# Generalized Likelihood Ratio Test for Voiced-Unvoiced Decision in Noisy Speech Using the Harmonic Model

Etan Fisher, Joseph Tabrikian, *Senior Member, IEEE*, and Shlomo Dubnov

***Abstract*—In this paper, a novel method for voiced-unvoiced decision within a pitch tracking algorithm is presented. Voiced-unvoiced decision is required for many applications, including modeling for analysis/synthesis, detection of model changes for segmentation purposes and signal characterization for indexing and recognition applications. The proposed method is based on the generalized likelihood ratio test (GLRT) and assumes colored Gaussian noise with unknown covariance. Under voiced hypothesis, a harmonic plus noise model is assumed. The derived method is combined with a maximum *a-posteriori* probability (MAP) scheme to obtain a pitch and voicing tracking algorithm. The performance of the proposed method is tested using several speech databases for different levels of additive noise and phone speech conditions. Results show that the GLRT is robust to speaker and environmental conditions and performs better than existing algorithms.**

***Index Terms*—Generalized likelihood ratio test (GLRT), harmonic model, likelihood ratio test (LRT), maximum a-posteriori probability, noisy speech, pitch tracking, voice activity detection (VAD), voiced-unvoiced decision.**

## I. Introduction

GROWING demand for advanced speech and audio applications requires new processing methods that are both flexible and robust to acoustical, environmental and system errors. As the demand for variable-rate speech coding applications increases, the role of voicing detection/decision is crucial for efficient bandwidth reduction. In speech, a decision is made between voiced and unvoiced speech phonemes. Correct voicing detection also allows for signal segmentation, reconstruction and denoising.

Due to the periodic nature of speech and most musical instruments, it is possible to closely represent the voiced signal of a speaking person, singing voice or musical instrument by a collection of sinusoidal oscillators. Sinusoidal modeling for speech applications was introduced by McAuley and Quatieri [1]. In [2] and [3], methods which consider the sinusoidal model with noise are presented.

Manuscript received March 4, 2003; revised March 15, 2005. This work was supported in part by the Bi-national Science Foundation (BSF). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ramesh A. Gopinath.

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

S. Dubnov is with the Department of Music, University of California at San Diego, La Jolla, CA 92093-0436 USA.

Digital Object Identifier 10.1109/TSA.2005.857806

The harmonic model assumes all sinusoidal components are harmonically related, i.e., the frequencies of the sinusoids are at integer multiples of the fundamental frequency. This approach reduces the number of parameters in the model and achieves more accurate estimates of signal of interest parameters than the sinusoidal model. Several approaches to this model have been developed. The harmonic model under noisy conditions has been implemented extensively in recent studies for speech synthesis [4] and analysis [5]–[7].

Recent studies on voicing decision implement various methods of sound modeling. In [8], a statistical model-based voiced activity detector (VAD) is presented. The decision rule is established from the geometric mean of the likelihood ratios for individual frequency bands. A first-order hidden Markov model (HMM) based hang-over scheme is applied. A method for voicing decision within a pitch-detection algorithm is presented in [9]. The decision is made by defining a threshold to the median values of the cepstral peaks, the zero crossing rate (ZCR) and a short-time energy decision. An auditory-based method for voicing decision within a pitch-tracking algorithm appears in [10]. In [11] the concept of dominance spectrum is used for voiced-unvoiced decision for added white and babble noise conditions. Phone speech voicing and parameter extraction are presented in [12]. A comparison of several pitch detection and voicing decision methods is presented in [13]. The comparison is carried out between a simplified inverse filter tracking (SIFT)-based method [14], a Frobenius norm based method [15], and bilinear time-frequency based methods [16], and is performed using the Keele University database [17].

In this paper, the voicing decision problem using the generalized likelihood ratio test (GLRT) is addressed. Assuming Markovian dynamics, maximum *a-posteriori* probability (MAP) tracking of a time-varying locally harmonic signal is performed. Voicing is considered as an additional state in the global likelihood function. The voiced log-likelihood, evaluated for estimated pitch, is compared to the unvoiced log-likelihood in every frame. The described GLRT is shown to be the relation between the projection of the signal upon the harmonic subspace and its projection upon the orthogonal, nonharmonic subspace.

The structure of the paper is as follows. In Section II, we present the problem formulation. In Section III, the GLRT for voicing decision is developed. Section IV describes the MAP-based decision tracking algorithm. In Section V, the performance of the proposed algorithm is evaluated under noisy conditions. Section VI presents our conclusions.

## II. Problem Formulation

Let $\mathbf{y} = [y(t_1), \ldots, y(t_L)]^T$ be a vector representing audio signal of a finite frame of $L$ samples. The harmonic model for the measurements of a given voiced frame is presented in [7] and can be written as

$$\mathbf{y} = \mathbf{A}(\omega_0)\mathbf{b} + \mathbf{n} \tag{1}$$

where $\mathbf{A}(\omega_0)$ is the harmonic matrix and $\mathbf{b}$ is the harmonic coefficient vector. The harmonic matrix, $\mathbf{A}(\omega_0)$, can be partitioned as $\mathbf{A}(\omega_0) = [\mathbf{A}^c(\omega_0)\ \mathbf{A}^s(\omega_0)]$ where

$$\begin{aligned}[\mathbf{A}^c(\omega_0)]_{lm} &= \cos(\omega_0 m t_l),\\ &\quad m = 0, \ldots, M, \quad l = 1, \ldots, L\\ [\mathbf{A}^s(\omega_0)]_{lm} &= \sin(\omega_0 m t_l),\\ &\quad m = 1, \ldots, M, \quad l = 1, \ldots, L\end{aligned}$$

where $M$ is the total number of harmonics in the signal and $\mathbf{b} \triangleq [b_0^c, \ldots, b_M^c, b_1^s, \ldots, b_M^s]^T$.

In this work, we assume that the noise covariance matrix, $\mathbf{R_n}$, is unknown. With no loss of generality $\mathbf{R_n}$ can be decomposed as $\mathbf{R_n} = \mathbf{\Phi} + \sigma_n^2\mathbf{I}$, where $\mathbf{\Phi}$ is an unknown nonnegative definite matrix, while $\sigma_n^2$ is known. This is not a limiting assumption, since in cases when the white noise variance, $\sigma_n^2$, is unknown, it can be assumed that $\sigma_n^2 \longrightarrow 0$, and $\mathbf{R_n} = \mathbf{\Phi}$ will be estimated.

Therefore, the problem is to decide between the following two hypotheses:

$$\begin{aligned}H_1 &: \mathbf{y} = \mathbf{A}(\omega_0)\mathbf{b} + \mathbf{n}\\ H_0 &: \mathbf{y} = \mathbf{n}.\end{aligned} \tag{2}$$

The first hypothesis, $H_1$, corresponds to the case of voiced speech. The signal is considered as harmonic with additive noise, modeled as zero-mean Gaussian. Hypothesis $H_0$ corresponds to the case of unvoiced speech or silence, in which the signal contains background noise only. Under this hypothesis, the signal is modeled as a colored, zero-mean Gaussian noise with unknown covariance matrix.

