Nonlinear Estimation of DEGG Signals with Applications to Speech Pitch Detection

Kenneth E. Barner

Applied Science and Engineering Laboratories
University of Delaware/A.I. duPont Institute
Wilmington, DE 19899

ABSTRACT

Speech pitch detection remains a fundamental problem due to importance in numerous aspects of speech processing. Current pitch detectors focus on determining the Glottal Closure Instant (GCI). Accurate GCI measures can be obtained from the Differentiated Electroglottograph (DEGG) signal. Unfortunately, DEGG signals are not available in most practical applications. A novel method of pitch detection is proposed here based on the nonlinear estimation of DEGG signals from the acoustic speech waveform. This method requires the DEGG signals only during optimization. In operation, the proposed pitch detector marks glottal closures based strictly on the acoustical speech waveform. In addition to the algorithm development, performance comparison results are presented.

1. INTRODUCTION

Pitch detection is an open fundamental problem in speech processing. This problem remains prominent because of its impact on other aspects of speech processing. The determination of pitch periods is confounded by their wide range. Not only do pitch period durations vary from speaker to speaker, but individual speakers vary pitch period durations according to pronunciation, emotion, and prosodic content. Consequently, as of yet, no pitch detection method exists that performs adequately over the required range of speakers and operating environments.

Recently developed pitch detectors have focused on determining the Glottal Closure Instant[1] (GCI). Such pitch detectors are referred to as event (glottal closing) based pitch detectors [1, 4, 6]. Although GCI event detectors have proved more effective at estimating pitch periods than classical methods, no completely satisfactory event detector method has yet been developed. One problem common to event detectors is determining an effective GCI indicator function. Current GCI indicator functions often produce many false alarms, resulting in numerous potential GCIs, or misses, resulting in missed GCIs and voiced sections of speech being classified as unvoiced. In contrast, a very accurate GCI indicator signal can be obtained with the use of an electroglottograph (EGG) [7]. The differentiated EGG (DEGG) marks the GCI with a sharp peak, allowing for simple determination of the GCI. The practical utility of the EGG, however, is limited since it requires the recording of a second channel. This restricts the use of the actual EGG signal to the laboratory, where individuals can be monitored.

In this paper we develop a method for estimating the DEGG signal from the acoustic waveform. Since the DEGG and acoustic signals are nonlinearly related, a nonlinear filtering approach must be employed. The filtering method proposed here is based on order statistics. This nonlinear estimation approach requires that the DEGG signal be available only during the optimization process. Once optimized, only the acoustic signal is needed.

2. EGG BASED PITCH DETECTION

2.1. Direct Use of EGG Signals

The EGG measures the impedance across an individuals glottis by placing pickups on either side of the throat at the level of the glottis. The recorded EGG has large shifts in bias which do not contain information on the GCIs. This signal must therefore be high-pass filtered to eliminate the bias shifts, leaving only the high-frequency, information bearing signal. Typically, a simple differentiator is sufficient for extracting the desired signal. The resulting DEGG signal, after appropriate phase shifting, marks the GCIs with sharp signal peaks, Fig. 1. The positive peeks in the DEGG clearly mark the GCIs. To differentiate between voiced and unvoiced speech, the DEGG can be thresholded. An appropriate threshold can be found as a function of the DEGG level observed during silence. For voiced sections, the speech can be broken up into frames (frames of 15 msec. were used here), and local peaks within the frames determined. These local peaks represent the GCIs. For continuity sake, the GCIs are marked in the speech as the nearest positive going zero crossing to the time indicated by the DEGG peaks.

[1] While glottal closure is a complex process, we take a simplified approach here and assume a specific instant can be identified as the time that the glottis closes.

