

COLORED L - ℓ FILTERS WITH APPLICATIONS TO SPEECH PITCH DETECTION

Kenneth E. Barner

Applied Science and Engineering Laboratories/
Department of Electrical Engineering
University of Delaware
Newark, DE 19716
barner@udel.edu

ABSTRACT

This paper develops colored L - ℓ filters and evaluates their performance in a fundamental speech processing problem: estimation of the glottal function for speech pitch detection. Colored L - ℓ filters are an extension of the temporal/rank order based L - ℓ filters in which the rank indexes are quantized (colored) and a bias is added to each weight. Quantizing the rank indexes reduces the number of filter parameters and allows the observation window size to grow beyond that previously practical. It is shown that the window size/rank quantization tradeoff has advantages in many applications.

1. INTRODUCTION

The rank ordering of data has proved advantageous in many applications. Indeed, rank ordering localized outliers in the extremes of the order set and is thus effective in applications where heavy tailed noise is common. In many applications, however, rank order alone is not sufficient. Thus, recent nonlinear filter design has focused on combining temporal and rank information [2, 10]. These filters effectively utilize the information in both orderings. However, their parameter space grows rapidly with window size, limiting the number of observation samples that can be used in practice.

The L - ℓ filter is a relatively simple formulation for utilizing temporal and rank order. Yet the L - ℓ parameter set grows as N^2 , where N is the number of observation samples. The number of parameters can be reduced to MN by quantizing the rank indexes to $M \leq N$ values. This quantization allows for the use of larger observation windows. It will be demonstrated that this tradeoff between window size and rank quantization has advantages in many applications. This paper develops and analyzes colored L - ℓ filters, derives optimization procedures, and demonstrates the advantages of colored L - ℓ filters in speech pitch detection.

2. L - ℓ TEMPORAL/RANK ORDER FILTERS

The L - ℓ filter utilizes both the temporal and rank order information by weighing each observation sample according to its temporal and rank index. Thus let $\mathbf{x} = [x_1, x_2, \dots, x_N]^T$ be the

(temporally) ordered input samples at a given instant. Let $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(N)}$ denote the rank ordered samples and r_i be the rank of the i th temporal sample, i.e., $x_i \equiv x_{(r_i)}$. The L - ℓ filtering operation can now be written as

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

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

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

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 $w_{i,1}, w_{i,2}, \dots, w_{i,N}$ 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, $F(\mathbf{x}) = \sum_{i=1}^N w_{i,r_i} x_i + b_{i,r_i}$. This, however, increases the number of weights to $2N^2$.

The nonlinear mapping capability of the L - ℓ filter often increases with window size. This is a result of the fact that knowledge of rank can be thought of as statistical quantization. For a stochastic input signal, each order statistic can be described by its distribution, or more simply, by its mean and variance. Let $\mu_{(k):N}$ and $\sigma_{(k):N}^2$ be the mean and variance (assuming existence) of the k th order statistic, respectively. The total number of samples is indicated by N and the index k is taken as the offset from the median, $-(N+1)/2 \leq k \leq (N+1)/2$. That is, $\mu_{-(N+1)/2:N}$, $\mu_{(0):N}$, and $\mu_{(N+1)/2:N}$ are the means of the minimum, median, and maximum samples, respectively. Then for many common input distributions, $\sigma_{(k):N}^2 \geq \sigma_{(k):N+2}^2$ and $|\mu_{(k):N} - \mu_{(k+1):N}| \geq |\mu_{(k):N+2} - \mu_{(k+1):N+2}|$. This is illustrated in Fig. 1 for an input consisting of Gaussian random variables of unit variance.

Thus the difference between successive order statistic means decreases with increasing window size. Similarly, the variance of each order statistic decreases with increasing window size. The quantization achieved through ranking thus becomes finer (means move closer together) and less noisy (variances decreased) with increasing window size. This improving quantization allows for more specific filter design.

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.