## III. GLRT for Voiced-Unvoiced Decision

The GLRT for decision between the two hypotheses, stated above, is

$$\text{GLRT} = \frac{\max\limits_{\omega_0, \mathbf{b}, \mathbf{R_n}} f(\mathbf{y}|\omega_0, \mathbf{b}, \mathbf{R_n}; H_1)}{\max\limits_{\mathbf{R_n}} f(\mathbf{y}|\mathbf{R_n}; H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{3}$$

where $f(\mathbf{y}|\omega_0, \mathbf{b}, \mathbf{R_n}; H_1)$ and $f(\mathbf{y}|\mathbf{R_n}; H_0)$ are the probability density functions (pdfs) under hypotheses $H_1$ and $H_0$, respectively. We now proceed to develop the likelihood functions for both hypotheses and present the resulting GLRT.

### A. $H_1$ : Harmonic + Noise

In order to derive the log-likelihood function under hypothesis $H_1$, we employ the results obtained in [18], in which the log-likelihood for an equivalent model to the $H_1$-hypothesis from (2) is developed. Consider the data model comprised of $K$ snapshots

$$\mathbf{y}_k = \mathbf{a}_\theta s_k + \mathbf{n}_k, \quad k = 1, \ldots, K \tag{4}$$

in which the complex amplitude vector $\mathbf{s} \triangleq (s_1, \ldots, s_K)^T$, is unknown deterministic and $\mathbf{a}_\theta$ is a known function of unknown deterministic vector parameter, $\boldsymbol{\theta}$, satisfying $\|\mathbf{a}_\theta\| = 1, \forall\, \boldsymbol{\theta}$. The noise vectors are an i.i.d. sequence with Gaussian distribution $\mathbf{n}_k \sim N(\mathbf{0}, \mathbf{R_n})$ where $\mathbf{R_n} = \mathbf{\Phi} + \sigma_n^2\mathbf{I}$ and $\mathbf{\Phi}$ is an unknown nonnegative-definite matrix. The vectors $\mathbf{y}$, $\mathbf{a}_\theta$ and $\mathbf{n}_k$ are of size $L \times 1$. Then in [18] it is shown that the log-likelihood function for estimating $\boldsymbol{\theta}$ (after maximization with respect to the nuisance parameters, $\mathbf{s}$ and $\mathbf{R_n}$) is given by[1]

$$\begin{aligned}L_1(\boldsymbol{\theta}) = &-\frac{1}{2}\log\left((2\pi)^L\sigma_n^2\right)\\ &-\frac{1}{2}\sum_{l=1}^{L-1}\left[\log\left(\max\left(\lambda_{\mathbf{K},l}, \sigma_n^2\right)\right) + \frac{\lambda_{\mathbf{K},l}}{\max\left(\lambda_{\mathbf{K},l}, \sigma_n^2\right)}\right]\end{aligned} \tag{5}$$

where $L$ is the size of the vector $\mathbf{y}_k$. $\{\lambda_{\mathbf{K},l}\}_{l=1}^{L-1}$ denote the eigenvalues of the matrix $\mathbf{K}_\theta = \mathbf{T}_\theta^T\mathbf{S}\mathbf{T}_\theta$, in which $\mathbf{S} = (1/K)\sum_{k=1}^K \mathbf{y}_k\mathbf{y}_k^T$ is the sample covariance matrix and $\mathbf{T}_\theta$ is an $L \times (L-1)$ matrix whose columns are orthogonal to $\mathbf{a}_\theta$ such that $\mathbf{E}_\theta \triangleq [\mathbf{a}_\theta\ \mathbf{T}_\theta]$ defines a complete orthonormal basis, which satisfies

$$\mathbf{E}_\theta^T\mathbf{E}_\theta = \mathbf{E}_\theta\mathbf{E}_\theta^T = \mathbf{a}_\theta\mathbf{a}_\theta^T + \mathbf{T}_\theta\mathbf{T}_\theta^T = \mathbf{I}. \tag{6}$$

In the Appendix, it is shown that in a single snapshot case, i.e., $K = 1$, the likelihood function is given by

$$\begin{aligned}\max_{\boldsymbol{\theta}} L_1(\boldsymbol{\theta}) = &-\frac{1}{2}\left[\log\left((2\pi)^L\sigma_n^{2(L-1)}\right) + 1\right]\\ &-\frac{1}{2}\log\left[\mathbf{y}^T\mathbf{y} - \max_{\boldsymbol{\theta}}\left|\mathbf{a}_\theta^T\mathbf{y}\right|^2\right]\end{aligned} \tag{7}$$

where $\mathbf{y} = \mathbf{y}_1$ is the measurement vector in the single snapshot case. The model under hypothesis $H_1$, presented in (2), is equivalent to (4), with a single snapshot, i.e., $K = 1$. The model in (2) can be rewritten as

$$\mathbf{y} = \underbrace{\frac{\mathbf{A}(\omega_0)\mathbf{b}}{\|\mathbf{A}(\omega_0)\mathbf{b}\|}}_{\mathbf{a}_\theta}\underbrace{\|\mathbf{A}(\omega_0)\mathbf{b}\|}_{s} + \mathbf{n} \tag{8}$$

where the unknown deterministic vector is $\boldsymbol{\theta} \triangleq \left(\omega_0, \mathbf{b}^T\right)^T$. For simplicity of notation, we omit the dependence of the matrix $\mathbf{A}$ on $\omega_0$. Thus, the term $\left|\mathbf{a}_\theta^T\mathbf{y}\right|^2$ from (7) can be expressed in terms of $\mathbf{A}$ and $\mathbf{b}$ as

$$\begin{aligned}\left|\mathbf{a}_\theta^T\mathbf{y}\right|^2 &= \mathbf{a}_\theta^T\mathbf{y}\mathbf{y}^T\mathbf{a}_\theta\\ &= \frac{\mathbf{b}^T\mathbf{A}^T}{\|\mathbf{A}\mathbf{b}\|}\mathbf{y}\mathbf{y}^T\frac{\mathbf{A}\mathbf{b}}{\|\mathbf{A}\mathbf{b}\|} = \frac{\mathbf{b}^T\mathbf{A}^T\mathbf{y}\mathbf{y}^T\mathbf{A}\mathbf{b}}{\mathbf{b}^T\mathbf{A}^T\mathbf{A}\mathbf{b}}.\end{aligned} \tag{9}$$

[1][18] is developed for complex signals. Equation (5) is obtained after simple modifications for a real signal model.

