

# Using stochastic complexity for ECG signal analysis

Rafael Aldea and Ciprian Doru Giurcăneanu\*

Institute of Signal Processing, Tampere University of Technology,  
P.O.Box 553, FIN-33101 Tampere, Finland  
Tel: +358 3 3653832; fax: +358 3 3653857  
e-mail: {aldea, cipriand}@cs.tut.fi

## ABSTRACT

Applying the MDL based digital signal segmentation method proposed in [1], a new QRS detection algorithm is designed and tested. It is experimentally shown that the algorithm performs better than other 11 algorithms, in terms of False Positive/False Negative estimations. The newly defined Weighted Diagnostic Distortion (WDD) measure[2] is computed for annotated files from MIT-BIH database to evaluate the accuracy of the detection algorithm and, also, to verify how well the P,T-waves are conserved by the broken line approximation.

## 1 Introduction

The interpolation is a classical technique applied in order to reduce the redundancy in ECG signals. For example, the well-known AZTEC (Amplitude Zone Time Epoch Coding) algorithm generates an approximation of the original ECG signal by alternating zero and first order interpolators which satisfy the constraint that the dynamic range of samples within a segment to be less than an experimentally found threshold[3]. The FAN method uses only first order interpolator for approximating the original signal with a broken line; a line segment approximates all that consecutive samples for which the maximum error does not exceed a given threshold[4]. The LADT (Linear Approximation Distance Thresholding) algorithm uses a similar condition on distance between the current point and the line segment which contains the approximation of the current sample[5]. A fast variant of the algorithm was introduced in [6] and recently was applied as a preprocessing stage for a neural network based QRS detection algorithm[7]. The LADT method is modified in [8] for the case when the total number of segments is chosen to ensure the given compression ratio. We consider in the sequel a segmentation algorithm which is based on MDL principle.

### 1.1 Stochastic complexity based segmentation algorithm

The segmentation problem we address in this paper is a particular case of the piecewise polynomial approximation problem treated in [1]. In the case we investigate, the degree of the polynomial for each segment is constrained to be 1 according to a wide accepted model of ECG signals[4].

For a given frame of ECG observations  $y_0, y_1, \dots, y_{N_f-1}$  we assume that at unknown time instants,  $T_1, \dots, T_{r-1}$ ,

\*This work has been supported by Academy of Finland, project No. 44876 (Finnish Centre of Excellence Program (2000-2005)).

some parameters in the model (polynomial coefficients or noise variance) change abruptly, while the noise mean is constantly zero. The observed samples in each segment  $i$  are represented as

$$y_t = \underline{\beta}'_i \underline{x}_t + \varepsilon_t, \quad T_{i-1} \leq t < T_i,$$

where  $\underline{\beta}_i$  is the 2-dimensional regressor vector, the regression vector is  $\underline{x}_t = [1 \ t - T_{i-1}]'$  for  $T_{i-1} \leq t < T_i$ ,  $r$  is the total number of segments, and  $T_0 \triangleq 0$ ,  $T_r \triangleq N_f$ . The symbol  $'$  denotes transposition and the observation vector is denoted  $\underline{y}_0^{N_f} \triangleq [y_0 \ y_1 \ \dots \ y_{N_f-1}]'$ . The  $n_i = T_i - T_{i-1}$  observations in the  $i$ -th segment can be written in vector form:

$$\underbrace{\begin{bmatrix} y_{T_{i-1}} \\ y_{T_{i-1}+1} \\ \vdots \\ y_{T_i-1} \end{bmatrix}}_{\underline{y}_{T_{i-1}}^{T_i}} = \underbrace{\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & n_i - 1 \end{bmatrix}}_{X_i} \underbrace{\begin{bmatrix} \beta_{1,i} \\ \beta_{2,i} \end{bmatrix}}_{\underline{\beta}_i} + \underbrace{\begin{bmatrix} \varepsilon_{T_{i-1}} \\ \varepsilon_{T_{i-1}+1} \\ \vdots \\ \varepsilon_{T_i-1} \end{bmatrix}}_{\underline{\varepsilon}_i}$$

We assume that the entries  $\varepsilon_{T_{i-1}}, \varepsilon_{T_{i-1}+1}, \dots, \varepsilon_{T_i-1}$  of the noise vector  $\underline{\varepsilon}_i$  are i.i.d. samples from a zero-mean Gaussian source with variance  $\sigma_i^2$ .

