INTERSPEECH 2009 BRIGHTON

Time-Varying Autoregressive Tests for Multiscale Speech Analysis

Daniel Rudoy[1,2], *Thomas F. Quatieri*[2], *Patrick J. Wolfe*[1]

[1]Harvard Engineering & Applied Sceinces, Cambridge MA, USA
[2]MIT Lincoln Laboratory, Lexington MA, USA
{rudoy, patrick}@seas.harvard.edu, quatieri@ll.mit.edu

10.21437/Interspeech.2009-725

Abstract

In this paper we develop hypothesis tests for speech waveform nonstationarity based on time-varying autoregressive models, and demonstrate their efficacy in speech analysis tasks at both segmental and sub-segmental scales. Key to the successful synthesis of these ideas is our employment of a generalized likelihood ratio testing framework tailored to autoregressive coefficient evolutions suitable for speech. After evaluating our framework on speech-like synthetic signals, we present preliminary results for two distinct analysis tasks using speech waveform data. At the segmental level, we develop an adaptive short-time segmentation scheme and evaluate it on whispered speech recordings, while at the sub-segmental level, we address the problem of detecting the glottal flow closed phase. Results show that our hypothesis testing framework can reliably detect changes in the vocal tract parameters across multiple scales, thereby underscoring its broad applicability to speech analysis.

Index Terms: TVAR models, hypothesis testing, GLRT, adaptive speech segmentation, glottal flow analysis

1. Introduction

It is widely accepted that speech is well-modeled as locally-stationary random process owing to the temporal variation of the glottal source and the vocal tract. Moreover, it is also known that explicitly taking advantage of this variability in front-end processing leads to superior algorithms in applications such as enhancement and recognition. Most analysis algorithms, however, simply break up the speech signal into $15-30$ ms short-time segments instead of taking advantage of the continuous evolution of vocal tract parameters or, more generally, the time-varying speech spectrum. Here, we take a first step in this direction by proposing a statistical model for vocal tract dynamics and using it to identify regions of speech in which the vocal tract configuration is not changing (e.g., steady-vowels and the glottal flow closed phase). This, in turn, can be applied to higher-resolution spectral estimates of steady-state harmonic content and improved inverse filtering algorithms, respectively [1, 2].

Motivated by the linear source-filter model of speech production, earlier work in this direction included hypothesis testing to find stationary speech segments by fitting an autoregressive (AR) process with piecewise constant parameters to the speech signal as described by [3] and references contained therein. However, in reality, the vocal tract configuration does not go through a sequence of abrupt jumps, but is instead slowly changing; hence modeling the resultant time-varying spectrum using time-varying autoregressive (TVAR) models (see, e.g., [4, 5]) is more appropriate. In this setting, the question of whether a signal is best described by an autoregressive or a TVAR process was partially addressed using the Rao test [6], but was not considered in the speech setting.

In this paper, we develop tests for spectral change based on a TVAR model, and use it to demonstrate improved speech analysis capabilities on both segmental and sub-segmental time scales. We model the temporal variation of the vocal tract through the use of TVAR models as described in Section 2, and propose in Section 3 to use a generalized likelihood ratio test (GLRT)—to detect changes in the vocal tract parameters. After evaluating the detection performance of the test on synthetic signals, we use it to design a novel adaptive analysis scheme based on the short-time Fourier transform (STFT) in Section 4.1, building on our earlier work in [1], and demonstrate its applicability to speech analysis on a segmental scale. In Section 4.2, we provide empirical results of using the GLRT on a sub-segmental scale—to detect the glottal flow closed phase.

