GENERALIZED LIKELIHOOD RATIO TEST FOR VOICED / UNVOICED DECISION USING THE HARMONIC PLUS NOISE MODEL

E. Fisher
Dept. of ECE
Ben-Gurion University
Beer-Sheva, Israel
fisher@bgumail.bgu.ac.il

J. Tabrikian
Dept. of ECE
Ben-Gurion University
Beer-Sheva, Israel
joseph@ee.bgu.ac.il

S. Dubnov
Dept. of CSE
Ben-Gurion University
Beer-Sheva, Israel
dubnov@bgumail.bgu.ac.il

ABSTRACT

In this paper, a novel method for voiced / unvoiced decision in speech and music signals is presented. Voiced / unvoiced decision is required for many applications, including better modeling for analysis/synthesis, detection of model changes for segmentation purposes and better 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 voiced unvoiced tracking algorithm. The performance of the proposed method is tested under the Keele University database for different signal-to-noise ratios (SNRs), and the results show that the algorithm performs well even under severe noise conditions.

1. 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.

Recent works on voicing decision implement various methods of sound modeling. In [1], a statistical model based on 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.

Another algorithm for voicing decision within a pitch-detection method is presented in [2]. The pitch detection is performed via a cepstrum-based method. Initial voicing decision is made by defining a threshold to the median values of the cepstral peaks. Further voicing considerations are made based upon the Zero Crossing Rate (ZCR) of the signal and a short-time energy decision.

A comparison of several pitch detection / voicing decision methods is presented in [3]. The comparison is carried out between a SIFT-based method [4], a Frobenius Norm based method [5], and bilinear time-frequency based methods [6] is performed using the Keele University database [9].

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. 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.

In this paper, the voicing decision problem is addressed using a Generalized Likelihood Ratio Test (GLRT). 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, non-harmonic subspace.

2. PROBLEM FORMULATION

Let $\mathbf{y}$ be a finite audio frame with L samples at t_l, $l = 1, 2, \cdots, L$. 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

$$[\mathbf{A}^c(\omega_0)]_{lm} = \cos(\omega_0 m t_l)\ , \quad m = 0, \cdots, M\ , \quad l = 1, \cdots, L\ ,$$

$$[\mathbf{A}^s(\omega_0)]_{lm} = \sin(\omega_0 m t_l)\ , \quad m = 1, \cdots, M, \quad l = 1, \cdots, L\ ,$$

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

In this work, we assume the noise covariance matrix, $\mathbf{R_n}$, is unknown. However, with no loss of generality it can be assumed that some prior knowledge on the variance of the noise component is available, i.e., $\mathbf{R_n} = \Phi + \sigma_n^2 \mathbf{I}$, where Φ is an unknown non-negative definite matrix, while σ_n^2 is known. This is not a limiting assumption, since in cases when σ_n^2 is unknown, the results can be obtained by looking at the limit $\sigma_n^2 \longrightarrow 0$.

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

$$\begin{array}{l} H_1 : \mathbf{y} = \mathbf{A}(\omega_0)\mathbf{b} + \mathbf{n}_1 \ , \\ H_0 : \mathbf{y} = \mathbf{n}_0 \ . \end{array} \tag{2}$$

The first hypothesis, H_1, corresponds to the case of voiced speech. The signal is considered as harmonic with additive noise. 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.

3. GLRT FOR VOICED / UNVOICED DECISION

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

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

We now proceed to develop the likelihood functions for both hypotheses and present the resulting GLRT.

3.1. H_1: Harmonic + Noise

In order to derive the log-likelihood function under hypothesis H_1, the results obtained in [8] are employed. Consider the data model

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

in which the signal $\mathbf{s} \triangleq \{s_k\}_{k=1}^K$, is unknown deterministic, and the noise vectors are an i.i.d. sequence with $\mathbf{n}_k \sim N^c(\mathbf{0}, \mathbf{R_n})$ where $\mathbf{R_n} = \Phi + \sigma_n^2\mathbf{I}$ and Φ is an unknown non-negative-definite matrix. Then in [8] it is shown that the log-likelihood function for estimating $\boldsymbol{\theta}$ (after maximization with respect to the nuisance parameters, $\mathbf{R_n}$, $\mathbf{s}$) is given by

