STANFORD ARTIFICIAL INTELLIGENCE LABORATORY
MEMO AIM- 207

STAN-CS- 73-378

THE OPTIMUM COMB METHOD OF PITCH PERIOD ANALYSIS OF CONTINUOUS DIGITIZED SPEECH

BY

JAMES ANDERSON MOORER

SUPPORTED BY

ADVANCED RESEARCH PROJECTS AGENCY

ARPA ORDER NO. 457

JULY 1973

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

STANFORD ARTIFICIAL INTELLIGENCE LABORATORY JULY 14, 1973
MEMO AIM-207

COMPUTER SCIENCE DEPARTMENT
REPORT CS-378

THE OPTIMUM COMB METHOD OF PITCH PERIOD ANALYSIS OF CONTINUOUS DIGITIZED SPEECH

by

James Anderson Moorer

ABSTRACT: A new method of tracking the fundamental frequency of voiced speech is described. The method is shown to be of similar accuracy as the Cepstrum technique. Since the method involves only additions, no multiplication, it is shown to be faster than the SIFT algorithm.

This research was supported in part by the Advanced Research Projects Agency of the Office of the Secretary of Defense under Contract No. SD-183.

LIST OF FIGURES

lower plot is the cepstrum of this segment. There were 512 input data points in this cepstrum.

Figure 8 - Equation (4) when evaluated using the speech waveform in figure 7 shows an obvious minimum. There are, of course, examples of the reverse case, where the cepstrum gives clean results and equation (4) does not.

Figure 9 - Comparison of the optimum comb method with the cepstrum technique. The circled points in the lower plot are from the cepstrum. The upper plot shows the speech waveform that was used as test data.

INTRODUCTION

The determination of the fundamental pitch period of voiced human speech is an important part of machine perception of speech. The non-trivial nature of the problem may be is reflected by the number of quite complex methods which have been reported [1-5]. It would seem that the most popular method is the Cepstrum technique [3]. This method uses two discrete fourier transforms, followed by a search for a significant maximum. The computational complexity of the Cepstrum technique thus is proportional to N*logN where N is the number of points in the window in question. The method to be presented here shows similar results to the Cepstrum technique but demonstrates a computational complexity proportional to N.

The core of the method is the comb filter. By way of review, the comb filter is defined by the recurrence relation

$$Yn \leftarrow Xn - Xn\text{-}m \qquad (1)$$

Where X is a discrete input sequence representing the input waveform sampled at time nT, Y is the output sequence, and m is a constant defining the characteristics of the filter. The magnitude-frequency response of the comb filter is

$$\sqrt{\sin^2(m\omega T) + [1-\cos(m\omega T)]^2} \qquad (2)$$

The comb filter has a zero of transmission at frequencies which are integral multiples of 1/mT Hertz. Thus, if the input waveform is a stationary signal consisting of nothing but frequencies which are multiples of 1/mT Hertz, the steady-state output of the filter will

be identically zero. This is the basis of the method.

THE METHOD

Basically, the method consists of taking a small window and determining the comb filter which when applied to the input sequence represented by this window produces the minimum output in a least squares sense. We seek to minimize the function

$$\sum_{i=0}^{k-1} (Xn+i - Xn+i-m)^2 \qquad (3)$$

With respect to m. The value of m which minimizes this function will be taken to be the pitch period.

The minimum is not unique. For a stationary input sequence, any integral multiple of m will also produce a minimum. We thus will accept only the largest value of m within a certain range. It is only necessary to search through the range of pitches represented by the human voice.

Since the function defined by (3) is not strictly unimodal, there is no simple technique for effecting the search besides trial and error, however there are several facts which tend to make the search more efficient. First, one does not need to take the sum of the squares of the differences as shown in equation (3). The absolute value is a perfectly acceptable distance function with much less computation than the square. The function to be minimized is then

$$\sum_{i=0}^{k-1} |Xn+i - Xn+i-m| \qquad (4)$$

The second simplifying fact is that the summation need only extend over one period of the input waveform. Since this period is not known at the time the summation is done, the period of the previous waveform may be used. The third simplifying fact is that the period does not change greatly from one period to the next, thus the search may begin with the last pitch period value found and proceed outward from there. Lastly, since the frequency resolution is much greater for the low frequency end of the scale (0.35 Hz at 70 Hz for 20Kc sampling rate), it is not necessary to compute (4) for all possible choices of m, but only for those values which provide sufficient frequency resolution. If we insist on a 1 Hz frequency resolution, we achieve a factor of three reduction in the number of values of m to be searched. By way of example, the expected number of summations per window was computed. A speech sample digitized at 20Kc was processed. The search was conducted over the frequency range 70 Hz (286 points) to 225 Hz (89 points). If the entire frequency range was searched at each window, one would expect 197 summations to be computed. Instead, only an average of 39.4 summations were computed at each window.

SOME EXAMPLES

Figure 1 shows a 26 millisecond segment of speech and the value of equation (4) computed for all values of m between 70 Hz and 225 Hz. We see a definite strong minimum at just over 100 Hz, and two other smaller minima, one at about 80 Hz and the other at about 190 Hz. This is a typical plot, comprising about 80% of the cases. The other 20% are like figures 2 and 3. In figure 2, the fundamental frequency is still the deepest minimum, but in figure 3, this is the case only by a slight margin. Sometimes (less than 2.5% of the cases) the deepest minimum is not related to the fundamental frequency. In these pathalogical cases, there is always a minimum at the fundamental frequency and it is always very close to the deepest minimum. Contextual information can easily be used to make the proper decision. In the author's program, the average of the pitch of the most recent periods is computed. When the situation becomes totally ambiguous, the proximity to the average is used to make the final decision. This simple heuristic seems to solve the problem adequately.