2. Modeling of Nonstationary Signals

2.1. Time-Varying Autoregressions: Modeling

An autoregressive, or linear predictive, formulation for speech time series forms the basis of many successful algorithms to date. Typically an AR model is fit to acoustic data on a per-segment basis (following application of a smooth window function), thus enabling piecewise variation in parameter estimates. However, a more flexible alternative is to let the autoregressive coefficients themselves vary independently of the analysis scale; to this end, a time-varying autoregressive model of order p is given by the following discrete-time difference equation:

$$x[n] = -\sum_{i=1}^{p} a_i[n]x[n-i] + w[n], \quad a_i[n] = \sum_{j=0}^{q} \alpha_{ij} f_j[n], \quad (1)$$

where $a_i[n]$ are the time-varying autoregressive coefficients whose *deterministic* temporal trajectory is represented in a predetermined basis of time-varying functions $f_j : \mathbb{N} \to \mathbb{R}$ for all $1 \leq i \leq p, 0 \leq j \leq q$ weighted by coefficients $\alpha_{ij} \in \mathbb{R}$. The input $w[n]$ is zero-mean white Gaussian noise with $\mathrm{E}(w[n]^2) = \sigma^2$ for all n. Finally, we assume that $f_0[n] = 1$ for all n—this implies that if for all $j > 0$ the coefficients α_{ij} are equal to zero, then the TVAR model of (1) reduces to an AR model as:

$$x[n] = -\sum_{i=1}^{p} \alpha_{i0} x[n-i] + w[n].$$

Lincoln Laboratory authors were supported by the Department of Defense under Air Force contract FA8721-05-C-0002. The opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government. Harvard authors were supported in part by the Defense Advanced Research Projects Agency under Grant No. HR0011-07-1-0007. Recordings of audio and EGG data were collected at the Center for Laryngeal Surgery and Voice Rehabilitation at Massachusetts General Hospital.

The choice of basis functions should ideally reflect prior knowledge about the smoothness class of coefficient trajectories; in practice, a variety of basis functions including Legendre and Fourier polynomials have been considered [5]. We discuss their relative merits below.

2.2. Time-Varying Autoregressions: Estimation

Let the vector $\boldsymbol{\alpha}_j = (\alpha_{1j}, \ldots, \alpha_{pj})^T$ for $0 \leq j \leq q$ contain all the weights of the j^{th} basis function and define the vector $\boldsymbol{\theta} = \left(\boldsymbol{\alpha}_0^T, \boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_q^T\right)^T$ to contain all the $p(q+1)$ regression coefficients. Here, we describe how to estimate $\boldsymbol{\theta}$ and σ^2, given a vector of N observations $\boldsymbol{x} = (x[0], x[1], \ldots, x[N-1])$. The least-squares estimator of $\boldsymbol{\theta}$ is obtained by minimizing the prediction error:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left(x[n] + \sum_{i=1}^{p} \sum_{j=0}^{q} \alpha_{ij} f_j[n] x[n-i] \right)^2. \quad (2)$$

A detailed derivation is omitted for brevity and can be found in [4, 5]. Since $w[n]$ is white Gaussian noise, then the above estimator is also the (approximate) maximum likelihood (ML) estimator of $\boldsymbol{\theta}$ for a finite number of observations and converges to the ML estimator asymptotically. To estimate the noise variance σ^2, first note that:

$$E(x[n]x[n]) = -\sum_{i=1}^{p} \sum_{j=0}^{q} \alpha_{ij} E\left(f_j[n]x[n]x[n-i]\right) + \sigma^2.$$

Once we have solved for $\widehat{\boldsymbol{\theta}}$, we may estimate the gain by:

$$\widehat{\sigma}^2 = E(x[n]x[n]) + \sum_{i=1}^{p} \sum_{j=0}^{q} \widehat{\alpha}_{ij} E\left(f_j[n]x[n]x[n-i]\right). \quad (3)$$

In correspondence with the covariance of linear predictions both expectations are approximated by a sum over $(x[p-1], \ldots, x[N-1])$. We prefer the covariance method to the autocorrelation method, since windowing dramatically affects the estimated time-varying AR trajectories, especially for short data records [4].