Given  $y_0, y_1, \dots, y_{N_f-1}$ , the stochastic complexity based algorithm introduced in [1] estimates the number of segments ( $\hat{r}$ ), their boundaries ( $\hat{T}_1, \dots, \hat{T}_{r-1}$ ), the coefficients  $\hat{\beta}_{1,1}, \hat{\beta}_{2,1}, \dots, \hat{\beta}_{1,r}, \hat{\beta}_{2,r}$ , and the noise variance for each segment. To encode these information, we compute for each segment  $i$ ,  $i \in \{1, 2, \dots, r\}$ :

$$\begin{aligned} \hat{y}_{\hat{T}_{i-1}} &= \text{round}(\hat{\beta}_{1,i}) \\ \hat{y}_{\hat{T}_i-1} &= \text{round}(\hat{\beta}_{1,i} + (\hat{n}_i - 1)\hat{\beta}_{2,i}) \end{aligned}$$

where  $\hat{n}_i = \hat{T}_i - \hat{T}_{i-1}$  and  $\text{round}(\cdot)$  states for the approximation to the nearest integer. For each segment, the data sent to the decoder are the integer-valued numbers  $\hat{y}_{\hat{T}_{i-1}}, \hat{y}_{\hat{T}_i-1}, \hat{n}_i$ , consequently, the samples of the "straight line" which approximates the segment  $i$  are given by:

$$\hat{y}_j = \text{round} \left( \frac{\hat{y}_{\hat{T}_i-1} - \hat{y}_{\hat{T}_{i-1}}}{\hat{n}_i - 1} (j - \hat{T}_{i-1}) + \hat{y}_{\hat{T}_{i-1}} \right),$$

where  $j \in \{\hat{T}_{i-1}, \hat{T}_{i-1} + 1, \dots, \hat{T}_i - 1\}$ . If the integer-valued errors  $y_j - \hat{y}_j$  are sent to the decoder, then the original signal is losslessly reconstructed at the decoder site. However, beginning from this segmentation procedure, one can

design a lossy to lossless compression scheme for ECG signals. We will investigate in the sequel how the broken line resulting after segmentation conserves the features of ECG signals that are relevant from medical point of view by computing the newly introduced Weighted Diagnostic Distortion (WDD) measure which is more meaningful than the traditional PRD or SNR measures[2]. These measures are computed when the linear approximation algorithm is applied for different files from MIT-BIH database available at <http://www.physionet.org/physiobank/archives.shtml>.

Since the coding algorithm operates frame-wise, we choose the frame length ( $N_f$ ) in such a way to have at least one QRS complex in each frame. The normal human pulse being 69 beats/minute, a frame of about 0.9 second will contain at least one QRS. Taking into account the sampling frequency, 250 Hz in our experiments, at least 218 samples will be necessary, but we choose conservatively the frame length of  $N_f = 250$ , and we also treat the unlikely cases of frames without QRS complexes.

The broken line approximation is found with the constraint that each segment contains at least two samples; it is allowed only for the last segment of each frame to have, eventually, only one sample.

In the next section we introduce a new QRS detection algorithm which is based only on the information provided by the piecewise linear approximation of the original ECG signal. We will show experimentally that, at a moderate complexity, the proposed algorithm performs better than other 11 algorithms which are tested.

## 2 QRS-complex detection using the piecewise linear approximation of the ECG signal

The main idea of the algorithm is related by the property of the R-wave to be associated with the peak of the QRS complex. This reduces the R-wave detection to finding two consecutive segments, the first one having large positive slope, and the second one with large negative slope. At the first step, the algorithm computes the slopes  $\ell_1, \ell_2, \dots, \ell_{\hat{r}}$  of the segments within the current frame, where

$$\ell_i = \frac{\hat{y}_{\hat{T}_i-1} - \hat{y}_{\hat{T}_i-1}}{\hat{n}_i - 1}, \quad i \in \{1, 2, \dots, \hat{r}\}$$