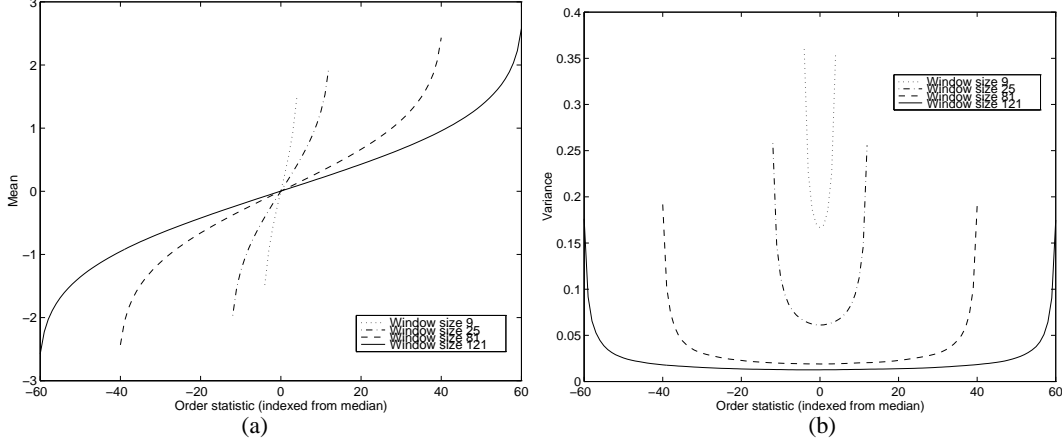


Figure 1: Order statistic (a) means and (b) variances (indexed from the median) for zero mean, unit variance Gaussian random variables.

The advantages of increasing window size and tap bias are illustrated in Fig. 2. This figure show a sigmoid function that is to be approximated by a single tap, i.e., only the center sample in the observation window is weighted, $F(\mathbf{x}) = w_{i,r_i} x_i + b_{i,r_i}$ where $i = (N + 1)/2$. The figure also show the N (MSE optimal) input/output lines for this tap when $N = 5, 15$ and 45 . The figures illustrate that optimal tap functions do not pass through the origin. Also, increasing N improves the approximation. This can be more clearly seen by examining the expected value of tap input/output function, Fig. 2 (d). Similar results hold for higher dimension functions.

While increasing the number of observation samples often results in improved function approximation and filter performance, the growth in the number of filter parameters ($2N^2$ for a l - ℓ filter with bias) limits the window size. Coloring is developed next as a way of limiting the number of parameters by quantizing the rank indexes.

3. COLORING OF RANK

Although increasing the window size can improve filter performance, 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 particularly important to know if a sample lies in certain rank regions, e.g., the extremes. The regions, therefore, may be nonuniform. This partitioning of the ranks can be accomplished through coloring [3], which is a method for quantizing (temporal or rank) order information.

To split the ranks into M ranges, define the N integer element vector $\mathbf{q} = [q(1), \dots, q(N)]^T$, where $1 = q(1) \leq \dots \leq q(N) = M$ and $q(i+1) - q(i) \in \{0, 1\}$. 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, $\tilde{\mathbf{x}} = [\tilde{x}_{1,1}, \dots, \tilde{x}_{1,M}, \tilde{x}_{2,1}, \dots, \tilde{x}_{N,M}]^T$, where now the interleaving is defined by

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

Given these definitions, the estimate can be expressed as

$$F(\mathbf{x}) = \sum_{i=1}^N w_{i,q(r_i)} x_i = \mathbf{w}^T \tilde{\mathbf{x}}, \quad (4)$$

where $\mathbf{w} = [w_{1,1}, w_{1,2}, \dots, w_{N,M}]^T$ is the weight matrix. 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, \dots, s_{M-1} be the transition points, i.e., $q(s_j - 1) = q(s_j) - 1$ for $j = 1, \dots, M - 1$. Set $s_0 = 1$ and $s_M = N$, and write $\mathbf{s}(M) = [s_0, \dots, s_M]$.

