Maximum *A-Posteriori* Probability Pitch Tracking in Noisy Environments Using Harmonic Model

Joseph Tabrikian, *Senior Member, IEEE*, Shlomo Dubnov, and Yulya Dickalov

***Abstract*—Modern speech processing applications require operation on signal of interest that is contaminated by high level of noise. This situation calls for a greater robustness in estimation of the speech parameters, a task which is hard to achieve using standard speech models. In this paper, we present an optimal estimation procedure for sound signals (such as speech) that are modeled by harmonic sources. The harmonic model achieves more robust and accurate estimation of voiced speech parameters. Using maximum *a posteriori* probability framework, successful tracking of pitch parameters is possible in ultra low signal to noise conditions (as low as -15 dB). The performance of the method is evaluated using the Keele pitch detection database with realistic background noise. The results show best performance in comparison to other state-of-the-art pitch detectors. Application of the proposed algorithm in a simple speaker identification system shows significant improvement in the performance.**

***Index Terms*—Cramer-Rao bound, harmonic model, MAP estimator, Markov model, maximum likelihood, noisy speech, PDA, pitch detection, pitch tracking, speech denoising.**

I. Introduction

MODERN speech processing applications require robustness to environmental noise and interference. There are many approaches toward increasing the robustness of speech processing systems. A common approach is to apply transformations, such as spectral subtraction [1] and spectral normalization, in order to achieve acoustic robustness under additive noise and linear distortion. These methods attempt to derive a reliable estimate of the noise spectrum usually using suboptimal criteria. Some methods assume an Autoregressive (AR) model for the speech and the noise. In such a case Bayesian estimates are used to optimally match the signal and environmental parameters [2]. The main drawback of these approaches is that small errors in the all-pole model are magnified at low signal-to-noise ratios (SNRs), which might cause artificial resonances in the restored speech. Another line of research uses sine-wave analysis-synthesis model of speech signals for noise suppression applications. In the sinusoidal model the speech signal is represented as the sum of a small number of sinusoids (harmonics, partials) with time-varying amplitudes and frequencies [4]. For clean speech, the signal parameters are estimated from peaks of the local periodogram, i.e., amplitudes of short-time Fourier transform. Considering that periodogram is a poor estimator of frequency in low SNRs, additional processing is required for estimation of noisy speech parameters [5], [6]. The optimal use of sinusoidal modeling for speech enhancement draws much research interest in the recent years and still remains an open question at large [7].

In this paper,[1] a statistical method for pitch tracking assuming a harmonic model is presented. The harmonic model could be regarded as a special case of a sinusoidal speech model where all sinusoidal components are assumed to be harmonically related, i.e., the frequencies of the sinusoids are at integer multiples of the fundamental frequency. This assumption reduces the number of parameters in the model and achieves much more accurate pitch estimates than the sinusoidal model. Assuming Markovian dynamics, Maximum *A-posteriori* Probability (MAP) tracking of time-varying harmonic signal is performed without prior knowledge of noise variance. The performance of the method was tested against a large pitch detection database, giving very good results under severe noise conditions: at SNRs ranging from -15 dB in the presence of white, stationary noise, and 0 dB in the presence of colored, nonstationary babble noise.

In the following we review some recent works on pitch detection methods that do not explicitly assume a harmonic model. It should be noted that pitch estimation is commonly used as a preprocessing step for sinusoidal analysis in order to "guide" the sinusoidal estimation procedure toward salient periodogram peaks. In our approach all model parameters (harmonic amplitudes, phases and fundamental frequency) are estimated in one common, statistically optimal framework.

The structure of the paper is as follows. In the next section, we review some recent state-of-the-art algorithms for pitch detection and their significance for harmonic models for speech. Section III presents the harmonic model and the details of the optimal parameter estimation procedure for MAP pitch tracking. The performances of the proposed method for pitch detection and tracking are demonstrated in Section IV. Finally, our conclusions are presented in Section V.

II. Background

A. Previous Work on Robust Pitch Detection

Pitch determination is an important part of speech coding, recognition and speech processing applications in general. Pitch detection algorithms (PDAs) [8] can be classified in two main categories: short-term analysis and time-domain based PDAs. Short-term analysis algorithms rely on transformations in the preprocessing step, which reveal approximate signal

Manuscript received July 10, 2002; revised August 19, 2003. This work was supported in part by the Israeli Science Foundation (ISF). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ramesh A. Gopinath.

The authors are with the Department of Electrical and Computer Engineering, Ben Gurion University of the Negev, Beer Sheva 84105, Israel (e-mail: tabrikian@ieee.org).

Digital Object Identifier 10.1109/TSA.2003.819950

[1] A preliminary version of this work was presented in [3].

1063-6676/04$20.00 © 2004 IEEE

similarity over pitch period shifts of the signal. The pitch period is then estimated using various correlation methods. Part of the short-term spectral methods exploit the harmonic structure of the signal, looking for spectral peaks at an equal spacing. This is equivalent to correlating the spectrum with an n-pulse template, where n is the number of included harmonics. Accordingly, such methods are sometimes called "pattern matching" methods. Time-domain pitch determination algorithms are in principle designed to track the signal period-by-period. They provide a sequence of period boundaries that mark local signal periods. The main difference between the two approaches is that in the first one, an average fundamental frequency over the analysis frame is estimated, while the second approach provides pitch *markers* in a "pitch synchronous" manner. The method presented in this paper belongs to the first broad class of PDAs.

Various pitch detection algorithms have been proposed, but they are sensitive to noise and interference and their performance significantly deteriorates at low SNRs. There are two main sources of errors in PDAs: false pitch estimates due to noise and signal distortion that occur in real environments and errors in voicing decision. This second type of errors becomes even more severe when operating in bad signal conditions. A common solution to noisy pitch estimates is by application of pitch tracking. Limiting the possible choice of pitch estimates among several noisy candidates according to earlier pitch estimates is done in [9]. Recently, more global approaches using dynamic programming (DP) have been suggested, offering an optimal decision over several frames. Wang and Seneff [10] developed a spectral domain score function (DLFT) using "template-frame" and "cross-frame" spectral correlation functions. The same function is applied both in the voiced and unvoiced regions of a speech signal, and the DP search is forced to find a pitch value for every frame, including unvoiced regions. Time-domain MAP pitch tracking was suggested by Droppo and Acero [11]. Their method requires knowledge of the noise variance, which is set as a free parameter in the algorithm. To handle the voicing problem, they train a separate two-state Hidden Markov Model (HMM), where each state is a mixture of Gaussians of a two-dimensional vector composed by frame energy and cross-correlation. The algorithms proposed in [10] and [11] are considered to be state of the art. A detailed comparison of performance between these two algorithms and the proposed method are presented in Section IV. Experimental results indicate that our algorithm performs as good as the above methods for clean speech and is significantly superior according to all analysis criteria in extreme noise situations.

B. Harmonic Speech Model

Speech analysis-synthesis methods that are based on a sinusoidal representation [4] assume that speech signal can be described as a sum of time-varying sinusoids. The amplitudes, frequencies and phases of the sinusoids are derived from short-time analysis of the salient spectral peaks of the speech signal. Several major contributions improved this initial model. In [12], an adaptive filtering algorithm for harmonic signal enhancement was presented. Adding a harmonic constraint for the frequencies in voiced speech was done in [9], [13]. This constraint assumes that the frequencies of the sinusoids are exact multiples of the fundamental frequency (the frequency of the lowest harmonic). The main motivation for using the harmonic assumption in these works is that it enables high-quality speech synthesis without coding the separate sine-wave frequencies. Recently, Stylianou [14] introduced a "Harmonic plus Noise" model (HNM) by imposing harmonicity in the analysis step of the sinusoidal part and also introduced the notion of Maximal Voicing Frequency, above which the signal is modeled as filtered noise. An HNM model is useful for speech synthesis manipulation and low bit-rate coding.

As will be discussed later in the paper, the harmonic assumption has an important effect on the precision of the frequency estimates and allows an optimal estimation of amplitude and phase parameters in the presence of noise. It should also be noted that the harmonic model falls in close relation to Harmonic regression. This is a linear regression model in which the predictor variables are trigonometric functions of a single variable, usually a time-related variable [15], [16]. Irizarry [17] used such models to determine the number of harmonic components in musical sounds.

III. Estimation of Harmonic Model Parameters

A. Problem Formulation and Modeling

Let $\{y(t_l)\}_{l=1}^{L}$ denote the samples of an audio signal, $y(t)$, measured by a single microphone at times $t_1, t_2, \ldots, t_L$. The model for the measurements of a given voiced frame can be written as

$$y(t_l) = b_{c0} + \sum_{m=1}^{M} \left(b_{cm}\cos(\omega_0 m t_l) + b_{sm}\sin(\omega_0 m t_l)\right) + \text{noise} \qquad l = 1, \ldots, L \quad (1)$$

where M denotes the number of harmonics and ω_0 stands for the fundamental angular frequency of the signal. The coefficients b_{cm} and b_{sm} carry the information on the intensity and phase of the mth harmonic of the signal.

