to-interference ratio
SNR - Signal-to-noise ratio
SVD - Singular value decomposition
Notations

$(\cdot)^{-1}$ - Matrix inverse

$(\cdot)^T$ - Matrix transpose

$(\cdot)^*$ - Conjugate

$(\cdot)^H$ - Conjugate transpose (Hermitian)

$E(\cdot)$ - Expectation

diag$(\cdot)$ - Diagonal matrix with the same diagonal as its matrix argument

$| \cdot |$ - Absolute value of a scalar or determinant of a matrix

$\| \cdot \|_F$ - Frobenius norm

$I_L$ - Identity matrix of size $L$

$[A]_{ij}$ - $(i, j)$th element of matrix $A$

$n$ - Discrete time index

t - Continuous time index

$\hat{\theta}$ - Estimation of $\theta$

$L$ - Number of subframe samples

$N$ - Number of frame samples

$K$ - Number of real harmonics

$M$ - Number of harmonic sources

$Q$ - Number of frames

$J$ - Number of subframes
$Z$ - Length of frequency grid

$\gamma$ - Vector of fundamental frequencies for a single harmonic source

$A(\gamma)$ - $N \times 2K$ harmonic matrix with fundamental frequency $\gamma$

$\tilde{A}(\gamma)$ - $L \times 2K$ harmonic matrix with fundamental frequency $\gamma$

$A_m(\gamma_m^{(q)})$ - $N \times 2K$ harmonic matrix at the $m$th iteration

$\tilde{A}_m(\gamma_m^{(q)})$ - $L \times 2K$ harmonic matrix at the $m$th iteration

$x_m^{(q)}$ - Frame measurement vector for the $q$th frame at the $m$th iteration

$\tilde{X}_m^{(q)}$ - Subframes matrix for the $q$th frame at the $m$th iteration

$f_s(\cdot)$ - Probability density function of the vector $s$

$P_{\tilde{A}_m}^\perp(\gamma_m^{(q)})$ - $L \times L$ harmonic projection matrix for the $q$th frame at the $m$th iteration

$P_{\tilde{A}_m}^\perp(\gamma_m^{(q)})$ - $N \times N$ harmonic projection matrix for the $q$th frame at the $m$th iteration

$b_m$ - $2K \times 1$ vector of complex amplitudes for the $m$th harmonic source

$e_m^{(q)}$ - $N \times 1$ interference vector for the $q$th frame at the $m$th iteration

$R_o$ - $L \times L$ covariance matrix for the interference vector

$L''_{x_m^{(q)}}(\gamma^{(q)})$ - Log-likelihood for the $q$th frame

$f_{\gamma^{(q)}|\gamma^{(q-1)}}$ - transitional $a$-priori pdf for $\gamma^{(q)}$ given $\gamma^{(q-1)}$

$G_m(\gamma_m^{(q)})$ - $N \times 2Km$ concatenated harmonic matrix

$P_{G_m}^\perp(\gamma_m^{(q)})$ - $N \times N$ harmonic projection matrix

$P$ - $Z \times Z$ probability transition matrix

$L_{x_m}$ - $Z \times Q$ log-likelihood matrix at the $m$th iteration

$W_m$ - $Z \times Q$ tracking log-likelihood matrix at the $m$th iteration

$U_{\tilde{A}}$ - orthonormal basis for subspace spanned by the columns of $\tilde{A}$
$T_{\tilde{A}}$ - Orthonormal basis for complement of the subspace spanned by the columns of $\tilde{A}$
List of Figures

2.1 Multi-pitch estimation with MUSIC. Two synthetic signals were mixed with AWGN with an SNR of 100dB. Plots (A) and (B) show the amplitude spectra of sources A and B, respectively, each showing 5 harmonics. Plot (C) displays the MUSIC spectrum, exhibiting peaks at the correct frequencies of 150Hz and 210Hz. ........................................ 13

2.2 Multi-pitch estimation with MUSIC. Two synthetic signals were mixed with AWGN with an SNR of 5dB. Plots (A) and (B) show the amplitude spectra of sources A and B, respectively, each showing 5 harmonics. Plot (C) displays the MUSIC spectrum, exhibiting lower peaks than the case with high SNR. ..................................................... 14

2.3 Multi-pitch estimation with MUSIC using real signals. Plots (A) and (B) show the amplitude spectra of sources A (female) and source B (male), respectively. Plot (C) displays the MUSIC spectrum, exhibiting peaks around 96Hz and 193Hz, which are close to the location of the peaks in the spectrograms. .......................................................... 15
2.4 Multi-pitch estimation with MUSIC. Two synthetic signals were mixed with AWGN with an SNR of 100dB. Plots (A) and (B) show the amplitude spectra of sources A and B, respectively, where the 3rd harmonic source B is attenuated by 30dB. Plot (C) displays the MUSIC spectrum, and the peak corresponding to source B is more than 50dB lower than the other peak. .................................................. 17

2.5 Multi-pitch estimation with MUSIC using real signals. Plots (A) and (B) show the amplitude spectra of sources A (female) and source B (male), respectively. Plot (C) displays the MUSIC spectrum, and the peak corresponding to source B is low, since the source exhibits weak harmonicity. . 18

2.6 Equalized versus regular MUSIC. Two synthetic signals were mixed with AWGN with an SNR of 100dB. Plot (A) displays the regular MUSIC spectrum and plot (B) displays the equalized MUSIC spectrum. The dots indicate the peak locations. .................................................. 21

4.1 Two synthetic signals were mixed with AWGN at an SNR of 10dB. Plots (A) and (B) display the amplitude spectra of the signals, plot (C) displays the log-likelihood and plot (D) displays the projected log-likelihood after maximizing the log-likelihood. The dots indicate the true fundamental frequencies. .................................................. 43
4.2 Two synthetic signals were mixed with AWGN at an SNR of 10dB. Plots (A) and (B) display the amplitude spectra of the signals, plot (C) displays the log-likelihood and plot (D) displays the projected log-likelihood after maximizing the log-likelihood. The dots indicate the true fundamental frequencies.

4.3 Analysis of male and female speakers. The spectrograms are shown in row 1, the log-likelihood functions are shown in row 2 and the tracking log-likelihood functions are shown in row 3. The pitch tracking lines are shown for each plot. The tracking log-likelihood takes values within the range specified by the pdf window.

4.4 Spectrogram for female speaker with pitch reference (solid line). The estimated pitch (open circles) is shown: by maximizing the log-likelihood in plot (A), by using forward tracking in plot (B), by using forward-backward tracking in plot (C).

4.5 Spectrogram for male speaker with pitch reference (solid line). The estimated pitch (open circles) is shown: by maximizing the log-likelihood in plot (A), by using forward tracking in plot (B), by using forward-backward tracking in plot (C).

4.6 Male-Female analysis using the proposed method and the reference method (Klapuri). The GER and RMSE for each source are shown as a function of SIR, averaged over 400 independent simulations.
4.7 Male-Male analysis using the proposed method and the reference method (Klapuri). The GER and RMSE for each source are shown as a function of SIR, averaged over 400 independent simulations. .......................... 51

4.8 Female-Female analysis using the proposed method and the reference method (Klapuri). The GER and RMSE for each source are shown as a function of SIR, averaged over 400 independent simulations. .......................... 52

4.9 Music analysis using the proposed method and the reference method (Klapuri). The GER and RMSE for each source are shown as a function of SIR, averaged over 67 independent simulations. .......................... 53
# List of Tables

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Simulation Parameters for Synthetic Signals</td>
<td>42</td>
</tr>
<tr>
<td>4.2</td>
<td>Simulation Parameters for Example 1</td>
<td>45</td>
</tr>
<tr>
<td>4.3</td>
<td>Simulation Parameters for Example 2</td>
<td>49</td>
</tr>
<tr>
<td>4.4</td>
<td>Simulation Parameters for Example 3</td>
<td>49</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Introduction

Estimation of multiple fundamental frequencies is of interest in a wide range of practical problems, as knowledge of the fundamental frequency, commonly perceived as pitch, is of crucial importance in speech and audio signal processing. Methods for the estimation of the fundamental frequency in the presence of a single harmonic source have been the subject of extensive research, resulting in multiple classes of algorithms, including time-domain algorithms [1], [2], [3] such as methods based on the autocorrelation function and methods based on zero-crossings, and frequency-domain algorithms, such as those based on cepstral analysis [4]. Algorithms designed for single-pitch estimation usually assume the presence of a single harmonic source, and therefore generally do not yield meaningful results in the case of multiple harmonic sources.

Estimating each pitch of mixed periodic signals has proved to be a more difficult problem than single-pitch estimation. Early work in the field of multiple-pitch estimation sought to accomplish automatic transcription of polyphonic music, such as duets and more complex compositions, by identifying the fundamental frequencies of the individual instrumental voices [5]. More recent pitch estimation methods for speech and music rely on psychoacoustical theory and techniques which employ preprocessing in order to
mimic human aural perception, deriving inspiration from the human facility to effectively discriminate between different harmonic voices or sounds. This class of algorithms [6] is based on multichannel bandpass filtering of the signal, performing processing and periodicity extraction at the channel level, and finally combining the data over multiple channels to form a period estimate. This type of perceptual model can account for several psychoacoustical phenomena, such as pitch perception in the absence of a fundamental frequency [7]. An iterative-based approach, in which the fundamental frequency of the predominant periodic signal is estimated and removed prior to estimation of the next fundamental frequency, was combined with a perceptual model in [8]. This process is repeated for all harmonic sources in the signal. A simplification of the multichannel model was proposed in [9] and uses only two channels, lowering the overall complexity of the algorithm. Recent algorithms based on a perceptual model include Klapuri [10], where an iterative cancellation technique is also used, while relying on the assumption that amplitudes of adjacent harmonics have are highly correlated. This assumption may be valid for many musical signals, though time-frequency analysis of speech signals shows that this is not the case in general. A subsequent method proposed by Klapuri [11] shares some of the characteristics of the former and is suitable for speech as well as music. A further paper by the same author [12] describes a similar method which is simpler to implement and was evaluated for music samples. Another class of methods relies on a probabilistic framework, typically aided by the harmonic model [13]. Such methods were proposed by Christensen et al. [14]. Harmonic MUSIC (MUltiple SIgnal Classification) was proposed [15] along with the harmonic model for multiple fundamental frequency estimation, although this approach suffers from a lack of robustness to model mismatch. Bayesian models were used in [16] for tracking harmonic components based on a smoothness assumption.
Chapter 2

Background

In this chapter, we discuss probabilistic approaches that rely on a measurements model for the signal samples (both using the harmonic model). Specifically, multidimensional maximum-likelihood and MUSIC-based algorithms are discussed. It will be shown that ultimately neither are successful at performing multiple pitch estimation from a practical point of view.

2.1 Harmonic Model - Review

Assume a harmonic real signal $x(t)$ with $K$ real harmonics and no DC level sampled at a rate of $F_s$ so that $x[n] = x(nT_s)$ where $T_s = \frac{1}{F_s}$. The vector of length $N$ containing the signal samples $x = [x[0], x[1], \ldots, x[N-1]]^T$ may be expressed as

$$x = A(\gamma_0)b$$

where $A(\gamma_0) = [a_{-K}(\gamma_0), \ldots, a_{-1}(\gamma_0), a_1(\gamma_0), \ldots, a_K(\gamma_0)]^T$ is the $N \times 2K$ harmonic matrix with fundamental frequency $\gamma_0$, and $b = [b_{-K}, \ldots, b_{-1}, b_1, \ldots, b_K]^T$ is the vector of complex amplitudes. The $k$th column of the harmonic matrix $A(\gamma_0)$ is defined as

$$a_k(\gamma_0) \triangleq [1, e^{j2\pi\gamma_0 kT_s}, \ldots, e^{j2\pi\gamma_0(N-1)kT_s}]^T$$

(2.2)
where $\gamma_0$ is the fundamental frequency.

It can be pointed out that the harmonic model in (2.1) has redundant degrees of freedom. When dealing with real $x$, the vector of complex amplitudes $b$ satisfies conjugate symmetry so that $[b]_i = [b]_{2K-i+1}^*$, $i = 1, \ldots, K$. The harmonics corresponding to negative frequencies could theoretically be discarded, resulting in a more compact model [17]. In practice, reducing the frequency information by expunging information at negative frequencies leads to distortion of the spectrum. This is due to the fact that the signal is with finite support, so complex harmonic at positive frequencies contain some spectral energy in the negative half of the spectrum as well.

### 2.2 Problem Definition Using the Harmonic Model

Consider an audio frame with $M$ harmonic sources. Using the harmonic model, the vector of signal samples may be represented as

$$x = \sum_{m=1}^{M} A(\gamma_m) b_m$$

(2.3)

where $b_m$ is the vector of complex amplitudes for the $m$th harmonic source. The objective is to estimate the vector of fundamental frequencies $\gamma \triangleq [\gamma_1, \ldots, \gamma_M]^T$. This applies to a short frame of input data where the fundamental frequency is nearly constant. In practice, the analysis is performed on large streams of input data which are split into short frames, then obtaining fundamental frequency estimates for each frame.