Thus, the likelihood function under hypothesis $H_1$ is given by

$$\begin{aligned} L_1 &= \max_{\omega_0,\mathbf{b}} L_1(\omega_0,\mathbf{b}) \\ &= -\frac{1}{2}\left[\log\left((2\pi)^L\sigma_n^{2(L-1)}\right)+1\right] \\ &\quad -\frac{1}{2}\log\left[\mathbf{y}^T\mathbf{y}-\max_{\omega_0,\mathbf{b}}\frac{\mathbf{b}^T\mathbf{A}^T\mathbf{y}\mathbf{y}^T\mathbf{A}\mathbf{b}}{\mathbf{b}^T\mathbf{A}^T\mathbf{A}\mathbf{b}}\right]. \end{aligned} \tag{10}$$

Let the square matrices $\mathbf{G}$ and $\mathbf{H}$ of size $2M+1$ be defined as $\mathbf{G} \triangleq \mathbf{A}^T\mathbf{y}\mathbf{y}^T\mathbf{A}$ and $\mathbf{H} \triangleq \mathbf{A}^T\mathbf{A}$. Then, maximization of $((\mathbf{b}^T\mathbf{A}^T\mathbf{y}\mathbf{y}^T\mathbf{A}\mathbf{b})/(\mathbf{b}^T\mathbf{A}^T\mathbf{A}\mathbf{b}))$ with respect to $\mathbf{b}$ can be performed by generalized eigen-decomposition of $(\mathbf{G},\mathbf{H})$. Let $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_{2M+1}$ and $(\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_{2M+1})$ denote the generalized eigenvalues and eigenvectors of $(\mathbf{G},\mathbf{H})$, respectively

$$\mathbf{G}\mathbf{u}_i = \gamma_i\mathbf{H}\mathbf{u}_i. \tag{11}$$

Then, $\max_{\mathbf{b}}((\mathbf{b}^T\mathbf{A}^T\mathbf{y}\mathbf{y}^T\mathbf{A}\mathbf{b})/(\mathbf{b}^T\mathbf{A}^T\mathbf{A}\mathbf{b})) = \gamma_1$. By substituting the terms for $\mathbf{G}$ and $\mathbf{H}$ in (11), and left-multiplying by $(\mathbf{A}^T\mathbf{A})^{-1}$, one obtains

$$(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}\mathbf{y}^T\mathbf{A}\mathbf{u}_i = \gamma_i\mathbf{u}_i. \tag{12}$$

Since $\operatorname{rank}((\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}\mathbf{y}^T\mathbf{A}) = 1$, then $\gamma_2 = \gamma_3 = \cdots = \gamma_{2M+1} = 0$ and $\gamma_1$ can be obtained by left multiplying (12) by $\mathbf{y}^T\mathbf{A}$, which yields

$$\gamma_1 = \max_{\mathbf{b}}\frac{\mathbf{b}^T\mathbf{A}^T\mathbf{y}\mathbf{y}^T\mathbf{A}\mathbf{b}}{\mathbf{b}^T\mathbf{A}^T\mathbf{A}\mathbf{b}} = \mathbf{y}^T\mathbf{P}_{\mathbf{A}}(\omega_0)\mathbf{y} \tag{13}$$

where $\mathbf{P}_{\mathbf{A}}(\omega_0) \triangleq \mathbf{A}(\omega_0)\left(\mathbf{A}^T(\omega_0)\mathbf{A}(\omega_0)\right)^{-1}\mathbf{A}^T(\omega_0)$ denotes the harmonic projection matrix. The resulting log-likelihood function under hypothesis $H_1$ is given by

$$\begin{aligned} L_1 = L_1(\hat{\omega}_0,\hat{\mathbf{b}}) = -\frac{1}{2}\Bigg[&\log\left((2\pi)^L\sigma_n^{2(L-1)}\right) \\ &+\log\left(\mathbf{y}^T(\mathbf{I}-\mathbf{P}_{\mathbf{A}}(\hat{\omega}_0))\mathbf{y}\right)+1\Bigg] \end{aligned} \tag{14}$$

where $\hat{\omega}_0$ is given by: $\hat{\omega}_0 = \arg\max_{\omega_0} L_1(\omega_0,\hat{\mathbf{b}}) = \arg\max_{\omega_0}\|\mathbf{P}_{\mathbf{A}}(\omega_0)\mathbf{y}\|^2$.

### B. $H_0$ : Noise Only

We now develop the likelihood function under hypothesis $H_0$, which represents the case of colored Gaussian noise with unknown covariance matrix, $\mathbf{R}_{\mathbf{n}}$. The log-likelihood function under hypothesis $H_0$ is given by

$$\begin{aligned} L_0 &= \max_{\mathbf{R}_{\mathbf{n}}}\log f(\mathbf{y}|\mathbf{R}_{\mathbf{n}};H_0) \\ &= \max_{\mathbf{R}_{\mathbf{n}}}\left\{-\frac{1}{2}\left[\log|2\pi\mathbf{R}_{\mathbf{n}}|+\mathbf{y}^T\mathbf{R}_{\mathbf{n}}^{-1}\mathbf{y}\right]\right\} \end{aligned} \tag{15}$$

where the maximization is performed with the constraint $\mathbf{R}_{\mathbf{n}} = \mathbf{\Phi}+\sigma_n^2\mathbf{I}$ assuming nonnegative definite matrix, $\mathbf{\Phi}$. Without this constraint, i.e., when $\sigma_n^2 = 0$, the ML estimate of $\mathbf{R}_{\mathbf{n}}$ is $\hat{\mathbf{R}}_{\mathbf{n}} = \mathbf{y}\mathbf{y}^T$. In [18] it is shown that the constrained ML estimate of $\mathbf{R}_{\mathbf{n}}$ is obtained by the sample covariance matrix after thresholding its eigenvalues by $\sigma_n^2$. Thus, the ML estimate of the covariance matrix is given by

$$\hat{\mathbf{R}}_{\mathbf{n}} = \left(1-\frac{\sigma_n^2}{\mathbf{y}^T\mathbf{y}}\right)\mathbf{y}\mathbf{y}^T+\sigma_n^2\mathbf{I} \tag{16}$$

with eigenvalues $q_1 = \mathbf{y}^T\mathbf{y}, q_2 = \cdots = q_L = \sigma_n^2$, and therefore

$$\log|\hat{\mathbf{R}}_{\mathbf{n}}| = \sum_{l=1}^{L-1}\log q_l = \log(\mathbf{y}^T\mathbf{y})+\log\sigma_n^{2(L-1)}. \tag{17}$$

The matrix $\hat{\mathbf{R}}_{\mathbf{n}}^{-1}$ can be calculated from (16) using the Bartlett identity, and it can be verified that $\mathbf{y}^T\hat{\mathbf{R}}_{\mathbf{n}}^{-1}\mathbf{y} = 1$. The resulting log-likelihood function under hypothesis $H_0$ is

$$L_0 = -\frac{1}{2}\left[\log\left((2\pi)^L\sigma_n^{2(L-1)}\right)+1+\log(\mathbf{y}^T\mathbf{y})\right]. \tag{18}$$