Figures 4, 5, and 6 show actual speech waveforms with plots of the pitch computed by the author's program. Figure 4 shows that the method is somewhat sensitive to gross changes in the waveform. Figure 5 demonstrates that the pitch is successfully tracked when the waveform changes slowly. Figure 6 shows the behavior of the pitch tracking as the speech goes from voiced to unvoiced.

ON COMPUTATIONAL COMPLEXITY

Let us assume a 10Kc sampling rate for purposes of computing the number of arithmetic operations. At this rate, an average of 22.4 evaluations of equation (4) are done at each point a pitch estimate is desired. If we conjecture that the average pitch is about 150 Hz, then approximately 66 points are in the summation. This means that at each point approximately 3000 arithmetic operations are done, all integer additions or subtractions. Markel [5] calculated that the SIFT algorithm required 1750 multiplies and 1625 additions to compute the pitch estimate. Clearly then, the optimum comb provides a computational advantage over the SIFT algorithm, although only by a narrow margin with the additional disadvantage that the optimum comb method does not readily yield the voiced-unvoiced decision. Markel also estimated that the Cepstrum method as described by Schafer and Rabiner [6] requires at least 20000 multiplications and 30000 additions to produce similar results, although it is not clear that a smaller cepstrum would not suffice. Schafer and Rabiner used a 1024-point FFT.

If the speech is digitized in 12 binary bits, it is clear that equation (4) could be computed on a machine with a 16-bit word length. Some scaling of the partial sums is required, but 16 bits is more than enough accuracy to assure usable results.

COMPARISON WITH THE CEPSTRUM

Neither the cepstrum nor the optimum comb method of pitch periodanalysisis 100% accurate. Some pathological conditions the optimum comb method exhibts were shown in figures 2 and 3. In figure 7, we see such a quirk for the cepstrum and in figure 8, it is shown that equation (4) does not exhibit such a quirk on this particular waveform. In figure 7, we see that the highest peak is not necessarily a good estimate of the pitch period, nor is the next highest. This shows that when the two methods fail, they seem to fail under different circumstances.

Figure 9 shows a speech waveform and a plot of the output of the optimum-comb pitch detector and the cepstrum pitch detector. One can see that except for occaisional gross errors by the cepstrum, the pitch estimates agree quite closely.

CONCLUSIONS

The optimum comb technique is a fast and useful technique for the extraction of pitch period data from continuous speech. The method is similar in accuracy to the cepstrum and is somewhat faster than the SIFT algorithm. It is certainly deserving of further study.

REFERENCES

[1] C.M. Harris, M.R. Weiss, "Pitch Extraction by Computer Processing of High Resolution Fourier Analysis Data," J. Acoust. Soc. Amer., vol 35, p339, 1963.

[2] Bernard Gold, "Computer Program for Pitch Extraction," J. Acoust. Soc. Amer., vol 34, p916, 1962.

[3] A. Michael Noll, "Cepstrum Pitch Determination," J. Acoust. Soc. Amer., vol 41, #2, p293, 1967.

[4] Bernard Gold, Lawrence R. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," J. Acoust. Soc. Amer, vol 46, #2, p442, 1969.

[5] J.D. Markel, "The SIFT Algorithm for Fundamental Frequency Extimation," IEEE Trans. on Audio and Electroacoustics, Vol AU-20, #5, pp367-377, December 1972

[6] Ronald W. Schafer, Lawrence R. Rabiner, "System for Automatic Formant Analysis of Voiced Speech," J. Acoust. Soc. Amer., vol 47, #2, p634, 1970.