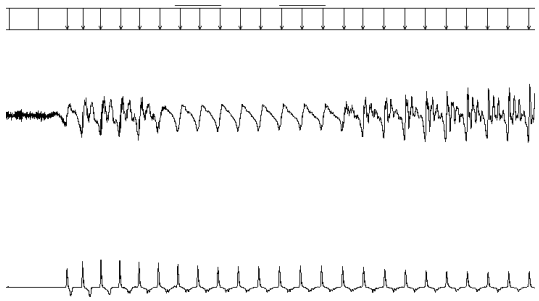

Figure 1: Example of recorded speech (middle) and DEGG (bottom) signals. The GCIs determined from the DEGG signal are marked by arrows at the top of the figure. Markers without arrow heads indicate unvoiced frames.

2.2. Estimation of DEGG Signals

Although direct use of EGG signals results in accurate localization of GCIs, the recording of the EGG channel is often impractical. We therefore propose a scheme that utilizes the EGG signal only during optimization. A diagram of the proposed pitch detector is shown in Fig. 2. During the optimization, the nonlinear filter is adaptively optimized based on the speech input and the true EGG. In operation, only the speech input is used and the GCI is determined from the estimated DEGG signal.

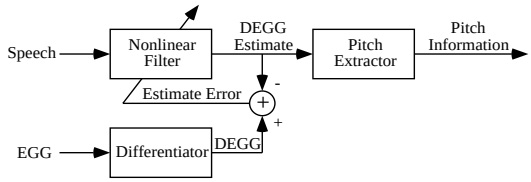

Figure 2: Block diagram of proposed pitch tracker.

The nonlinear nature of both speech and EGG signals necessitates the use of nonlinear techniques, such as those based on amplitude [3, 5]. Here, we focus on rank as an indicator of amplitude. The reliance on rank has inherent advantages in that it allows the processing of signals at different scales without the need for normalization. Proper normalization is often problematic, especially for short data sets.

To estimate the DEGG, we propose a modified L-ℓ filter. The L-ℓ filter [9] weights each observation sample according to its temporal and rank index. Thus let $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$ be the (temporally) ordered input samples at a given instant. Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}$ denote the rank ordered samples and r_i be the rank of the ith temporal sample, i.e., $x_i \equiv x_{(r_i)}$. The L-ℓ filtering operation can now be written as

$$F(\mathbf{x}) = \sum_{i=1}^{N} w_{i,r_i} x_i = \mathbf{w}^T \tilde{\mathbf{x}}, \tag{1}$$

where $\tilde{\mathbf{x}} = [\tilde{x}_{1,1}, \ldots, \tilde{x}_{1,N}, \tilde{x}_{2,1}, \ldots, \tilde{x}_{N,N}]^T$ is the expanded observation vector, $\mathbf{w} = [w_{1,1}, w_{1,2}, \ldots, w_{N,N}]^T$ is the weight vector, and

$$\tilde{x}_{i,j} = \begin{cases} x_i & \text{if} \quad r_i = j \\ 0 & \text{else} \end{cases} \tag{2}$$

is the interleaving operation.

The standard L-ℓ filter formulation applies a tap weight to each input sample. The weight applied to each sample is a function of the sample rank. Thus, the weighted sample $w_{i,r_i} x_i$ lies on one of N lines depending on which weight $w_{i,1}, w_{i,2}, \ldots, w_{i,N}$ is used. Note that each line is restricted to pass through the origin. This restriction can cause discontinuities as samples change ranks. This restriction is easily lifted by associating a bias with each weight.

Note that in an L-ℓ structure, it is often not necessary to know the exact rank of each sample. It is often sufficient to simply know what region of the ordered set each sample lies in. Moreover, it may be more important to know if a sample lies in certain rank regions, e.g., the extremes, than others. The regions, therefore, may be nonuniform. This partitioning of the ranks can be accomplished through coloring [2], which is a method for quantizing (temporal or rank) order information.