Equation (1) in matrix notation can be written as

$$\mathbf{y} = \mathbf{A}(\omega_0)\mathbf{b} + \mathbf{n} \quad (2)$$

where $\mathbf{b} \triangleq [b_{c0}, \ldots, b_{cM}, b_{s1}, \ldots, b_{sM}]^T$, and the matrix $\mathbf{A}(\omega_0)$ can be partitioned as: $\mathbf{A}(\omega_0) = [\mathbf{A}_c(\omega_0) \quad \mathbf{A}_s(\omega_0)]$. The elements of $\mathbf{A}(\omega_0)$ are given by

$$\mathbf{A}_{clm}(\omega_0) = \cos(\omega_o m t_l), \quad m = 0, \ldots, M, \quad l = 1, \ldots, L$$
$$\mathbf{A}_{slm}(\omega_0) = \sin(\omega_o m t_l), \quad m = 1, \ldots, M, \quad l = 1, \ldots, L.$$

The noise vector, $\mathbf{n}$, is assumed to be zero-mean, Gaussian with known covariance matrix, $\mathbf{R_n}$. With no loss of generality, we can assume that the covariance matrix of the noise is diagonal, $\mathbf{R_n} = \sigma_n^2\mathbf{I}$, since the signal can be pre-whitened. The whitening operation is performed by left-multiplication of (2) by $\mathbf{R_n}^{-1/2}$

$$\underbrace{\tilde{\mathbf{y}}}_{\mathbf{R_n}^{-1/2}\mathbf{y}} = \underbrace{\tilde{\mathbf{A}}(\omega_0)}_{\mathbf{R_n}^{-1/2}\mathbf{A}(\omega_0)}\mathbf{b} + \underbrace{\tilde{\mathbf{n}}}_{\mathbf{R_n}^{-1/2}\mathbf{n}}. \quad (3)$$

The modified noise vector, $\tilde{\mathbf{n}}$ is also Gaussian distributed: $\tilde{\mathbf{n}} \sim N(\mathbf{0}, \mathbf{I})$. We will consider the noise intensity as an additional free parameter which could be either known or unknown.

For simplicity of notation the tilde sign, $\tilde{(\cdot)}$, will be omitted in the following. The pre-whitening procedure, described above, is efficient only if the noise covariance matrix is available and accurate. In practice, the noise covariance matrix $\mathbf{R_n}$ can be estimated from silence periods which are available and correctly identified. However, in the presence of rapidly time-varying background noise statistics, the noise covariance matrix cannot be estimated, and therefore, the whitening operation is not effective. In these cases, the data should be windowed in order to obtain a robust pitch tracking algorithm which is not sensitive to out of band noise energy. This can be obtained by left-multiplying the data and the harmonic matrix, $\mathbf{A}(\omega_0)$, by a diagonal matrix $\mathbf{W}$ whose diagonal is the desired window function.

The unknown parameters in the model described in (2), are: the fundamental frequency, ω_0, the vector of harmonics, $\mathbf{b}$, and the noise intensity, σ_n^2. The number of harmonics, M, is assumed to be known.[2] Our goal here is to estimate the fundamental frequency, ω_0, and the vector of harmonics, $\mathbf{b}$, depending on the application. For denoising and signal reconstruction purposes, for example, both parameters ω_0 and $\mathbf{b}$ are required. For speech recognition applications where spectral envelope estimation is required, only the amplitudes of the harmonics, $a_m = \sqrt{b_{cm}^2 + b_{sm}^2}$, are considered. The noise variance, σ_n^2, is a nuisance parameter which is also required to be estimated.

B. Single-Frame Maximum-Likelihood Estimator

Under the assumptions stated above, the conditional probability density function (pdf) of the measurement vector, $\mathbf{y}$, given the unknown parameters, is given by

$$f_{\mathbf{y}|\omega_0,\mathbf{b},\sigma_n^2}(\mathbf{y}|\omega_0,\mathbf{b},\sigma_n^2) = \frac{1}{(2\pi\sigma_n^2)^{1/2}} e^{-1/2\sigma_n^2 \|\mathbf{y}-\mathbf{A}(\omega_0)\mathbf{b}\|^2} . \tag{4}$$

The Maximum-Likelihood (ML) estimator is obtained by maximizing the likelihood function from (4) (or its logarithm) with respect to the unknown parameters

$$(\hat{\omega}_0, \hat{\mathbf{b}}, \hat{\sigma}_n^2)_{ML} = \arg \max_{\omega_0,\mathbf{b},\sigma_n^2} L_{\mathbf{y}}(\omega_0,\mathbf{b},\sigma_n^2) \tag{5}$$

where $L_{\mathbf{y}}(\cdot,\cdot,\cdot)$ is the log-likelihood function defined as

$$\begin{aligned} L_{\mathbf{y}}(\omega_0,\mathbf{b},\sigma_n^2) &\stackrel{\triangle}{=} \log f_{\mathbf{y}|\omega_0,\mathbf{b},\sigma_n^2}(\mathbf{y}|\omega_0,\mathbf{b},\sigma_n^2) \\ &= \left[-\frac{1}{2}\log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2}\|\mathbf{y}-\mathbf{A}(\omega_0)\mathbf{b}\|^2 \right] \end{aligned} \tag{6}$$

and $\log$ denotes the natural logarithm function. Note that if the assumptions on noise statistics are not satisfied, maximizing (6) is no longer the ML estimator of the unknown parameters, but it can be considered as the least-squares (LS) estimator. By maximizing the log-likelihood function in (6) with respect to $\mathbf{b}$, one obtains

$$\hat{\mathbf{b}} = (\mathbf{A}^T(\omega_0)\mathbf{A}(\omega_0))^{-1}\mathbf{A}^T(\omega_0)\mathbf{y} \tag{7}$$

[2]The number of harmonics can either be estimated using information theoretic criteria based techniques, such as MDL, [18], or AIC [19], or be set to the maximum according to the signal bandwidth, B: $M = \lfloor 2\pi B/\omega_{omin} \rfloor$.

where ^{T} denotes the matrix transpose operation. Accordingly, the log-likelihood function can be rewritten as

$$L_{\mathbf{y}}(\omega_0,\hat{\mathbf{b}},\sigma_n^2) = -\frac{1}{2}\log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2}\mathbf{y}^T\left[\mathbf{I}-\mathbf{P_A}(\omega_0)\right]\mathbf{y} \tag{8}$$

where $\mathbf{P_A}(\omega_0) \stackrel{\triangle}{=} \mathbf{A}(\omega_0)(\mathbf{A}^T(\omega_0)\mathbf{A}(\omega_0))^{-1}\mathbf{A}^T(\omega_0)$ is the projection matrix into the subspace spanned by the columns of $\mathbf{A}(\omega_0)$. Therefore, the ML estimator of the fundamental angular frequency for known noise variance, σ_n^2, is obtained by maximizing $L_{\mathbf{y}}(\omega_0,\hat{\mathbf{b}},\sigma_n^2)$, that is

$$\hat{\omega}_{0ML} = \arg\max_{\omega_0} \|\mathbf{P_A}(\omega_0)\mathbf{y}\|^2 . \tag{9}$$

For unknown noise variance case, the log-likelihood function at (8) needs to be maximized also with respect to σ_n^2. Maximization of (8), with respect to σ_n^2, using the projection matrix properties $\mathbf{P_A}\mathbf{P_A} = \mathbf{P_A}$ and $\mathbf{P_A}^T = \mathbf{P_A}$, results in

$$\hat{\sigma}_n^2 = \mathbf{y}^T\left(\mathbf{I}-\mathbf{P_A}(\omega_0)\right)\mathbf{y} = \|\left(\mathbf{I}-\mathbf{P_A}(\omega_0)\right)\mathbf{y}\|^2 . \tag{10}$$

By substitution of (10) into (8), one obtains

$$L_{\mathbf{y}}(\omega_0,\hat{\mathbf{b}},\hat{\sigma}_n^2) = -\frac{1}{2}\left[\log(2\pi\,\|(\mathbf{I}-\mathbf{P_A}(\omega_0))\,\mathbf{y}\|^2) + 1\right] \tag{11}$$

which implies that the ML estimator of ω_0 is identical to the known noise variance case, presented in (9).

As mentioned above, in order to obtain a pitch estimation method that is robust to mismatch in noise statistics, one requires to window the data and the harmonic matrix. In this case, the estimator is similar to the one described above

$$L_{\mathbf{y}}(\omega_0,\hat{\mathbf{b}},\hat{\sigma}_n^2) = -\frac{1}{2}\left[\log(2\pi\,\|(\mathbf{I}-\mathbf{P_{WA}}(\omega_0))\,\mathbf{W}\mathbf{y}\|^2) + 1\right] \tag{12}$$

where $\mathbf{P_{WA}}(\omega_0) \stackrel{\triangle}{=} \mathbf{WA}(\omega_0)(\mathbf{A}^T(\omega_0)\mathbf{W}^2\mathbf{A}(\omega_0))^{-1}\mathbf{A}^T(\omega_0)\mathbf{W}$ and $\mathbf{W}$ is a diagonal matrix whose diagonal elements represent the windowing function.