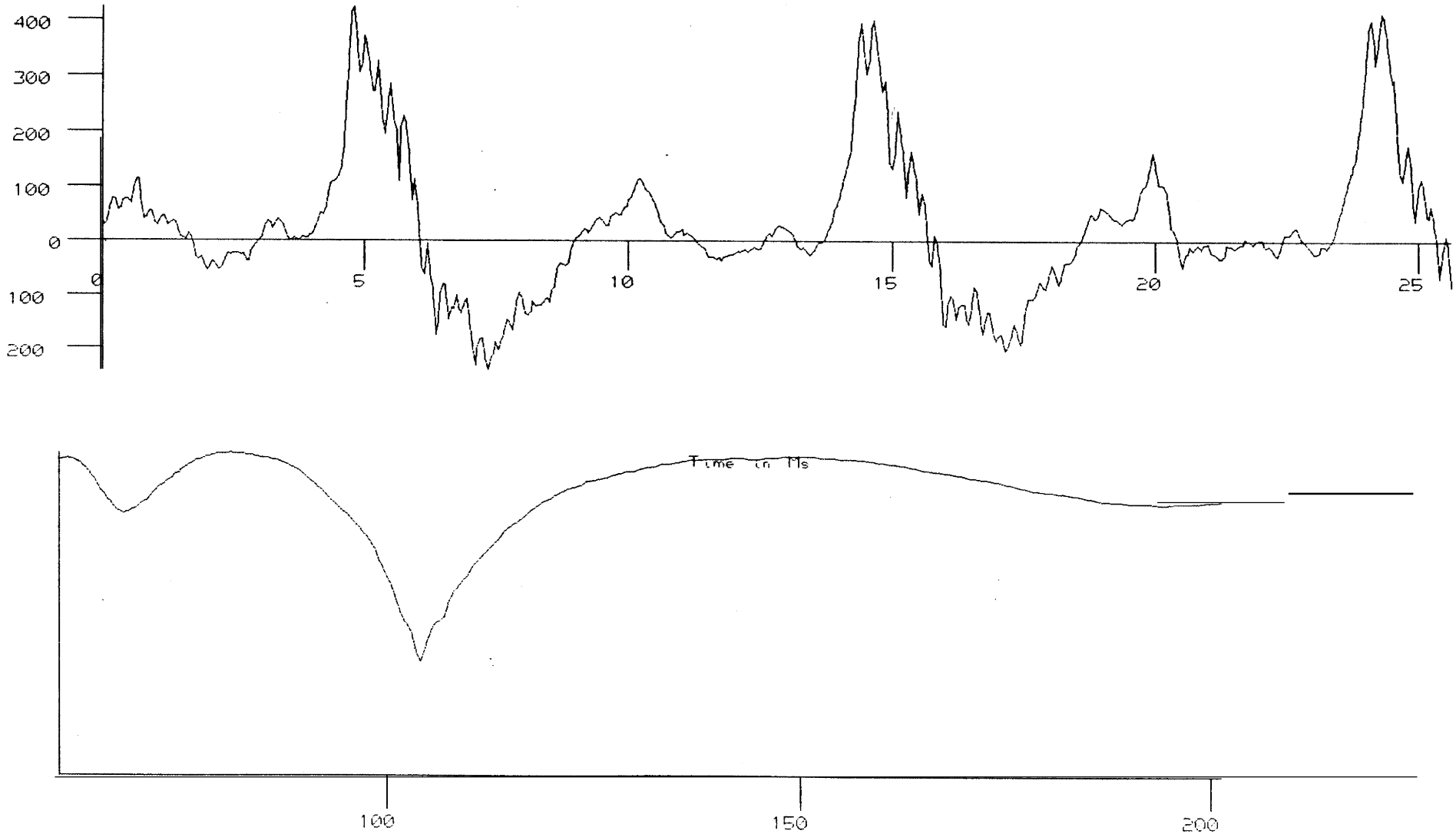

Figure 1

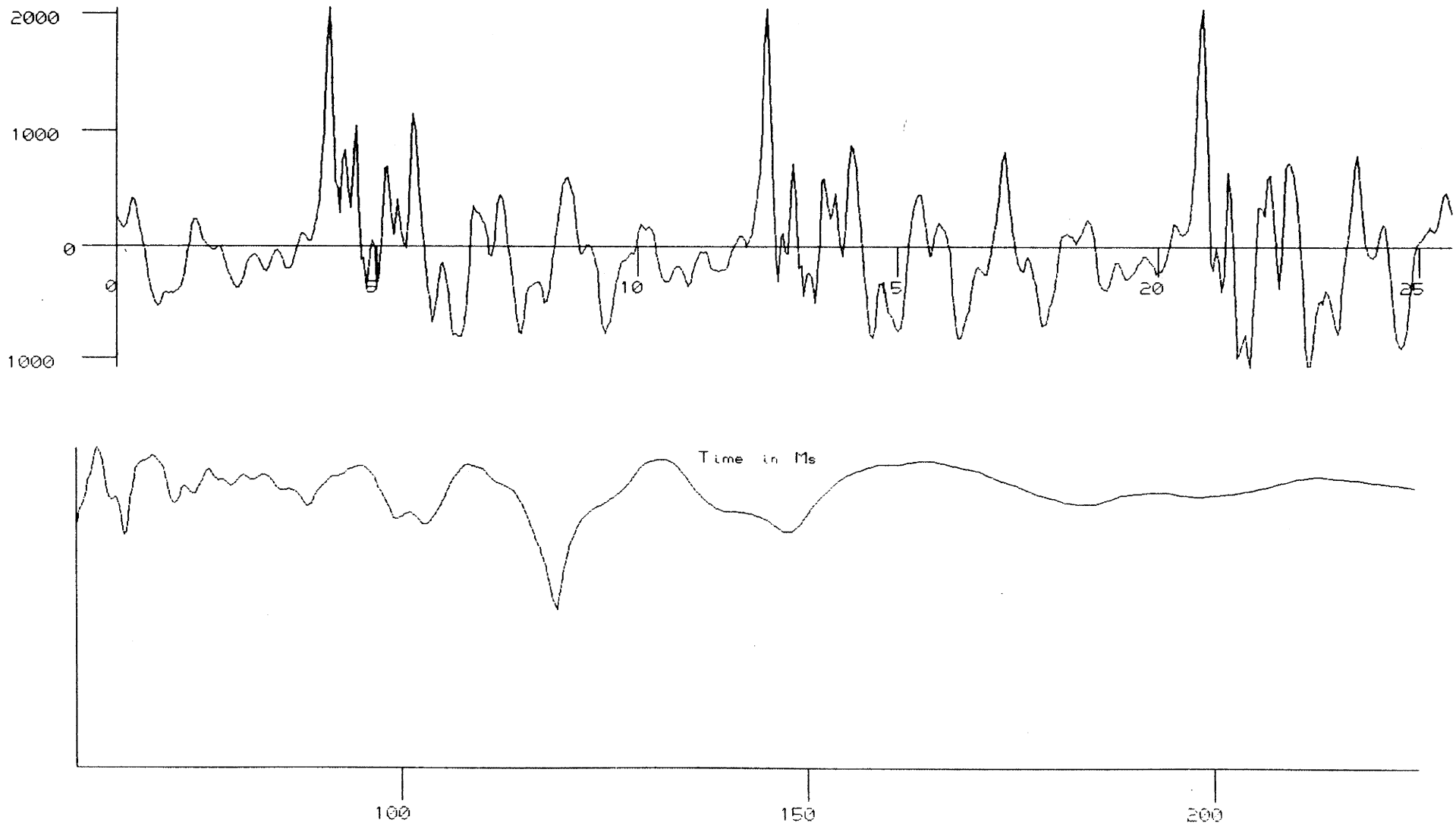

Figure 2

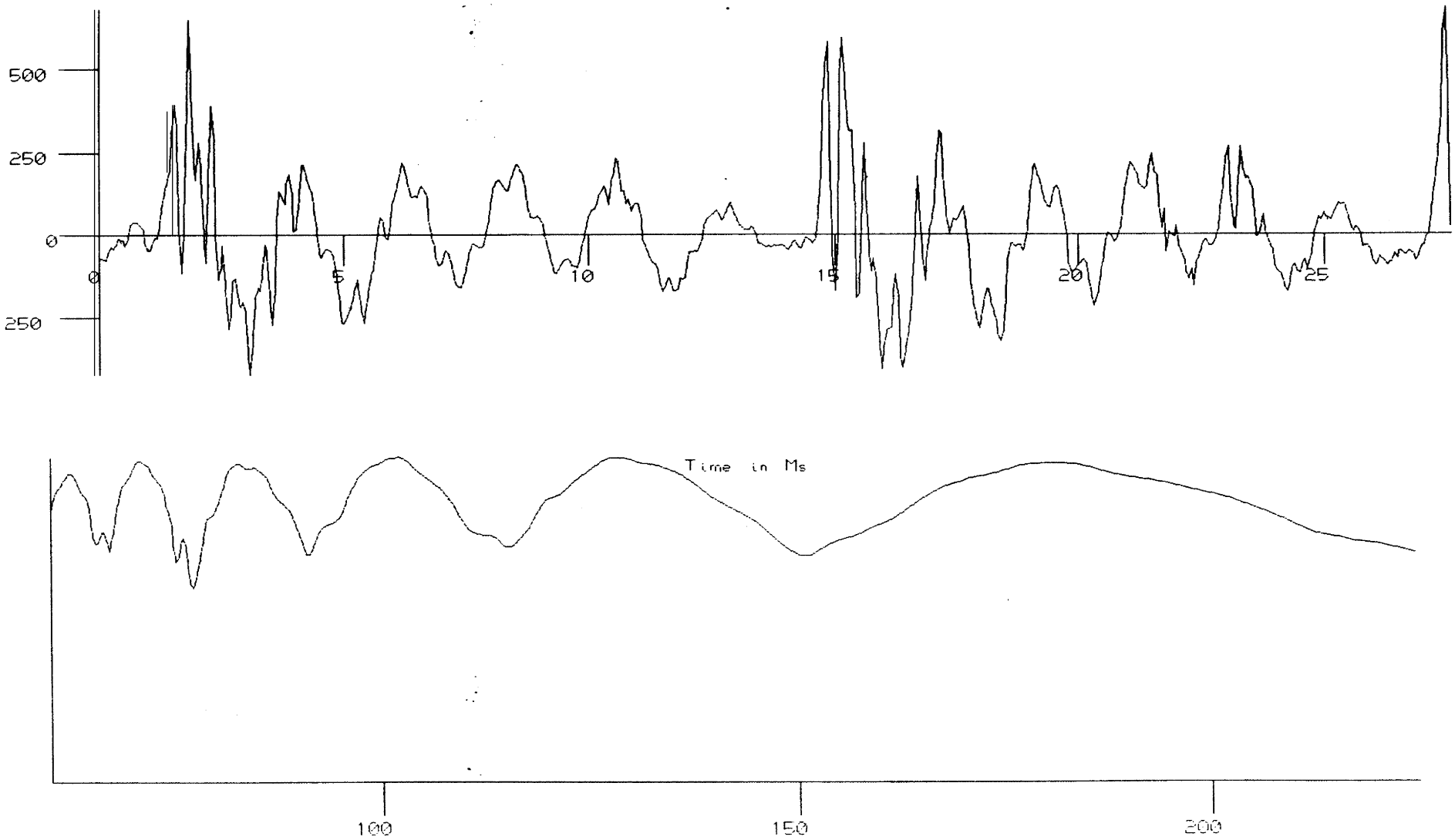

Figure 3