3. Testing for Nonstationarity

Assuming that nonstationarity in speech is well-modeled by time-varying autoregressions, we turn to the problem of testing the null hypothesis that a data record $\boldsymbol{x} = (x[0], \ldots, x[N-1])$ came from an AR(p) process against the alternative that it came from a TVAR(p) process. This enables us to find stationary speech segments by testing whether or not the AR coefficients, and consequently the vocal tract resonances, are changing. Since a TVAR process reduces to an AR process when $\boldsymbol{\alpha}_j = \mathbf{0}$ for all $j > 0$, checking whether a segment is nonstationary may be accomplished by testing the hypothesis that all these parameters are zero. Regrouping the TVAR model parameters $\{\boldsymbol{\theta}, \sigma^2\}$ according to: $\left\{\boldsymbol{\theta}_1^T; \boldsymbol{\alpha}_0, \sigma^2\right\} = \left\{\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_q^T; \boldsymbol{\alpha}_0, \sigma^2\right\}$, the hypothesis test of interest is given by:

$$\begin{aligned} \mathcal{H}_0 &: \left(\boldsymbol{\theta}_1; \boldsymbol{\alpha}_0, \sigma^2\right) = \left(\mathbf{0}_{pq\times 1}; \boldsymbol{\alpha}_0, \sigma^2\right) \\ \mathcal{H}_1 &: \left(\boldsymbol{\theta}_1; \boldsymbol{\alpha}_0, \sigma^2\right) \neq \left(\mathbf{0}_{pq\times 1}; \boldsymbol{\alpha}_0, \sigma^2\right). \end{aligned} \quad (4)$$

A number of different test statistics may be used to realize the hypothesis test of (4). When parameter estimates under $\mathcal{H}_1$ are difficult to compute (e.g., time-varying σ^2), the Rao test can be utilized [6]. However, since σ^2 is time-invariant and ML estimates of all model parameters are available, we employed the GLRT because it tends to outperform the Rao test on finite data records—a fact we have empirically confirmed—even though the tests are asymptotically equivalent. Moreover, the GLRT provides estimates of the time-varying AR coefficients, which are useful in many speech analysis tasks.

3.1. Generalized Likelihood Ratio Test

The generalized likelihood ratio test statistic is given by:

$$L_G(\boldsymbol{x}) = \frac{p(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\alpha}}_0, \widehat{\sigma}^2, \mathcal{H}_1)}{p(\boldsymbol{x}; \widehat{\boldsymbol{\alpha}}_0, \widehat{\sigma}^2, \mathcal{H}_0)}, \quad (5)$$

where, under $\mathcal{H}_1$, the maximum likelihood estimates $(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\alpha}}_0)$ are obtained using (2) and $\widehat{\sigma}^2$ is given by (3). Under $\mathcal{H}_0$, the estimates $\widehat{\boldsymbol{\alpha}}_0$ and σ^2 are found using the well-known covariance method of linear prediction—the relevant equations are obtained by setting $q = 0$ in (2) and (3). In this composite hypothesis testing setting, $l_G(\boldsymbol{x}) \triangleq 2 \log L_G(\boldsymbol{x})$ is asymptotically distributed (in the data length N) according to a chi-squared density which is given by:

$$l_G(\boldsymbol{x}) \sim \begin{cases} \chi_d^2 & \text{under} \quad H_0 \\ \chi_d^2(\lambda) & \text{under} \quad H_1, \end{cases}$$

with $d = pq$ degrees of freedom and a non-centrality parameter λ that depends on the coefficients α_{ij}, the basis functions f_j and is found in [6]. It is critical to note that, *under* H_0, the asymptotic distribution of $l_G(\boldsymbol{x})$ depends only on the order of the AR model being fitted and the number of time-varying basis functions in the expansion of each AR coefficient. It *does not* depend on the coefficients α_{ij} or the basis functions f_j, thereby allowing us to set a constant false alarm rate (CFAR) threshold γ (e.g., 5% false alarm rate). Thus, using the test statistic given in (5), we reject H_0 if $l_G(\boldsymbol{x}) > \gamma$.

This analysis elucidates two factors which impact the test sensitivity. Selecting basis functions that most closely model the temporal trajectory of the TVAR coefficients tends to increase the power of the test. However, this may require increasing the number of basis functions q (and, therefore, the degrees of freedom), leading to a greater overlap between the distributions of the test statistic under both $\mathcal{H}_0$ and $\mathcal{H}_1$, thereby reducing the power of the test. Similarly, if the number of AR coefficients p is selected to be larger than what is necessary to model the signal with high fidelity, the power of the test is reduced.