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

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 $\mathbf{K}_\theta = \mathbf{T}_\theta^H \mathbf{S} \mathbf{T}_\theta$, in which $\mathbf{S}$ 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

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

In our problem, the unknown deterministic vector is $\boldsymbol{\theta} \triangleq (\omega_0, \mathbf{b}^T)^T$, $\mathbf{a}_\theta \triangleq \frac{\mathbf{A}(\omega_0)\mathbf{b}}{\|\mathbf{A}(\omega_0)\mathbf{b}\|}$ and $s \triangleq \|\mathbf{A}(\omega_0)\mathbf{b}\|$. A single snapshot is available in each frame and therefore the sample covariance matrix is given by $\mathbf{S} = \mathbf{y}\mathbf{y}^H$. The matrix $\mathbf{K}_\theta$ can be rewritten as

$$\mathbf{K}_\theta = \mathbf{T}_\theta^H \mathbf{y}\mathbf{y}^H \mathbf{T}_\theta \ . \tag{7}$$

The rank of $\mathbf{K}_\theta$ is 1 and therefore all of its eigenvalues are zero except the first one, λ_1. It can be shown that the eigenvalues of the matrix $\mathbf{K}_\theta$ are given by $(\lambda_1, 0, \cdots, 0)$:

$$\lambda_1 = \mathbf{y}^H \mathbf{T}_\theta \mathbf{T}_\theta^H \mathbf{y} \ . \tag{8}$$

Assuming $\lambda_1 > \sigma_n^2$, the log-likelihood function is [1]

$$L_1(\boldsymbol{\theta}) = -\frac{1}{2}\left[\log\left((2\pi)^L \sigma_n^2\right) + \log\lambda_1 + 1 + (L-2)\log\sigma_n^2\right] \ . \tag{9}$$

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

$$\lambda_1 = \mathbf{y}^H(\mathbf{I} - \mathbf{a}_\theta \mathbf{a}_\theta^H)\mathbf{y} = \mathbf{y}^H\mathbf{y} - |\mathbf{a}_\theta^H \mathbf{y}|^2, \tag{10}$$

Maximization of $-\log\lambda_1$ is achieved when $|\mathbf{a}_\theta^H \mathbf{y}|^2$ is maximal. Returning to the original model parameters $\boldsymbol{\theta} = (\omega_0, \mathbf{b})$, it follows:

$$|\mathbf{a}_\theta^H \mathbf{y}|^2 = \frac{\mathbf{b}^H \mathbf{A}^H}{\|\mathbf{A}^H \mathbf{b}\|} \mathbf{y}\mathbf{y}^H \frac{\mathbf{A}\mathbf{b}}{\|\mathbf{A}\mathbf{b}\|} = \frac{\mathbf{b}^H \mathbf{A}^H \mathbf{y}\mathbf{y}^H \mathbf{A}\mathbf{b}}{\mathbf{b}^H \mathbf{A}^H \mathbf{A}\mathbf{b}}. \tag{11}$$

Maximization of (11) with respect to $\mathbf{b}$ is given by the maximum generalized eigenvalue of $(\mathbf{G}, \mathbf{H})$, where $\mathbf{G} = \mathbf{A}^H \mathbf{y}\mathbf{y}^H \mathbf{A}$ and $\mathbf{H} = \mathbf{A}^H \mathbf{A}$ are matrices of size $(2M+1) \times (2M+1)$. Let $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_{2M+1}$ and $(\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_{2M+1})$ denote the generalized eigenvalues and eigenvectors of $(\mathbf{G}, \mathbf{H})$, respectively:

$$\mathbf{G}\mathbf{u}_i = \gamma_i \mathbf{H}\mathbf{u}_i \ . \tag{12}$$

Then, $\max_{\mathbf{b}} \frac{\mathbf{b}^H \mathbf{A}^H \mathbf{y}\mathbf{y}^H \mathbf{A}\mathbf{b}}{\mathbf{b}^H \mathbf{A}^H \mathbf{A}\mathbf{b}} = \gamma_1$. By substituting the terms for $\mathbf{G}$ and $\mathbf{H}$ in Eq. (12), and left-multiplying by $(\mathbf{A}^H\mathbf{A})^{-1}$, one obtains

$$(\mathbf{A}^H \mathbf{A})^{-1} \mathbf{A}^H \mathbf{y}\mathbf{y}^H \mathbf{A}\mathbf{u}_i = \gamma_i \mathbf{u}_i \ . \tag{13}$$