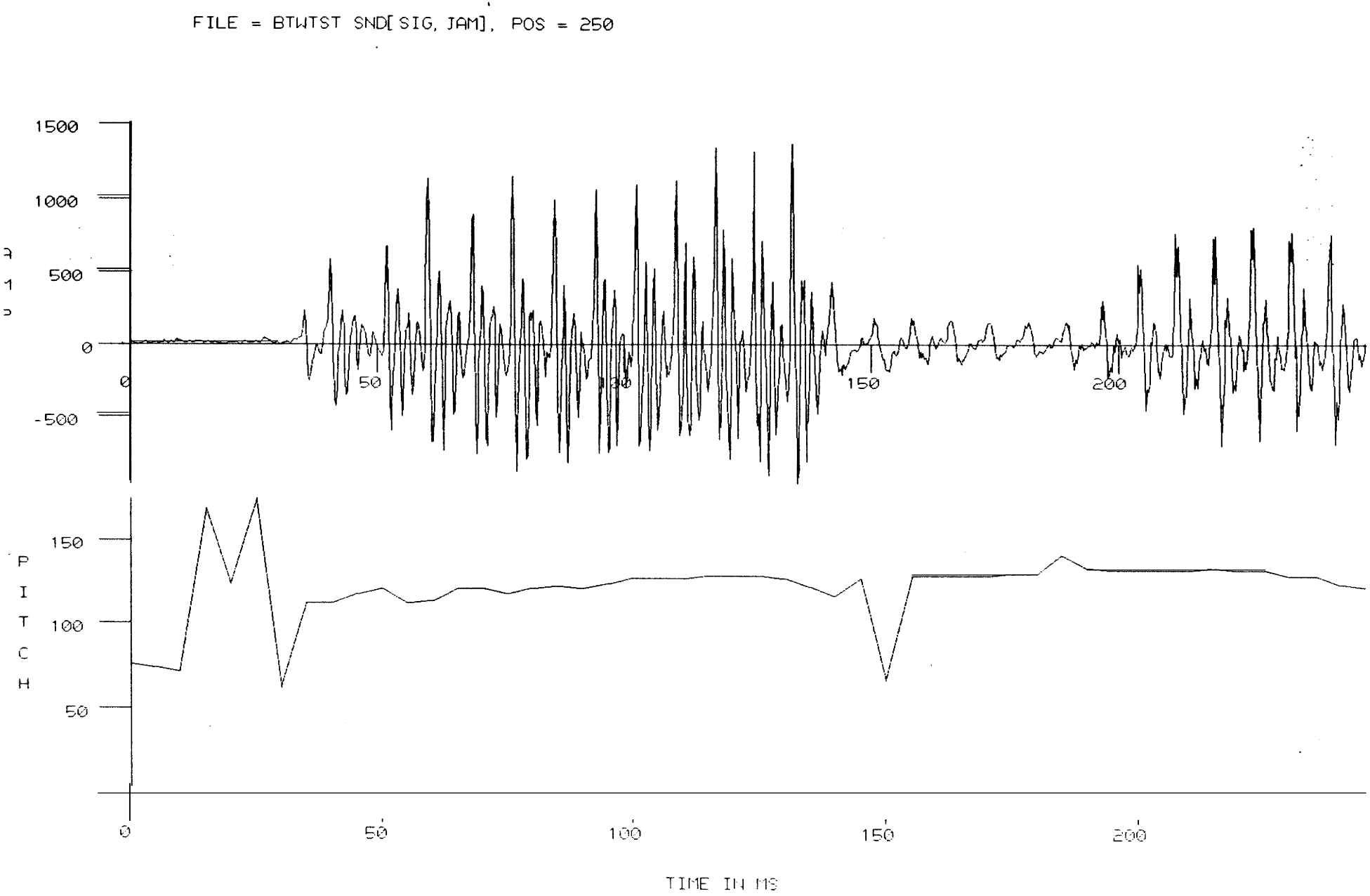
FILE = BTWTST SND[SIG, JAM], POS = 250
1500
1000
500
0
-500
0
50
100
150
200
PITCH
150
100
50
0
50
100
150
200
TIME IN MS

Figure 4