Since the algorithm is designed to work in conjunction with a frame-wise encoding scheme, no information from the next frame can be used in order to detect the R-waves in the current frame, but we can use the information from the previous frame. If  $\hat{r}'$  denotes the number of segments in the previous frame, then  $\ell_0$  is the slope corresponding to  $(\hat{r}')$ -th segment when this segment contains more than one sample; otherwise,  $\ell_0$  is the slope for the  $(\hat{r}' - 1)$ -th segment. We extend the set of slopes of the current frame to  $\{\ell_0, \ell_1, \ell_2, \dots, \ell_{\hat{r}}\}$  and consider all products  $|\ell_j \ell_{j+1}|$ ,  $0 \leq j < \hat{r}$ , which verify simultaneously  $\ell_j > 0$  and  $\ell_{j+1} < 0$ . The two largest products  $\pi_1, \pi_2$  are accepted as candidates only if they are larger than  $\alpha\pi$  where  $\alpha \in [0, 1)$  is a parameter determined experimentally, while  $\pi$  denotes the magnitude for the product of slopes corresponding to the position of the last R-wave found in the previous frame. If  $\pi_1 < \alpha\pi$  and  $\pi_2 < \alpha\pi$ , then we conclude that no R-wave was detected in the current frame. We consider for the comparison only the two largest products since we limit the possible number of QRS complexes within a frame to maximum two. Once a product  $|\ell_j \ell_{j+1}|$  accomplishes all conditions, the position of R-wave is given by the position of the first sample in segment  $j + 1$ . A supplementary condition used in order to validate the current position

is on the time interval  $\tau$  since the last R-wave:  $\tau > 500$  msec if  $RR_{av} > 700$  msec and  $\tau > 0.7RR_{av}$  if  $RR_{av} \leq 700$  msec, where  $RR_{av}$  states for the mean of last three R-R intervals. The requirement was inspired by the algorithm for T-wave detection described in [9], namely the right boundary of the region in which the T-wave is searched.

The value of parameter  $\alpha$  was determined by using the annotated files cu01.dat, channel 1 (Creighton University Ventricular Tachyarrhythmia Database), x\_04936.dat, channel 2 (MIT-BIH Atrial Fibrillation/Flutter Database), x\_slpdb.dat, channel 1 (MIT-BIH Polysomnographic Database), for which the MIT-BIH database provides the markers of R-wave locations: 204 for the first file, 2001 for the second one and 710 for the last one. The value of  $\alpha$  was varied between 0 and 0.9. The case  $\alpha = 0$  is equivalent with eliminating the comparisons between the actual slope products and the previous product, for each record. The aim of the experiments is to choose  $\alpha$  which ensures the best accuracy in the location of R-wave. For each value of  $\alpha$ , the estimated positions  $\hat{R}$  for R-wave are compared with the original annotations. The estimation  $\hat{R}$  is reported as *False Positive* (FP), if there is no true location  $R^*$  such that  $|\hat{R} - R^*| < \Theta_R$ . During this experiment we use the threshold value  $\Theta_R = 20$ . If there are several annotated locations within  $\Theta_R$  samples of the estimated location we associate the estimated location to the closest annotated location. A *False Negative* (FN) will be reported if for an annotated location  $R^*$  does not exist any estimated location within  $\Theta_R$  samples.

As it is shown in Figure 1, the best results are obtained for  $\alpha = 0.1$ . We name the proposed algorithm MDL-PLA and we compare in Figure 3 its performances measured in terms of FP/FN estimations (for  $\alpha = 0.1$ ), with other 11 algorithms satisfying the low-complexity goal. Following the classification and notations from [10] we consider the first 9 QRS detection algorithms, which are *algorithms based on amplitude and first derivative* (AF1,AF2,AF3), *algorithms based on first derivative only* (FD1,FD2), *algorithms based on first and second derivative* (FS1,FS2) and *algorithms based on digital filters* (DF1,DF2). The algorithms Pan-Tompkins (PT)[11] and Murthy-Rangaraj (MR)[12] are also considered.

MDL-PLA is the best in terms of FP/FN estimations, the only algorithm which has similar results is AF3. We compare in Figure 4 the shape of the histograms of distances from estimated location of R-wave to the annotated position, for AF3, FS2, MDL-PLA and MR; the only algorithm which performs better than MDL-PLA is MR.

The main drawback of the MDL-PLA algorithm is connected with the detection of abnormal beats which appear for some subjects. It was pointed out in [9] that these beats appear periodically or randomly within the signal. We plot in Figure 2 an example of abnormal beat. In this case, the MDL-PLA algorithm may fail in detecting the R-wave since the R position is associated to a notch, not to a peak.