These observations have two crucial implications for applying the hypothesis test of (4) to speech. First, the power of the test may be increased if the speech signal were bandpass filtered in order to reduce the number of AR coefficients required to model the signal spectrum. (This amounts to using side information as statistical "prior knowledge.") For instance, when identifying regions of change in the first two formants, it is helpful to resample the signal down to 4 kHz and use 4 AR coefficients instead of using 10 coefficients and working at the original sampling rate of 10 kHz. Second, using a small set of basis functions such as a line with a slope and constant offset $(q = 2)$ may be sufficient to detect change, even if not enough to accurately model the coefficient trajectories.

3.2. GLRT Detection Performance

Consider an N-sample synthetic signal generated by filtering white Gaussian noise through a two-pole bandpass filter with a

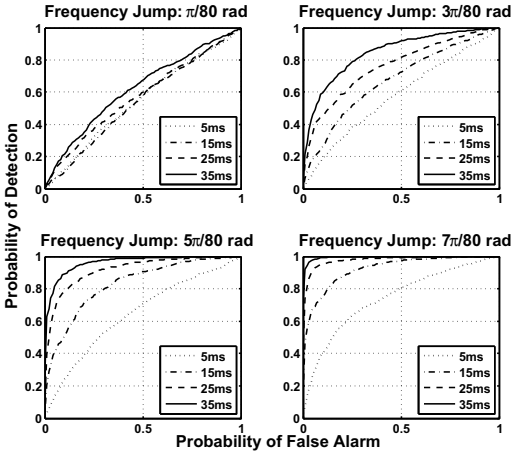


Figure 1: *ROC curves summarizing GLRT performance for different sizes of frequency jumps δ and signal lengths N.*

time-varying center frequency set to $\pi/4$ rad for the first $N/2$ samples and raised by δ rad for the last $N/2$ samples; the bandwidth was kept unchanged. A number of alternate hypotheses corresponding to $\delta \in (\pi/80, 3\pi/80, 5\pi/80, 7\pi/80)$ rad were explored and δ was set to 0 rad in order to generate data under the null hypothesis (i.e., no change). Two time-varying poles ($p = 2$), each expanded in a five-element Legendre polynomial basis ($q = 5$), were used to conduct the hypothesis test, while N was varied from 80 to 560 samples in 160 sample (10 ms) increments. One thousand Monte Carlo simulations were done for each combination of δ and N from which ROC curves shown in Figure 1 were computed. In agreement with our statistical intuition, when δ is increased while N is fixed the detection performance improves and vice versa—simply put, larger changes and those occurring over longer intervals are easier to detect.

4. Applications to Speech Analysis

4.1. Segmental Analysis: Adaptive STFT

We recently proposed an adaptive STFT analysis-synthesis scheme and applied it to signal and speech enhancement [1]. The adaptive STFT was obtained by merging certain adjacent windows of an initial wideband fixed-resolution STFT resulting in a variable-length window tiling of the signal. A nonparametric time-frequency concentration measure was used in [**?**] to decide which neighboring short-time segments to merge; here, we use the parametric hypothesis test of (4) instead.

In order to implement this idea, the GLRT statistic of (5) is computed for the joined segment to test if it is nonstationary, and if the null hypothesis is rejected, the short-time segments *are not* merged. However if the null is not rejected, then the neighboring segments *are* merged. Thus, if the signal spectrum is evolving due vocal tract changes (e.g., formant motion), and assuming a constant f_0, then neighboring windows will not be merged so that the estimated signal spectrum is not smeared. But when the formants are constant (e.g., sustained vowel), neighboring windows are merged resulting in longer windows which improves spectral resolution. In the context of speech denoising (e.g., via Wiener filtering), such adaptation leads to reduction of musical noise in enhanced speech [1].