Since $rank((\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H\mathbf{y}\mathbf{y}^H\mathbf{A}) = 1$ then $\gamma_2 = \gamma_3 = \cdots = \gamma_{2M+1} = 0$ and γ_1 can be obtained by left multiplying (13) by $\mathbf{y}^H\mathbf{A}$, which yields $\gamma_1 = \mathbf{y}^H \mathbf{P_A}(\omega_0)\mathbf{y}$ with $\mathbf{P_A}(\omega_0) \triangleq \mathbf{A}(\omega_0)(\mathbf{A}^H(\omega_0)\mathbf{A}(\omega_0))^{-1}\mathbf{A}^H(\omega_0)$ denoting the harmonic projection matrix. Thus:

$$\max_{\mathbf{b}} \lambda_1 = \mathbf{y}^H \left(\mathbf{I} - \mathbf{P_A}(\omega_0)\right)\mathbf{y} \ , \tag{14}$$

and the resulting log-likelihood function under hypothesis H_1 is given by maximization of :

$$L_1(\omega_0, \hat{\mathbf{b}}) = -\log\left(\pi^L \sigma_n^{2(L-1)}\right) - \log\left(\mathbf{y}^H(\mathbf{I} - \mathbf{P_A}(\omega_0))\mathbf{y}\right) - 1 \tag{15}$$

with respect to ω_0: $L_1 = \max_{\omega_0} L_1(\omega_0, \hat{\mathbf{b}}) = L_1(\hat{\omega}_0, \hat{\mathbf{b}})$.

[1] Note that in this problem the data vector is real.

3.2. 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, $\mathbf{R_n}$. The log-likelihood function under hypothesis H_0 is given by:

$$\begin{aligned} L_0 &= \max_{\mathbf{R_n}} \log f(\mathbf{y}|\mathbf{R_n}; H_0) \\ &= \max_{\mathbf{R_n}} \left\{ -\frac{1}{2} \left[\log |2\pi \mathbf{R_n}| + \mathbf{y}^H \mathbf{R_n}^{-1} \mathbf{y} \right] \right\} , \quad (16) \end{aligned}$$

where the maximization is performed with the constraint $\mathbf{R_n} = \Phi + \sigma_n^2 \mathbf{I}$ assuming non-negative definite matrix, Φ. Without this constraint, i.e. when $\sigma_n^2 = 0$, the ML estimate of $\mathbf{R_n}$ is $\hat{\mathbf{R}}_\mathbf{n} = \mathbf{y}\mathbf{y}^H$. In [8] it is shown that the constraint ML estimate of $\mathbf{R_n}$ is obtained by the sample covariance matrix after thresholding its eigenvalues by σ_n^2. The sample covariance matrix in this case is given by $\mathbf{S} = \mathbf{y}\mathbf{y}^H$, and by thresholding its eigenvalues the ML estimate of the covariance matrix is given by

$$\hat{\mathbf{R}}_\mathbf{n} = \left(1 - \frac{\sigma_n^2}{\mathbf{y}^H \mathbf{y}}\right) \mathbf{y}\mathbf{y}^H + \sigma_n^2 \mathbf{I} , \quad (17)$$

with eigenvalues $q_1 = \mathbf{y}^H \mathbf{y}$, $q_2 = \cdots = q_{L-1} = \sigma_n^2$, and therefore

$$\log |\hat{\mathbf{R}}_\mathbf{n}| = \sum_{l=1}^{L-1} \log q_l = \log(\mathbf{y}^H \mathbf{y}) + (L-1) \log \sigma_n^2 . \quad (18)$$

Using the Bartlett identity, it can be shown that $\mathbf{y}^H \hat{\mathbf{R}}_\mathbf{n}^{-1} \mathbf{y} = 1$. The resulting log-likelihood function under hypothesis H_0 is

$$L_0 = -\frac{1}{2} \left[\log \left((2\pi)^L \sigma_n^{2(L-1)} \right) + \log(\mathbf{y}^H \mathbf{y}) + 1 \right] . \quad (19)$$