### 2.1 The detection of boundaries for QRS complex

We explain briefly how the MDL-PLA algorithm can be improved in order to find Q-wave, S-wave and also the boundaries  $QRS_{on}$ ,  $QRS_{off}$ . Assume that the R position corresponds to the first sample in the segment  $i$ . It results from how the MDL-PLA is designed that the slopes  $\ell_{i-1}$  and  $\ell_i$  verify:  $\ell_{i-1} > 0 > \ell_i$ . We find an integer  $q > 0$  such that  $\ell_{i-q}, \ell_{i-q+1}, \dots, \ell_{i-1} \geq 0$  and  $\ell_{i-q-1} < 0$ . The Q position is  $\hat{T}_{i-q-1} - 1$  when  $\hat{y}_{\hat{T}_{i-q-1}-1} < \hat{y}_{\hat{T}_{i-q-1}}$  and is  $\hat{T}_{i-q-1}$  oth-

erwise. At the next step we search for an integer  $t > 0$  such that  $l_{i-t}, \dots, l_{i-q-1} \leq 0$  and  $l_{i-t-1} > 0$ . The  $QRS_{on}$  is given by  $\hat{T}_{i-t-1}$ . The procedure for detecting S-wave and  $QRS_{off}$  is similar. We introduce a penalty in choosing  $QRS_{on}$  and  $QRS_{off}$  in order to limit the total duration of the QRS complex to maximum 130 msec[9].

## 2.2 Partial WDD for measuring the accuracy of QRS detection algorithm

In [2], 18 diagnostic features are considered for computing the WDD. We select from these only four features, namely  $RR_{int}$  (the time duration, in msec, between the current and the previous location of the R waves),  $QRS_{dur}$  (the time duration, in msec, between the onset and the offset of the QRS complex),  $QRS_{amp}^+$  (the maximum amplitude, in mm, of the QRS complex), and  $QRS_{amp}^-$  (the minimum amplitude, in mm, of the QRS complex). Denote by  $\phi_i$  the annotated value for feature  $i$ ,  $i \in \{1, 2, 3, 4\}$  and by  $\hat{\phi}_i$  the value of the same feature estimated by the MDL-PLA algorithm. The distance between  $\phi_i$  and  $\hat{\phi}_i$  is given by[2]:

$$\Delta\phi_i = \frac{|\phi_i - \hat{\phi}_i|}{\max(|\phi_i|, |\hat{\phi}_i|)}.$$

Since the features included in this reduced set are related only on the QRS complex shape, we call  $WDD_{QRS}$  the partial QRS computed by weighting  $\Delta\phi_1, \dots, \Delta\phi_4$  as indicated in [2]. It is obvious that  $WDD_{QRS}$  measures the accuracy of QRS detection algorithm: smaller  $WDD_{QRS}$  means better accuracy.  $WDD_{QRS}$  was computed for five annotated files from MIT-BIH database, namely QT database, and the results are shown in the third row of Table 1. Note that for this experiment was not considered any file used to determine the "optimal" value of parameter  $\alpha$ . As we expect, the  $WDD_{QRS}$  takes larger values for the records containing abnormal beats (the name of these files are marked with \* in Table 1).

## 3 Does the broken line approximation conserve the important features of ECG signal ?

The most important drawback related by methods as AZTEC or FAN is the damage produced on P-wave and T-wave by lossy compression[4][9]. In order to evaluate how the proposed segmentation method conserves the P and T waves we selected 5 files from MIT-BIH database. The records are from the annotated QT database which is dedicated for testing the algorithms that perform P and T waves detection. In fact, we use the same files as in the experiments that aimed to evaluate the accuracy of QRS detection algorithm.

The idea of the actual experiment it is to run some algorithms which detect P,T-waves and then to compare the results obtained on the original signal and on the corresponding broken line by computing a partial WDD called  $WDD_{PT}$ . Since the QRS boundaries play a crucial role in detecting P,T-waves, we use the annotated values of these boundaries when running the detection algorithms for the original signal, and also for the broken line. We have implemented the detection algorithms as they are described in [9]. Note that for computing  $WDD_{PT}$  we consider 11 features as they are defined in [2]: the time duration between  $QRS_{on}$  and  $T_{off}$  (msec), the time duration between  $QRS_{on}$  and  $T_p$  (msec), the time duration between  $P_{on}$  and  $P_{off}$  (msec), the time duration between  $P_{on}$  and  $QRS_{on}$  (msec), the shape of T-wave (3 types), the shape of P-wave (9 types), the shape of ST segment (5 types), the amplitude of P-wave (mm), the

amplitude of T-wave (mm), the ST elevation (mm), the ST slope (mm/sec). Note that  $T_p$  denotes the position of the peak in T-wave.