C. Multiple Frames Pitch Tracking

In the previous subsection, the ML estimator was developed for a single frame, providing an independent parameter estimation for every frame. In real speech, the harmonic parameters have a smooth behavior over time. Moreover, it is reasonable to assume that the noise statistics are piece-wise constant, i.e., unchanging for at least several speech frames. Incorporating the smoothness of speech parameters into the estimation procedure is presented in [20], in which an ad-hoc smoothness cost function is defined. In this subsection, the MAP estimator for the fundamental frequency, based on measurements collected over several frames, is developed. The resulting algorithm is implemented by a dynamic programming procedure, and it is an exact implementation of the MAP algorithm.

The parameters which are required to be estimated, are given by the vectors $\mathbf{\Omega} \stackrel{\triangle}{=} (\omega_1,\ldots,\omega_K)^T$ and $\mathbf{B} \stackrel{\triangle}{=} (\mathbf{b}_1^T,\ldots,\mathbf{b}_K^T)^T$ where K is the number of the given frames. Here we assume that $\{\omega_k\}_{k=1}^K$ is a Markov sequence, and $\{\mathbf{b}_k\}_{k=1}^K$ an unknown deterministic sequence of vectors. In fact, the amplitudes, $\{\mathbf{b}_k\}_{k=1}^K$ could also be considered as a Markov sequence, but for pitch estimation, this information is weak comparing to the information on the Markovian property of $\{\omega_k\}_{k=1}^K$. The

MAP estimator of these parameters is obtained by maximizing the joint pdf of the data vector, $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_K^T)^T$, and the unknown random vector parameter, $\mathbf{\Omega}$, conditioned on the unknown nonrandom parameters $\mathbf{B}$

$$\left(\hat{\mathbf{\Omega}}, \hat{\mathbf{B}}\right)_{MAP} = \arg\max_{\mathbf{\Omega},\mathbf{B}} f_{\mathbf{y},\mathbf{\Omega};\mathbf{B}}(\mathbf{y}, \mathbf{\Omega}; \mathbf{B}). \quad (13)$$

For simplicity of notation in the following derivations, we will omit the dependence of the pdfs on the random variables, i.e., use the notation f_x instead of $f_x(x)$. The random vector parameter $\mathbf{\Omega}$ is assumed to be independent of $\mathbf{B}$, and therefore,

$$f_{\mathbf{y},\mathbf{\Omega};\mathbf{B}} = f_{\mathbf{y}|\mathbf{\Omega};\mathbf{B}} \cdot f_{\mathbf{\Omega}} \ . \quad (14)$$

The vector of fundamental frequencies, $\mathbf{\Omega}$ is treated as a first order Markov chain and therefore

$$f_{\mathbf{\Omega}} = \prod_{k=1}^{K} f_{\omega_k|\omega_{k-1}} \quad (15)$$

where $f_{\omega_1|\omega_0} \triangleq f_{\omega_1}$ is the prior pdf of the fundamental frequency at the first frame. Since the noise is assumed to be independent between frames, then the data at different frames given the vector of fundamental frequencies, $\mathbf{\Omega}$, are statistically independent, and therefore

$$f_{\mathbf{y}|\Omega;\mathbf{B}} = \prod_{k=1}^{K} f_{\mathbf{y}_k|\omega_k;\mathbf{b}_k} \ . \quad (16)$$

Substituting (15) and (16) into (14), the pdf in (13) is given by

$$f_{\mathbf{y},\mathbf{\Omega};\mathbf{B}} = \prod_{k=1}^{K} \left[f_{\mathbf{y}_k|\omega_k;\mathbf{b}_k} f_{\omega_k|\omega_{k-1}} \right] \ . \quad (17)$$

Taking the logarithm of the pdf in (13), and substituting (17) into (13) gives

$$\left(\hat{\mathbf{\Omega}}, \hat{\mathbf{B}}, \hat{\sigma}_n^2\right) = \arg\max_{\mathbf{\Omega},\mathbf{B},\sigma_n^2} \sum_{k=1}^{K} \left[\log f_{\mathbf{y}_k|\omega_k;\mathbf{b}_k} + \log f_{\omega_k|\omega_{k-1}} \right]. \quad (18)$$

The transition probability density functions $\{f_{\omega_k|\omega_{k-1}}\}_{k=1}^{K}$, are assumed to be *a-priori* known. Maximization of (18) with respect to $\mathbf{B}$ is identical to the single frame case from (7)

$$\hat{\mathbf{b}}_k = (\mathbf{A}^T(\omega_k)\mathbf{A}(\omega_k))^{-1}\mathbf{A}^T(\omega_k)\mathbf{y}_k \ . \quad (19)$$

Accordingly, the fundamental frequency estimator in different frames becomes

$$(\hat{\mathbf{\Omega}}, \hat{\sigma}_n^2) = \arg\max_{\mathbf{\Omega},\sigma_n^2} \sum_{k=1}^{K} \left[L_{\mathbf{y}_k}(\omega_k, \hat{\mathbf{b}}_k, \sigma_n^2) + \log f_{\omega_k|\omega_{k-1}} \right] \quad (20)$$

where $L_{\mathbf{y}}(\cdot,\cdot,\cdot)$ is given in (8).

1) Known Noise Variance: The MAP estimate of $\mathbf{\Omega}$ for known noise variance can be obtained by (20) where no maximization with respect to σ_n^2 is required, that is

$$\hat{\mathbf{\Omega}}_{MAP} = \arg\max_{\mathbf{\Omega}} \sum_{k=1}^{K} \left[L_{\mathbf{y}_k}(\omega_k, \hat{\mathbf{b}}_k, \sigma_n^2) + \log f_{\omega_k|\omega_{k-1}} \right] \ . \quad (21)$$

Assuming that each one of the elements of the unknown random parameter vector, $\mathbf{\Omega}$, is discrete on a given pre-determined grid $(r_1, \ldots, r_J)$, the MAP estimator in (21) can be implemented using a fast recursive dynamic programming technique [21]. The choice of the actual grid for discretization of the frequency parameter depends on the required frequency resolution. Moreover, it is possible to work in an adaptive manner, using a coarse grid for initial frequency estimation and a refined grid for more precise pitch detection, performed over a smaller range of candidate pitches.

The required maximization with respect to $\mathbf{\Omega}$ can be performed using the algorithm described in the following. Let us define the following terms:

$$V_i(k) \triangleq L_{\mathbf{y}_k}(\omega_k = r_i, \hat{\mathbf{b}}_k, \sigma_n^2), \ k=1,\ldots,K, \ i=1,\ldots,J \quad (22)$$

$$B_{j,i}(k) \triangleq \log p(\omega_k = r_i | \omega_{k-1} = r_j), k=1,..,K, \ i,j=1,\ldots,J. \quad (23)$$

The transition probabilities $p(\omega_k = r_i|\omega_{k-1} = r_j)$ are used instead of the transition pdfs, because of using discrete values for the elements of $\mathbf{\Omega}$. Then the estimator of ω_k is given by

$$\hat{\omega}_{k_{MAP}} = r_{\hat{m}(k)} \quad (24)$$

$$\hat{m}(k) = \arg\max_{i=1,\ldots,J} W_i(k) \quad (25)$$

where $W_i(k)$ can be calculated by the recursive equation

$$W_i(k) = \max_{j=1,\ldots,J} [W_j(k-1) + B_{j,i}(k)] + V_i(k) \qquad k = 1,\ldots,K, \ \ i = 1,\ldots,J \quad (26)$$

with the initial condition: $W_i(0) = 0$ for $i = 1, \ldots, J$.

The above procedure provides the MAP estimate of the fundamental frequency ω_k, given the measurements $\mathbf{y}_1, \ldots, \mathbf{y}_k$. In order to obtain the MAP estimate of ω_k, based on all measurements, $\mathbf{y}_1, \ldots, \mathbf{y}_k, \ldots, \mathbf{y}_K$, a similar, backward procedure [21] can be implemented.

The estimator assumes knowledge of the transition probabilities $p(\omega_k = r_i|\omega_{k-1} = r_j)$, and the additive noise variance, σ_n^2.

2) Unknown Noise Variance: In the case where the noise variance is unknown, it needs to be estimated. By maximization of (20) with respect to σ_n^2 one obtains

$$\hat{\sigma}_n^2 = \frac{1}{K} \sum_{k=1}^{K} \left\| (\mathbf{I} - \mathbf{P}_{\mathbf{A}}(\omega_k)) \mathbf{y}_k \right\|^2 \ . \quad (27)$$

It can be seen that substitution of this estimate into (20) results in a complex expression which cannot be maximized using a fast dynamic programming search method as described above. Another solution for this problem is a joint search over $\mathbf{\Omega}$ and σ_n^2, where the likelihood function is calculated over a predetermined grid on σ_n^2. For each value of σ_n^2, the fast recursive method, described above for calculation of the likelihood function, can be implemented. This solution involves a large amount of computations in comparison to the case when the noise variance is known. A suboptimal solution can be obtained by a recursive procedure as follows.