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, $\mathbf{s}^i(M + 1) = [s_0^i, \dots, s_{M+1}^i]$ where

$$s_j^i = \begin{cases} s_j & \text{if } j < i \\ \text{round}((s_j + s_{j-1})/2) & \text{if } j = i \\ s_{j-1} & \text{if } j > i \end{cases} \quad (5)$$

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 $i = 1, \dots, k - 1$.
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 $i = 1, 2, \dots, k - 1$.
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 Section 5 we employ the AIC [9] to determine an optimal number of partitions.

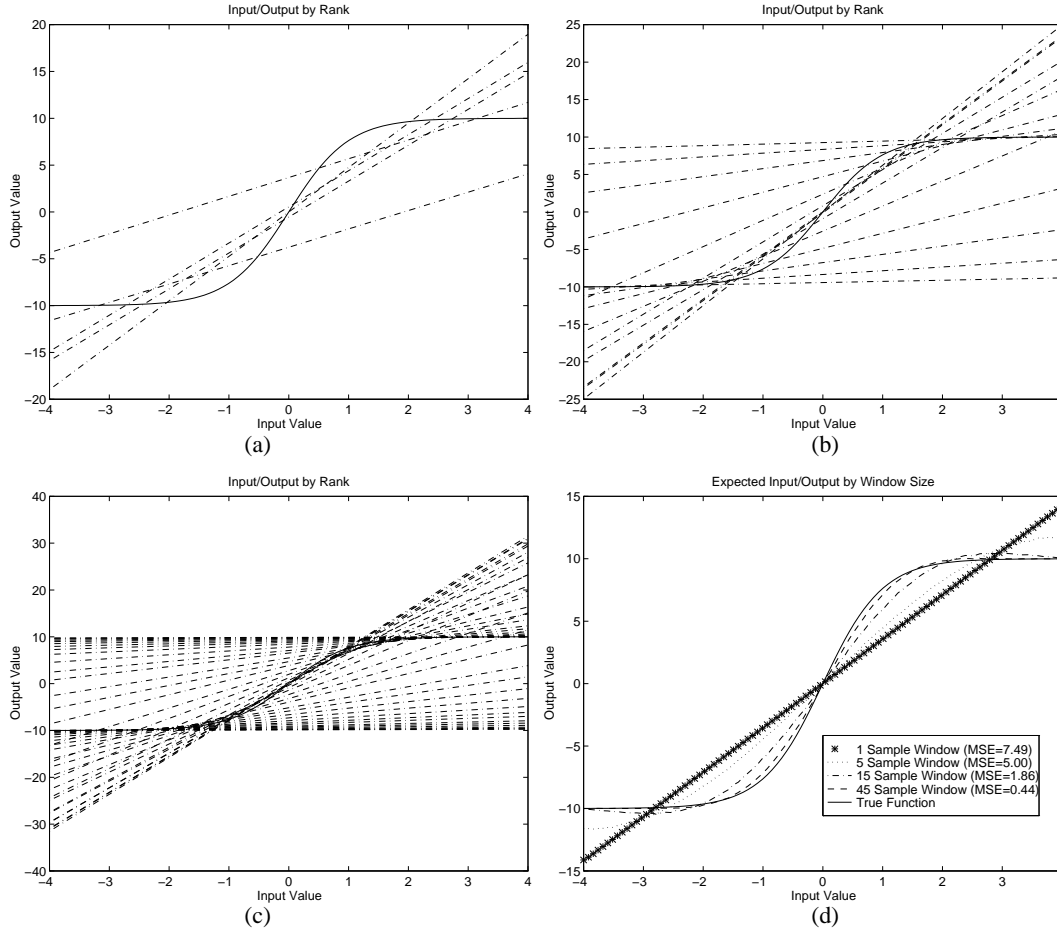


Figure 2: The input/output lines of a single tap L - ℓ (with bias) approximation of a sigmoid function for (a) $N = 5$, (b) $N = 15$, and (c) $N = 45$. The expected input/output (for Gaussian input) functions are shown in (d). For comparison, the true sigmoid function is shown in each figure.