The results of this experiment are reported in Table 1, last row, where we can observe that the values of WDD are small, which shows the capabilities of the segmentation method in conserving the P,T-waves of the original signal. In order to have an intuitive idea about how the estimation works, we plot in Figure 5 the beat of the sel104.dat, channel 1, with the smallest  $WDD_{PT}$  and in Figure 6 the beat of the file sel100.dat, channel 1, for which  $WDD_{PT}$  takes the largest value.

## References

- [1] C.D. Giurcăneanu, I. Tăbuș, and J. Rissanen. MDL based digital signal segmentation. In *Proc. EUSIPCO 2000, 10th European Signal Processing Conference*, volume 1, pages 339–342, Tampere, Finland, Sep. 4-8 2000.
- [2] Y. Zigel, A. Cohen, and A. Katz. The weighted diagnostic distortion (WDD) measure for ECG signal compression. *IEEE Transactions on Biomedical Engineering*, 47(11):1422–1430, Nov. 2000.
- [3] J.R. Cox, F.M. Nolle, H.A. Fozzard, and G.C. Oliver. AZTEC, a preprocessing program for real-time ECG rhythm analysis. *IEEE Transactions on Biomedical Engineering*, pages 128–129, Apr. 1968.
- [4] S.M.S. Jaleldine, C.H. Hutchens, R.D. Strattan, and W.A. Coberly. ECG data compression techniques - a unified approach. *IEEE Transactions on Biomedical Engineering*, 37(4):329–343, Apr. 1990.
- [5] J.K. Udupa and I.S.N. Murthy. Syntactic approach to ECG rhythm analysis. *IEEE Transactions on Biomedical Engineering*, BME-27(7):370–375, Jul. 1980.
- [6] L. Gang, F. Jing, L. Ling, and Y. Qilian. Fast realization of the LADT ECG data compression method. *IEEE Engineering in Medicine and Biology*, pages 255–258, Apr.-May 1994.
- [7] L. Gang, Y. Wenyu, L. Ling, Y. Qilian, and Y. Xuemin. An artificial-intelligence approach to ECG analysis. *IEEE Engineering in Medicine and Biology*, pages 95–100, Mar.-Apr. 2000.
- [8] A. Koski and M. Juhola. Segmentation of digital signals based on estimated compression ratio. *IEEE Transactions on Biomedical Engineering*, 43(9):928–938, Sep. 1996.
- [9] Y. Zigel. ECG signal compression. Master's thesis, Ben Gurion Univ. of the Negev, Aug. 1998.
- [10] G.M. Friesen, Th.C. Jannett, M.N. Jadallah, S.L. Yates, St.R. Quint, and H.T. Nagle. A comparison of the noise sensitivity of nine QRS detection algorithms. *IEEE Transactions on Biomedical Engineering*, 37(1):85–98, Jan. 1990.
- [11] J. Pan and W.J. Tompkins. A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236, Mar. 1985.
- [12] I.S. Murthy and M.R. Rangaraj. New concepts for PVC detection. *IEEE Transactions on Biomedical Engineering*, BME-26(7):409–416, Jul. 1979.

File	sel100.dat	sel103.dat	sel103.dat	sel104.dat*	sel114.dat*
Channel	1	1	2	1	1
$WDD_{QRS}$ [%]	0.25	0.36	0.29	2.25	4.09
$WDD_{PT}$ [%]	6.85	3.90	4.95	1.30	1.80

Table 1:  $WDD_{QRS}$  and  $WDD_{PT}$  computed for some files from QT database.

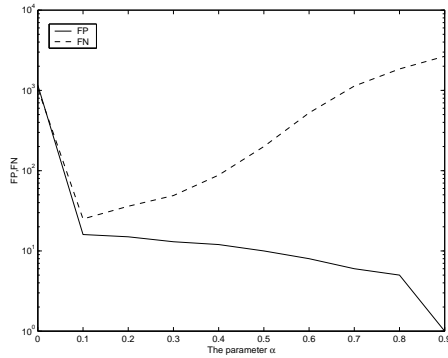


Figure 1: How the number of FP/FN estimations depends on  $\alpha$  for the proposed algorithm.