• *Initialization*: Find the ML estimates of $\{\omega_k\}_{k=1}^{K}$ for each frame, ignoring the prior statistical information expressed by the transition pdfs, $f_{\omega_k|\omega_{k-1}}$

$$\hat{\omega}_k = \arg\max_{\omega_k} \|\mathbf{P}_{\mathbf{A}}(\omega_k)\mathbf{y}_k\|^2 \ .$$

• *Iteration:*
1. Estimate the noise variance, σ_n^2, according to **(27)** using the ML estimates of $\{\omega_k\}_{k=1}^K$. This is an approximation to the MAP estimate of the noise variance.
2. Calculate the MAP estimate of the fundamental frequencies using **(21)**, and the newly estimated noise variance **(27)**.
3. If the increase in the likelihood is greater than a given threshold, return to step 1. Otherwise, stop.

Note that at each step the likelihood function is increased, and therefore the convergence of this procedure at a local maximum is guaranteed.

D. Computational Aspects

Implementation of the pitch tracking algorithm stated above, involves calculation of the log-likelihood function $L_{\mathbf{y}_k}(\omega_k, \hat{\mathbf{b}}_k, \sigma_n^2)$, on a pre-determined grid on the fundamental frequencies axis, at each frame, and the MAP tracking algorithm. The log-likelihood function $L_{\mathbf{y}_k}(\omega, \hat{\mathbf{b}}_k, \sigma_n^2)$ from (8) can be written as follows:

$$L_{\mathbf{y}_k}(\omega_k, \hat{\mathbf{b}}_k, \sigma_n^2) = \text{const} - \frac{1}{2\sigma_n^2}\left[\|\mathbf{y}_k\|^2 - \|\mathbf{T}(\omega_k)\mathbf{y}_k\|^2\right] \tag{28}$$

where the matrix $\mathbf{T}(\omega) \triangleq \left(\mathbf{A}^T(\omega)\mathbf{A}(\omega)\right)^{-1/2}\mathbf{A}^T(\omega)$ is a transformation matrix of size $(2M+1)\times L$, which can be calculated off-line. From (28) we can see that calculation of the log-likelihood function involves this transformation and the norm operation. This method of computation is efficient since typically $M \ll L$. The tracking algorithm involves $(J+1)J$ summations and a maximization over a vector of size J at each frame, where J is the grid size over the fundamental frequency axis ω.

The computation time of the proposed algorithm enables its implementation in real-time processing systems. Comparing the computational complexity of the proposed method to other existing state-of-the-art pitch tracking algorithms, such as Droppo-Acero [11] and DLFT [10], [22], shows that all the three algorithms involve the same order of computations. The tracking stage in the three methods require similar amount of computations. While in the proposed method, the log-likelihood function involves a linear transformation from a space of size L to a space of size $2M+1$, the other two methods require auto-correlation/cross-correlation calculation at different lags. The number of lags in both cases is of the same size as $2M+1$.

E. Advantages of the Harmonic Model

The Cramer–Rao bound (CRB) for frequency estimation, based on a general stochastic sinusoidal model is well known [23]. In the harmonic case, the number of parameters is reduced, resulting in a single frequency parameter ω_0 rather than M separate frequencies. By using the results of the CRB for multiple sinusoids, one is able to express the respective part of the harmonic Fisher Information Matrix (FIM), $J^{(H)}_{\omega_0,\omega_0}$, using the FIM for multiple individual sinusoids as J_{ω_m,ω_m}

$$J^{(H)}_{\omega_0,\omega_0} = E\left(\frac{\partial \log f^{(H)}_{\mathbf{y}|\omega_0}}{\partial \omega_0}\right)^2 = E\left(\sum_{m=1}^{M}\frac{\partial \log f^{(H)}_{\mathbf{y}|\omega_0}}{\partial \omega_m}\frac{\partial \omega_m}{\partial \omega_0}\right)^2 = \sum_{m=1}^{M} m^2 J_{\omega_m,\omega_m} \tag{29}$$

where we used the fact that $\omega_m = m\omega_0$. Considering the asymptotic expression of CRB for multiple sinusoids, we obtain

$$\text{CRB}^{(H)}(\omega_0) \approx \frac{24\sigma_n^2}{L^3}\left\{\sum_{m=1}^{M} m^2|a_m|^2\right\}^{-1} \tag{30}$$

with σ_n^2 being the noise variance, and a_m denotes the mth harmonic amplitude. In particular, (30) implies that for a true harmonic signal, the variance of errors of the proposed method is expected to be significantly lower comparing to other procedures which first estimate individual harmonics and then use these separate estimates to estimate the pitch. The relation between the variances of individual frequencies estimates of both approaches can be approximated as

$$\frac{\text{CRB}(\hat{\omega}_m)}{\text{CRB}(m\cdot\hat{\omega}_0)} \approx 1 + \frac{1}{|a_m|^2}\Sigma_{l=1 l\neq m}^{M} l^2|a_l|^2\,. \tag{31}$$

This indicates that a purely harmonic model is much more robust under the assumption of harmonic sources. Note that using a harmonic model, the mth partial reduces the variance of the estimate error by a factor of m^{-2}. Thus, for instance, the pitch estimate for a signal that has only two partials—a fundamental and one high partial—has significantly lower variance than considering the fundamental partial alone.

Another major concern in our estimator is the question of windowing. There are two main reasons to apply windowing:

1) avoid leakage of energy between adjacent sinusoids;
2) give a lesser weighting to future and past samples because the signal is time-varying.

When L is large enough, but the pitch can still be considered as stationary, the sinusoidal and cosinusoidal matrices become practically orthogonal

$$\mathbf{A}^T\mathbf{A} \approx \frac{L}{2}\mathbf{I}_M \tag{32}$$

where $\mathbf{I}_M$ is an identity matrix of size M. Then we can write a simplified version of our nonlinear estimator from (8) as

$$\hat{\omega} = \arg\max_{\omega} L_{\mathbf{y}}(\omega, \hat{\mathbf{b}}, \sigma_n^2) \tag{33}$$

and it can be approximated by

$$\begin{aligned} L_{\mathbf{y}}(\omega, \hat{\mathbf{b}}, \sigma_n^2) &\approx \text{const} + \frac{1}{L}\mathbf{y}^T\mathbf{A}\mathbf{A}^T\mathbf{y} \\ &= \text{const} + \frac{1}{L}\sum_{m=1}^{M}\left|\sum_{l=1}^{L} y(t_l)e^{-jm\omega t_l}\right|^2 \\ &= \text{const} + \sum_{m=1}^{M} P(m\omega) \end{aligned} \tag{34}$$

where $P(m\omega)$ is the periodogram spectral estimator evaluated at harmonically related frequencies. We can see that "standard" sinusoidal peak picking methods are approximately optimal for large frame length, L, and in case that the signal amplitude is large relative to noise and there is no interference between the harmonic components.

In the proposed method, we solve directly the nonlinear maximum likelihood estimation equation, and thus there are no difficulties that arise in the analysis stage due to spectral leakage of the signal harmonics (we are not trying to pick them separately). Moreover, in estimation of the amplitude and phase parameters, the orthogonal projection matrix causes cancelation of the side-lobe disturbances.

In the case of unknown noise statistics or nonstationary noise, one may consider windowing of the data and the harmonic matrix $\mathbf{A}(\omega_0)$. As mentioned above, in the presence of nonstationary noise, the whitening procedure is not efficient to suppress the effect of noise, because the assumed noise covariance matrix is not accurate. Therefore, the algorithm needs to cope with noise whose statistics are not available. There are two types of pitch estimation errors: local error and nonlocal error. The nonlocal errors can be caused by noise from side-lobe frequencies. Local errors are caused by noise from the main-lobe frequency. The side-lobes can be attenuated if one windows the data. However, the windowing procedure widens the main-lobe, which increases the pitch estimation error variance. In order to consider the question of nonstationarity, we need to evaluate the effect of windowing on our estimator. It is known from literature [24], [25], [17] that the effect of windowing can be taken into account as

$$Var_w(\omega) \approx \frac{W_0^2 U_2 - 2W_0 W_1 U_1 + W_1^2 U_0}{(W_0 W_2 - W_1^2)^2} Var_R(\omega) \quad (35)$$

where Var_w and Var_R are variances with window $w(t)$ and rectangular window, respectively, where W_n and U_n are defined as

$$W_n = \int_0^1 t^n w(t) dt$$

and

$$U_n = \int_0^1 t^n w(t)^2 dt \,.$$

Widening of the main-lobe causes more noise energy to be collected from the main lobe and thus increases the variance of the estimator. The increase in variance can now be readily calculated from (35). Evaluating for different window types, we find an increase in variance by scale factor of 1.64 for Hamming window and 2.33 for Hanning window compared to rectangular window. This increase in variance is significant for low SNRs, and application of such a window requires doubling the frame length to avoid overlap between spectral peaks. Therefore, simple rectangular windowing is recommended in cases where the background noise statistics can be precisely obtained.