4. PITCH DETECTION

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 Instant¹ (GCI). Such pitch detectors are referred to as event (glottal closing) based pitch detectors [1, 5, 7]. 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 pro-

¹While 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.

duce 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) [8]. The differentiated EGG (DEGG) marks the GCI with a sharp peak, allowing for simple determination of the GCI.

The EGG measures the impedance across an individual's 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. 3. The positive peaks 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

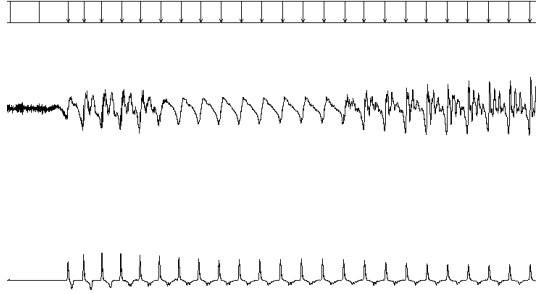


Figure 3: Example of recorded speech (middle) and DEGG (bottom) signals. The GCIs determined from the DEGG signal are marked by arrows at the top of the figure. Markers without arrow heads indicate unvoiced frames.

speech as the nearest positive going zero crossing to the time indicated by the DEGG peaks.

Although direct use of EGG signals results in accurate localization of GCIs, the practical utility of the EGG 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. We therefore propose a scheme that utilizes the EGG signal only during optimization. A diagram of the proposed pitch detector is shown in Fig. 4. 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.

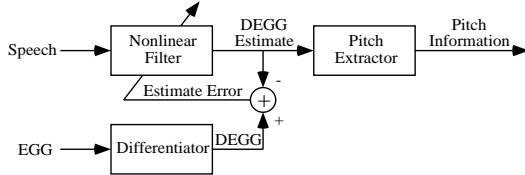


Figure 4: Block diagram of proposed pitch tracker.

The nonlinear nature of both speech and EGG signals necessitates the use of nonlinear techniques, such as those based on amplitude [4, 6]. Here, we focus on rank, as formulated in the colored L - l filter, 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.

5. RESULTS

The results presented here are for speech sample at 16 kHz. The proposed method and that based on wavelet decomposition [7] 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.

To evaluate the effectiveness of determining GCIs from estimated DEGG signals, two male speakers were recorded (both audio and EEG) speaking the “My Grandfather” paragraph. The colored L - l filter was optimized using the second speaker. The

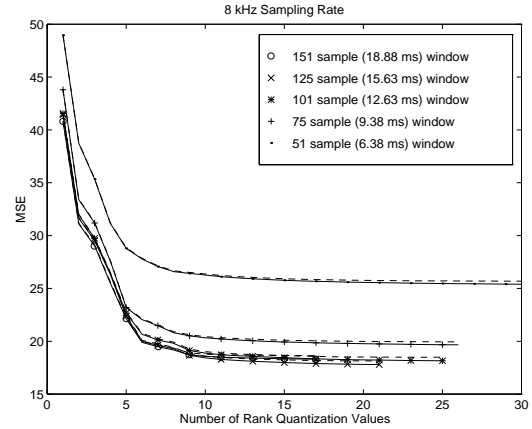


Figure 5: Estimation error as a function of the number of rank ranges (colors) for various window sizes.

GCIs were estimated for both speakers using the true and estimated DEGG signals, as well as the wavelet based [7] approach.

Consider first the filter optimization results. Figure 5 shows the estimation MSE as a function of the number of colors (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 6 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. 7. 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, Fig. 8.