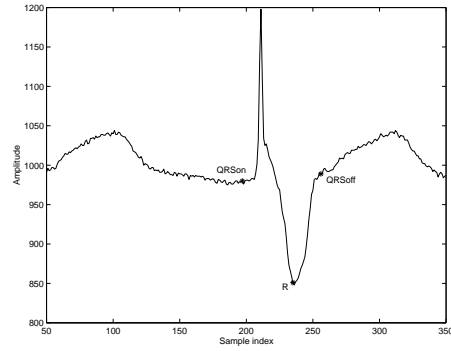


Figure 2: An example of abnormal beat from the file sel104.dat, channel 1, QT database.

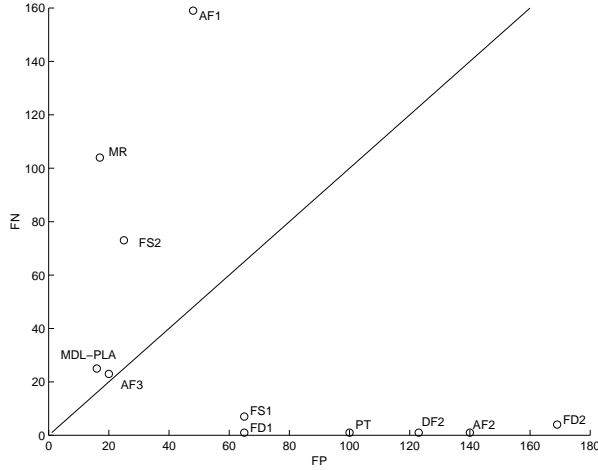


Figure 3: Comparison of False Positive/False Negative estimations for MDL-PLA algorithm and other 11 QRS detection algorithms

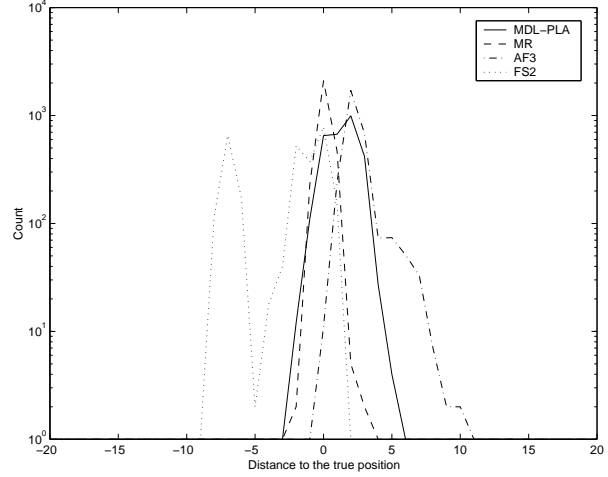


Figure 4: The histogram of distances from estimated location of R-wave to the true position. The content of counters is represented on logarithmic scale.

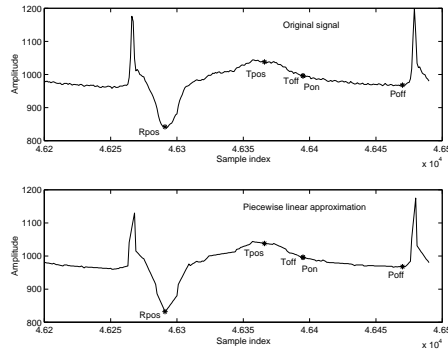


Figure 5: The beat of file sel104.dat, channel 1, for which  $WDD_{PT}$  has the smallest value.

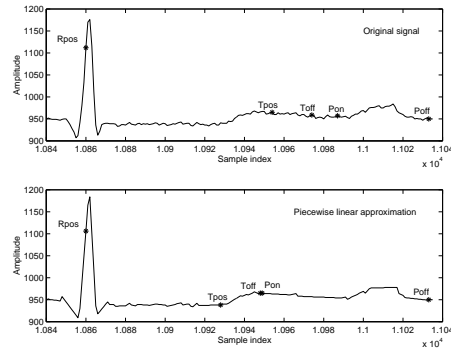


Figure 6: The beat of file sel100.dat, channel 1, for which  $WDD_{PT}$  has the largest value.