### 2.3 Multidimensional Maximum-Likelihood

A straightforward way of attempting to solve the multipitch problem would be to derive the log-likelihood function for the signal samples, then maximize it with respect to the unknown parameters including the fundamental frequencies. The following model is
assumed, equivalent to (2.3) with additive noise

\[ \mathbf{x} = \mathbf{A}_0(\gamma)\mathbf{b}_0 + \mathbf{n} \]  

(2.4)

where \( \mathbf{A}_0(\gamma) \triangleq [\mathbf{A}(\gamma_1), \ldots, \mathbf{A}(\gamma_M)] \) and \( \mathbf{b}_0 \triangleq [\mathbf{b}_1, \ldots, \mathbf{b}_M]^T \). The vector \( \mathbf{n} \) denotes white Gaussian noise with zero mean and variance \( \sigma^2 \). The vectors \( \mathbf{b}_0 \) and \( \gamma \) are treated as deterministic unknowns. The log-likelihood function in this case is

\[
    f_{\mathbf{x};\gamma;\mathbf{b}_0} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}((\mathbf{x}-\mathbf{A}_0(\gamma)\mathbf{b}_0)^T(\mathbf{x}-\mathbf{A}_0(\gamma)\mathbf{b}_0)}
\]

(2.5)

The ML estimate of \( \mathbf{b}_0 \) is \( \hat{\mathbf{b}}_0 = (\mathbf{A}_0^H(\gamma)\mathbf{A}_0(\gamma))^{-1}\mathbf{A}_0^H(\gamma) \). After substituting \( \hat{\mathbf{b}}_0 \) in (2.5), we arrive at the following ML estimator for \( \gamma \)

\[
    \hat{\gamma}_{\text{ML}} = \arg\max_\gamma \| \mathbf{P}_{\mathbf{A}_0}(\gamma)\mathbf{x} \|
\]

(2.6)

where \( \mathbf{P}_{\mathbf{A}_0}(\gamma) \triangleq \mathbf{A}_0(\gamma) (\mathbf{A}_0^H(\gamma)\mathbf{A}_0(\gamma))^{-1}\mathbf{A}_0^H(\gamma) \). Evaluation of \( \hat{\gamma}_{\text{ML}} \) requires a multidimensional search over the coordinates of \( \gamma \) to find the global maximum. This is not a computationally feasible task, so this is not a viable method of performing multiple pitch estimation.

### 2.4 MUSIC

In this section, an approach based on the MUSIC algorithm will be examined. First, the MUSIC estimator will be introduced from the context of the direction-of-arrival (DOA) problem in array processing. The harmonic MUSIC estimator for fundamental frequency estimation will then be developed, followed by various modifications, including the multiple fundamental frequency version.
2.4.1 MUSIC Criterion

The MUSIC algorithm was conceived as a high-resolution DOA estimator for narrowband signals [18]. Consider the following signal model for an \( N \)-sensor arbitrary geometry array

\[
x_j = V(\theta)s_j + n_j \quad j = 1, \ldots, J
\]  

(2.7)

where \( J \) is the total number of snapshots, \( x_j \) is the \( j \)th \( N \times 1 \) signal snapshot vector, \( V(\theta) = [v(\theta_1), \ldots, v(\theta_M)] \) where \( v(\theta_m) \) is the \( m \)th steering vector, \( \theta = [\theta_1, \ldots, \theta_M]^T \) is the vector of unknown DOA’s, \( s_j = [s_{1j}, \ldots, s_{Mj}]^T \) is the signal vector and is assumed to be random zero mean wide-sense stationary with a full-rank autocorrelation matrix \( R_s = E[s_j s_j^H] \), and the noise vectors \( \{n_j\}_{j=1}^J \) are a random zero mean i.i.d. sequence with an autocorrelation matrix \( E[n_j n_j^H] = \sigma^2 I_N \) where \( I_N \) is an identity matrix of size \( N \) so that \( n_j \) and \( s_j \) are uncorrelated.

The autocorrelation matrix of the measurements vector is

\[
R_x = E[x_j x_j^H] = V(\theta)R_s V(\theta)^H + \sigma^2 I_N.
\]  

(2.8)

The left element in the sum may be decomposed as

\[
V(\theta)R_s V(\theta)^H = [U_S U_N] \Lambda \begin{bmatrix} U_S^H \\ U_N^H \end{bmatrix}
\]  

(2.9)

where \( \Lambda \) is a diagonal matrix consisting of the eigenvalues of \( V(\theta)R_s V(\theta)^H \), \( U_S \) is a matrix whose columns are the eigenvectors corresponding to \( M \) nonzero eigenvalues, and \( U_N \) is a matrix whose columns are the eigenvectors corresponding to \( N - M \) zero eigenvalues (in accordance with the requirement that the number of sensors exceeds the number of sources). In this manner, \( U_S, U_N \) each form an orthonormal basis for the signal and noise subspaces, respectively. Therefore, for an eigenvector of \( V(\theta)R_s V(\theta)^H \), denoted as \( u_k \), it
holds that

\[ V(\theta)R_s V^H(\theta) u_k = (U_S \Lambda U_S^H + U_N \Lambda U_N^H) u_k = 0 \cdot u_k = 0 \quad (2.10) \]

\[ k = M + 1, \ldots, N. \]

Thus

\[ V(\theta)R_s V^H(\theta)U_N = 0. \quad (2.11) \]

Since \( R_s \) is full rank, and under the assumption that the columns of \( V(\theta) \) are linearly independent, \( V(\theta)R_s \) is also full rank. Left-multiplying both sides of (2.11) by \( R_s^{-1} (V^H(\theta)V(\theta))^{-1} V^H(\theta) \) and taking the conjugate transpose yields

\[ U_N^H V(\theta) = 0 \quad (2.12) \]

which implies that

\[ U_N^H V(\theta_m) = 0, \quad m = 1, \ldots, M. \quad (2.13) \]

In other words, each steering vector representing an actual source is orthogonal to the noise subspace. Since \( U_N \) is estimated from the measurements, the orthogonality with the estimated noise subspace in (2.13) holds only approximately.

The MUSIC estimator is therefore defined as the peaks of the function

\[ P_{MU}(\theta) = \frac{1}{\left\| \hat{U}_N^H a(\theta) \right\|^2} \quad (2.14) \]

where \( \hat{U}_N \) is the estimated noise subspace which is calculated by selecting eigenvectors corresponding to the \( N - M \) smallest eigenvalues of the estimated autocorrelation matrix,

\[ \hat{R}_x = \frac{1}{J} \sum_{j=1}^{J} x_j x_j^H. \]
2.4.2 Single-Pitch Estimation

In this section, a MUSIC-based pitch estimation algorithm previously published in [19] is described. Assume a noisy harmonic signal which fits the model in (2.1). The signal may be expressed as

\[ x = \sum_{k=-K, k \neq 0}^{K} a_k(\gamma)b_k + n \]  \hspace{1cm} (2.15)

which is essentially equivalent to the narrowband model described in Appendix (A.1) with \(2K\) sources. The complex amplitudes are now assumed to be random zero-mean with a covariance matrix \(R_b = E[bb^H]\). In order to formulate the MUSIC algorithm in the context of the harmonic model, the signal autocorrelation matrix must be estimated.

**Estimation of the snapshot autocorrelation matrix \(R_x\):**

The samples vector \(x\) is split into \(P\) overlapping subframes \(x_0, \ldots, x_{P-1}\), each of length \(L\), where

\[ x_p \triangleq \begin{pmatrix} x[(p-1)P] \\ \vdots \\ x[(p-1)P + L - 1] \end{pmatrix}, \quad p = 1, \ldots, P. \]  \hspace{1cm} (2.16)

Each subframe is characterized by a time lag relative to the first subframe. Since a time lag for a harmonic signal is essentially a phase shift, the model may be expressed for each subframe as

\[ x_p = A(\gamma)\Gamma_p(\gamma)b + n_p \]  \hspace{1cm} (2.17)

where \(A(\gamma)\) is the harmonic matrix of dimensions \(L \times 2K\) and \(\Gamma_p(\gamma)\) incorporates the phase shift, defined as

\[ \Gamma_p(\gamma) \triangleq \text{diag} \left( [e^{-j2\pi K\gamma p T}, \ldots, e^{-j2\pi K\gamma p T}, e^{j2\pi K\gamma p T}, \ldots, e^{j2\pi K\gamma p T}] \right). \]  \hspace{1cm} (2.18)

The noise vector is dropped from the model, allowing for a straightforward development of estimator.
The autocorrelation matrix is then estimated from the subframe vectors as

\[ \hat{R}_x = \frac{1}{P} \sum_{p=1}^{P} x_p x_p^H. \]  

(2.19)

It follows that

\[ \hat{R}_x \approx A(\gamma)\overline{R}_b A_0^H(\gamma) + \frac{1}{P} \sum_{p=1}^{P} n_p n_p^H \]

where

\[ \overline{R}_b = \frac{1}{P} \sum_{p=1}^{P} \Gamma_p(\gamma) b b^H \Gamma_p^H(\gamma). \]  

(2.20)

The matrix \( \overline{R}_b \) may be regarded as analogous to \( R_s \) in (2.7). The factors \( \Gamma_p(\gamma) \) are introduced by the division into subframes and serve to increase the rank of \( \hat{R}_x \) so that it is full-rank, which is requisite for MUSIC.

**Harmonic MUSIC estimator**

As in (2.9), the noise subspace \( U_N \) is chosen to be the eigenvectors corresponding to the \( L - 2K \) smallest eigenvalues of \( A_0 R_b A_0^H \). Selecting the eigenvectors corresponding to the \( L - 2K \) smallest eigenvalues of \( \hat{R}_x \) yields an estimate of the noise subspace, \( \hat{U}_N \) [19].

Assuming \( \overline{R}_b \) is full rank, it therefore holds (according to (2.13))

\[ U_N^H A(\gamma) = 0. \]  

(2.21)

The MUSIC estimator of the fundamental frequency is therefore defined [19] as the peak of the function

\[ P_{MU}(\gamma) = \frac{1}{\| \hat{U}_N^H A(\gamma) \|_F^2} = \frac{1}{\sum_{k=-K}^{K} \| \hat{U}_N^H a_k(\gamma) \|_F^2}. \]  

(2.22)

where \( \| \cdot \|_F \) denotes the Frobenius norm. Each element of the sum in the denominator of (2.22) tests the orthogonality of the estimated noise subspace to a single complex harmonic. Assuming orthogonality holds at a certain frequency for all harmonics, the sum is approximately zero and the MUSIC spectrum tends to infinity. If orthogonality
does not hold for a one or more harmonics, the sum is larger than zero and the MUSIC spectrum does not exhibit a strong peak for that frequency.

Model order estimation

The orthogonality defined by (2.21) holds for the correct frequency $\gamma$ and for the correct model order $2K$, which is the number of complex harmonics. The process outlined until here assumed in the calculation of both $\hat{U}_N$ and $A(\gamma)$ that $2K$ is known. When the model order is unknown, define the joint estimator for both fundamental frequency and model order as

$$P_{MU}(f, K) = \frac{1}{\|\hat{U}_N^H(K)A(\gamma, K)\|_F^2}$$

(2.23)

where $\hat{U}_N(K)$ is a matrix whose columns are the eigenvectors corresponding to the $L-2K$ smallest eigenvalues of $\hat{R}_x$ and $A(\gamma, K)$ is an $L \times 2K$ harmonic matrix with $K$ real harmonics (the number of real harmonics is treated as an additional variable). Evaluating $P_{MU}(\gamma, K)$ over a 2-dimensional grid and selecting the maximum yields an estimate for both the frequency and the model order [19].

2.4.3 Multi-Pitch Estimation

A signal comprising of multiple harmonic sources may be described by (2.3) where the $m$th signal has $2K_m$ complex harmonics. Splitting the samples vector into $P$ overlapping subframes, the model for the $p$th subframe is

$$x_p = \sum_{m=1}^{M} A(\gamma_m) \Gamma_p(\gamma_m)b_{(m)}$$

(2.24)
where $A(\gamma_m)$ has $2K_m$ columns corresponding to the $2K_m$ complex harmonics of the $m$th signal. Let $\Gamma_p^{(M)}$ be defined as

$$
\Gamma_p^{(M)} \triangleq \begin{pmatrix}
\Gamma_p(\gamma_1) & 0 \\
0 & \ddots \\
0 & \Gamma_p(\gamma_M)
\end{pmatrix}.
$$

The data model for the $p$th subframe is given by

$$
x_p = A^{(M)}(\gamma) \Gamma_p^{(M)} b^{(M)}. \tag{2.26}
$$

In this case, the estimated covariance matrix is

$$
\hat{R}_x \cong A(\gamma) \overline{R}_b^{(M)} A^{H}(\gamma) + \frac{1}{P} \sum_{p=1}^{P} n_p n_p^{H}
$$

where

$$
\overline{R}_b^{(M)} = \frac{1}{P} \sum_{p=1}^{P} \Gamma_p^{(M)} b^{(M)} (b^{(M)})^{H} (\Gamma_p^{(M)})^{H}. \tag{2.27}
$$

The noise subspace $U_N$ is chosen to be the eigenvectors corresponding to the $L-2 \sum_{m=1}^{M} K_m$ smallest eigenvalues of $A^{(M)} \overline{R}_b (A^{(M)})^{H}$. The estimated noise subspace $\hat{U}_N$ is the collection of eigenvectors corresponding to the $L-2 \sum_{m=1}^{M} K_m$ smallest eigenvalues of $\hat{R}_x$.