F. Signal Reconstruction

In voiced frames, the MAP estimate of the fundamental frequency at k-th frame, ω_k, and the corresponding vector of harmonics, $\mathbf{b}_k$, can be used in order to reconstruct the signal by

$$\hat{\mathbf{s}}_k = \mathbf{A}(\hat{\omega}_k)\hat{\mathbf{b}}_k \,, \quad k = 1, \ldots, K \quad (36)$$

where the vector $\hat{\mathbf{s}}_k$ denotes the reconstructed signal at kth frame. By substitution of $\hat{\mathbf{b}}_k$ from (7) into (36), one obtains

$$\hat{\mathbf{s}}_k = \mathbf{A}(\hat{\omega}_k)[\mathbf{A}^T(\hat{\omega}_k)\mathbf{A}(\hat{\omega}_k)]^{-1}\mathbf{A}^T(\hat{\omega}_k)\mathbf{y}_k = \mathbf{P}_{\mathbf{A}}(\hat{\omega}_k)\mathbf{y}_k \,. \quad (37)$$

In other words, the optimal signal enhancement procedure in voiced frame data is obtained by projection of the data into the subspace spanned by the columns of the matrix of harmonics with the estimated fundamental frequency $\mathbf{A}(\hat{\omega}_k)$.

IV. Experimental Results

In this section, performance of the proposed algorithm is evaluated and compared to other existing methods. The Keele pitch extraction reference database [26] was chosen for this test. This database provides a reference pitch, which is obtained from a simultaneously recorded laryngograph trace, and is referred to as the "ground truth." Pitch values are provided at a 100 Hz frame rate using 25.6 ms window. Unvoiced frames are signed with zero pitch value and negative values are used for uncertain frames. The database includes five male and five female speakers, each speaking a short story of about 35 s length, sampled at a rate of 20 kHz. In order to allow performance evaluation using this database, the same analysis parameters (frame rate and window size) were chosen by the proposed method. The tests have been carried out on the data with different noise levels added to the signal. The signal-to-noise ratio (SNR) is calculated by: $SNR = \sum_l |s(t_l)|^2 / \sum_l |n(t_l)|^2$, where $s(t_l)$ and $n(t_l)$ are the signal and noise samples, respectively. The of speech to silence ratio in the examined database was about 58% in average for female speakers and 61% for male speakers.

The images in Figs. 1 and 2 refer to recorded data of a female and male speakers, respectively. The data in these figures correspond to an excerpt of 3 s from the reference database, which was contaminated by an additive white Gaussian noise to provide an SNR of 0 dB. Upper images of these figures present the spectrograms of the noisy data. The normalized log-likelihood functions for both cases were calculated from (8) with $\sigma_n^2 = 1$ and appear in middle of Figs. 1 and 2. Bottom images of these figures present the normalized log-likelihood functions with forward and backward tracking, i.e., after taking into account the prior pdfs of the fundamental frequencies of the recorded interval. Forward tracking results are important when one requires an immediate estimate of the fundamental frequency based on the measurements in the previous and current frames. Backward tracking improves the accuracy of forward tracking results by taking into account also measurements from future frames. Backward tracking mode can be used only when one can afford delays, or when working off-line with batch data so that an immediate estimate is not required. The transition pdf of the fundamental frequencies vector is assumed to be Gaussian: $(\omega_k|\omega_{k-1}) \sim N(\omega_{k-1}, \sigma_\omega^2)$ where $\sigma_f = \sigma_\omega / 2\pi$

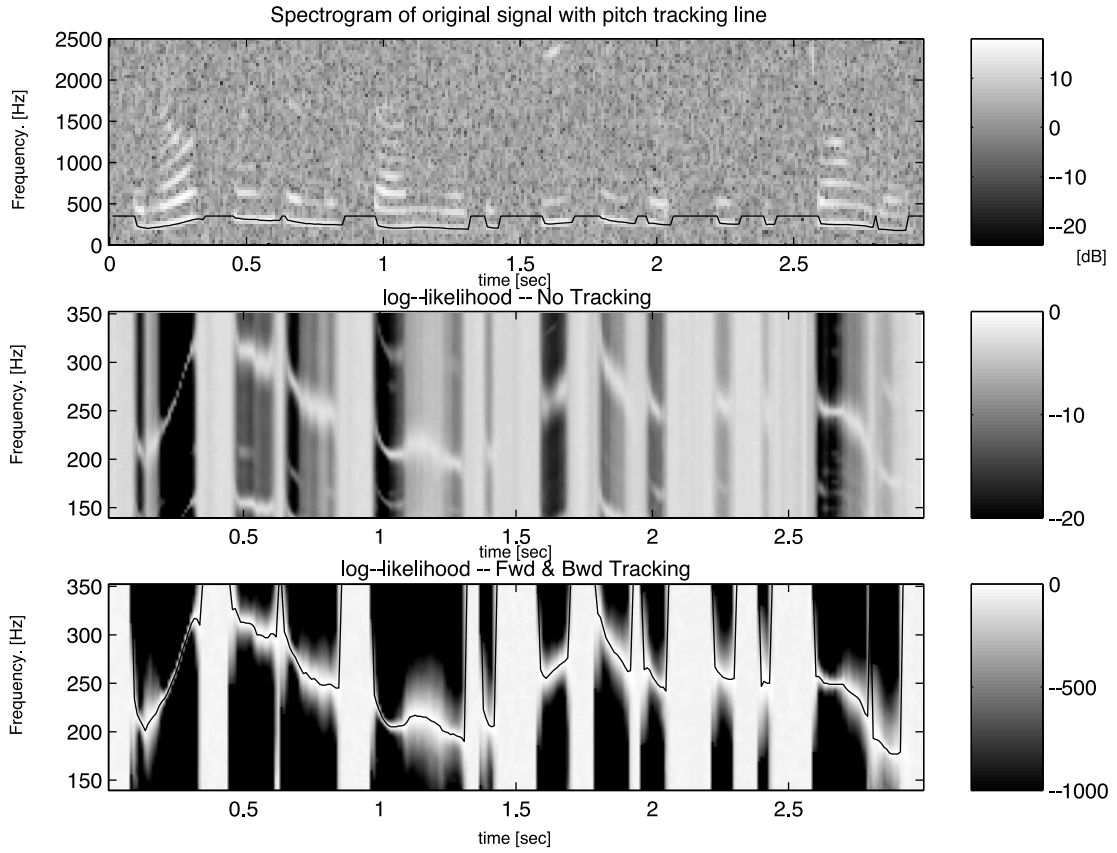


Fig. 1. Spectrogram, log-likelihood function, and log-likelihood with backward tracking—female speaker.

was set to 5 Hz. One can observe that these figures demonstrate the different advantages of pitch tracking. In order to handle frames which are not voiced (unvoiced and silence), we used the voiced/unvoiced decision algorithm described in [27]. In order to measure the sensitivity of the proposed tracking method to the pitch variation prior, the performance of the algorithm was evaluated for different values of σ_f. The results, shown in Fig. 3, indicate that the proposed method is relatively robust to variations in this parameter for different SNRs. Accordingly, it is not required to determine a precise value of this parameter. Note that for Gaussian priors, the likelihood function in (21) is insensitive to the ratio between σ_n^2 and σ_ω^2. Therefore, if σ_ω^2 is set to an assumed value σ_0^2, then the algorithm estimates a modified noise variance which is equal to $\sigma_n^2\sigma_0^2/\sigma_\omega^2$. Moreover, the pitch estimate is not expected to be affected if there is an error in the assumed value of the pitch variation prior, σ_ω^2. However, in practice, the algorithm may converge to a local maximum, if the initial value for the modified noise level is far from its true value. This is the reason that the algorithm is robust to the choose of the pitch variation prior within a relatively wide range.

We have compared the proposed pitch tracking method to a MAP-based (Droppo-Acero) [11] and the DLFT [10], [22] pitch tracking algorithms, which were implemented and tested using Keele database for different SNRs. Two computable measures according to [28] are used for performance evaluation: Gross Error Rate (GER), which is a measure for the amount of "gross pitch period errors," and the RMSE of "fine pitch errors." The GER is measured as the percentage of the pitch period estimation errors which are greater than 1 ms in their absolute values. The Droppo-Acero pitch tracking algorithm was implemented with parameter $\gamma = 10^{-8}$ which results in optimal GER performance at high SNRs. The GER and RMSE of the three methods for both female and male speakers are presented in Fig. 4. This figure shows that at low SNRs the other methods fail while the proposed method performs well both in GER and RMSE. In order to demonstrate the contribution of the prior statistical information used in MAP pitch tracking, the performance of the ML pitch estimator was also evaluated and its performance is presented in Fig. 4. It can be seen that the MAP tracking improves the results at low SNRs while at high SNRs the tracking algorithm does not contribute to the performance. Note that the although ML estimator for the pitch of the female speaker at high SNRs has lower RMSE than the MAP estimator, its GER is higher. That is, the tracking improves the GER, but it may reduce local error estimation accuracy.

Robustness of the proposed method to background noise mismatch was tested using babble noise, which represents nonstationary, non-Gaussian and colored noise.[3] In order to evaluate the nonstationarity of the babble noise, the short-term noise power over segments of 0.01 s is calculated over a period of 5 s and compared to white, stationary, Gaussian noise in Fig. 5. In Fig. 6, the performance of the proposed algorithm with Droppo-Acero and DLFT pitch tracking methods for both male and fe-

[3]The babble noise file was downloaded from http://spib.rice.edu/spib/data/signals/noise/babble.html. More information about the babble noise can be found at this site.