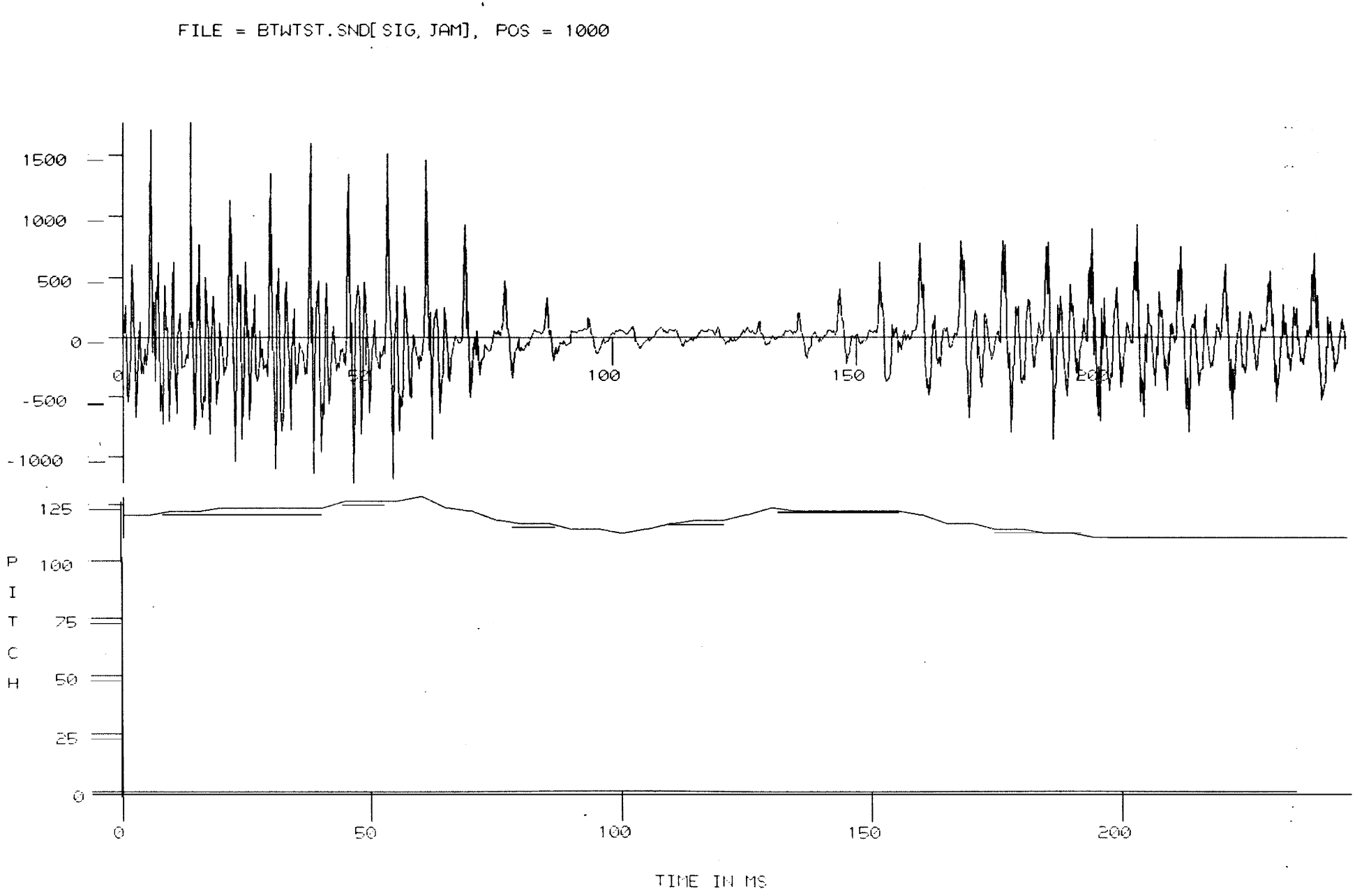
FILE = BTWTST.SND[SIG,JAM], POS = 1000
1500
1000
500
0
-500
-1000
0
50
100
150
200
125
100
75
50
25
0
PITCH
0
50
100
150
200
TIME IN MS