The orthogonality condition (2.13) in this case is

$$
U_N^{H} A^{(M)}(\gamma) = 0 \tag{2.28}
$$

which implies that

$$
U_N^{H} A(\gamma_m) = 0, \quad m = 1, \ldots, M. \tag{2.29}
$$

If the total number of harmonics contained in all the sources is overestimated, $U_N^{H}$ will contain at least one vector which belongs to the signal subspace. This violates the orthogonality defined in (2.28), resulting in poor estimation results. If the total number of harmonics is underestimated, only part of the noise subspace is used which may result in more spurious peaks appearing in the MUSIC spectrum corresponding to ratios of the fundamental frequency.
The MUSIC estimator of the multiple fundamental frequencies is therefore defined as the $M$ highest peaks of the function

$$R_{MU}(\gamma) = \frac{1}{\| \hat{U}_N^H A(\gamma) \|_F^2}. \quad (2.30)$$

Since the estimator uses the same harmonic matrix to estimate all the sources regardless of the number of harmonics, the number of columns $A(\gamma)$ in (2.30) must not exceed twice the number of harmonics contained in the source with the least harmonics.

An example of estimation of two fundamental frequencies belonging to two synthetic harmonic sources with 5 harmonics each and fundamental frequencies of 150Hz and 210Hz is displayed in Fig. 2.1. A frame size of 30ms was used with a sampling rate of 10kHz. A subframe size of 18ms was used with 98% overlap between subframes. The estimator assumes a-priori knowledge of the total number of harmonics belonging to the two sources. A similar example in low SNR conditions is displayed in Fig. 2.2, where it can be seen that the MUSIC spectrum is degraded due to the additive noise. An example using real speech signal segments 60ms in duration from the Keele database [20] is displayed in Fig. 2.3. A subframe size of 36ms was used with 98% overlap between subframes. It was assumed that each signal contains 4 harmonics.

### 2.4.4 Root-MUSIC

Root-MUSIC is an implementation of MUSIC for uniform linear arrays which replaces the grid search with a root-finding procedure on a polynomial function [18]. A corresponding implementation in the context of the harmonic model is now developed. According to (2.21)

$$\| U_N^H a_k(\gamma) \|^2 = a_k^H(\gamma) U_N U_N^H a_k(\gamma) = 0, \quad k = -K, \ldots, -1, 1, \ldots K. \quad (2.31)$$
Define $\mathbf{\Gamma} = \mathbf{U}_N \mathbf{U}_N^H$, $z = e^{j2\pi\gamma T}$, $\mathbf{a}_k(\gamma) = [1, z, z^2, \ldots, z^L]^T$. Equation (2.31) may be rewritten as

$$
\sum_{m=0}^{L-1} \sum_{n=0}^{L-1} [\mathbf{U}_N \mathbf{U}_N^H]_{mn} z^{k(n-m)} = \sum_{n=-(L-1)}^{L-1} c_n z^{kn} \quad (2.32)
$$

where

$$
c_n = \sum_{k=|n|}^{L-|n|-1} [\mathbf{U}_N \mathbf{U}_N^H]_{k,n+k} \quad (2.33)
$$

Define $P(z)$ as a $2L-2$ order polynomial with coefficients $\{c_n\}_{n=-(L-1)}^{(L-1)}$. The orthogonality condition for the $k$th harmonic (2.31) is equivalent to $P(z^k) = 0$.

The frequency may be estimated from the $k$th harmonic by the following procedure:

1. Calculate the polynomial coefficients (2.33) from the estimated noise subspace.
Figure 2.2: Multi-pitch estimation with MUSIC. Two synthetic signals were mixed with AWGN with an SNR of 5dB. Plots (A) and (B) show the amplitude spectra of sources A and B, respectively, each showing 5 harmonics. Plot (C) displays the MUSIC spectrum, exhibiting lower peaks than the case with high SNR.

2. Calculate the \((2L - 2) \cdot K\) roots of \(P(z^k)\).

3. Divide the phase of the root which is closest to the unit circle by \(2\pi T\) for the frequency estimate.

The roots of \(P(z^k)\) may be calculated in two stages: first calculating the \((2L - 2)\) roots of \(P(z)\) as previously, then calculating the \(k\) roots of each result.

The orthogonality condition for all harmonics (2.21) is equivalent to finding the roots of the sum of polynomials

\[
\sum_{k=-K, \ K \neq 0}^{K} P(z^k) = 0
\]

(2.34)

which is equivalent to a polynomial of degree \(2K \cdot (2L - 2)\).
Figure 2.3: Multi-pitch estimation with MUSIC using real signals. Plots (A) and (B) show the amplitude spectra of sources A (female) and source B (male), respectively. Plot (C) displays the MUSIC spectrum, exhibiting peaks around 96Hz and 193Hz, which are close to the location of the peaks in the spectrograms.

2.4.5 MUSIC - Discussion

Drawbacks

- The harmonic MUSIC estimator suffers from a sensitivity to model mismatch. In general, MUSIC relies on the orthogonality of the estimated noise subspace to the assumed signal subspace - the harmonic matrix. If the harmonic matrix contains a harmonic vector which is weak or nonexistent in the signal, it cannot be considered to belong to the signal subspace, therefore it belongs to the noise subspace and is not orthogonal to the noise subspace. Therefore, a harmonic matrix with the inclusion of a signal nonexistent harmonic violates the orthogonality (2.21). Since there is no simple way to know which harmonics are present without knowledge of the fundamental frequency, the harmonic matrix may be incorrectly constructed.
The problem is compounded in the multi-pitch case if each signal has a different harmonic matrix order. It should be noted that it is desirable to include as many higher-order harmonics as possible in the construction of the harmonic matrix, since they are more significant in terms of decreasing the Cramér-Rao bound for the fundamental frequency estimation (see Section (2.3)).

- Assuming the estimated noise subspace is approximately orthogonal to the harmonic matrix, there are often spurious peaks at half or twice the fundamental frequency (or at other simple ratios) since some of the harmonic vectors remain the same in these cases. This is especially a problem in the multiple pitch case, since the second-highest peak (2.30) may represent a ratio of the fundamental frequency belonging to signal represented by the highest peak, thus the second fundamental frequency will be incorrectly estimated. This problem is characteristic of many pitch detection algorithms.

- Since the fundamental frequencies over successive frames are usually highly correlated, it can be desirable to implement a tracking algorithm following the estimation method. One such way to accomplish this is to define a conditional \textit{a-priori} probability density function (pdf) for the fundamental frequencies and to implement a maximum \textit{a-posteriori} probability (MAP) estimator. Such a procedure it fairly straightforward for an ML estimator (as will be detailed in Subsection 3.2.3). In the case of MUSIC, there is no clear method of defining such a tracking scheme.

An example of how missing harmonics have a detrimental effect on the MUSIC spectrum can be seen in Fig. 2.4. Two synthetic harmonic sources with 4 harmonics each and fundamental frequencies of 150Hz and 210Hz were mixed with an SNR of 100dB and a frame size of 30ms was used with a sampling rate of 10kHz. A subframe size of 18ms was used with 98% overlap between subframes. The MUSIC spectrum is degraded due to a weak harmonic in one of the sources. An example with real signals from the Keele database is shown in Fig. 2.5 where low harmonics and/or weak harmonicity contributes to a degraded MUSIC spectrum.
Figure 2.4: Multi-pitch estimation with MUSIC. Two synthetic signals were mixed with AWGN with an SNR of 100dB. Plots (A) and (B) show the amplitude spectra of sources A and B, respectively, where the 3rd harmonic source B is attenuated by 30dB. Plot (C) displays the MUSIC spectrum, and the peak corresponding to source B is more than 50dB lower than the other peak.

Model order estimation

In the single-pitch case, the model order may be jointly estimated along with the fundamental frequency (2.23) (assuming no model mismatch, i.e. all harmonics are present in the signal) by a 2-dimensional grid search, which was detailed in [19]. This method does not scale well to the multiple-pitch case, since the estimated noise subspace depends on the total number of complex harmonics. For example, in the case of two harmonic signals, a 4-dimensional grid search is required since the orthogonality holds only if both model orders and both harmonic matrices are correct.
Figure 2.5: Multi-pitch estimation with MUSIC using real signals. Plots (A) and (B) show the amplitude spectra of sources A (female) and source B (male), respectively. Plot (C) displays the MUSIC spectrum, and the peak corresponding to source B is low, since the source exhibits weak harmonicity.

**Noise subspace estimation**

Recall that the noise subspace is estimated as $L - K_{\text{tot}}$ eigenvectors where $K_{\text{tot}}$ is the total number of complex harmonics belonging to all harmonic sources. If $K_{\text{tot}}$ is underestimated, the estimated noise subspace includes at least one signal subspace vector, which will violate the orthogonality in (2.21). Methods for estimating the number of complex harmonics, including unstructured MDL [21], are not impervious to underestimation.

**Root-MUSIC**

Since Root-MUSIC is based on the MUSIC orthogonality, it suffers from the same drawbacks as MUSIC (an advantage over MUSIC is the absense of a grid search). An additional issue in this case is the possibility of high order polynomials, as the orthogonality for $2K$
harmonics results in a polynomial of degree $2K \cdot (2L - 2)$.

2.4.6 Equalized MUSIC

Over the course of the research presented in this thesis, the problem with harmonic MUSIC was perceived to be, in part, the assumption that the harmonics defined by the columns of the harmonic matrix $A(\gamma)$ are present in the signal measurements. In practice, however, no such assumptions can be made, since real audio signals often have missing or weak harmonics. The orthogonality requirement in (2.21) was considered to be too strict, and a weaker form of orthogonality was sought in order to lead to a robust estimator. The modification introduced in this section was made in attempt to mitigate the consequences of weak harmonics in the case of a single harmonic source as defined in (2.15), then extending to the multiple harmonic source case. Instead of requiring the full orthogonality in (2.21) to hold, the following orthogonality will be assumed

$$U_N^H A(\gamma) b = 0.$$  \hspace{1cm} (2.35)

In this manner, a linear combination of the harmonic vectors is assumed to be orthogonal to the subspace, where for any weak harmonics the corresponding element in vector $b$ will be close to zero, causing the harmonic’s effect on the entire sum to be negligible. After normalization of the vector $A(\gamma)b$, the MUSIC spectrum is given as

$$P_{MU}(\gamma) = \max_b \frac{\|A(\gamma)b\|^2}{\|U_N^H A(\gamma)b\|^2} = \max_b \frac{b^H A^H(\gamma) A(\gamma)b}{b^H A^H(\gamma) U_N U_N^H A(\gamma)b}.$$  \hspace{1cm} (2.36)

Define

$$G(\gamma) = A^H(\gamma) A(\gamma)$$  \hspace{1cm} (2.37)

$$H(\gamma) = A^H(\gamma) U_N U_N^H A(\gamma).$$
The MUSIC spectrum can be written as

\[ P_{\text{MU}}(\gamma) = \max_b b^H G(\gamma) b \]

where the solution for the maximization over \( b \) is given by \( \lambda_{\text{max}}(\gamma) \) obtained from the following generalized eigenvalue problem

\[ G(\gamma) b = \lambda_{\text{max}}(\gamma) H(\gamma) b \]

where \( \lambda_{\text{max}}(\gamma) \) is the maximum generalized eigenvalue of the matrix pair \((G(\gamma), H(\gamma))\).

Divide both sides by \( \lambda_{\text{max}}(\gamma) \), and since \( G(\gamma) \) is nonsingular, by left-multiplying both sides by \( G^{-1/2}(\gamma) \) one obtains

\[ \frac{1}{\lambda_{\text{max}}(\gamma)} G^{1/2}(\gamma) b = G^{-1/2}(\gamma) H(\gamma) G^{-1/2}(\gamma) G^{1/2}(\gamma) b \]

so that

\[ \frac{1}{\lambda_{\text{max}}(\gamma)} \tilde{b} = G^{-1/2}(\gamma) H(\gamma) G^{-1/2}(\gamma) \tilde{b} \]

where \( \tilde{b} = G^{1/2}(\gamma) b \). The estimated MUSIC spectrum can therefore be expressed as

\[ P_{\text{MU}}(\gamma) = \lambda_{\text{min}} \left( (A^H(\gamma) A(\gamma))^{-1/2} A^H(\gamma) \hat{U}_N \hat{U}_N^H A(\gamma) (A^H(\gamma) A(\gamma))^{-1/2} \right) \]

where \( \lambda_{\text{min}} (\cdot) \) denotes the minimal eigenvalue of a matrix.