### C. Decision Between Hypotheses

The GLRT from (3) can be expressed by

$$\text{GLRT} = \frac{\exp(L_1)}{\exp(L_0)} = \exp(L_1-L_0) \tag{19}$$

where $L_1$ and $L_0$ are the log-likelihood functions under hypotheses $H_1$ and $H_0$, respectively, derived in the previous subsections. By subtracting the two log-likelihoods, $L_1$ and $L_0$, from (14) and (18) we obtain

$$L_1-L_0 = \frac{1}{2}\log\frac{\mathbf{y}^T\mathbf{y}}{\mathbf{y}^T(\mathbf{I}-\mathbf{P}_{\mathbf{A}}(\hat{\omega}_0))\mathbf{y}}. \tag{20}$$

The matrix $\mathbf{P}_{\mathbf{A}}^{\perp}(\hat{\omega}_0) = \mathbf{I}-\mathbf{P}_{\mathbf{A}}(\hat{\omega}_0)$ is also a projection matrix satisfying $\mathbf{P}_{\mathbf{A}}^{\perp}(\hat{\omega}_0)\mathbf{P}_{\mathbf{A}}^{\perp}(\hat{\omega}_0) = \mathbf{P}_{\mathbf{A}}^{\perp}(\hat{\omega}_0)$ and therefore

$$\begin{aligned} L_1-L_0 &= \log\left(\frac{\mathbf{y}^T\mathbf{y}}{\mathbf{y}^T(\mathbf{I}-\mathbf{P}_{\mathbf{A}}(\hat{\omega}_0))\mathbf{y}}\right)^{1/2} \\ &= \log\left(\frac{\mathbf{y}^T\mathbf{y}}{\left\|\mathbf{P}_{\mathbf{A}}^{\perp}(\hat{\omega}_0)\mathbf{y}\right\|^2}\right)^{1/2}. \end{aligned} \tag{21}$$

Thus, from (19) and (21) we obtain

$$\begin{aligned} \text{GLRT} &= \left(\frac{\|\mathbf{y}\|^2}{\left\|\mathbf{P}_{\mathbf{A}}^{\perp}(\hat{\omega}_0)\mathbf{y}\right\|^2}\right)^{1/2} \\ &= \left(\frac{\|\mathbf{P}_{\mathbf{A}}(\hat{\omega}_0)\mathbf{y}\|^2+\left\|\mathbf{P}_{\mathbf{A}}^{\perp}(\hat{\omega}_0)\mathbf{y}\right\|^2}{\left\|\mathbf{P}_{\mathbf{A}}^{\perp}(\hat{\omega}_0)\mathbf{y}\right\|^2}\right)^{1/2} \underset{H_0}{\overset{H_1}{\gtrless}} \eta_0. \end{aligned} \tag{22}$$

Finally the test can be rewritten as

$$\frac{\|\mathbf{P}_{\mathbf{A}}(\hat{\omega}_0)\mathbf{y}\|^2}{\left\|\mathbf{P}_{\mathbf{A}}^{\perp}(\hat{\omega}_0)\mathbf{y}\right\|^2} \underset{H_0}{\overset{H_1}{\gtrless}} \eta_0^2-1 = \eta. \tag{23}$$

The GLRT for voicing decision proposes to measure the ratio between the energy of the harmonic part of the signal matching $\hat{\omega}_0$ and the energy of the complement signal. If the energy of the harmonic part of the signal is large compared to the rest of the signal, the frame is decided to be voiced.

The threshold, $\eta$, is determined to minimize the probability of error. As will be explored in the experiments, the optimal value of the threshold depends on the SNR.

## IV. Multiple Frame Tracking

A forward-backward Viterbi-like tracking algorithm is applied to the multi-framed signal. In [7] the MAP estimator for the fundamental frequency based on the harmonic model is presented. The pitch is tracked based on measurements collected over several frames using the MAP approach. In this method, a grid of possible states for the fundamental frequency, $\omega_0$, is determined, and the likelihood function is calculated for each frame. The tracking algorithm estimates and tracks the fundamental frequency using the likelihood function at each frame and the transition probability matrix, introducing the prior statistical knowledge on the fundamental frequency dynamics.

A similar tracking algorithm for MAP-based voicing decision is implemented with an additional unvoiced state. The log-likelihood under hypothesis $H_0$ is calculated in addition to $L_1(\omega_0)$. The transition probability matrix is extended to include transition to and from the unvoiced state between adjacent frames. This algorithm simultaneously tracks the pitch and decides between the two hypotheses.

The tracking algorithm calculates the *a-posteriori* cumulative probabilities for each hypothesis at each step. Therefore, tracking is performed on the input matrix comprised of the log-likelihood functions, $[L_1\ L_0 + \log(\eta_0)]$, where $\log(\eta_0)$ is the actual threshold value used for the test.

## V. Experimental Results

The results of the proposed GLRT decision method were tested under conditions of both additive noise and phone speech conditions. Additive noise experiments were carried out using the Keele pitch database and the well-known TIMIT speech database. Phone speech results were achieved using the NTIMIT phone speech database which consists of the TIMIT speech files re-recorded through an actual telephone network [19].

### A. Additive Noise

The Keele University pitch database was developed for the purpose of comparing pitch extraction algorithms [17]. The database consists of two types of signals: an acoustic signal digitized at a sampling rate of 20 KHz and a laryngograph of the acoustic signal. Five female and five male speakers were recorded reading the same passage of English text. The recordings were performed in low ambient noise conditions using a sound-proof room. The database includes reference files containing voiced-unvoiced segmentation and a pitch estimate for 25.6 msec segments overlapping every 10 msec. The reference files also mark uncertain pitch and voicing decision. For the results presented, uncertain frames and transient frames were marked and not included in the statistics. Transient frames were defined as the frames in which a change occurs in the database voicing decision and the immediately adjacent frames. Analysis was performed for a frequency range of 80–360 Hz with a resolution of 1 Hz. Pitch tracking results have been previously published and appear in [7].

The GLRT was tested for two types of additive noise: zero-mean white Gaussian noise (WGN) and babble (cocktail party) noise. The performance of the proposed decision method was tested at SNR's from 0 dB to 25 dB. The calculation of the error decision probabilities is comprised of unvoiced frames detected as voiced frames, $P_e(H_0)$, and voiced frames detected as unvoiced frames, $P_e(H_1)$.

The test threshold, $\eta$, can be optimized to minimize the probability of error. The optimal threshold value depends on the SNR. Fig. 1 shows the results of the GLRT at threshold levels, optimized for SNR's of 15 dB and 25 dB and for clean speech. The lower bound curve is obtained by adaptive threshold setting in which the threshold is adjusted for each SNR. This lower bound curve can be achieved by adaptive noise estimation. The threshold has a relatively large variance and achieves best results at the threshold extracted at an SNR of 15 dB. At high SNR's, there are less missed voiced frames, but the number of false detected voiced frames rises. For a low level of additive noise (around 15 dB SNR) the error is smallest. In the analysis of clean speech, there is a bias toward voiced detection. This implies voiced phonemes are detected better than unvoiced phonemes. Adding a low level of noise to the data signal would solve this problem.