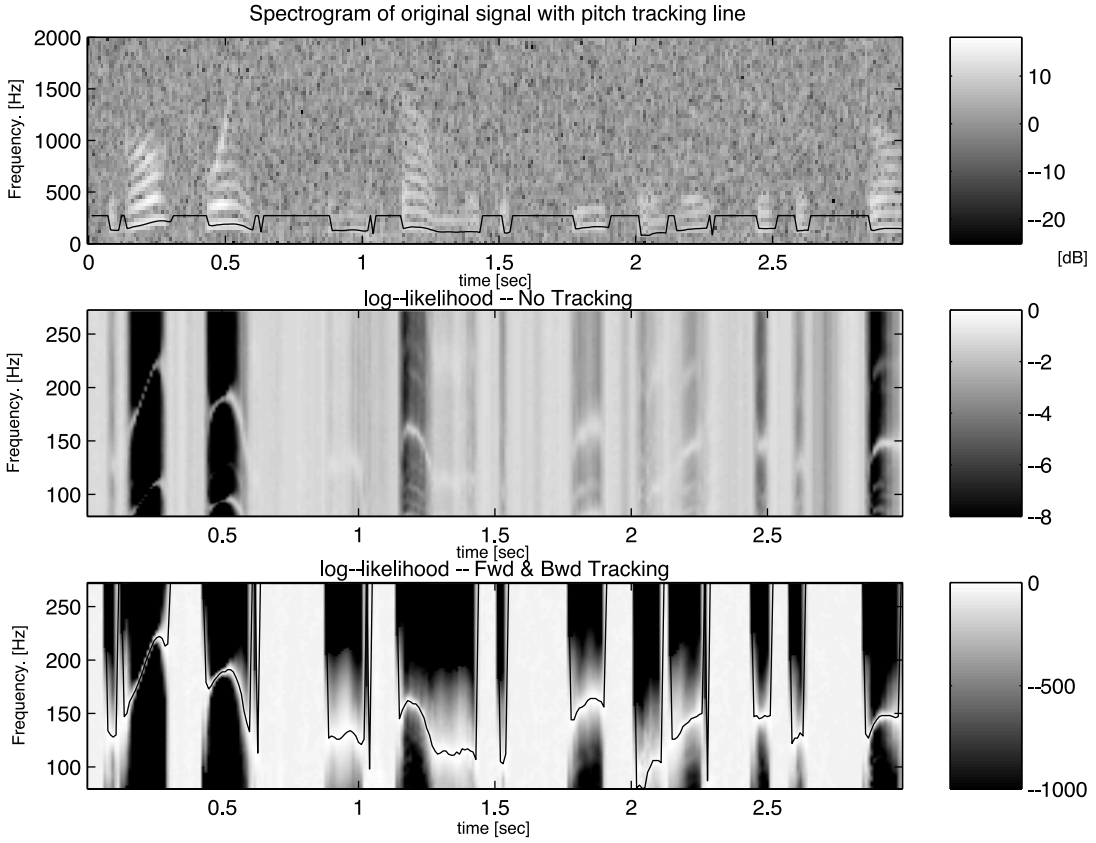


Fig. 2. Spectrogram, log-likelihood function and log-likelihood with backward tracking—male speaker.

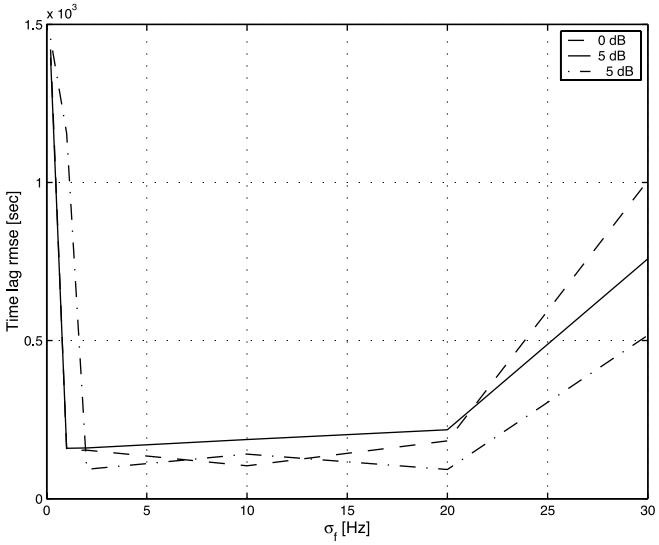


Fig. 3. Sensitivity of the time lag estimation RMSE to the pitch variation prior (transition probability between adjacent frames).

male speakers are compared. It can be seen that in the presence of babble noise the performance degrades, presumably because babble noise contains harmonic components. When these components are not weak comparing to the signal of interest, then the algorithm may track them in silence or unvoiced segments of the signal of interest. The proposed algorithm performs better than the other methods also with this type of noise.

Finally, we have evaluated the advantage of the proposed pitch tracking method, when used in a two-speaker identification algorithm. The implemented algorithm is based on Gaussian Mixture Model (GMM) for distribution of tenth-order cepstrum coefficients [29] and 16 Gaussians. The performance of the algorithm with data contaminated with white, Gaussian noise, was evaluated in terms of probability of false identification, which is the average between the probability of false identification under the hypotheses of the two speakers. This algorithm was trained with our pitch tracking and reconstruction method as presented in Section III-F, which was implemented as a pre-processing stage both in train and test modes. The performance of this speaker identification system was compared to the performance of another system which does not apply our pitch tracking algorithm in train and test modes. Both systems were trained with speech signals of two male speakers. The training length of each speaker was 10 s. Silence detection was performed using energy criterion, that is, frames with energy less than 10% were not considered for training. Speaker identification is made based on speech segments of 0.5 s. The probability of false identification, P_e with and without harmonic MAP pitch tracking is shown in Fig. 7, and the results show that by using the proposed method for pitch tracking combined with signal reconstruction, the speaker identification performance dramatically improves. The performance of the system without pitch tracking algorithm can be improved at low SNRs if one uses noisy data in the training mode. The performance of this system with 10 dB SNR in the

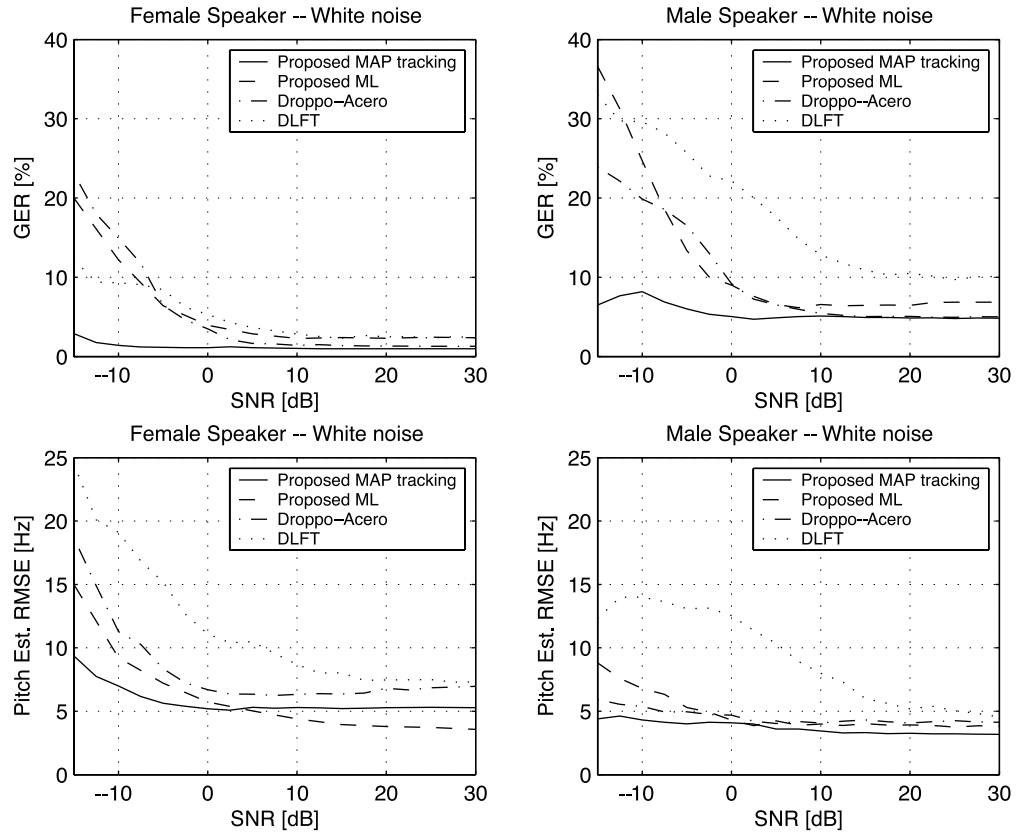


Fig. 4. Gross error rate and RMSE as a function of SNR for female/male signals white background noise.