This method was found to produce a spectrum which does not appear to correlate with the signal, therefore does not appear to represent a solution to the lack of robustness which characterizes harmonic MUSIC. One reason for this problem is that the maximization w.r.t. \( b \) allows for too many degrees of freedom, which leaves the signal model ill-defined. For example, if a harmonic is missing from the signal, its corresponding coefficients in the vector \( b \) may assume any values without changing the overall representation of the signal. This problem could be solved by proposing a scheme to limit the range of values...
can assume within the context of the above modification. An example of the equalized MUSIC spectrum is given in Fig. 2.6. Two synthetic harmonic sources with 9 harmonics each and fundamental frequencies of 150Hz and 210Hz were mixed with an SNR of 100dB and a frame size of 30ms was used with a sampling rate of 10kHz. A subframe size of 18ms was used with a 98% overlap between subframes. It can be seen that while the regular MUSIC spectrum has peaks in the locations of correct frequencies, the equalized MUSIC spectrum has a large number of spurious peaks.

In the next chapter, the proposed method will be introduced using a more robust measure of the presence of harmonic sources than the reliance upon the assumed structure of the harmonic matrix, which is a prime drawback of MUSIC-based methods.

Figure 2.6: Equalized versus regular MUSIC. Two synthetic signals were mixed with AWGN with an SNR of 100dB. Plot (A) displays the regular MUSIC spectrum and plot (B) displays the equalized MUSIC spectrum. The dots indicate the peak locations.
Chapter 3

The Proposed ML-MAP Method

In the previous chapter, methods based on MUSIC and multidimensional ML were shown to be inadequate for the purpose of multiple pitch estimation. In this chapter, a more robust method is proposed, based on the log-likelihood function for a measurement model consisting of a harmonic source in the presence of unknown random interference and on a MAP tracking algorithm. The measurement model is similar to the narrowband model (A.1) used in array processing, and the resulting ML estimator is analogous to the minimum variance distortionless response (MVDR) estimator from that original context. In Section 3.1 the MVDR criterion and the resulting estimator are reviewed, and it is subsequently shown that the MVDR estimator is equivalent to an ML estimate under certain conditions (including the particular definition of the measurements model). In Section 3.2 the proposed estimator is developed for the single pitch case, first by developing the log-likelihood function, then by implementing a MAP tracking algorithm. The estimator is extended to the multiple pitch case in Section 3.3.
3.1 MVDR / ML Estimator

3.1.1 MVDR Criterion

Assume a collection of array output samples given according to the narrowband model (A.1) with a zero-mean noise vector \( \mathbf{n} \) and an autocorrelation matrix \( \mathbf{R_x} = \mathbb{E}[\mathbf{x}\mathbf{x}^H] \). In a beamforming scenario, the MVDR beamformer attempts to attenuate all incident signals arriving from directions other than the array look direction without signal power loss in the look direction \( \theta_0 \). The beamformer weights are the result of the following constrained optimization problem:

\[
\mathbb{E}\left[|\mathbf{w}^H\mathbf{x}|^2\right] \xrightarrow{\mathbf{w}} \min
\]

subject to : \( |\mathbf{w}^H\mathbf{v}(\theta_0)| = 1 \).

The solution for \( \mathbf{w} \) is:

\[
\mathbf{w}_0 = \frac{\mathbf{R}_x^{-1}\mathbf{v}(\theta_0)}{\mathbf{v}^H(\theta_0)\mathbf{R}_x^{-1}\mathbf{v}(\theta_0)}
\]

and assuming \( ||\mathbf{v}(\theta_0)|| = 1 \), the resulting beamformer average output power is:

\[
\mathbb{E}\left[|\mathbf{w}_0^H\mathbf{x}|^2\right] = \frac{1}{\mathbf{v}^H(\theta_0)\mathbf{R}_x^{-1}\mathbf{v}(\theta_0)}.
\]

In a DOA estimation scenario, once the autocorrelation matrix has been estimated, (3.3) may be maximized over \( \theta_0 \) for the MVDR estimate of \( \theta_0 \). It will be shown in the next section that this is actually a maximum-likelihood estimate of \( \theta_0 \) for a single source with unknown Gaussian interference.
3.1.2 ML DOA Estimation in the Presence of Unknown Gaussian Interference

Consider the following signal model for the problem of DOA estimation:

\[
x_j = v(\theta)s_j + i_j + n_j, \quad j = 1, \ldots, J
\]

where \(J\) denotes the number of snapshots, \(\theta\) denotes the source DOA and \(v(\cdot)\) is the array steering vector. Also, \(\{i_j\}_{j=1}^J\) represents i.i.d. interference satisfying \(i_j \sim \mathcal{N}(0, R_i)\) where \(R_i\) is unknown and \(\{n_j\}_{j=1}^J\) is the i.i.d. sequence of white noise vectors with \(n_j \sim \mathcal{N}(0, \sigma^2 I)\). Also, \(n_j\) and \(i_j\) are statistically independent for \(j = 1, \ldots, J\). The signal \(\{s_j\}_{j=1}^J\) is assumed to be deterministic unknown. In this manner, the model assumes a single source while other harmonic sources that may be present are modeled as random Gaussian interference. The resulting ML estimator (in the case of unknown \(\sigma^2\)) for \(\theta\) is given as [22]

\[
\hat{\theta}_{\text{ML}} = \arg \min_{\theta} |T^H_{\theta} \hat{R}_x T_{\theta}|
\]

where \(|\cdot|\) is the determinant operation, \(T_{\theta}\) is an \(N \times (N - 1)\) matrix such that \([v(\theta) \ T_{\theta}]\) forms a complete orthonormal basis, and \(\hat{R}_x\) is the sample covariance matrix, defined as:

\[
\hat{R}_x = \frac{1}{J} \sum_{j=1}^J x_j x_j^H.
\]

In [22] it is shown that under certain conditions, the estimator in (3.5) is equivalent to:

\[
\hat{\theta}_{\text{ML}} = \arg \max_{\theta} \frac{1}{v^H(\theta) \hat{R}_x^{-1} v(\theta)}
\]

which means that maximization of the MVDR beamformer output power is an ML estimate assuming the model (3.4).

In the case of multiple sources, estimates for \(M\) DOAs are given as the \(M\) highest peaks of the expression \(\left(v^H(\theta)\hat{R}_x^{-1} v(\theta)\right)^{-1}\). This way, each harmonic source is treated in
turn as the source in the model (3.4), while the rest are regarded as random interference, depending on the value of \( \theta \). The estimates may be regarded as “locally-ML”, since the likelihood function is maximized within a specific range for each source.

### 3.2 Single Source Pitch Tracking

In this section, a new method will be derived for the case of a single harmonic source in the presence of unknown random interference. The input measurements are given as a sampled audio stream of limited duration, consisting of multiple harmonic sources, which is subsequently split into \( Q \) overlapping frames

\[
X \triangleq [x^{(1)}, \ldots, x^{(Q)}].
\]  

A Bayesian MAP estimation technique, detailed in Subsection 3.2.3, is used to estimate the fundamental frequency of the dominant harmonic source in all of the input frames. The ML-MAP technique is implemented via a forward-backward tracking algorithm which relies on the partially-maximized log-likelihood function for each frame (developed in Subsection 3.2.2), in addition to a-priori information for the fundamental frequencies. This assumed information is a transitional pdf, where the fundamental frequency for the current frame is distributed symmetrically around the estimation result for the previous frame. In this manner, the estimation result for a single frame relies on previous and upcoming frames, therefore the algorithm is designed to process the input data in batches.

#### 3.2.1 Single Harmonic Source + Interference Model

The proposed method is based on the following model

\[
x = A(\gamma_0)b + e.
\]
In (3.9), the dominant harmonic source is modeled by $A(\gamma_0)b$ and the other harmonic sources, as well as non-harmonic sources and additive noise, are treated as unknown random interference and denoted by $e$. This description fits several scenarios including the single-pitch case, with or without other voiced/unvoiced sources.

### 3.2.2 Log-likelihood for Single Source + Interference Model

The log-likelihood corresponding to the model presented in (3.9) for a single frame of data is now developed, along with maximization with respect to the unknown parameters excluding the fundamental frequency. The frame vector $x$ is assumed to be stationary, corresponding to a single frame of the input audio stream. The superscript corresponding to the frame index is suppressed here. In the case of speech, the acceptable duration is 20-30msec, and may be longer in the case of music. The vector $x$ is split into a collection of $J$ overlapping subframes of length $L$, which for the sake of analytic tractability are assumed to be statistically independent (an assumption not satisfied in practice). The $j$th subframe is described using the following model

$$\tilde{x}_j = \tilde{A}(\gamma)\tilde{b}_j^{(0)} + \tilde{e}_j, \ j = 1, \ldots, J$$

(3.10)

where $\tilde{A}(\gamma) = [\tilde{a}_{-K}(\gamma), \ldots, \tilde{a}_1(\gamma), \tilde{a}_1(\gamma), \ldots, \tilde{a}_K(\gamma)]^T$ is an $L \times 2K$ harmonic matrix with fundamental frequency $\gamma$, and $\tilde{b}_j^{(0)}$ and $\tilde{e}_j$ are the $2K \times 1$ vector of complex amplitudes and the $L \times 1$ vector of interference at the $j$th subframe, respectively. The columns of the harmonic matrix $\tilde{A}(\gamma)$ are defined as

$$\tilde{a}_k(\gamma) = [1, e^{j2\pi\gamma kT_s}, \ldots, e^{j2\pi\gamma(L-1)kT_s}]^T.$$ (3.11)

We assume that $L \geq 2K$. The vector $\tilde{b}_j^{(0)}$ is assumed to be deterministic unknown. The interference vector $\tilde{e}_j$ is assumed to be zero-mean Gaussian, and with no loss of generality, it can be decomposed into two statistically independent Gaussian components: $\tilde{e}_j = \tilde{i}_j + \tilde{n}_j$ where $\tilde{n}_j$ represents the white noise component with known variance, $\sigma^2$. 

26
while $\tilde{i}_j$ denotes the colored component with an unknown covariance matrix. If no part of the noise is assumed to be unknown, then $\sigma^2 = 0$. The interference may not be Gaussian in general (e.g. used to represent other harmonic sources), but this assumption is used for the sake of tractability. The interference vector is decomposed as

$$\tilde{i}_j = \tilde{A}(\gamma)\alpha_j + T_{\tilde{A}}(\gamma)\beta_j$$  \hspace{1cm} (3.12)$$

where $T_{\tilde{A}}(\gamma)$ is an $L \times (L - 2K)$ matrix whose columns form an orthonormal basis for the subspace complementary to the subspace spanned by the columns of $\tilde{A}(\gamma)$. The vector $\tilde{A}(\gamma)\alpha_j$ represents the interference component in the signal subspace and is joined with $\tilde{A}(\gamma)\tilde{b}_j^{(0)}$, defining $\tilde{b}_j = \tilde{b}_j^{(0)} + \alpha_j$, and $\beta_j$ is a random zero-mean Gaussian vector of size $L - 2K$ with an unknown covariance matrix $\Psi = \mathbb{E}[\beta_p\beta_p^H]$. Accordingly, the probability distribution of the $j$th subframe vector is

$$\tilde{x}_j \sim \mathcal{N}(\tilde{A}(\gamma)\tilde{b}_j, R_e)$$  \hspace{1cm} (3.13)$$

where

$$R_e = T_{\tilde{A}}(\gamma)\Psi T_{\tilde{A}}^H(\gamma) + \sigma^2 I_L$$  \hspace{1cm} (3.14)$$

and is an unknown covariance matrix and $I_L$ is an identity matrix of size $L$.