Figure 5

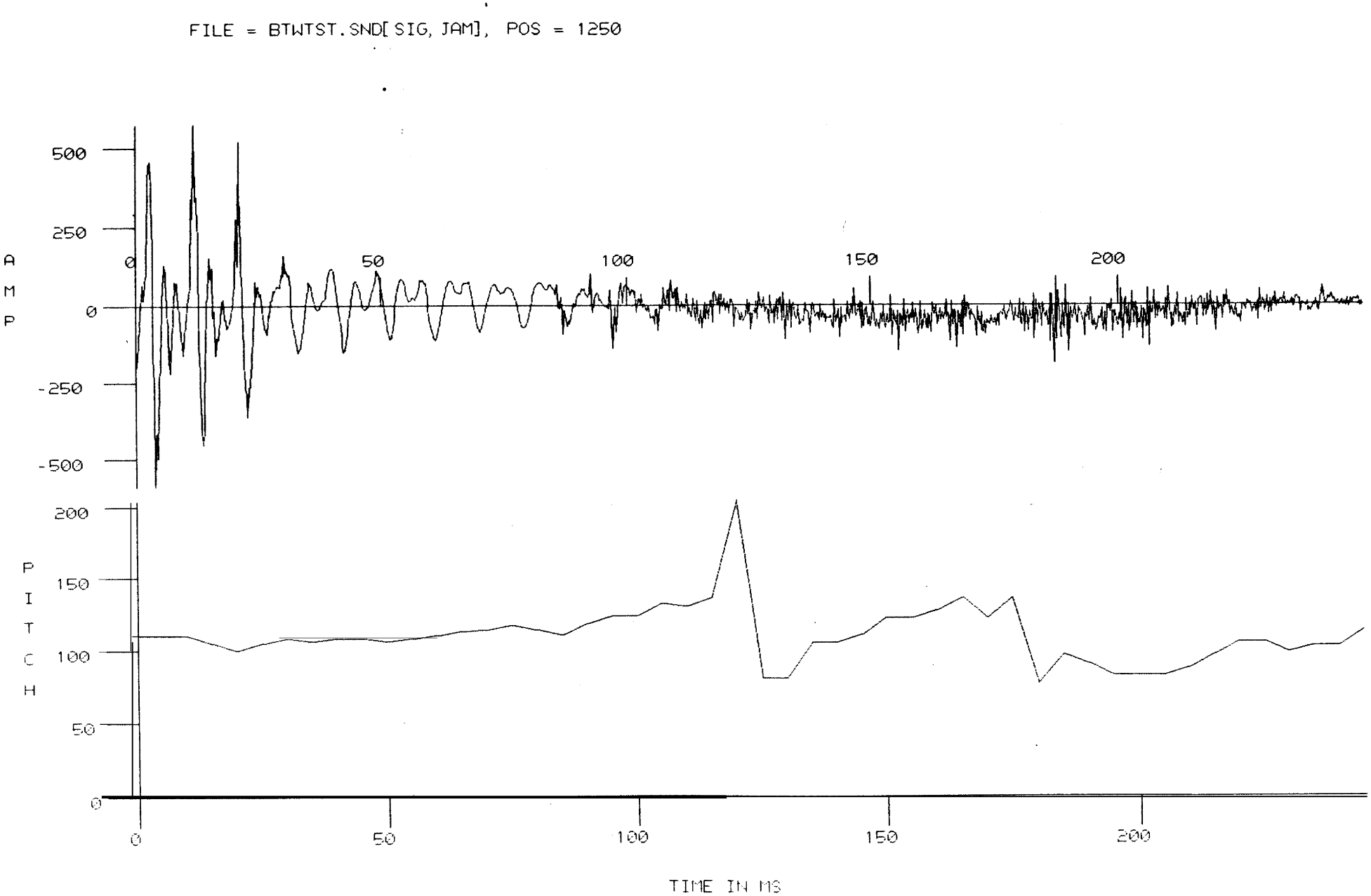
FILE = BTWTST.SND[SIG,JAM], POS = 1250
AMP
500
250
0
-250
-500
0
50
100
150
200
PITCH
200
150
100
50
0
0
50
100
150
200
TIME IN MS