The GLRT was also tested on similar data from the TIMIT speech database. Five female speakers and five male speakers were chosen randomly. The voicing decision was extracted from the phonetic transcription accompanying the speech files. Transient frames were not included in the statistics. Results for this test appear in Fig. 2. These results are very close to the results of Fig. 1 and demonstrate the robustness of the algorithm to speaker and environmental conditions.

Fig. 3 compares the GLRT error results for additive WGN and babble noise using the threshold optimized for $\mathrm{SNR} = 15\ \mathrm{dB}$. The harmonic nature of the babble noise causes a larger $P_e(H_1)$ error for this case and therefore, a greater decision error. It should be stated that the minimum threshold values for babble noise vary only slightly and, therefore, have little or no dependence on SNR.

Fig. 4 presents decision error tradeoff (DET) curves, i.e., $P_e(H_0)$ as a function of $P_e(H_1)$ for added WGN at SNR values of 5, 10 and 15 dB. A comparison of several different pitch detection and voicing detection methods was tested against the Keele pitch database and presented in [13]. Fig. 5 presents the DET curves for the methods compared in [13] using clean signal. A comparison between the performance of the proposed GLRT decision method (Fig. 4) and the methods presented in [13] (Fig. 5) shows that the GLRT provides much better voicing decision performance. For example, at a high SNR (15 dB), the GLRT obtains $P_e(H_1) = P_e(H_0) = 1.7\%$, improving on the methods presented in Fig. 5. The best corresponding result (for the BTFR method) is at $P_e(H_1) = P_e(H_0) = 12.5\%$. The DET curves

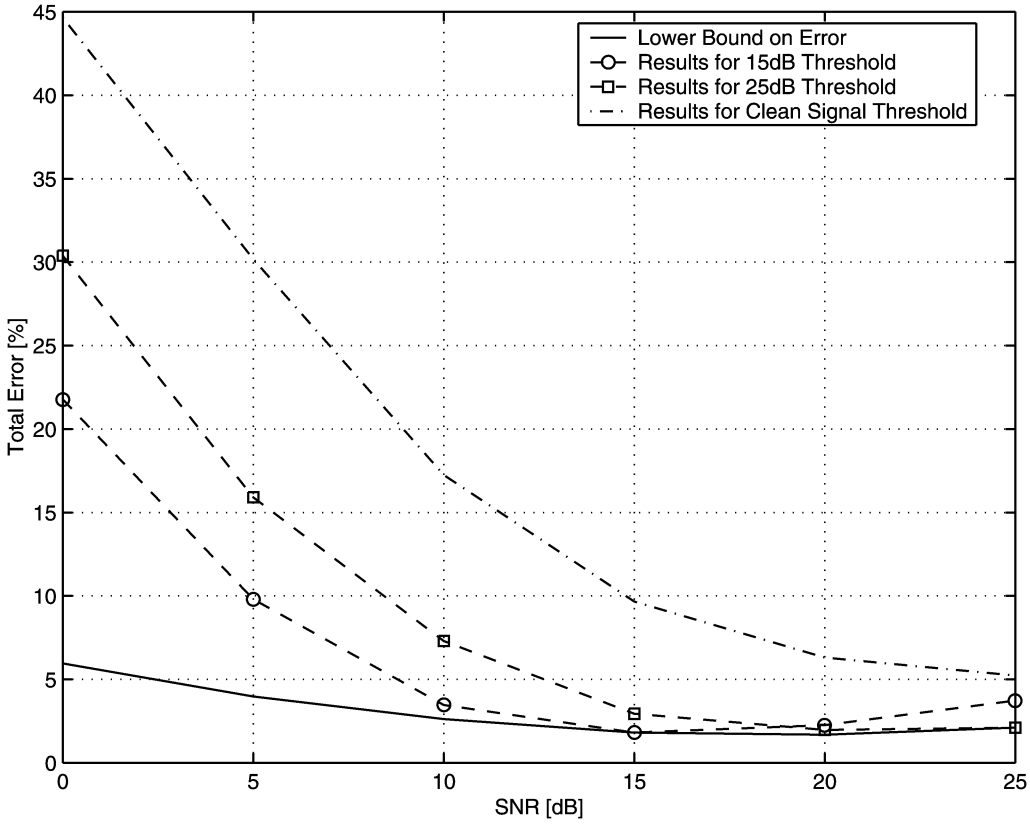


Fig. 1. Total decision error results for WGN at various threshold levels using Keele database.

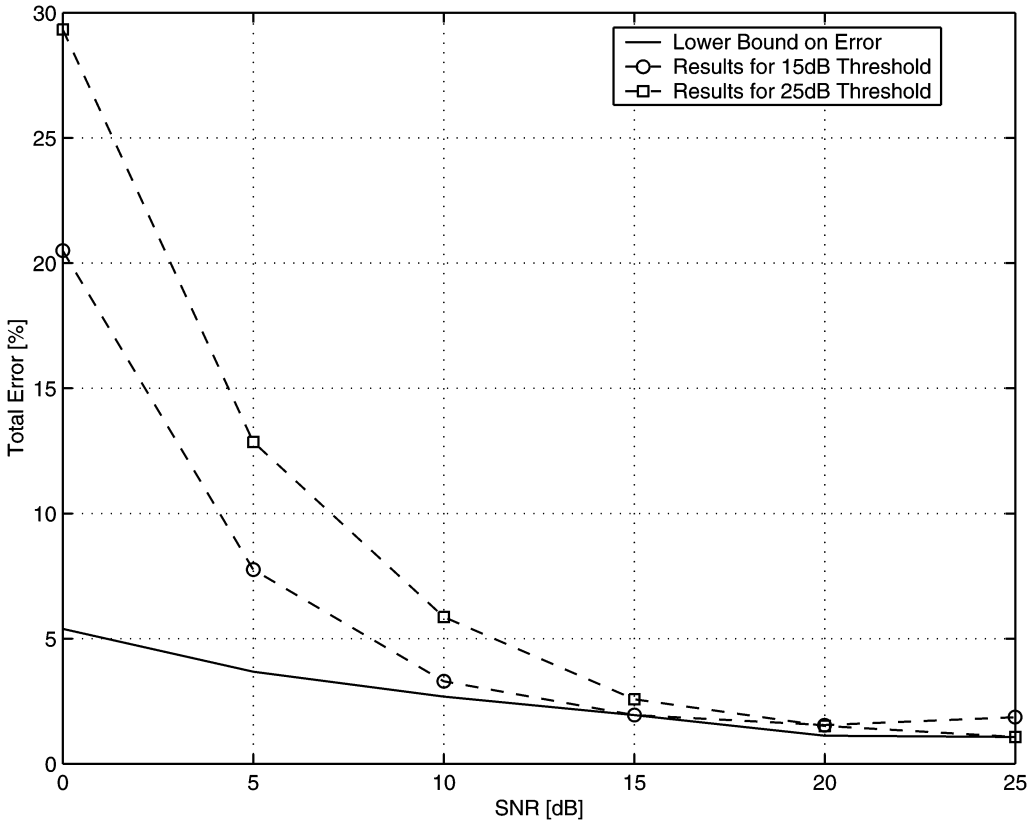


Fig. 2. Total decision error results for additive WGN using TIMIT database.