To split the ranks into M ranges, define the N integer element vector $\mathbf{q} = [q(1), \ldots, q(N)]^T$, where $1 = q(1) \leq \cdots \leq q(N) = M$. The term $q(r_i)$ gives the rank range that x_i lies in. Effectively, we have quantized (or colored) the ranks to M values. The observation vector can now be expanded to include the knowledge of which rank range each sample lies in, $\tilde{\mathbf{x}} = [\tilde{x}_{1,1}, \ldots, \tilde{x}_{1,M}, \tilde{x}_{2,1}, \ldots, \tilde{x}_{N,M}]^T$, where now the interleaving is defined by

$$\tilde{x}_{i,j} = \begin{cases} x_i & \text{if} \quad q(r_i) = j \\ 0 & \text{else} \end{cases} . \tag{3}$$

Given these definitions, the estimate can be expressed as $F(\mathbf{x}) = \sum_{i=1}^{N} w_{i,q(r_i)} x_i = \mathbf{w}^T \tilde{\mathbf{x}}$, where $\mathbf{w} = [w_{1,1}, w_{1,2}, \ldots, w_{N,M}]^T$. Note that as in the previous case, a bias can be associated with each weight resulting in $F(\mathbf{x}) = \sum_{i=1}^{N} w_{i,q(r_i)} x_i + b_{i,q(r_i)}$.

The filtering operation is thus a function of the filter weights, $\mathbf{w}$, and the rank quantization vector $\mathbf{q}$. Due to their nonlinear coupling, the joint optimization of $\mathbf{w}$ and $\mathbf{q}$ is not tractable. Therefore, a suboptimal two step recursive approach is taken. This approach is based on the fact that given $\mathbf{q}$, the estimate is a linear function of $\mathbf{w}$ that can be optimized in a MSE sense. To optimize $\mathbf{q}$, a progressive partitioning method is used, which requires the following definitions. Since the elements of $\mathbf{q}$ are integers that increase monotonically from 1 to M, $\mathbf{q}$ can be represented by its transition points. Let $s_1, \ldots, s_{M-1}$ be the transition points, i.e., $q(s_j - 1) = q(s_j) - 1$ for $j = 1, \ldots, M-1$. Set $s_0 = 1$ and $s_M = N$, and write $\mathbf{s}(M) = [s_0, \ldots, s_M]$.

Each of the M rank ranges represented by $\mathbf{s}(M)$ can be split to produce a $M+1$ range partition. This generates M possible $M+1$ rank range partitions, $\mathbf{s}^i(M+1) = [s_0^i, \ldots, s_{M+1}^i]$

where

$$s_j^i = \begin{cases} s_j & \text{if} \quad j < i \\ round((s_j + s_{j-1})/2) & \text{if} \quad j = i \\ s_{j-1} & \text{if} \quad j > i \end{cases} . \qquad (4)$$

Given the initialization $k = 2$ and starting partition $\mathbf{s}(1) = [1, N]$, the filter optimization proceeds as follows:

1. Generate $\mathbf{s}^i(k)$, $\mathbf{w}^i$ (optimal weight matrix given $\mathbf{s}^i(k)$) and the residual estimate error e^i for $i = 1, \ldots, k-1$.
2. Set $\mathbf{s}(k) = \mathbf{s}^{min}(k)$ and $\mathbf{w} = \mathbf{w}^{min}$ where min is the index satisfying $e^{min} \leq e^i$ for $i = 1, 2, \ldots, k-1$.
3. If $k = M$ stop. Else increment k and go to 1.

Rather than using a hard stop, information criteria can be used to set the number of partition. In the next section, we employ the AIC [8] to determine an optimal number of partitions.

3. RESULTS

The results presented here are for speech sample at 16 kHz. The proposed method and that based on wavelet decomposition [6] are compared using GCIs determined with direct use of EGG signals as a reference. Under the proposed method, the speech signal is down-sampled prior to DEGG estimation. Several down-sampling ratios and filter window sizes are investigated.