The log-likelihood function of the collection of subframes $\tilde{x}_1, \ldots, \tilde{x}_J$ is given by

$$L_x\left(\gamma, \left\{\tilde{b}_j\right\}_{j=1}^J, R_e\right) = -\frac{J}{2} \log |2\pi R_e|$$  \hspace{1cm} (3.15)$$

$$- \frac{1}{2} \sum_{j=1}^J \left(\tilde{x}_j - \tilde{A}(\gamma)\tilde{b}_j\right)^H R_e^{-1} \left(\tilde{x}_j - \tilde{A}(\gamma)\tilde{b}_j\right).$$

In the following, we will analytically maximize the log-likelihood in (3.16) w.r.t. $\tilde{b}_1, \ldots, \tilde{b}_J$ and $R_e$. For the sake of simplicity, the dependence on $\gamma$ will be omitted in the following.
According to the singular value decomposition

\[ \tilde{\mathbf{A}} = \mathbf{U}\tilde{\mathbf{A}} \mathbf{V}^H \tilde{\mathbf{A}} \]  

(3.16)

where \( \mathbf{D} \) is a diagonal matrix and \( \mathbf{U}\tilde{\mathbf{A}} \) and \( \mathbf{V}\tilde{\mathbf{A}} \) satisfy \( \mathbf{U}^H\tilde{\mathbf{A}}\mathbf{U}\tilde{\mathbf{A}} = \mathbf{I}_{2K} \) and \( \mathbf{V}^H\tilde{\mathbf{A}}\mathbf{V}\tilde{\mathbf{A}} = \mathbf{I}_{2K} \). Since \( \mathbf{U}\tilde{\mathbf{A}} \) is a matrix of size \( L \times 2K \) whose columns form an orthonormal basis for the signal subspace, the columns of the matrix \( \mathbf{E} = [\mathbf{U}\tilde{\mathbf{A}} \mathbf{T}\tilde{\mathbf{A}}] \) form a complete orthonormal basis. The covariance matrix, \( \mathbf{R}_e \) from (3.14), may now be expressed as

\[ \mathbf{R}_e = \mathbf{T}\tilde{\mathbf{A}} \mathbf{\Psi}\mathbf{T}^H\tilde{\mathbf{A}} + \sigma^2 (\mathbf{T}\tilde{\mathbf{A}}\mathbf{T}^H\tilde{\mathbf{A}} + \mathbf{U}\tilde{\mathbf{A}}\mathbf{U}^H\tilde{\mathbf{A}}) = \mathbf{T}\tilde{\mathbf{A}} \mathbf{\Gamma}\mathbf{T}^H\tilde{\mathbf{A}} + \sigma^2 \mathbf{U}\tilde{\mathbf{A}}\mathbf{U}^H\tilde{\mathbf{A}} \]  

(3.17)

where \( \mathbf{\Gamma} \triangleq \mathbf{\Psi} + \sigma^2 \mathbf{I}_{L-2K} \). Since \( \mathbf{T}\tilde{\mathbf{A}} \) and \( \mathbf{U}\tilde{\mathbf{A}} \) are orthonormal, the inverse and determinant of \( \mathbf{R}_e \) may be written as

\[ \mathbf{R}_e^{-1} = \mathbf{T}\tilde{\mathbf{A}} \mathbf{\Gamma}^{-1}\mathbf{T}^H\tilde{\mathbf{A}} + \frac{1}{\sigma^2} \mathbf{U}\tilde{\mathbf{A}}\mathbf{U}^H\tilde{\mathbf{A}}. \]  

(3.18)

and

\[ |\mathbf{R}_e| = |\mathbf{T}\tilde{\mathbf{A}} \mathbf{\Gamma}\mathbf{T}^H\tilde{\mathbf{A}} + \sigma^2 \mathbf{U}\tilde{\mathbf{A}}\mathbf{U}^H\tilde{\mathbf{A}}| = \sigma^{4K} |\mathbf{\Gamma}|, \]  

(3.19)

respectively.

Substitution of (3.18) and (3.19) into (3.16) yields

\[ \mathbf{L}_x^{\prime} \left( \mathbf{\Gamma}, \{ \tilde{\mathbf{b}}_j \}_{j=1}^J \right) = -\frac{J}{2} 2K \log(2\pi\sigma^2) - \frac{J}{2} \log |\mathbf{\Gamma}| \]  

(3.20)

\[ - \frac{1}{2} \sum_{j=1}^J \bar{\mathbf{x}}_j^H\mathbf{T}\tilde{\mathbf{A}} \mathbf{\Gamma}^{-1}\mathbf{T}^H\tilde{\mathbf{A}} \bar{\mathbf{x}}_j - \frac{1}{2\sigma^2} \sum_{j=1}^J \left\| \mathbf{U}\tilde{\mathbf{A}} \left( \bar{\mathbf{x}}_j - \mathbf{A}\tilde{\mathbf{b}}_j \right) \right\|^2. \]

The ML estimate of \( \tilde{\mathbf{b}}_j \) given \( \gamma \) is obtained by differentiating w.r.t. \( \tilde{\mathbf{b}}_j \) and equating to zero, which results in

\[ \tilde{\mathbf{b}}_j(\gamma) = \left( \mathbf{\tilde{A}}^H(\gamma)\mathbf{\tilde{A}}(\gamma) \right)^{-1} \mathbf{\tilde{A}}^H(\gamma)\bar{\mathbf{x}}_j \]  

(3.21)

which cancels the rightmost expression in (3.20). The log-likelihood after maximizing
w.r.t. $\tilde{b}_j$, $j = 1, \ldots, J$ and after applying the trace operator to the scalar expression inside the sum is

$$
L'_x \left( \gamma, \left\{ \tilde{b}_j \right\}_{j=1}^J, \Gamma \right) = -\frac{J}{2} 2K \log(2\pi\sigma^2) - \frac{J}{2} \log |\Gamma| - \frac{1}{2} \sum_{j=1}^{J} \text{tr} (\tilde{x}_j^HT_A \Gamma^{-1} T_A^H \tilde{x}_j) \\
= -\frac{J}{2} \left[ 2K \log(2\pi\sigma^2) + \log |\Gamma| + \text{tr} \left( \Gamma^{-1} T_A^H \hat{R}_x T_A \right) \right] 
$$

(3.22)

where $\hat{R}_x \triangleq \frac{1}{J} \sum_{j=1}^{J} \tilde{x}_j \tilde{x}_j^H$ is the sample autocorrelation matrix. The ML estimate of $\Gamma$ given $\gamma$ in the case of unknown $\sigma^2$ is obtained by maximizing (3.22) w.r.t. $\Gamma$ [22]

$$
\hat{\Gamma}_{ML} = T_A^H \hat{R}_x T_A. 
$$

(3.23)

Substitution of (3.23) into (3.22) yields

$$
L'_x \left( \gamma, \left\{ \tilde{b}_j \right\}_{j=1}^J, \hat{\Gamma} \right) = -\frac{J}{2} \left[ 2K \log(2\pi\sigma^2) + (L - 2K) + \log \left| T_A^H \hat{R}_x T_A \right| \right]. 
$$

(3.24)

Note that the rank of $T_A^H \hat{R}_x T_A$ may be less than or equal to $J$, in accordance with the rank of $\hat{R}_x$. This requires at least $L - 2K$ subframes for the matrix $T_A^H \hat{R}_x T_A$ to be nonsingular. This requirement is circumvented by discarding the eigenvalues of $T_A^H \hat{R}_x T_A$ which are equal to zero, which leads to a simpler form for (3.24). Define $\hat{X} \triangleq [\hat{x}_1, \ldots, \hat{x}_J]$. The sample autocorrelation matrix may be expressed as $\hat{R}_x = \frac{1}{J} \hat{X} \hat{X}^H$ and let $C \triangleq T_A^H \hat{X}$. It follows that

$$
T_A^H \hat{R}_x T_A = \frac{1}{J} T_A^H \hat{X} \hat{X}^H T_A = \frac{1}{J} CC^H. 
$$

(3.25)

The eigenvector equation of $\frac{1}{J} CC^H$ is

$$
\frac{1}{J} CC^H u_i = \lambda_i u_i. 
$$

(3.26)
Left-multiplication of (3.26) by \( C^H \) gives

\[
\frac{1}{J} C^H C \left( C^H u_i \right) = \lambda_i \left( \frac{1}{J} C^H u_i \right).
\] (3.27)

Therefore, the non-zero eigenvalues of the \( J \times J \) matrix \( \frac{1}{J} C^H C \) are identical to the non-zero eigenvalues of the \( (L - 2K) \times (L - 2K) \) matrix \( \frac{1}{J} C C^H = T_A^H \hat{R}_x T_A \). Denote the eigenvalues of \( T_A^H \hat{R}_x T_A \) (in descending order) as \( \lambda_1, \ldots, \lambda_{L-2K} \) and observe that

\[
\log \left| T_A^H \hat{R}_x T_A \right| = \sum_{i=1}^{J} \log \lambda_i + \chi = \log \left| \frac{1}{J} C^H C \right| + \chi
\] (3.28)

where \( \chi \triangleq \sum_{i=J+1}^{L-2K} \log \lambda_i \) is a singular constant since the \( (L-2K-J) \) smallest eigenvalues of \( T_A^H \hat{R}_x T_A \) are equal to zero. Further simplification leads to

\[
\log \left| T_A^H \hat{R}_x T_A \right| = \log \left| \frac{1}{J} C^H C \right| + \chi = \log \left| \frac{1}{J} \tilde{X}^H T_A T_A^H \tilde{X} \right| + \chi
\] (3.29)

where \( \tilde{X} \) is the \( L \times L \) projection matrix to the subspace complementary to the subspace spanned by the columns of \( \tilde{X} \). The log-likelihood function can now be written as

\[
L_x' \left( \gamma, \left\{ \hat{b}_j \right\}_{j=1}^{J}, \tilde{X} \right) = -\frac{J}{2} \left[ 2K \log(2\pi\sigma^2) + (L - 2K) - J \log J \right]
+ \log \left| \tilde{X}^H P_A^\perp(\gamma) \tilde{X} \right| + \chi.
\]

### 3.2.3 Forward-Backward Tracking

In this subsection, the rationale for using forward-backward tracking will be given, followed by a description of the tracking procedure. Once the log-likelihood function for an audio frame has been developed, a straightforward way to obtain a frequency estimate is to use an ML estimator, i.e. maximize the log-likelihood function w.r.t. \( \gamma \) over a predetermined one-dimensional grid. Upon considering the entire audio stream, how-
ever, in many cases it is possible to assume that the fundamental frequency varies slowly from frame to frame (as in the case of continuous voiced speech). This assumption may be incorporated by treating the vector of fundamental frequencies over $Q$ overlapping frames, denoted as $\mathbf{\gamma} = [\gamma^{(1)}, \ldots, \gamma^{(Q)}]^T$, as a first-order Markov sequence, while defining an \textit{a-priori} transitional pdf.

The following ML-MAP estimator is introduced

$$\hat{\mathbf{\gamma}} = \arg \max_{\mathbf{\gamma}} \max_{\mathbf{B}} f_{\mathbf{X}, \mathbf{\gamma}; \mathbf{B}} (\mathbf{X}, \mathbf{\gamma}; \mathbf{B})$$  \hspace{1cm} (3.30)

where $f_{\mathbf{X}, \mathbf{\gamma}; \mathbf{B}}$ is the joint pdf of the measurement frames and the set of fundamental frequencies $\mathbf{\gamma}$, where the measurements are described by the model in (3.9), parameterized by $\mathbf{B}$. The collection $\mathbf{B}$ includes all the unknown deterministic parameters, namely the complex amplitude vectors and interference statistics. The measurement frames matrix was defined in (3.8). The \textit{a-priori} pdf of the random parameters $\mathbf{\gamma}$ are assumed to be independent of the deterministic parameters $\mathbf{B}$, therefore

$$f_{\mathbf{X}, \mathbf{\gamma}; \mathbf{B}} = f_{\mathbf{X} \mid \mathbf{\gamma}; \mathbf{B}} \cdot f_{\mathbf{\gamma}}.$$  \hspace{1cm} (3.31)

It is assumed that the measurements for the $q$th frame are dependent on $\mathbf{\gamma}$ only through $\gamma^{(q)}$, and that the measurement vectors are statistically independent (which is an assumption not satisfied in practice, in part due to the overlap). The conditional likelihood function $f_{\mathbf{X} \mid \mathbf{\gamma}; \mathbf{B}}$ can be therefore written as

$$f_{\mathbf{X} \mid \mathbf{\gamma}; \mathbf{B}} = \prod_{q=1}^{Q} f_{\mathbf{x}^{(q)} \mid \mathbf{\gamma}_{\mathbf{B}}^{(q)}} = \prod_{q=1}^{Q} f_{\mathbf{x}^{(q)} \mid \mathbf{\gamma}^{(q)}; \mathbf{B}}.$$  \hspace{1cm} (3.32)

Assuming that the sequence $\{\gamma^{(q)}\}_{q=1}^{Q}$ is a 1st order Markov process, and defining the set of conditional pdfs $\{f_{\gamma^{(q)} \mid \gamma^{(q-1)}}\}_{q=2}^{Q}$ and $f_{\gamma^{(1)}}$, the joint pdf of the fundamental
frequencies can be expressed as

\[ f_\gamma = f_{\gamma(1)} \prod_{q=2}^{Q} f_{\gamma(q) | \gamma(q-1)} \]  

(3.33)

and by maximizing the logarithm of the likelihood function and substituting (3.31), (3.32) and (3.33) in (3.30), one obtains (after analytic maximization w.r.t. \( B \))

\[ \hat{\gamma} = \arg \max_\gamma \left\{ L''_{x(1)} + \log f_{\gamma(1)} + \sum_{q=2}^{Q} \left[ L''_{x(q)} + \log f_{\gamma(q) | \gamma(q-1)} \right] \right\} \]

(3.34)

where \( L''_{x(q)} \) is the log-likelihood in (3.30) up to a constant, that is

\[ L''_{x(q)} (\gamma(q)) \triangleq L''_{x(1)} (\gamma(1)), P_{\gamma(q) | \gamma(q-1)} \triangleq \log f_{\gamma(q) | \gamma(q-1)} (\gamma(q), \gamma(q-1)) \text{ and } P_{\gamma(1)} \triangleq \log f_{\gamma(1)} (\gamma(1)). \]