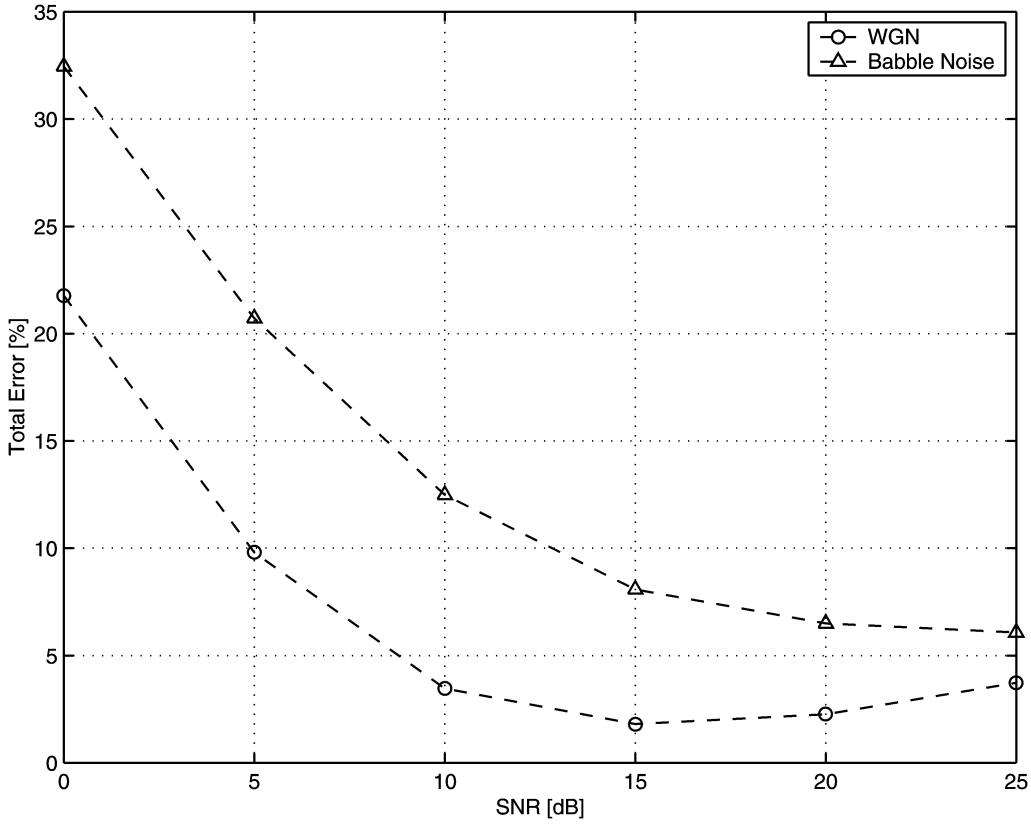


Fig. 3. Total decision error results for WGN and babble noise using Keele database. The threshold was optimized for $\mathrm{SNR} = 15$ dB.

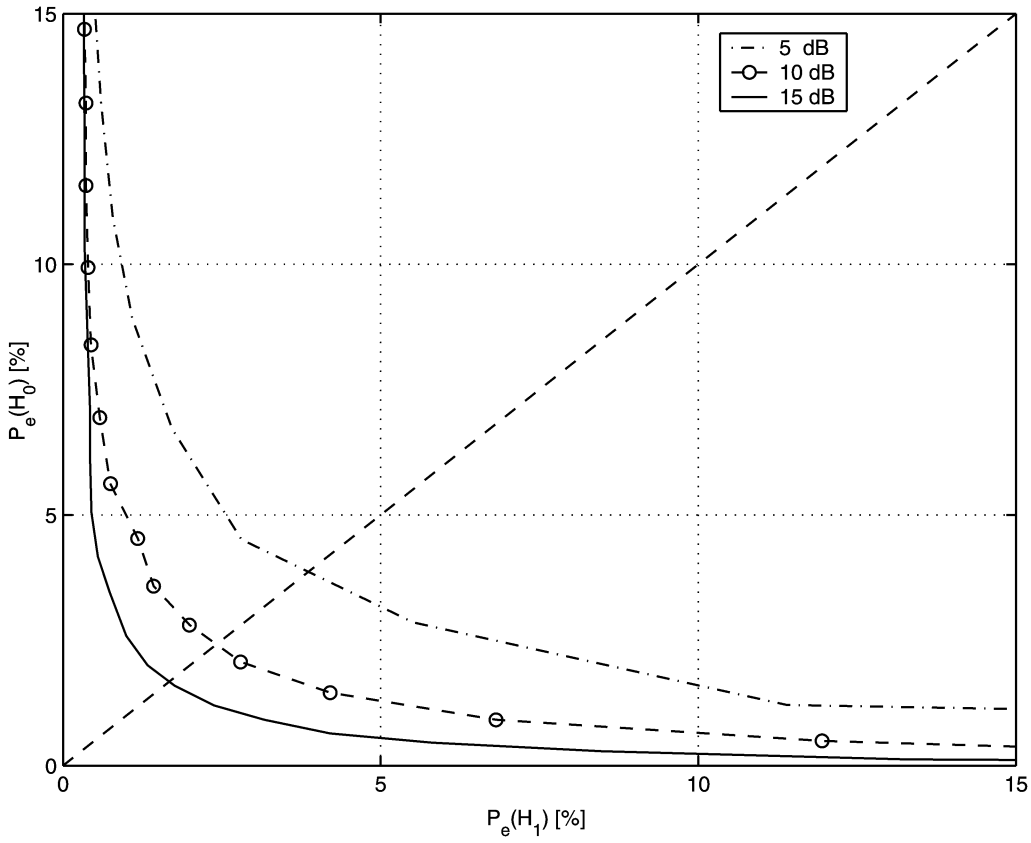


Fig. 4. DET curve for WGN at 5, 10, and 15 dB using Keele database.