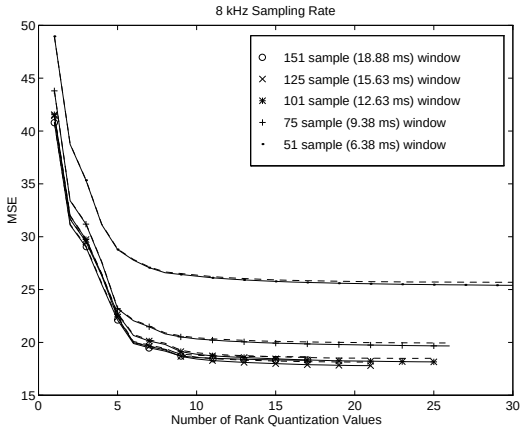

Figure 3: Estimation error as a function of the number of rank ranges (quantization values) for various window sizes.

Consider first the filter optimization results. Figure 3 shows the estimation MSE as a function of the number of rank ranges. Results are shown for several window sizes in the down-sampling by two case. The curve for each window size is terminated at the point where the AIC (dashed line in plot) increases. Note that a substantial decrease in error results after only a few rank ranges are added. Figure 4 shows the optimization produced partitioning for the first six steps of window size $N = 51$ case. Note that the extreme ranks are more finely quantized than the central ranks. Thus, the extremes of the ordered set provide the most valuable information. The optimization produced weight and bias values are plotted in Fig. 5. An examination of the weights reveals that those corresponding to the extreme rank ranges have the most variation. In fact, the weights have the structure of a differentiator, where the level of differentiation is controlled by the rank ranges. Similar partition and weight structures were observed for all sampling rates and window sizes. Thus, the filter can be intuitively interpreted as differentiating the input speech signal when the center of the observation window contains samples that are in the extremes of the ordered set. This results in a sharp peak in the filter output at the GCI.

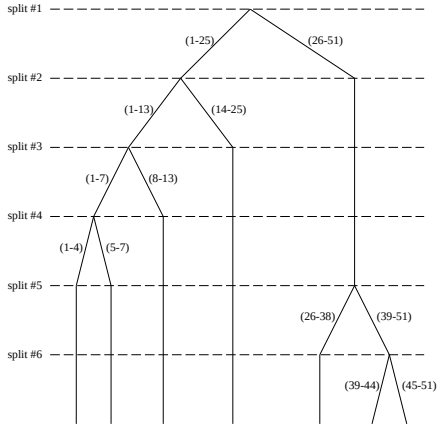

Figure 4: For $N = 51$ the first 6 splitting partitions generated during the optimization.

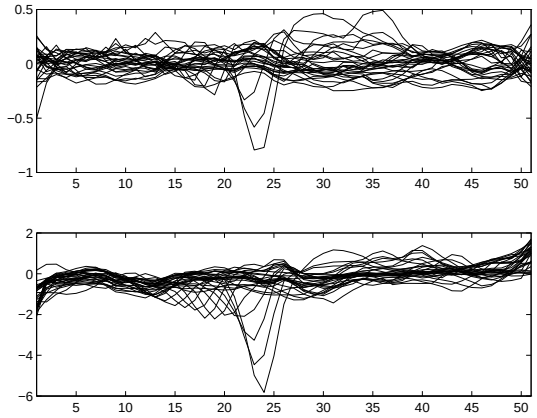

Figure 5: For $N = 51$ the filter tap weights (top) and biases (bottom) generated during the optimization. The optimization resulted in 29 rank bins, with each window location having a unique weight and bias for each bin.

To evaluate the effectiveness of determining GCIs from estimated DEGG signals, two male speakers were recorded (both audio and EEG) speaking the following "My Grandfather" paragraph:

	Speaker #1				Speaker #2			
Method/SR	Time	Matches	Insertions	Deletions	Time	Matches	Insertions	Deletions
Wavelet	222.4	94.88	7.35	5.12	203.6	86.39	13.10	13.61
Est. DEGG (8K)	84.5	96.69	10.91	3.31	72.8	93.82	10.74	6.18
Est. DEGG (4K)	46.5	96.23	8.37	3.77	42.6	91.92	6.49	8.08
Est. DEGG (2K)	27.0	95.95	7.57	4.05	21.9	93.75	6.52	6.25

Table 1: The required processing time (seconds) and the percentage of matches, insertions, and deletions for GCIs determined by the wavelet and estimated DEGG methods. The estimation filter utilize 101 observation samples and 10 rank ranges. Three down-sampling ratios were investigated.