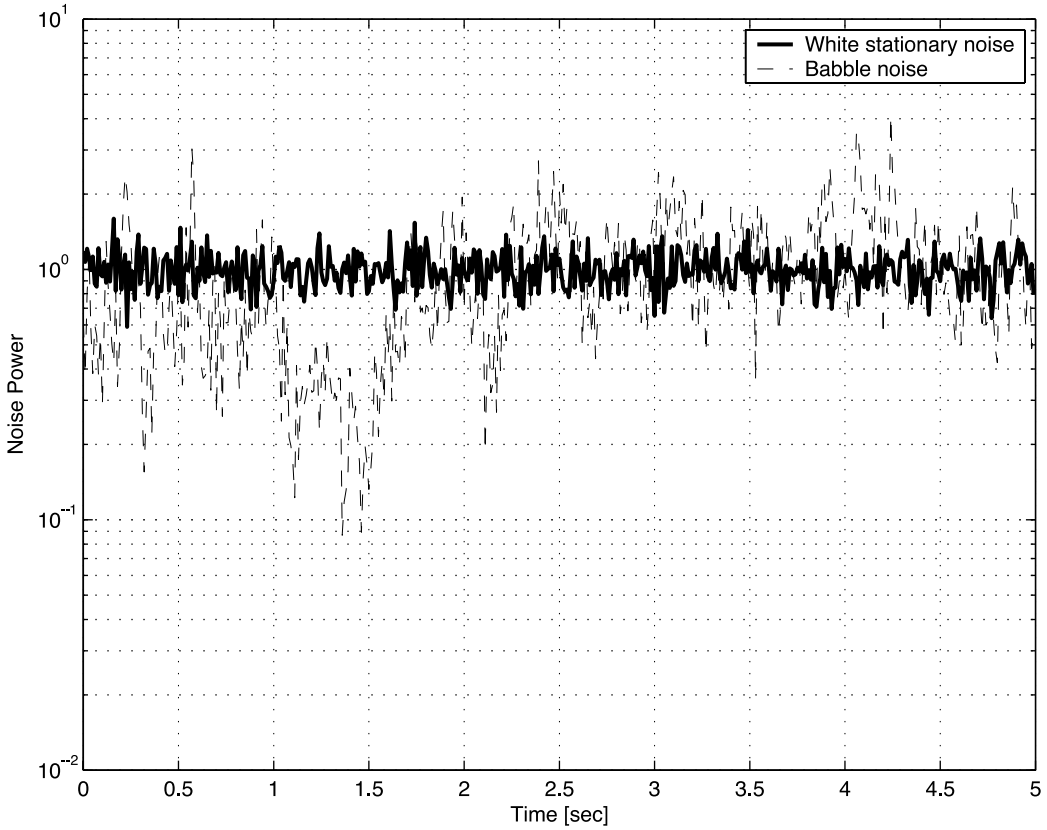


Fig. 5. Short-term (0.01 s) babble noise power versus time.

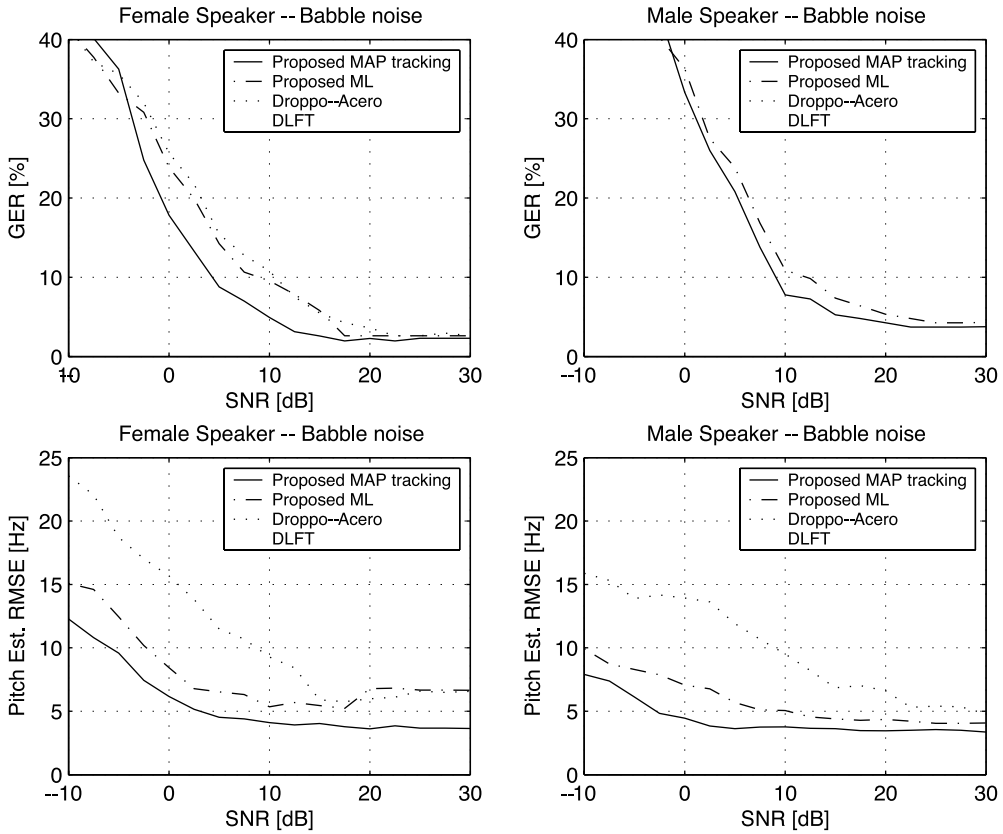


Fig. 6. Gross error rate and RMSE as a function of SNR for female/male signals with babble background noise.

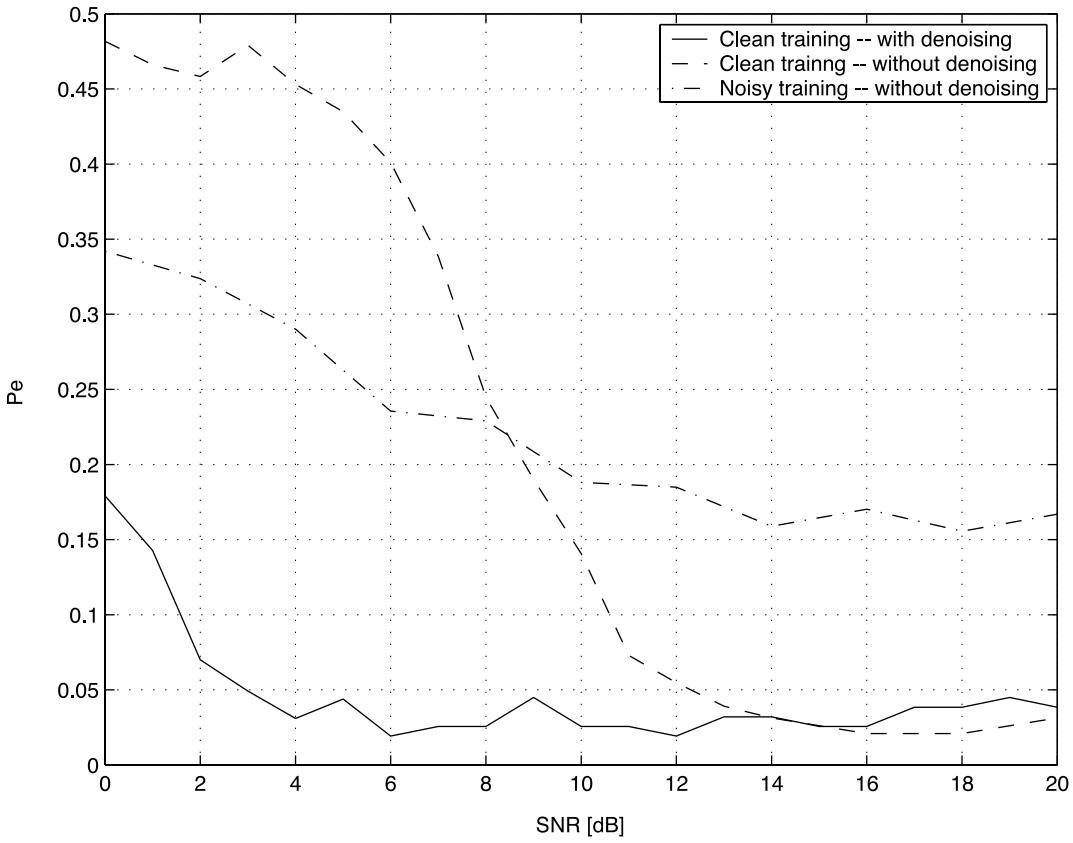


Fig. 7. Performance of the GMM-based speaker identification algorithm algorithm under white, stationary noise conditions.