Upon examining the estimator for \( \gamma^{(Q)} \), the order of maximization can be rearranged such
that

\[
\hat{\gamma}_Q = \arg \max_{\gamma(Q)} \max_{\gamma(1), \ldots, \gamma(Q-1)} \left\{ L_{\gamma(1)} + P_{\gamma(1)} + L_{\gamma(q)}/\gamma(q-1) \right\} \\
+ \sum_{q=2}^{Q-1} \left[ L_{\gamma(q)} + P_{\gamma(q)}/\gamma(q-1) \right]
\]

\[
= \arg \max_{\gamma(Q)} \left\{ L_{\gamma(Q)} + \max_{\gamma(1), \ldots, \gamma(Q-1)} \left\{ L_{\gamma(1)} + P_{\gamma(1)} + P_{\gamma(Q)/\gamma(Q-1)} + \sum_{q=2}^{Q-1} \left[ L_{\gamma(q)} + P_{\gamma(q)/\gamma(q-1)} \right] \right\} \right\}.
\]

Define

\[
W_Q \triangleq L_{\gamma(Q)} + \max_{\gamma(1), \ldots, \gamma(Q-1)} \left\{ L_{\gamma(1)} + P_{\gamma(1)} + P_{\gamma(Q)/\gamma(Q-1)} + \sum_{q=2}^{Q-1} \left[ L_{\gamma(q)} + P_{\gamma(q)/\gamma(q-1)} \right] \right\}
\]

(3.38)

and observe that

\[
W_Q = L_{\gamma(Q)} + \max_{\gamma(Q-1)} \left\{ P_{\gamma(Q)/\gamma(Q-1)} + L_{\gamma(Q-1)} \right\}
\]

(3.39)

\[
+ \max_{\gamma(1), \ldots, \gamma(Q-2)} \left\{ L_{\gamma(1)} + P_{\gamma(1)} + P_{\gamma(Q)/\gamma(Q-1)} + \sum_{q=2}^{Q-2} \left[ L_{\gamma(q)} + P_{\gamma(q)/\gamma(q-1)} \right] \right\}
\]

\[
= L_{\gamma(Q)} + \max_{\gamma(Q-1)} \left\{ W_{Q-1} + P_{\gamma(Q)/\gamma(Q-1)} \right\}
\]

where

\[
W_{Q-1} \triangleq L_{Q-1} + \max_{\gamma(Q-2)} \left\{ L_{\gamma(1)} + P_{\gamma(1)} + P_{\gamma(Q-1)/\gamma(Q-2)} + \sum_{q=2}^{Q-2} \left[ L_{\gamma(q)} + P_{\gamma(q)/\gamma(q-1)} \right] \right\}
\]

(3.40)

By generalization of (3.41), define

\[
W_q = L_{\gamma(q)} + \max_{\gamma(q-1)} \left\{ W_{q-1} + P_{\gamma(q)/\gamma(q-1)} \right\}, \quad q = 2, \ldots, Q
\]

(3.41)

\[
W_1 = L_{\gamma(1)} + \max_{\gamma(1)} P_{\gamma(1)}.
\]
Using (3.37) and (3.38), it holds that

$$\hat{\gamma}^{(Q)} = \arg \max_{\gamma^{(Q)}} W_q.$$  \hspace{1cm} (3.42)

In order to develop the estimator for $\gamma^{(Q-1)}$, the order of maximization is changed in (3.36)

$$\hat{\gamma}^{(Q-1)} = \arg \max_{\gamma^{(Q-1)}} \max_{\gamma^{(1)}, \ldots, \gamma^{(Q-2)}} \max_{\gamma^{(Q)}} \{ L_{\gamma^{(1)}} + P_{\gamma^{(1)}} + L_{\gamma^{(Q)}} 
+ P_{\gamma^{(Q)}/\gamma^{(Q-1)}} + \sum_{i=2}^{Q-1} \left[ L_{\gamma^{(i)}} + P_{\gamma^{(i)}/\gamma^{(q-1)}} \right] \} \n= \arg \max_{\gamma^{(Q-1)}} \max_{\gamma^{(1)}, \ldots, \gamma^{(Q-2)}} \{ L_{\gamma^{(1)}} + P_{\gamma^{(1)}} + L_{\hat{\gamma}^{(Q)}} + P_{\hat{\gamma}^{(Q)}/\gamma^{(Q-1)}} 
+ \sum_{q=2}^{Q-1} \left[ L_{\gamma^{(q)}} + P_{\gamma^{(q)}/\gamma^{(q-1)}} \right] \} \n= \arg \max_{\gamma^{(Q-1)}} \{ P_{\hat{\gamma}^{(Q)}/\gamma^{(Q-1)}} + W_{Q-1} \} \n= \arg \max_{\gamma^{(Q-1)}} \{ P_{\hat{\gamma}^{(Q)}/\gamma^{(Q-1)}} + W_{Q-1} \} \hspace{1cm} (3.43)

where

$$L_{\hat{\gamma}^{(Q)}} \triangleq L''_{\gamma^{(Q)}}(\hat{\gamma}^{(Q)})$$

$$P_{\hat{\gamma}^{(Q)}/\gamma^{(Q-1)}} \triangleq f_{\gamma^{(Q)}|\gamma^{(Q-1)}}(\hat{\gamma}^{(Q)}, \gamma^{(Q-1)}).$$

We may generalize (3.43) to obtain recursive estimators for all coordinates of $\gamma$

$$\hat{\gamma}^{(Q)} = \arg \max_{\gamma^{(Q)}} W_q$$

$$\hat{\gamma}^{(q)} = \arg \max_{\gamma^{(q)}} \{ P_{\hat{\gamma}^{(q+1)}/\gamma^{(q)}} + W_q \}, \quad q = 1, \ldots, Q - 1$$

$$P_{\hat{\gamma}^{(q+1)}/\gamma^{(q)}} \triangleq f_{\gamma^{(q+1)}|\gamma^{(q)}}(\hat{\gamma}^{(q+1)}, \gamma^{(q)}).$$

The relations in (3.45) yield estimators such that the estimate for the $q$th frame depends on the estimate for the $(q+1)$th frame, and is therefore dependent on the measurements for all
following frames. Implementation of a forward-only estimator for $\gamma$ may be accomplished by defining the following estimator which maximizes only according to past frames

$$
\hat{\gamma}^{(q)} = \arg \max_{\gamma^{(1)}, \ldots, \gamma^{(q)}} \left\{ L_{\gamma^{(1)}} + P_{\gamma^{(1)}} + \sum_{q=2}^{Q} \left[ L_{\gamma^{(q)}} + P_{\gamma^{(q)}/\gamma^{(q-1)}} \right] \right\},
$$

$$
q = 1, \ldots, Q
$$

(3.46)

which is equivalent to

$$
\hat{\gamma}^{(q)} = \arg \max_{\gamma^{(q)}} W_q, \quad q = 1, \ldots, Q.
$$

(3.47)

The forward-only estimator given by (3.48) uses less information than (3.45) as it discards data from upcoming frames and consequently exhibits inferior performance. A discrete implementation of (3.45) is now given. Assume that each element of $\gamma$ takes values from a grid $r_1, \ldots, r_Z$. Define the following terms

$$
V_i(q) \triangleq L''_{x(q)} (\gamma^{(q)} = r_i), \quad q = 1, \ldots, Q, \quad i = 1, \ldots, Z
$$

(3.48)

$$
P_{j,i}(q) \triangleq \log p(\gamma^{(q)} = r_i | \gamma^{(q-1)} = r_j), \quad q = 1, \ldots, Q, \quad i, j = 1, \ldots, Z.
$$

(3.49)

The conditional probabilities $p(\gamma^{(q)} = r_i | \gamma^{(q-1)} = r_j)$ are used instead of the conditional pdfs, since the elements of $\gamma$ are evaluated over a discrete grid. The conditional probabilities are normalized so that their sum equals unity. Now define the following

$$
W_i(q) \triangleq \max_{j=1, \ldots, Z} \left[ W_j(q-1) + P_{j,i}(q) \right] + V_i(q) \quad q = 1, \ldots, Q, \quad i = 1, \ldots, Z
$$

(3.50)

with the initial condition: $W_i(0) = 0$ for $i = 1, \ldots, Z$. The MAP estimator of $\gamma^{(q)}$ is given
by

\[ \hat{\gamma}^{(q)} = r_{\hat{m}(q)}, \quad q = 1, \ldots, Q \]  

\[ \hat{m}(q) = \arg \max_{j=1,\ldots,Z} \left[ W_j(q) + \log p(\gamma^{(q+1)} = r_{\hat{m}(q+1)} | \gamma^{(q)} = r_j) \right]. \]  

It is easy to see that the frequency estimate for the \( q \)th frame depends on the estimation result of the following frame.

The transition probability matrix is chosen according to prior considerations. Since the fundamental frequency commonly changes slowly over time, the transition distribution should be centered around the estimated frequency from the previous frame with a small standard deviation. In order to further constrain the algorithm to “look” only within a small range of frequencies centered around the previous result, the distribution may be set to zero outside a chosen range. This greatly improves the consistency of the estimated frequency to source association (i.e. correctly matching the estimated frequencies to their respective sources) when extending the algorithm to deal with multiple harmonic sources.

### 3.3 Multiple Pitch Tracking With Projection

In this section, the proposed method is extended to accommodate additional harmonic sources using an iterative cancellation method with the aid of harmonic projection matrices. The vector \( \hat{\gamma}_1 = [\hat{\gamma}_1^{(1)}, \ldots, \hat{\gamma}_1^{(Q)}]^T \) represents the fundamental frequency estimates for the dominant harmonic source in the input audio stream, and is obtained with the procedure detailed in Subsection 3.2.3. In order to obtain frequency estimates for the next-dominant source, the dominant harmonic source is removed from the input audio stream. This is accomplished by left-multiplying the frame measurement vector by an appropriate harmonic projection matrix. The data model for the \( q \)th frame at the first iteration can be represented as (3.9)

\[ x_1^{(q)} = A(\gamma_1^{(q)})b_1^{(q)} + e_1^{(q)} \]  

36
where the added subscript indicates the iteration number. The frame interference can be expanded as

$$e_1^{(q)} = A(\gamma_2^{(q)})b_2^{(q)} + \bar{e}_2^{(q)}$$ (3.54)

where $A(\gamma_2^{(q)})b_2^{(q)}$ is the second-dominant harmonic source and $\bar{e}_2^{(q)}$ represents any further harmonic sources as well as noise. The harmonic projection matrix onto the subspace complementary to the subspace spanned by the columns of the matrix $A(\hat{\gamma}_1^{(q)})$ is defined as

$$P_{A}(\hat{\gamma}_1^{(q)}) = I_N - \frac{1}{A(\hat{\gamma}_1^{(q)})A(\hat{\gamma}_1^{(q)})^H}A(\hat{\gamma}_1^{(q)})A(\hat{\gamma}_1^{(q)})^H.$$ (3.55)

Multiplication of the $q$th frame measurement vector by (3.55) and substituting (3.54) gives

$$x_2^{(q)} \triangleq P_{A}(\hat{\gamma}_1^{(q)})x_1^{(q)} = \Delta + P_{A}(\hat{\gamma}_2^{(q)})e_1^{(q)} = A_2(\gamma_2^{(q)})b_2^{(q)} + e_2^{(q)}$$ (3.56)

where $A_2(\gamma_2^{(q)}) \triangleq P_{A}(\hat{\gamma}_1^{(q)})A(\gamma_2^{(q)})$ and $e_2^{(q)} \triangleq P_{A}(\hat{\gamma}_1^{(q)})\bar{e}_2^{(q)} + \Delta$ and $\Delta \triangleq P_{A}(\hat{\gamma}_1^{(q)})A(\gamma_1^{(q)})P_{A}(\hat{\gamma}_1^{(q)})$ is the residual of the dominant harmonic source (this does not carry any consequences, since the interference statistics are estimated anyway). The assumption here is that the estimate $\hat{\gamma}_1^{(q)}$ is close to $\gamma_1^{(q)}$ so that the dominant harmonic source is mostly removed. In other words, $\Delta$ decreases to zero as $\hat{\gamma}_1^{(q)} \rightarrow \gamma_1^{(q)}$. Therefore, in order to obtain the measurement vectors at the second iteration, harmonic projection matrices are calculated for each frame, yielding the set of $Q$ matrices $P_{A}(\hat{\gamma}_1^{(1)}), \ldots, P_{A}(\hat{\gamma}_1^{(Q)})$. The vectors of measurement frames are projected out of the estimate of the dominant source’s harmonic subspace using the set of projection matrices such that $x_2^{(q)} \triangleq P_{A}(\hat{\gamma}_1^{(q)})x_1^{(q)}$, $q = 1, \ldots, Q$ is the collection of frame vectors at the second iteration. The log-likelihood function for the next iteration reflects the projected model (3.56)

$$L_{x_2^{(q)}}(\gamma_2^{(q)}) = -\frac{J}{2} \log \left| \left( \hat{X}_2^{(q)} \right)^H P_{A_2}(\gamma_2^{(q)}) \hat{X}_2^{(q)} \right|$$ (3.57)
where \( \tilde{A}_2(\gamma_2^{(q)}) \triangleq P_A^{\perp}(\hat{\gamma}_1^{(q)}A(\gamma_2^{(q)})) \) and
\[
P_A^{\perp}(\gamma_2^{(q)}) = \tilde{A}_2(\gamma_2^{(q)}) \left[ \tilde{A}_2^H(\gamma_2^{(q)})\tilde{A}_2(\gamma_2^{(q)}) \right]^{-1} \tilde{A}_2^H(\gamma_2^{(q)}).
\]
(3.58)

The columns of the matrix \( \tilde{X}_2^{(q)} \) contain the subframe vectors for the \( q \)th frame at the second iteration. The upper \( \tilde{\cdot} \) indicates a subframe-sized matrix or vector. Once the collection of log-likelihood functions \( L''_{x_2^{(q)}}(1), \ldots, L''_{x_2^{(q)}}(Q) \) have been calculated, the forward-backward tracking procedure is used to obtain the frequency estimates for the second-dominant harmonic source \( \tilde{\gamma}_2 = \left[ \hat{\gamma}_2^{(1)}, \ldots, \hat{\gamma}_2^{(Q)} \right] \).