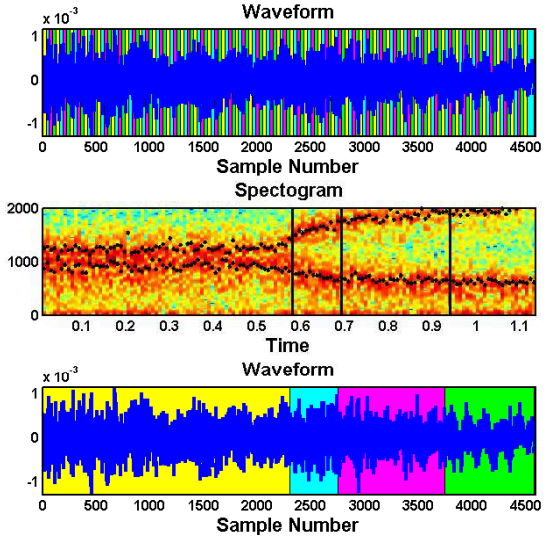
We apply the TVAR-based adaptive analysis scheme to two whispered utterances: a vowel followed by a diphthong ([a aI]) and a steady-state vowel followed by a plosive ([i t]), both recorded at 16 kHz and containing a slowly-varying and rapidly-varying spectrum, respectively. Note that whispered speech is consistent with the assumption in the TVAR model of a white innovations sequence. The results of applying the TVAR-based adaptive STFT analysis scheme to the first utterance resampled down to 4 Khz are shown in Figure 2(a). The adaptive analysis shows that the GLRT is sensitive to formant motion. Regions in which the first two formants are not moving, such as the vowel at the beginning of the waveform, are analyzed using long windows, whereas the region containing formant transitions, at the beginning of the diphthong, is analyzed separately. The relatively slow change in motion of the first two formants explains why a number of windows are joined in the transitory region of the diphthong, however, change is detected once enough data are observed. Thus, the framework performs as expected, detecting the slowly changing formants only once enough data has been examined. We have empirically observed this segmentation to be *robust* to not only reasonable choices of p and q, but also to sampling rate and the size of the initial window length—we have chosen to show the result for a sampling rate of 4 kHz for clarity.

A second, equally important, example shows that the GLRT framework can be effectively applied to detecting plosives. Even though the onset of spectral change is fast, we are able to detect it with high temporal resolution since the change is large (as compared to, e.g., the relatively slow change in the spectrum of the diphthong). The results of applying the adaptive scheme to the whispered utterance [i t] are shown in Figure 2(b). Not only does the vowel /i/ get a long window and the plosive /t/ a short one—to improve spectral resolution and prevent smearing, respectively—but also the silence before and the aspiration after the plosive are isolated.

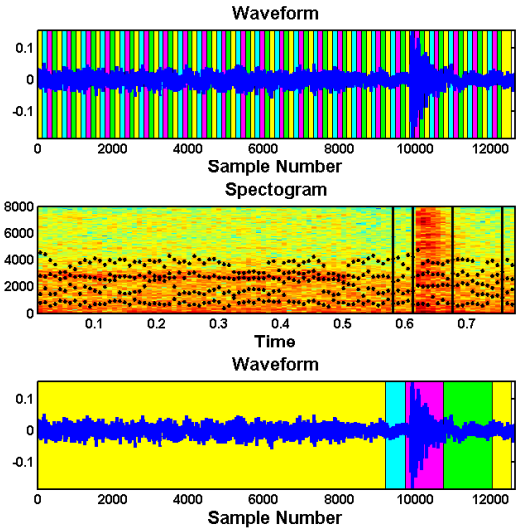
4.2. Sub-Segmental Analysis: Closed Phase ID

Detecting spectral change is easier with more data (see e.g., Section 3.2), however, speech analysis tasks at a sub-segmental scale, such as finding the glottal closed phase—useful in inverse filtering and speaker identification [2]—often suffer from a paucity of data. However, we demonstrate that the GLRT can be effective even on this scale by using it as part of a novel approach to identifying the closed phase. Specifically, it is well-known that while the vocal tract parameters are relatively constant during the closed phase, they undergo change at the beginning of the open phase [2]; therefore, we can test the ability of the GLRT to detect this boundary. We evaluate the approach on four vowels, spoken by a male with an average fundamental frequency of 109 Hz, synchronously recorded with electroglottograph (EGG) signal, which we use to obtain "ground truth."