3.3. Decision Between Hypotheses

For deriving the GLRT, the two log-likelihood functions are subtracted:

$$L_1 - L_0 = \log \frac{\mathbf{y}^H \mathbf{y}}{\mathbf{y}^H (\mathbf{I} - \mathbf{P_A}(\hat{\omega}_0)) \mathbf{y}} . \quad (20)$$

In terms of the harmonic projection matrix $\mathbf{P_A}(\hat{\omega}_0)$ and its complement, $\mathbf{P_A^\perp}(\hat{\omega}_0) \triangleq \mathbf{I} - \mathbf{P_A}(\hat{\omega}_0)$, the GLRT is expressed as:

$$GLRT = \frac{\|\mathbf{y}\|^2}{\|\mathbf{P_A^\perp y}\|^2} = \frac{\|\mathbf{P_A y}\|^2 + \|\mathbf{P_A^\perp y}\|^2}{\|\mathbf{P_A^\perp y}\|^2} \underset{H_0}{\overset{H_1}{\gtrless}} \eta . \quad (21)$$

Finally the test can be rewritten as

$$\frac{\|\mathbf{P_A y}\|^2}{\|\mathbf{P_A^\perp y}\|^2} \underset{H_0}{\overset{H_1}{\gtrless}} \eta - 1 = \eta' . \quad (22)$$

The GLRT for voicing decision proposes to measure the ratio between the energies in the harmonic part of the signal and the non-harmonic part. If the energy of the harmonic part of the signal is large compared to the non-harmonic part, then voicing is decided.

4. MULTIPLE FRAME TRACKING

A forward / backward Viterbi-like tracking algorithm is applied to the multi-framed signal. The MAP estimator for the fundamental frequency (pitch) was presented in [7], and is based on measurements collected over several frames. In this method, a grid of possible states for the fundamental frequency, ω_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. Therefore, the log-likelihood under hypothesis H_0 is calculated in addition to $L_1(\omega_0)$. The transition probability matrix is also 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.

Tracking is performed on the input matrix comprised of the log-likelihood functions, $[L_1 \;\; L_0 + \eta_0]$, where $\eta_0 = \log(\eta)$ is the actual threshold value used for the test.

5. EXPERIMENTAL RESULTS

In this section, we evaluate the results of the proposed GLRT decision method. The tests were performed using the Keele University pitch database, developed for the purpose of comparing pitch extraction algorithms [9]. The database consists of two types of signal: 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.

The proposed decision method was tested in varying noise conditions. White Gaussian noise was added to the signals at SNR's of 0 dB to 25 dB. 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 total error decision probability for 60 sec data of a single speaker as a function of SNR can be seen in Fig. 1. The average voicing decision error appears in Table 1: Fig. 2 presents $P_e(H_0)$ as a function of $P_e(H_1)$ for SNR values of 5, 10 and 20 dB.

SNR [dB]	Female Data	Male Data	Full Database
5	16	21	19
10	9	11	10
15	6	6.4	6
20	4.1	4.8	4.5
25	3.2	4.7	4

Table 1: Average GLRT decision error percentage

In [3], a comparison of several different pitch detection and voicing detection methods was tested against the Keele pitch database. Fig. 3 presents the UV-V / V-UV error results ($P_e(H_1)$ versus $P_e(H_0)$) for the methods compared in this study. No additive noise was added to the signal. Comparison between the performance of the proposed GLRT decision method (Fig. 2) and the methods presented in [9] (Fig. 3) shows that the proposed method provides better voicing decision performance. For example, at SNR = 10 dB, the GLRT obtains $P_e(H_1) = P_e(H_0) = 0.1$, better than the methods presented in Fig. 3 (noise free tests).