In order to generalize the procedure for the \( m \)th iteration, define the following
\[
G_m^{(q)} \triangleq \begin{bmatrix} A(\hat{\gamma}_1^{(q)}), \ldots, A(\hat{\gamma}_m^{(q)}) \end{bmatrix}
\]
(3.59)
\[
P_{G_m^{(q)}} \triangleq I_N - G_m^{(q)} \left[ (G_m^{(q)})^H G_m^{(q)} \right]^{-1} (G_m^{(q)})^H.
\]
(3.60)

In this manner, a projection matrix onto the subspace complementary to the first \( m \) estimated harmonic subspaces is constructed. The model for the \( q \)th frame vector at the \( m \)th, \( m > 1 \) iteration is given by
\[
x_m^{(q)} \triangleq P_{G_m^{(q)}} = A_m(\gamma_m^{(q)})b_m^{(q)} + e_m^{(q)}
\]
(3.61)

where \( b_m^{(q)} \) is the vector of complex amplitudes for the \( m \)th harmonic source and \( A_m(\gamma_m^{(q)}) \triangleq P_{G_m^{(q)}} A(\gamma_m^{(q)}) \) and \( e_m^{(q)} \) represents any projected further harmonic sources, as well as the residuals for the first \( m - 1 \) harmonic sources and any additive noise.

In order to define the corresponding log-likelihood functions for the \( m \)th iteration, define the \( L \times L \) subframe projection matrices
\[
\tilde{G}_m^{(q)} \triangleq \begin{bmatrix} \tilde{A}(\hat{\gamma}_1^{(q)}), \ldots, \tilde{A}(\hat{\gamma}_m^{(q)}) \end{bmatrix}
\]
(3.62)
\[
P_{\tilde{G}_m^{(q)}} \triangleq I_L - \tilde{G}_m^{(q)} \left[ (\tilde{G}_m^{(q)})^H \tilde{G}_m^{(q)} \right]^{-1} (\tilde{G}_m^{(q)})^H.
\]
(3.63)
The corresponding log-likelihood functions are calculated for \( x_m^{(1)}, \ldots, x_m^{(Q)} \) using

\[
L''_{x_m^{(q)}}(\gamma_m^{(q)}) = -\frac{J}{2} \log \left| (\tilde{X}_m^{(q)})^H P_{\tilde{A}_m}^{\perp}(\gamma_m^{(q)}) \tilde{X}_m^{(q)} \right|
\]  

(3.64)

where \( \tilde{A}_m(\gamma_m^{(q)}) \triangleq P_{\tilde{G}_{m-1}}^{\perp} \tilde{A}(\gamma_m^{(q)}) \) and

\[
P_{\tilde{A}_m}^{\perp}(\gamma_m^{(q)}) = \tilde{A}_m(\gamma_m^{(q)}) \left[ \tilde{A}_m^{H}(\gamma_m^{(q)}) \tilde{A}_m(\gamma_m^{(q)}) \right]^{-1} \tilde{A}_m^{H}(\gamma_m^{(q)})
\]  

(3.65)

and the columns of \( \tilde{X}_m^{(q)} \) contain the subframe vectors for \( x_m^{(q)} \). The forward-backward procedure is used to calculate the fundamental frequency estimates \( \hat{\gamma}_m^{(1)}, \ldots, \hat{\gamma}_m^{(Q)} \).

### 3.3.1 Algorithm Outline

The following parameters are set:

- \( M \) - number of harmonic sources
- \( r_1, \ldots, r_Z \) - frequency grid
- \( N \) - length of input frames
- Overlap length of input frames
- \( L \) - length of subframes
- Overlap length of subframes
- \( K \) - number of real harmonics used in the harmonic matrix \( A \)

The values of the above parameters control the performance and execution time of the algorithm. For example, a denser frequency grid results in improved estimates, but requires more processing. Similarly, lowering the overlap length of subframes results in more subframes. This leads to a larger matrix in the determinant expression of (3.30), requiring more execution time. In Chapter 4, nominal values for these parameters have been empirically selected.
Algorithm 1 Calculate multiple fundamental frequencies

Determine frequency grid \( r = [r_1, \ldots, r_Z]^T \).

Load input stream and split into overlapping frames \( x_1^{(1)}, \ldots, x_1^{(Q)} \).

Calculate the \( Z \times Z \) probability transition matrix \( P \) where
\[
[P]_{i,j} = \log p(\gamma^{(q)} = r_i | \gamma^{(q-1)} = r_j) \text{ for } q = 2, \ldots, Q.
\]

\textbf{for} \( m = 1 \text{ to } M \) \textbf{do}

\textbf{if} \( m \neq 1 \) then

Calculate the harmonic projection matrices \( P_{G_m}^G_{(1)}, \ldots, P_{G_m}^G_{(q)} \) and \( P_{\tilde{G}_m}^G_{(1)}, \ldots, P_{\tilde{G}_m}^G_{(q)} \)

using (3.60) and (3.63), respectively.

\textbf{else}

\[
P_{G_m}^G_{(q)} = I_N, \quad P_{\tilde{G}_m}^G_{(q)} = I_L, \quad q = 1, \ldots, Q.
\]

\textbf{end if}

Calculate the \( Q \) frame vectors for the \( m \)th iteration using \( x_m^{(q)} = P_{G_m}^G_{(q)} x_1^{(q)}, \quad q = 1, \ldots, Q \) and split each into overlapping subframes and arrange in the columns of matrix, yielding \( \tilde{X}_m^{(1)}, \ldots, \tilde{X}_m^{(Q)} \).

Calculate the \( Z \times Q \) log-likelihood matrix \( L_{X_m} \) where \( [L_{X_m}]_{i,q} = L''_{x_m^{(q)}}(r_i) \) using (3.64).

Calculate the \( Z \times Q \) matrix \( W_m \) using
\[
[W_m]_{i,1} = \max_{k=1,\ldots,Z} \{ [P]_{k,i} + [L_{X_m}]_{i,1} \}
\]
\[
[W_m]_{i,q} = \max_{k=1,\ldots,Z} \left\{ [W_m]_{k,q-1} + [P]_{k,i} \right\} + [L_{X_m}]_{i,q}, \quad q = 2, \ldots, Q \quad (3.66)
\]

Obtain a vector of frequency estimates \( \hat{\gamma}_m = [\hat{\gamma}_m^{(1)}, \ldots, \hat{\gamma}_m^{(Q)}]^T \)
\[
\hat{m}_q = \max_{i=1,\ldots,Z} \{ [W_m]_{i,q} + [P]_{i,\hat{m}_{q+1}} \} \quad q = 2, \ldots, Q \quad (3.67)
\]
\[
\hat{m}_Q = \max_{i=1,\ldots,Z} [W_m]_{i,Q}
\]
\[
\hat{\gamma}_m^{(q)} = r_{\hat{m}_q}, \quad q = 1, \ldots, Q
\]

\textbf{end for}
Chapter 4

Simulation Results

In this chapter, MATLAB simulation results are presented for the proposed ML-MAP method for real and artificial signals. Individual signals were artificially mixed at different signal-to-interference ratios (SIRs) where one source is regarded as the signal and the other source is treated as interference. The SIR is defined as $\frac{\sigma_s^2}{\sigma_i^2}$ where $\sigma_s^2$ is the signal power and $\sigma_i^2$ is the interference power.

4.1 Synthetic Signals

In this section, a single audio frame 30ms in length sampled at 10kHz containing two harmonic sources with fundamental frequencies 150Hz and 210Hz was synthesized and the fundamental frequencies were estimated using the ML method (with only a single audio frame it is not possible to perform tracking). The parameters for the synthetic signal simulations are given in Table 4.1.

Fig. 4.1 shows the result for sources with 5 harmonics each mixed at an SNR of 10dB. The amplitude spectra of the individual sources are displayed along with the log-likelihood and the projected log-likelihood. The peaks fairly correspond to the actual frequencies indicated by the dots. It may be seen how the projection cancels the dominant harmonic source from the mixture.

Fig. 4.2 shows the result for a similar simulation where the first source is missing the
Table 4.1: Simulation Parameters for Synthetic Signals

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling rate</td>
<td>10kHz</td>
</tr>
<tr>
<td>Input stream length / samples</td>
<td>30msec / 300</td>
</tr>
<tr>
<td>Frame length / samples</td>
<td>30msec / 300</td>
</tr>
<tr>
<td>Number of frames</td>
<td>1</td>
</tr>
<tr>
<td>Subframe length / samples</td>
<td>18msec / 180</td>
</tr>
<tr>
<td>Subframe overlap length / samples</td>
<td>17.64msec / 176</td>
</tr>
<tr>
<td>Number of subframes</td>
<td>34</td>
</tr>
<tr>
<td>Frequency grid + resolution</td>
<td>55-250Hz @1Hz</td>
</tr>
<tr>
<td>SIR</td>
<td>0dB</td>
</tr>
<tr>
<td>Real harmonics ($K$)</td>
<td>5</td>
</tr>
</tbody>
</table>

second and fourth harmonic and the second source is missing the third harmonic. The amplitude spectra of the individual sources are displayed along with the log-likelihood and the projected log-likelihood, which are not much different from those displayed in 4.1-(D), where all harmonics are present. This example illustrates the ML estimator’s resilience to missing harmonics, in contrast to the harmonic MUSIC estimator.

4.2 Real Signals

In this section, a set of simulations on speech mixtures and music mixtures containing two sources were carried out in order to evaluate the performance of the proposed method. Music samples were taken from the University of Iowa website\(^1\), which offers a collection of music samples ordered by instrument and recorded in an anechoic chamber resulting in 16 bit, 44.1 kHz audio files. Speech samples were taken from the Keele University database [20] which consists of a phonetically balanced text read by male and female speakers, resulting in 16 bit, 20kHz audio files. The MATLAB simulations were run on a cluster powered by dual 2.5GHz/12M/1333MHz 80W Quad-Core Xeon processors with 16GB of memory. All mixtures were created synthetically at SIRs where one source is regarded as the signal and the other source is treated as interference. The SIR is defined as $\frac{\sigma_s^2}{\sigma_i^2}$ where $\sigma_s^2$ is the signal power and $\sigma_i^2$ is the interference power.

\(^1\)http://theremin.music.uiowa.edu/MIS.html
Figure 4.1: Two synthetic signals were mixed with AWGN at an SNR of 10dB. Plots (A) and (B) display the amplitude spectra of the signals, plot (C) displays the log-likelihood and plot (D) displays the projected log-likelihood after maximizing the log-likelihood. The dots indicate the true fundamental frequencies.

The proposed method was compared to a reference method (Klapuri) detailed in [12] using an implementation provided by the author (the reference method assumes a 44.1kHz sampling rate). Two types of error measures were examined - gross error rate (GER) and root mean square error (RMSE). The GER is a measure of “large” errors, defined as the percentage of estimated fundamental frequencies which differ from the true frequency by more than a given threshold which is determined as 5Hz. The RMSE is defined as square root of the average squared estimation error with estimation errors which are smaller than the GER threshold of 5Hz. In order to obtain “ground truth” values for the fundamental frequencies, the proposed method was used separately on the unmixed sources (i.e. in a single harmonic source case). Inherent errors due to estimation of double or half the
Figure 4.2: Two synthetic signals were mixed with AWGN at an SNR of 10dB. Plots (A) and (B) display the amplitude spectra of the signals, plot (C) displays the log-likelihood and plot (D) displays the projected log-likelihood after maximizing the log-likelihood. The dots indicate the true fundamental frequencies.

true frequency are not taken into account, as the estimation results are multiplied by the correcting factor so that the total estimation error is minimized.

4.2.1 Example 1 - Log-likelihood Illustration

In this example, two sources (female and male speakers from the Keele database) were mixed together and the fundamental frequencies were estimated using the proposed method. The spectrograms, log-likelihoods, and tracking log-likelihoods are displayed for each source in the first, second and third rows of Fig. 4.3, respectively, along with the separate pitch tracking lines. The estimated pitches using the log-likelihood (no tracking), forward tracking and forward-backward tracking for sources A and B are shown in Figs. 4.4 and
4.5, respectively. The simulation parameters are given in Table 4.2.