You wished to know all about my grandfather. Well, he is nearly ninety-three years old; he dresses himself in an ancient black frock coat, usually minus several buttons; yet he still thinks as swiftly as ever. A long flowing beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks, his voice is just a bit cracked and quivers a trifle. Twice each day he plays skillfully and with zest upon our small organ. Except in the winter when the ooze or snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less, but he always answers, "Banana oil!" Grandfather likes to be modern in his language.

The estimation method was optimized using the second speaker. The GCIs were estimated for both speakers using the true and estimated DEGG signals, as well as the wavelet based [6] approach. Using the GCIs determined from the recorded EEG as a benchmark, Table 1 reports the percentage of GCI matches, insertions, deletions, as well as processing time. Note that the estimated DEGG method produces slightly better results and requires significantly less processing time. The computational savings arise from the fact that the estimation filter, after sorting, is linear. The ranking processed adds only $O(\ln N)$ operations since samples are taken in serially.

The results indicate that further development of optimized pitch detectors is warranted and that the DEGG is one possible source for a "training" signal. This approach may be particularly useful for single user or adverse condition systems, such as a noisy environment or a dysarthric talker.

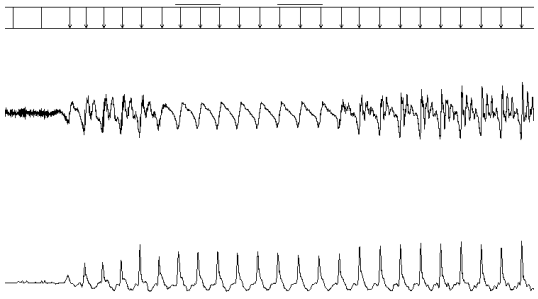

Figure 6: Speech, estimated DEGG, and marked GCIs.

4. ACKNOWLEDGEMENTS

This work is supported by The Rehabilitation Engineering Research Center on Augmentative and Alternative Communication (grant #H133E30010-96) of the National Institute on Disability and Rehabilitation Research, U.S. Department of Education, the National Science Foundation (grant #HRD-9450019), and the Nemours Research Programs.

5. REFERENCES

1. T. V. Ananthapadmanabha and B. Yegnanarayana. Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Trans. Acoust., Spch., and Sig. Proc.*, 27(8), Aug. 1979.

2. K. E. Barner and G. R. Arce. Order–statistic filtering and smoothing of time series: Part 2. *Handbook of Statistics-16 Order Statistics and their Applications*, 1996. Invited paper to appear.

3. K. E. Barner, R. C. Hardie, and G. R. Arce. On the permutation and quantization partitioning of r^n and the filtering problem. In *Proceedings of the 1994 CISS*, Princeton, New Jersey, Mar. 1994.

4. Y. M. Chen and D. O'Shaughnessy. Automatic and reliable estimation of glottal closure instant period. *IEEE Trans. Acoust., Spch., and Sig. Proc.*, 37(12), Dec. 1989.

5. J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner. Real time digital hardware pitch detector. *IEEE Trans. Acoust., Spch., and Sig. Proc.*, 24, Feb. 1976.

6. S. Kadambe and G. F. Boudreaux-Bartels. Application of the wavelet transform for pitch detection of speech signals. *IEEE Trans. Inf. Thry.*, 38(2), Mar. 1992.

7. A. K. Krishnamurthy and D. G. Childers. Two-channel speech analysis. *IEEE Trans. Acoust., Spch., and Sig. Proc.*, 34, 1986.

8. L. Ljung. *System identification : theory for the user.* Prentice-Hall, Englewood Cliffs, NJ, 1987.

9. F. Palmieri and C. G. Boncelet, Jr. L*l*-filters–a new class of order statistic filters. *IEEE Trans. Acoust., Spch., and Sig. Proc.*, 37(5), May 1989.