An excised segment of the vowel /a/ and the corresponding EGG derivative are shown in the top- and bottom-left panels of Figure 3. A 50-sample rectangular window left-aligned with the second peak in the EGG derivative and, at each iteration of the sequential testing scheme, was moved one sample to the right. The GLRT with $p = 4$, $q = 2$ was performed for each window location—the test statistic $l_G(\boldsymbol{x})$ is shown in the bottom-right panel of Figure 3 along with an 15% CFAR threshold γ. When $l_G(\boldsymbol{x})$ exceeds γ, the procedure is stopped and the region from the start of the first window (glottal closure) to the end of the last window (glottal opening), marked by a dashed black line in all the panels, is declared to contain the closed phase. As the top-right panel shows, the point at which the closed phase is determined to end corresponds to *marked change* in the AR coefficients due to *both* a change in the frequency/bandwidth of



(a) *Whispered waveform [a aI]*



(b) *Whispered waveform [i t]*

Figure 2: Fixed (top) and adaptive (bottom) segmentations of two waveforms shown along with their spectrograms, computed using 16 ms triangular windows with 50% overlap and overlaid with Wavesurfer formant tracks (middle). Varying widths of the superimposed rectangles correspond to the temporal extent of the underlying analysis windows; colored for visual contrast.

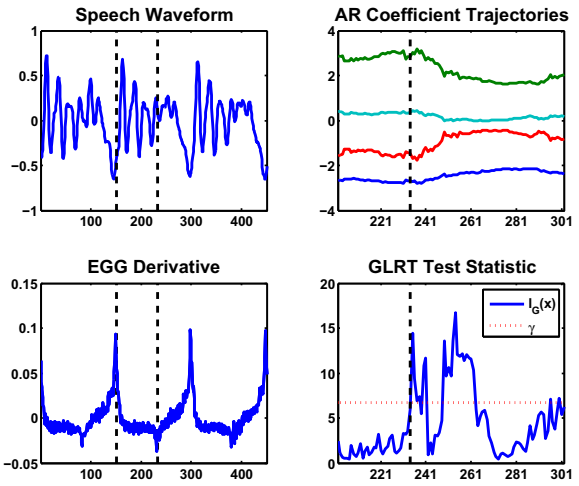


Figure 3: *Glottal flow analysis with TVAR-based GLRT. The dashed black line shows the determined instant of change in vocal tract parameters signifying the end of the closed phase.*

the first formant resulting from nonlinear source-filter interaction [2] as well as the increase in the glottal flow at the start of the open phase. Finally, the dip in the EGG derivative, often indicating the end of the closed phase [7], coincides with the output of the algorithm.

Further evaluation was done by computing the root mean-square error (RMSE) between the *detected* end of the closed phase and dips in the EGG derivative over 100 periods in each vowel; all parameters were as in the above example. The RMSE for the vowels /a/, /e/, /i/ and /o/ was found to be 0.69, 1.31, 0.95 and 1.17, respectively—their small magnitude (relative to the pitch period) underscores the promise of the approach.

5. Discussion

We have developed a hypothesis test, based on a TVAR model for the speech spectrum, to detect regions during which the vocal tract parameters are not changing and applied it to speech analysis problems on the segmental and sub-segmental scales. In future work, we will use the GLRT to detect vocal tract changes in voiced utterances as well as glottal closure instants.

6. References

[1] D. Rudoy, P. Basu, T. F. Quatieri, B. Dunn, and P. J. Wolfe, "Adaptive short-time analysis-synthesis for speech enhancement," *Proc. IEEE ICASSP*, vol. 32, pp. 4905–4908, 2008.

[2] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 7(5), pp. 569–586, 1999.

[3] R. Andre-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-36, pp. 29–40, 1988.

[4] M. G. Hall, A. V. Oppenheim, and A. S. Willsky, "Time-varying parametric modeling of speech," *Signal Process.*, vol. 5, pp. 267–285, 1983.

[5] Y. Grenier, "Time-dependent ARMA modeling of nonstationary signals," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-31, pp. 899–911, 1983.

[6] S. M. Kay, "A new nonstationarity detector," *IEEE Trans. Signal Process.*, vol. 56(4), pp. 1440–1451, 2008.

[7] D. G. Childers and J. N. Larar, "Electroglottography for laryngeal function assessment and speech analysis," *IEEE Trans. on Biomed. Eng.*, vol. 31(12), pp. 807–817, 1984.