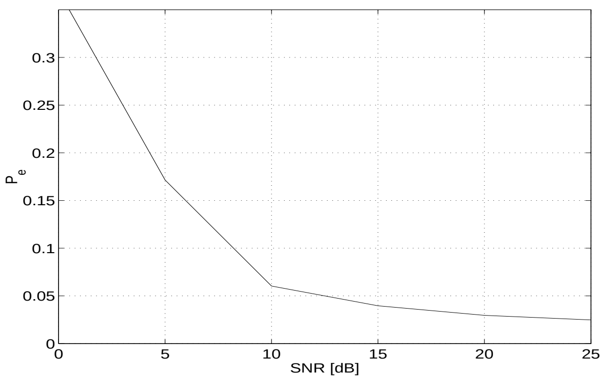

Figure 1: GLRT error results for single speaker

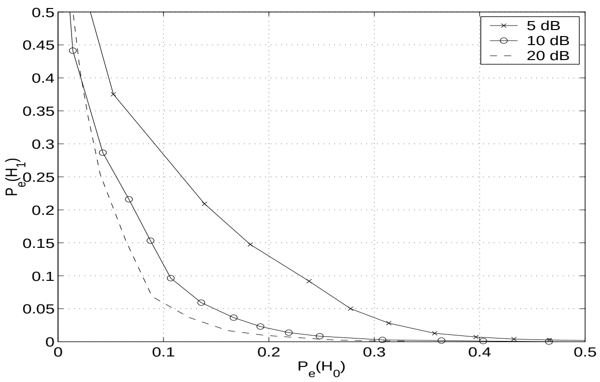

Figure 2: V-UV error ($P_e(H_1)$) vs. UV-V ($P_e(H_0)$) error for SNR = 5, 10, 20 dB.

6. CONCLUSIONS

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 results show better performance comparing to other existing methods for voice / unvoiced decision algorithms. The proposed decision method is robust to high noise level, and performs well even at low SNR's.

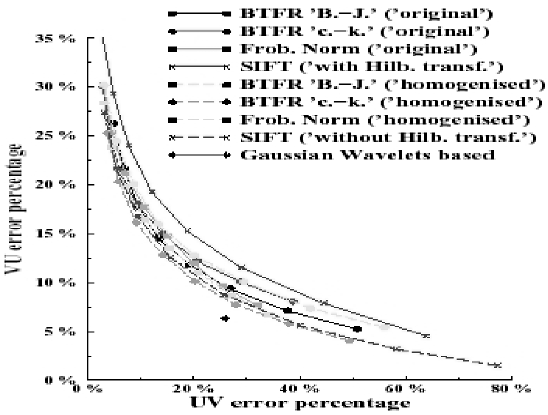

Figure 3: Error graph for methods presented in [3].

Acknowledgment

The authors would like to thank E. Mousset, W. A. Ainsworth and J. A. R. Fonollosa for providing us with their results (Fig. 3). This study was partially funded by the Bi-national Science Foundation (BSF).

REFERENCES

[1] J. Sohn, N. S. Kim, W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, Jan. 1999.

[2] S. Ahmadi and A. S. Spanias, "Cepstrum-Based Pitch Detection Using a New Statistical V/UV Classification Algorithm," *IEEE Trans. on Speech and Audio Proc.*, 1999.

[3] E. Mousset, W.A. Ainsworth and J.A.R. Fonollosa, "A comparison of several recent methods of fundamental frequency and voicing decision estimation," *The Fourth Int. Conf. on Spoken Language Processing*, vol. 2, pp. 1273-1276, 1996.

[4] G. F. Meyer, F. Plante and W. A. Ainsworth, "Pitch detection: auditory model versus inverse filtering," *Proc. I.O.A.*, 16(5) pp.81-88, 1994.

[5] Y. Kamp, C. X. Ma and L. F. Willems, "A Frobenius-norm approach to glottal closure detection from the speech signal," *IEEE Trans. on Speech and Audio Processing*, 2(2) pp.258-265, 1994.

[6] J. L. Navarro and I. Esquerra, "A time-frequency approach to epoch detection," *Proc. EUROSPEECH 95*, pp. 405-407, 1995.

[7] J. Tabrikian, S. Dubnov and Y. Dickalov. "Tracking speech in a noisy environment using the harmonic model," Submitted for publication in *IEEE Trans. Speech and Audio Processing.*

[8] K. Harmanci, J. Tabrikian and J. L. Krolik "Relationships between adaptive minimum variance beamforming and optimal source localization,", *IEEE Trans. on Signal Processing*, January 2000.

[9] F. Plante, G. F. Meyer and W.A. Ainsworth, "A pitch extraction reference database," *ESCA Eurospeech '95*, 1995.