Table 4.2: Simulation Parameters for Example 1

<table>
<thead>
<tr>
<th>Source A</th>
<th>f1nw0000 @ 1.37sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source B</td>
<td>m3nw0000 @ 3.2sec</td>
</tr>
<tr>
<td>Sampling rate</td>
<td>20kHz</td>
</tr>
<tr>
<td>Input stream length / samples</td>
<td>306.5msec / 6131</td>
</tr>
<tr>
<td>Frame length / samples</td>
<td>23.22msec / 464</td>
</tr>
<tr>
<td>Frame overlap length / samples</td>
<td>9.3msec / 185</td>
</tr>
<tr>
<td>Number of frames</td>
<td>41</td>
</tr>
<tr>
<td>Subframe length / samples</td>
<td>21.3msec / 426</td>
</tr>
<tr>
<td>Subframe overlap length / samples</td>
<td>20.8msec / 416</td>
</tr>
<tr>
<td>Number of subframes</td>
<td>10</td>
</tr>
<tr>
<td>Frequency grid + resolution</td>
<td>55-230Hz @1Hz</td>
</tr>
<tr>
<td>SIR</td>
<td>0dB</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>25Hz</td>
</tr>
<tr>
<td>PDF window length (samples)</td>
<td>15</td>
</tr>
<tr>
<td>Real harmonics (K)</td>
<td>5</td>
</tr>
</tbody>
</table>

The spectrogram plots in Fig. 4.3 show the estimated pitch line following the spectrogram peak. The dominant source in this case is Source A (female speaker). The log-likelihoods and tracking log-likelihoods with the estimated pitch lines are also displayed. It can be observed that the log-likelihood sometimes has a significant peak not at the true frequency. Forward-backward tracking uses *a-priori* information to mitigate errors that can arise where the highest peak in the log-likelihood is in the wrong location. The tracking log-likelihood takes numeric values within a limited range near the estimated frequency due to the windowed conditional probability matrix.

In Fig. 4.4-(A), many estimated pitches (open circles) are far from the actual pitches, indicating the inadequacy of ML estimation without any added tracking. The situation is improved in plot (B) where forward tracking is employed, but a few of the estimated pitches are far from the real values. Forward-backward tracking is used to estimate the pitches in plot (C), which shows the best results. It should be noted that zeroing the conditional pdf to constrain the look window for estimating successive pitches is effective only when forward-backward tracking is used, and thus results in a relatively smooth
Figure 4.3: Analysis of male and female speakers. The spectrograms are shown in row 1, the log-likelihood functions are shown in row 2 and the tracking log-likelihood functions are shown in row 3. The pitch tracking lines are shown for each plot. The tracking log-likelihood takes values within the range specified by the pdf window.

pitch tracking line. The improvement by using forward-backward tracking in the case of the male speaker in Fig. 4.5 is less marked.

4.2.2 Example 2 - Speaker Combinations

These simulations show the GER and RMSE for each of the two individual sources as a function of SIR. The GER and MSE were averaged over 400 independent simulations. Different combinations of male and female speakers are examined: male-female (Fig. 4.6), male-male (Fig. 4.7) and female-female (Fig. 4.8). The data for each source was selected
Figure 4.4: Spectrogram for female speaker with pitch reference (solid line). The estimated pitch (open circles) is shown: by maximizing the log-likelihood in plot (A), by using forward tracking in plot (B), by using forward-backward tracking in plot (C).

from a collection of a small number of different speakers. The simulation parameters are shown in Table 4.3. Results using the reference method (Klapuri) are also shown for comparison, using an upsampled version of the source mixture. Runtime of our MATLAB implementation using the values given in Table 4.3 is approximately 70 times slower than realtime.

In all cases, it is seen that after a certain threshold, the GER declines, with a simultaneous increase in the RMSE. Note that the reference method yields inferior results, indicating that it most likely was not designed to analyze speech signals.
Figure 4.5: Spectrogram for male speaker with pitch reference (solid line). The estimated pitch (open circles) is shown: by maximizing the log-likelihood in plot (A), by using forward tracking in plot (B), by using forward-backward tracking in plot (C).

4.2.3 Example 3 - Music Combination

For this simulation, audio segments from the following instruments were mixed: violin, cello, flute, clarinet, oboe, trombone, tuba, trumpet and saxophone. The simulation parameters are given in Table 4.4. Downsampling to 10kHz was performed to save execution time. The GER and RMSE for each of two individual sources as a function of SIR are displayed in Fig. 4.9. The GER and MSE were averaged over 62 independent simulations. The MATLAB execution was 320 times slower than realtime.

The GER decreases rapidly with SIR, and it may be seen that our method outperforms the reference method for SIR larger than -10dB. The RMSE is also lower with our method than with the reference method.
Table 4.3: Simulation Parameters for Example 2

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling rate</td>
<td>20kHz</td>
</tr>
<tr>
<td>Input stream length / samples</td>
<td>306.5msec / 6131</td>
</tr>
<tr>
<td>Frame length / samples</td>
<td>23.22msec / 464</td>
</tr>
<tr>
<td>Frame overlap length / samples</td>
<td>9.3msec / 185</td>
</tr>
<tr>
<td>Number of frames</td>
<td>31</td>
</tr>
<tr>
<td>Subframe length / samples</td>
<td>21.3msec / 426</td>
</tr>
<tr>
<td>Subframe overlap length / samples</td>
<td>20.9msec / 418</td>
</tr>
<tr>
<td>Number of subframes</td>
<td>6</td>
</tr>
<tr>
<td>Frequency grid + resolution</td>
<td>55-250Hz @1Hz</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>15Hz</td>
</tr>
<tr>
<td>PDF window length (samples)</td>
<td>9</td>
</tr>
<tr>
<td>Real harmonics ($K$)</td>
<td>4</td>
</tr>
<tr>
<td>Klapuri - frame size (samples)</td>
<td>1024</td>
</tr>
<tr>
<td>Klapuri - frame hop (samples)</td>
<td>409</td>
</tr>
</tbody>
</table>

Table 4.4: Simulation Parameters for Example 3

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling rate</td>
<td>10kHz</td>
</tr>
<tr>
<td>Input stream length / samples</td>
<td>500msec / 4852</td>
</tr>
<tr>
<td>Frame length / samples</td>
<td>46.43msec / 463</td>
</tr>
<tr>
<td>Frame overlap length / samples</td>
<td>18.6msec / 185</td>
</tr>
<tr>
<td>Number of frames</td>
<td>16</td>
</tr>
<tr>
<td>Subframe length / samples</td>
<td>42.7msec / 426</td>
</tr>
<tr>
<td>Subframe overlap length / samples</td>
<td>41.86msec / 418</td>
</tr>
<tr>
<td>Number of subframes</td>
<td>8</td>
</tr>
<tr>
<td>Frequency grid + resolution</td>
<td>55-1150Hz @1Hz</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>15Hz</td>
</tr>
<tr>
<td>PDF window length (samples)</td>
<td>25</td>
</tr>
<tr>
<td>Real harmonics ($K$)</td>
<td>5</td>
</tr>
<tr>
<td>Klapuri - frame size (samples)</td>
<td>2048</td>
</tr>
<tr>
<td>Klapuri - frame hop (samples)</td>
<td>1228</td>
</tr>
</tbody>
</table>
Figure 4.6: Male-Female analysis using the proposed method and the reference method (Klapuri). The GER and RMSE for each source are shown as a function of SIR, averaged over 400 independent simulations.
Figure 4.7: Male-Male analysis using the proposed method and the reference method (Klapuri). The GER and RMSE for each source are shown as a function of SIR, averaged over 400 independent simulations.
Figure 4.8: Female-Female analysis using the proposed method and the reference method (Klapuri). The GER and RMSE for each source are shown as a function of SIR, averaged over 400 independent simulations.
Figure 4.9: Music analysis using the proposed method and the reference method (Klapuri). The GER and RMSE for each source are shown as a function of SIR, averaged over 67 independent simulations.
Chapter 5

Discussion and Conclusions

5.1 Discussion and Conclusions

In this thesis, a new method for multiple pitch estimation was presented. The proposed ML-MAP method is based on the harmonic model with unknown random interference, for which the log-likelihood function is developed for a single frame of data, including maximization w.r.t. the deterministic nuisance parameters. A MAP estimator is then derived for the collection of fundamental frequencies over all data frames, using the partially maximized log-likelihood function (hence the name “ML-MAP”). In this manner, the fundamental frequencies for the dominant harmonic source are estimated, after which the dominant source is removed from the measurements using harmonic projection matrices. The fundamental frequencies for the next dominant harmonic source are then estimated in a similar manner using a projected harmonic model. This procedure is repeated for any remaining harmonic sources. Advantages of this approach include the use of a robust data model in which relatively little assumptions are made regarding the harmonic sources and any non-harmonic interference/noise. This has the added benefit of describing multiple scenarios involving one or more harmonic sources. Furthermore, the tracking algorithm ensures a smooth pitch tracking line (under the assumption that the harmonic sources remains voiced for the duration of the input stream) by substituting \textit{a-priori} knowledge
when information from the measurements is weak. Tracking also provides rudimentary
fundamental frequency to source association. However, the use of projection matrices
and of the log-likelihood for the considered model result in high computation complexity.
Methods for fundamental frequency estimation based on MUSIC were also examined and
found to be generally inferior, primary due to a lack of robustness inherent to the MUSIC
criterion (2.21) and to the lack of an integrated tracking method.

The determinant expression in the log-likelihood (3.30) may be interpreted as a mea-
ure analogous to the signal power outside the harmonic subspace for the chosen grid
frequency. In other words, the subframe spectrum is nullified at the specified pitch and at
a limited number of harmonics, then cross-correlated with itself, after which the determi-
nant is calculated. When the determinant quantity is small, it provides an indication of a
harmonic source at the chosen frequency and a candidate for the dominant fundamental
frequency. Calculation of the log-likelihood at a simple ratio of the true frequency nullifies
part of the correct harmonics, lowering the energy outside the harmonic subspace. This
generalized energy quantity is a more robust measure than the MUSIC criterion which
checks the orthogonality to a harmonic matrix. The MUSIC criterion measures the or-
thogonality between the signal null subspace and the complete harmonic model. In other
words, it checks whether the signal includes all the harmonics components. Conversely,
the proposed ML-based method measures the orthogonality between the signal subspace
and the complementary space of the harmonic model, i.e. it measures the extent to which
the signal subspace can be represented by the assumed harmonic model. This renders
the log-likelihood less susceptible to model mismatch, since modeling errors are of less
consequence when examining the residual.

Some problematic issues inherent to multiple pitch estimation include frequency halv-
ing/doubling, association between harmonic sources and voiced/unvoiced detection. Fre-
quency halving or doubling refers to the detection of a simple ratio of the true fundamental
frequency, usually half or twice the actual value. Association between harmonic sources
means consistently matching the pitch estimates over time with their respective sources.
Shao and Wang proposed a sequential grouping algorithm [24] for cochannel source as-
sociation, which extends a previously introduced hidden Markov model-based multipitch estimation algorithm [25]. The proposed method includes a tracking algorithm which generally yields consistent association. Voiced/unvoiced detection is relatively simple in the single-pitch case [26], however the multiple combinations which arise in the more general case contribute to the complexity of the issue. A more general algorithm also would include some kind of voiced/unvoiced detection which would allow analysis of general speech, without being limited to voiced-only segments.

We believe that the algorithm proposed in this thesis represents a promising methodology for multiple fundamental frequency estimation based on the harmonic model. Further work in this area can include finding ways to lower the computational complexity in the matrix calculations for the log-likelihood. Other issues left for future research are voiced-unvoiced detection for the multiple source case as well as a method for determining the number of harmonic sources present in the measurements.
Bibliography


Appendix A

Narrowband model used in array processing

The narrowband array-processing model for an $N$-sensor array with $M$ sources is given as [18]

$$
x_j = a(\theta_1)s_{1j} + \cdots + a(\theta_M)s_{Mj} + n_j \tag{A.1}
$$

$$
= A(\theta)s_j + n_j \quad j = 1, \ldots, J
$$

where $J$ is the total number of snapshots, $x_j$ is the $j$th $N \times 1$ signal snapshot vector, $A(\theta) \triangleq [a(\theta_1), \ldots, a(\theta_M)]^T$ where $a(\theta_i)$ denotes the $i$th steering vector, $\theta \triangleq [\theta_1, \ldots, \theta_M]^T$. Each angle $\theta_m$, $m = 1, \ldots, M$ represents a direction-of-arrival (DOA) corresponding to
an emitting source. The random noise vector has the distribution

\[ \mathbf{n}_j \sim \mathcal{N}^c (0, \mathbf{R}_n) \]  \hspace{1cm} (A.2)

where the autocorrelation matrix \( \mathbf{R}_n \) may be deterministic unknown and \( \mathbf{n}_i, \mathbf{n}_j \) are uncorrelated for \( i \neq j \). The signal vectors \( \{ \mathbf{s}_j \}_{j=1}^J \) may be either random or deterministic unknown.