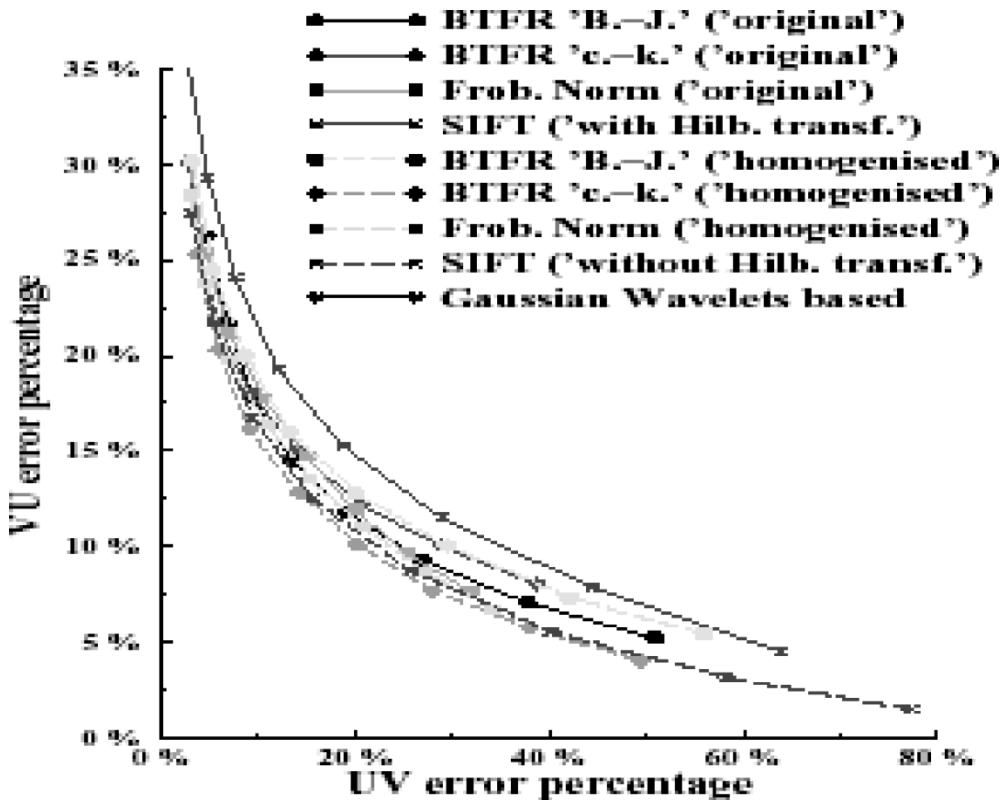


Fig. 5. DET curves for different methods published in [13] using clean signal from Keele database.

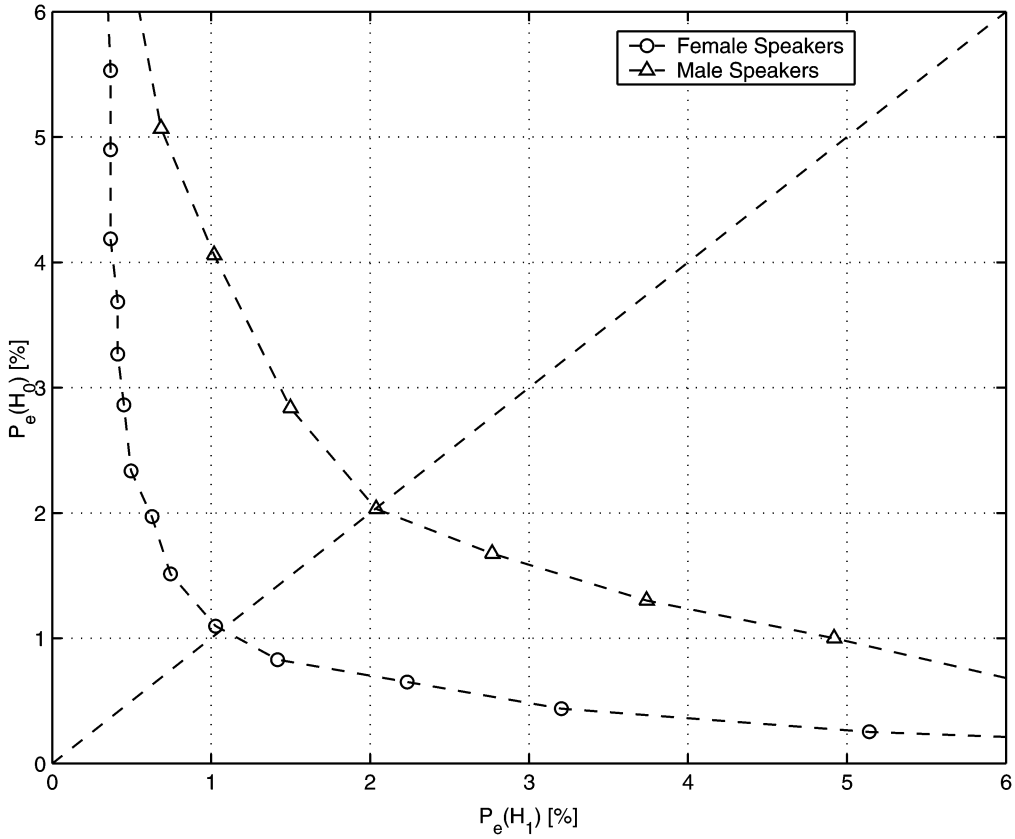


Fig. 6. Male and female speaker DET curves at 15 dB WGN using Keele database.

for female and male speakers at $\mathrm{SNR} = 15$ dB with WGN appear in Fig. 6. The GLRT obtains $P_e(H_1) = P_e(H_0) = 1.1\%$ for female speakers and $P_e(H_1) = P_e(H_0) = 2\%$ for male speakers.

TABLE I
DECISION ERROR RESULTS FOR CLEAN SPEECH (TIMIT) AND THE CORRESPONDING PHONE SPEECH (NTIMIT)

| | $P_e(H_0)$ [%] | $P_e(H_1)$ [%] | Total Error [%] |
|---|---|---|---|
| TIMIT | 0.66 | 1.77 | 0.95 |
| NTIMIT | 2.2 | 7.23 | 3.27 |

TABLE II
DECISION ERROR FOR VOICED PHONEMES FROM NTIMIT DATABASE

| Phoneme | eh | iy | aa | hv | y | n | Average |
|---|---|---|---|---|---|---|---|
| Error [%] | 0 | 0.67 | 0 | 3.3 | 2 | 0 | 0.995 |

TABLE III
DECISION ERROR FOR UNVOICED PHONEMES FROM NTIMIT DATABASE

| Phoneme | s | sh | d | dx | k | g | Average |
|---|---|---|---|---|---|---|---|
| Error [%] | 13.36 | 5.08 | 0 | 0 | 13.3 | 15 | 7.79 |

### B. Phone Speech

The GLRT was also tested on a database extracted from the NTIMIT phone speech database. The speakers were chosen to be the same as the speakers from the TIMIT database presented in the previous section. The phone speech files from the NTIMIT database were tested using the proposed GLRT and compared to the results for the corresponding clean TIMIT speech files. The results appear in Table I.

Tables II and III show the GLRT error for voiced and unvoiced phonemes from the NTIMIT database [19], respectively. These results imply there is a bias in the GLRT toward voiced decisions, as mentioned above. In the voiced case, semi-vowels and glides may cause errors. In the unvoiced case, the harmonic nature of some phonemes causes the relatively large error. For example, fricative 's' is frequently accompanied by a low amplitude whistle. The GLRT for phone speech performs better than other existing algorithms, such as ARTIFEX [12].

### C. Computational Aspects

The GLRT was tested on a Pentium 4, 1.4-GHz Linux server running MATLAB. The analysis time for a 60-s speech file was 56.5 s. The corresponding number of floating point operations (FLOPS) was under 1.45e10, i.e., about 250 MFlops/s. These results show the GLRT can be implemented in real-time. Further improvement could be achieved by real-time optimization and implementation in lower level programming languages.