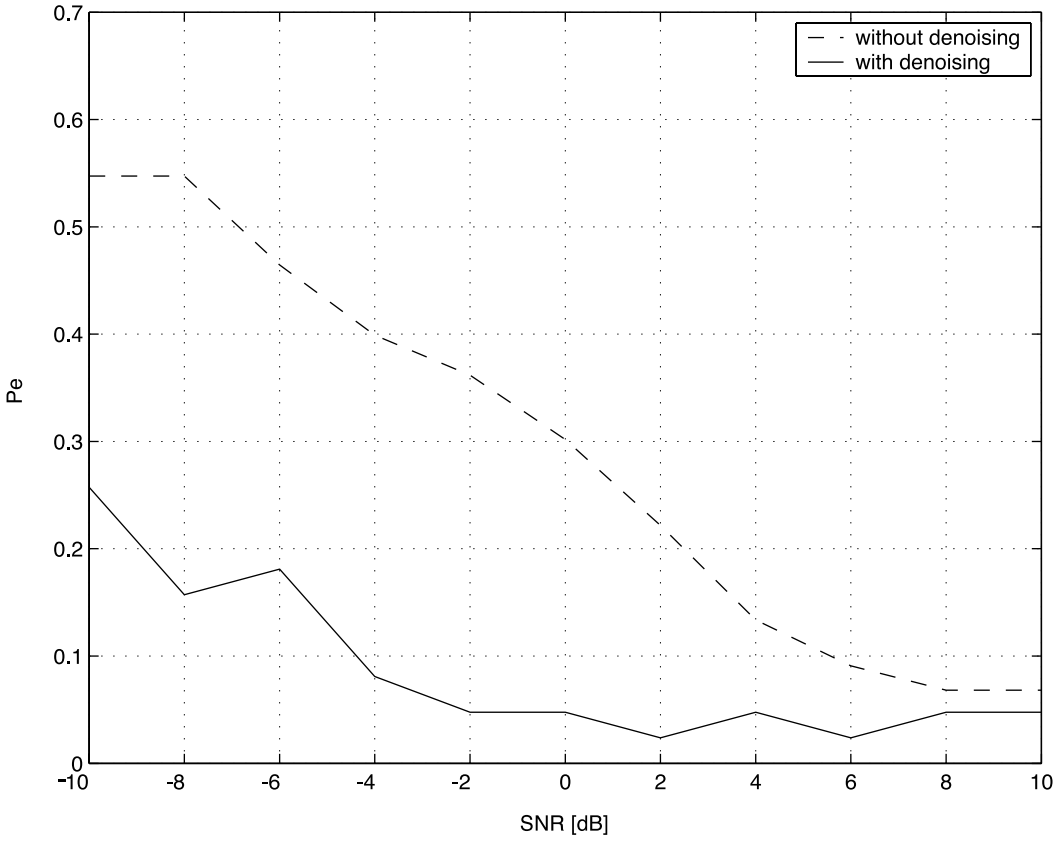


Fig. 8. Performance of the GMM-based speaker identification algorithm under babble noise conditions.

training data is also presented in the Fig. 7. It can be seen that the pitch-tracking-based system performance is better also than this system.

The performances of the two speaker identification systems were compared also under babble noise conditions. The probability of false identification, P_e with and without harmonic MAP pitch tracking is shown in Fig. 8, and again the results show that the proposed method for pitch tracking combined with signal reconstruction improves the speaker identification performance.

V. Conclusions

In this paper, we presented an optimal statistical procedure for estimation of speech parameters assuming a harmonic model. The proposed model and its parameter estimation method exhibit high robustness for ultra low signal to noise conditions, thus making it suitable for operation in difficult environments, such as mobile or wireless applications. Maximum likelihood estimation of pitch for a single frame is presented in the paper. The MAP estimation procedure was developed for tracking the locally harmonic signal in time. The MAP procedure achieves global optimal solution over larger signal segments. The performance of the method was evaluated in the paper using white Gaussian, and babble background environments in terms of pitch detection performance and estimation of model parameters. It was shown that our method outperforms other state-of-the-art pitch detection algorithms for severe noise conditions. Simulation results for pitch detection and tracking are reported and demonstrated in the paper. The proposed method does not require prior knowledge of the noise variance, and it is not sensitive to errors in the pitch variation priors.

Acknowledgment

The authors thank the reviewers for their constructive comments, which helped to improve the quality of the paper.

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 2, 1979.

[2] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*. Norwell, MA: Kluwer, 1992.

[3] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Speech enhancement by harmonic modeling via MAP pitch tracking," in *Proc. ICASSP*, 2002.

[4] R. J. McAulay and T. F. Quatieri, "Speech analysis-synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 744–754, 1986.

[5] T. F. Quatieri and R. J. McAulay, "Noise reduction using soft-decision sine-wave vector quantizer," in *Proc. ICASSP*, vol. 2, 1990.

[6] D. V. Anderson and M. A. Clements, "Audio signal noise reduction using multi-resolution sinusoidal modeling," in *Proc. ICASSP*, 1999.

[7] S. Dubost and O. Capp, "Analysis and enhancement of locally harmonic signals using adaptive multi-kernel methods," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1999.

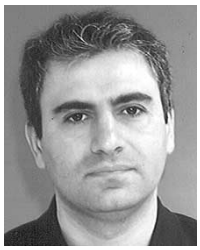
[8] W. J. Hess, *Pitch Determination of Speech Signals—Algorithms and Devices*. Berlin, Germany: Springer-Verlag, 1983.

[9] D. W. Griffin and J. S. Lim, "Multi-band excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 8, pp. 664–678, 1988.

[10] C. Wang and S. Seneff, "Robust pitch tracking for prosodic modeling in telephone speech," in *Proc. ICASSP*, 2000.

[11] J. Droppo and A. Acero, "Maximum a posteriori pitch tracking," in *Proc. ICSLP'98*, 1998, pp. 943–946.

[12] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1124–1138, 1986.
[13] R. J. McAulay and T. F. Quatieri, "Low-rate speech coding based on the sinusoidal model," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1991, pp. 165–208.
[14] Y. Stylianou, "Harmonic Plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification," Ph.D. dissertation, Ecole Nationale des Télécommunications, Paris, France, 1996.
[15] A. M. Walker, "On the estimation of a harmonic component in a time series with stationary independent residuals," *Biometrika*, vol. 58, pp. 21–36, 1971.
[16] E. J. Hannan, "The estimation of frequency," *J. Appl. Prob.*, vol. 10, pp. 510–519, 1973.
[17] R. A. Irizarry, "Statistics and Music: Fitting a Local Harmonic Model to Musical Sound Signals," Ph.D. dissertation, Univ. California, Berkeley, 1998.
[18] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.
[19] H. Akaike, "A new look at the statistical identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.
[20] H. Ney, "Dynamic programming algorithm for optimal estimation of speech parameter contours," *IEEE Trans. Syst., Man, Cybern,*, vol. SMC-13, no. 3, pp. 208–214, 1983.
[21] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1992.
[22] C. Wang and S. Seneff, "A study of tones and tempo in continuous Mandarin digit strings and their application in telephone quality speech recognition," in *Proc. 5th Int. Conf. Spoken Language Processing*, 1998.
[23] B. Porat, *Digital Signal Processing of Random Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[24] E. N. Brown, "A note on the asymptotic distribution of the parameter estimates for the harmonic model," *Biometrika*, vol. 77, pp. 653–656, 1990.
[25] B. G. Quinn, "Estimating the number of terms in a sinusoidal regression," *J. Time Series Anal.*, vol. 10, pp. 12–31, 1989.
[26] F. Plante, G. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *EUROSPEECH'95*, Madrid, Spain, 1995, pp. 827–840.
[27] E. Fisher, J. Tabrikian, and S. Dubnov, "Generalized likelihood ratio test for voiced/unvoiced decision using the harmonic plus noise model," in *Proc. ICASSP*, 2003.
[28] M. J. Cheng, L. R. Rabiner, and A. E. Rosenberg, "A comparative study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, pp. 399–417, 1976.
[29] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture models," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 72–83, 1995.
[30] Y. Ephraim, "Statistical model based speech enhancement systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 80, pp. 1526–1555, Oct. 1992.
[31] J. S. Lim, *Speech Enhancement*. Englewood Cliffs, NJ: Prentice-Hall, 1983.
[32] N. M. Laird, A. P. Dempster, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Ann. Roy. Statist. Soc.*, pp. 1–38, 1987.
[33] M. Feder, A. V. Oppenheim, and E. Weinstein, "Maximum likelihood noise cancellation using the EM algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 2, pp. 204–216, 1989.



Joseph Tabrikian (S'89–M'97–SM'98) received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the Tel-Aviv University, Tel-Aviv, Israel, in 1986, 1992, and 1997, respectively.

From 1996 to 1998, he was with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, as an Assistant Research Professor. He is now a Faculty Member with the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel. His research interests include statistical signal processing, source detection and localization, and speech and audio processing.

Dr. Tabrikian has served as an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING since 2001.

Shlomo Dubnov received the Ph.D. degree in computer science from Hebrew University of Jerusalem, Jerusalem, Israel.

He is a Lecturer in the Department of Communication Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel. His research interests include multimedia signal processing, audio engineering, and computer music.

Yulya Dickalov received the B.Sc. degree from the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel, in 2001.

She currently serves as an electrical engineer in the Israeli Army. Her research interest lies in the field of speech and signal processing.