The sharp peaks in the estimated DEGG signal can be used to easily determine the GCIs. Using the GCIs determined directly from the recorded EEG as a benchmark, Table 1 reports the percentage of GCI matches, insertions, deletions, as well as required processing time, for the GCIs in the “My Grandfather” paragraph determined from the DEGG estimation method and the wavelet approach. 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 the further development of optimized pitch detectors is warranted and that the DEGG is one possible source for a “training” signal. Additionally, the results indicate the coloring is an effective means of quantizing the rank indexes and constraining filter parameter growth, thus allowing the filter window to grow beyond that previously possible.

Method/SR	Speaker #1				Speaker #2			
	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 on a SPARC 20) 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.

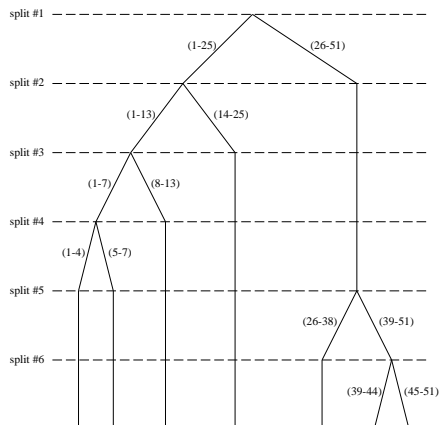


Figure 6: For $N = 51$ the first 6 splitting partitions generated during the optimization.

6. 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), August 1979.
- [2] K. E. Barner and G. R. Arce. Permutation filters: A class of non-linear filters based on set permutations. *IEEE Trans. Sig. Proc.*, 42(4), April 1994.
- [3] K. E. Barner and G. R. Arce. Design of permutation order statistic filters through group colorings. *IEEE Trans. Circ. and Syst.*, 44(7), July 1997.
- [4] K. E. Barner, R. C. Hardie, and G. R. Arce. On the permutation and quantization partitioning of \mathbf{R}^N and the filtering problem. In *Proceedings of the 1994 CISS*, Princeton, New Jersey, March 1994.
- [5] Yan Ming Chen and Douglas O'Shaughnessy. Automatic and reliable estimation of glottal closure instant period. *IEEE Trans. Acoust., Spch., and Sig. Proc.*, 37(12), December 1989.
- [6] J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner. Real time digital hardware pitch detector. *IEEE Trans. Acoust., Spch., and Sig. Proc.*, 24, February 1976.
- [7] Shuba Kadambe and G. Faye Boudreaux-Bartels. Application of the wavelet transform for pitch detection of speech signals. *IEEE Trans. Inf. Thry.*, 38(2), March 1992.

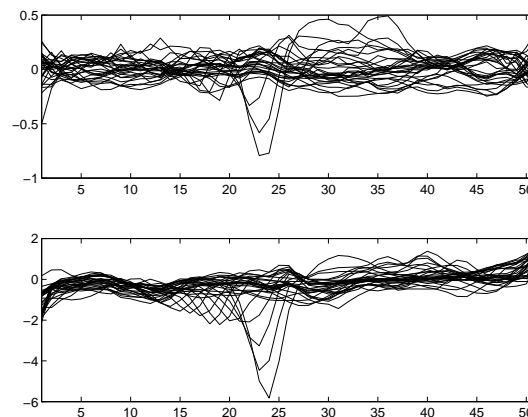


Figure 7: For $N = 51$ the filter tap weights (top) and biases (bottom) generated during the optimization. The optimization resulted in 29 rank ranges (colors), with each window location having a unique weight and bias for each bin.

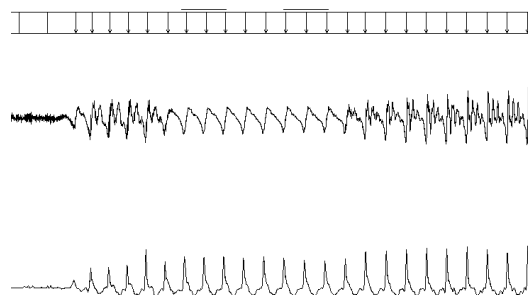


Figure 8: Speech, estimated DEGG, and marked GCIs.