## VI. CONCLUSION

The problem of voiced-unvoiced decision was addressed in this paper. A novel method based on the GLRT was derived where the voiced hypothesis was modeled by a harmonic signal and an additive Gaussian noise with unknown covariance. The unvoiced data model was a zero-mean, Gaussian vector with unknown covariance matrix. A MAP-based tracking algorithm was implemented. The GLRT was tested on the Keele pitch database and TIMIT and NTIMIT databases. The GLRT performs well for both noise and convolutional distortion such as phone speech. Results show better performance compared to other existing methods for voiced-unvoiced decision.

## APPENDIX
## DERIVATION OF THE LIKELIHOOD FUNCTION IN (7) FOR THE SINGLE SNAPSHOT CASE

In the single snapshot case where $K = 1$, the sample covariance matrix is given by $\mathbf{S} = \mathbf{y}\mathbf{y}^T$ and the matrix $\mathbf{K}_\theta$ can be rewritten as

$$\mathbf{K}_\theta = \mathbf{T}_\theta^T \mathbf{y}\mathbf{y}^T \mathbf{T}_\theta. \tag{24}$$

Since the matrix $\mathbf{K}_\theta$ is of rank one, then its eigenvalues are equal to zero except the first one, $\lambda_1$, given by

$$\lambda_1(\boldsymbol{\theta}) = \mathbf{y}^T \mathbf{T}_\theta \mathbf{T}_\theta^T \mathbf{y}. \tag{25}$$

According to (6), $\mathbf{T}_\theta \mathbf{T}_\theta^T = \mathbf{I} - \mathbf{a}_\theta \mathbf{a}_\theta^T$, and thus,

$$\lambda_1(\boldsymbol{\theta}) = \mathbf{y}^T \left(\mathbf{I} - \mathbf{a}_\theta \mathbf{a}_\theta^T\right) \mathbf{y} = \mathbf{y}^T\mathbf{y} - |\mathbf{a}_\theta^T \mathbf{y}|^2. \tag{26}$$

Assuming $\lambda_1 > \sigma_n^2$, the log-likelihood function from (5) is

$$L_1(\boldsymbol{\theta}) = -\frac{1}{2}\Big[\log\left((2\pi)^L \sigma_n^2\right) + (L-2)\log\sigma_n^2 + 1 + \log\lambda_1(\boldsymbol{\theta})\Big] \tag{27}$$

and the likelihood function is given by

$$L_1(\boldsymbol{\theta}) = -\frac{1}{2}\Big[\log\left((2\pi)^L \sigma_n^{2(L-1)}\right) + 1 + \log\left(\mathbf{y}^T\mathbf{y} - |\mathbf{a}_\theta^T\mathbf{y}|^2\right)\Big]. \tag{28}$$

## ACKNOWLEDGMENT

The authors would like to thank E. Mousset, W. A. Ainsworth and J. A. R. Fonollosa for providing us with their results (Fig. 5).

## REFERENCES

[1] T. F. Quatieri and R. J. McAulay, "Speech transformations based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 6, pp. 1449–1464 , Dec. 1986.
[2] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Proc. 11th IEEE Signal Processing Workshop on Statistical Signal Processing*, 2001, pp. 213–216.
[3] K. Fitz, L. Haken, and P. Christensen, "A new algorithm for bandwidth association in bandwidth-enhanced sinusoidal sound modeling," in *Proc. Int. Computer Music Conf.*, 2000.
[4] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 1, pp. 21–29, Jan. 2001.
[5] R. A. Irizarry, "The additive sinusoidal plus residual model: A statistical analysis," in *Proc. CNMAT*, Berkeley, CA, 1999.
[6] L. Para and U. Jain, "Approximate Kalman filtering for the harmonic plus noise model," in *Proc. IEEE Workshop App. Sig. Proc. to Audio and Acoust.*, 2001.
[7] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Tracking speech in a noisy environment using the harmonic model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 76–87, Jan. 2004.
[8] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, pp. 1–3, Jan. 1999.
[9] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 3, pp. 333–338, May 1999.

[10] J. Rouat, Y. C. Liu, and D. Morrisette, "A pitch determination and voiced/unvoiced decision algorithm for noisy speech," *Speech Commun.*, vol. 21, 1997.
[11] T. Nakatani, T. Irino, and P. Zolfaghari, "Dominance spectrum based V/UV classification and $F_0$ estimation," in *Proc. Eurospeech '03*, 2003.
[12] S. Chang, S. Greenberg, and M. Wester, "An elitist approach to articulatory-acoustic feature classification," in *Proc. Eurospeech '01*, 2001.
[13] E. Mousset, W. A. Ainsworth, and J. A. R. Fonollosa, "A comparison of several recent methods of fundamental frequency and voicing decision estimation," in *Proc. 4th Int. Conf. Spoken Language Processing*, vol. 2, 1996, pp. 1273–1276.
[14] G. F. Meyer, F. Plante, and W. A. Ainsworth, "Pitch detection: Auditory model versus inverse filtering," in *Proc. IOA*, vol. 16, 1994, pp. 81–88.
[15] Y. Kamp, C. X. Ma, and L. F. Willems, "A Frobenius-norm approach to glottal closure detection from the speech signal," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 258–265, Apr. 1994.
[16] J. L. Navarro and I. Esquerra, "A time-frequency approach to epoch detection," in *Proc. Eurospeech '95*, 1995, pp. 405–407.
[17] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. Eurospeech '95*, 1995.
[18] K. Harmanci, J. Tabrikian, and J. L. Krolik, "Relationships between adaptive minimum variance beamforming and optimal source localization," *IEEE Trans. Signal Process.*, vol. 48, no. 1, pp. 1–12, Jan. 2000.
[19] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proc. ICASSP '90*, 1990, pp. 109–112.

**Etan Fisher** received the B.Sc. and M.Sc. degrees in electrical engineering from Ben-Gurion University of the Negev, Beer-Sheva, Isreal, where he is currently working towards the Ph.D. degree.

His research interests include speech and audio processing, computer music, and acoustics.

**Joseph Tabrikian** (S'89–M'97–SM'98) received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Tel-Aviv University, Tel-Aviv, Isreal, 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 at the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Isreal. His research interests include statistical signal processing, source detection and localization, and speech and audio processing.

Dr. Tabrikian served as an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2001 to 2004.

**Shlomo Dubnov** received the B.A. degree in music composition from the Rubin Academy of Music and Dance, Jerusalem, Isreal, and the Ph.D. degree in computer science from Hebrew University of Jerusalem.

From 1996 to 1998, he was an Invited Researcher at IRCAM, Centre Pompidou, Paris, France. From 1998 to 2003, he was a Senior Lecturer in the Department of Communication Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Isreal. He is now an Associate Professor at the Department of Music and a Researcher at New Media Arts, CALIT2, University of California at San Diego, La Jolla.