Figure 6

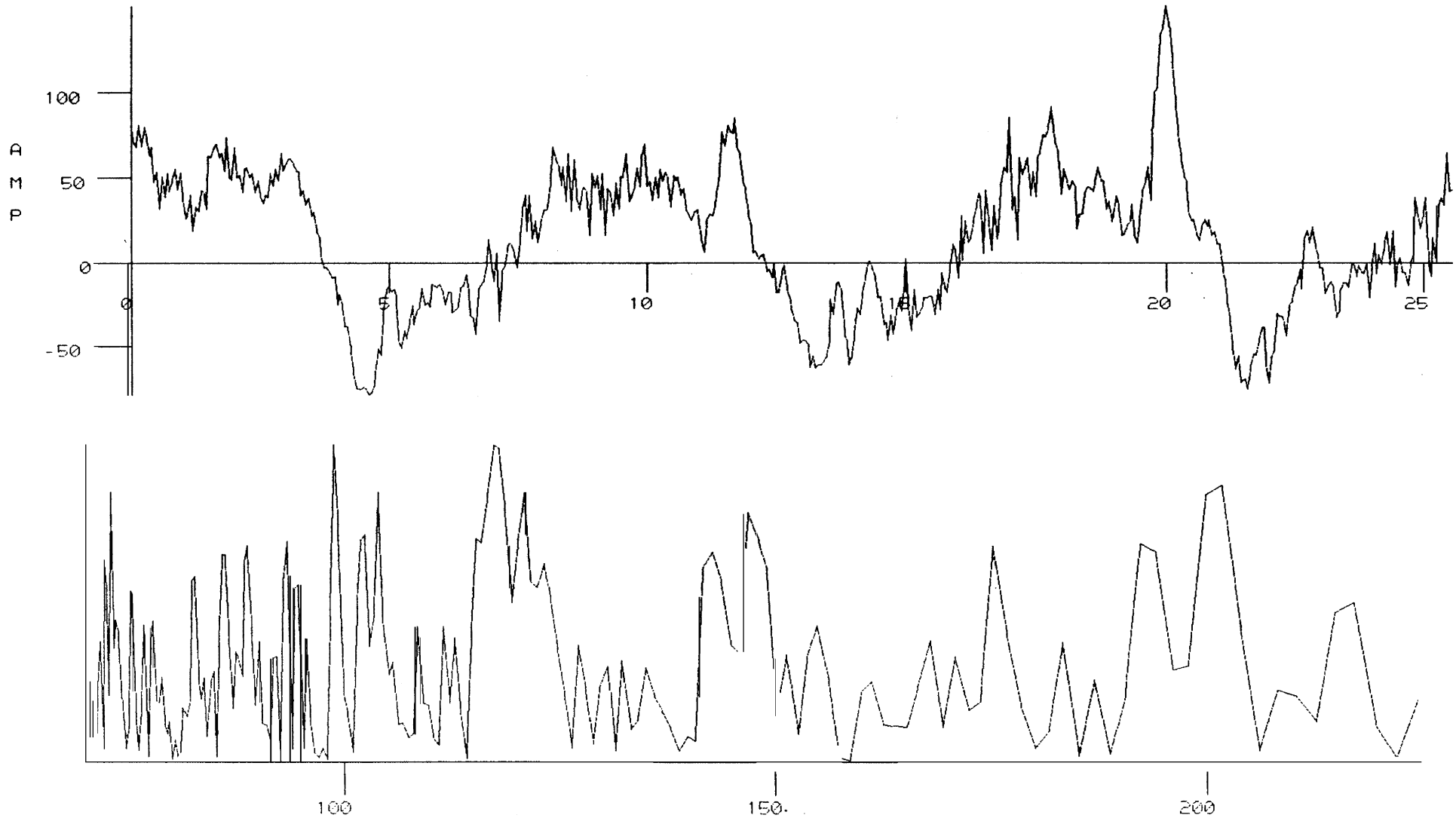

Figure 7

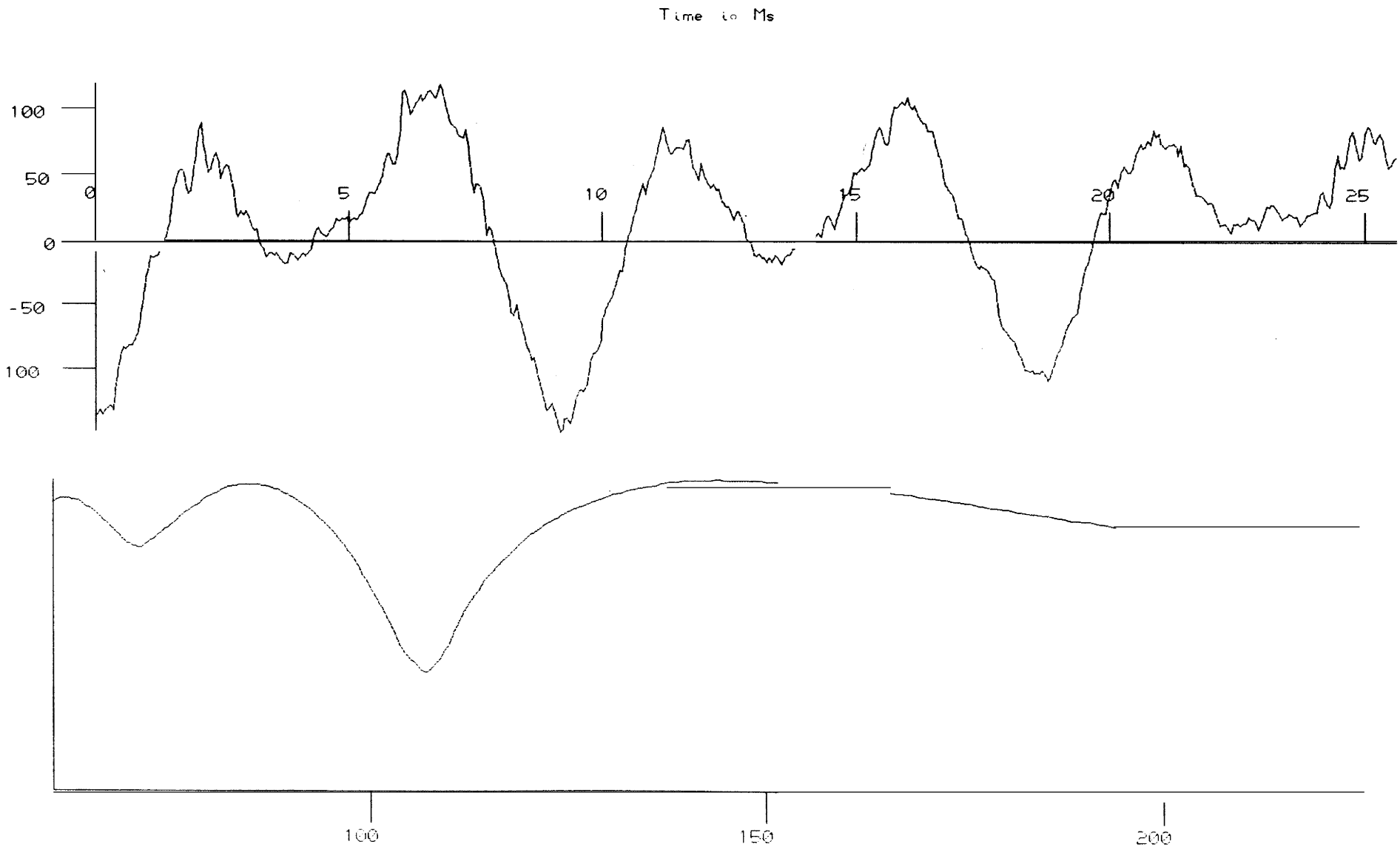

Figure 8

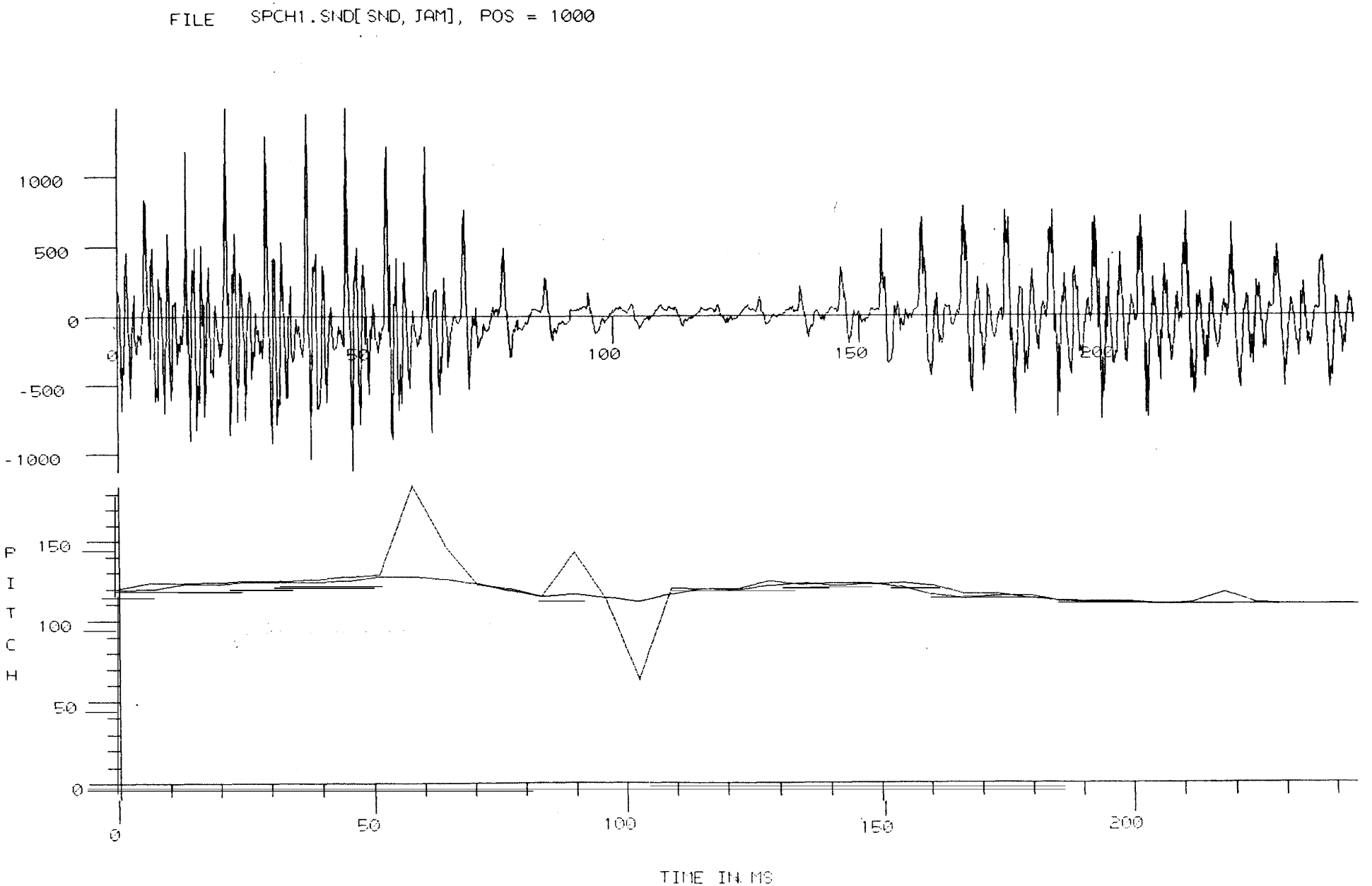
FILE SPCH1.SND[SND, JAM], POS = 1000
1000
500
0
-500
-1000
0
50
100
150
200
PITCH
150
100
50
0
0
50
100
150
200
TIME IN MS

Figure 9