

AD 634162

NRL Memorandum Report 1699

Pitch Variations in Vocoder Voice

W. M. JEWETT

Communication Branch
Radio Division

CLEARINGHOUSE FOR FEDERAL SCIENTIFIC AND TECHNICAL INFORMATION		
Hardcopy	Microfiche	
\$2.00	\$1.50	3.30 ad
ARCHE COPY		

May 20, 1966



D D C
RECEIVED
JUN 24 1966
RECEIVED
C

U.S. NAVAL RESEARCH LABORATORY
Washington, D.C.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

CONTENTS

Abstract	ii
Problem Status	ii
Authorization	ii
INTRODUCTION	1
PITCH VARIATIONS IN NORMAL SPEECH	1
KY-537/U VOCODER VOICING CIRCUITRY AND PITCH EXTRACTOR	2
EXPERIMENTAL PROCEDURE	3
RESULTS	4
SUMMARY	8
ACKNOWLEDGEMENTS	8
REFERENCES	9

ABSTRACT

In the analysis of the KY-537/U channel vocoder for the Bureau of Ships, a study was performed of the pitch variations in the vocoded digital data signal using three untrained male speakers. The results obtained have been compared to published data for the pitch perturbations in natural speech for trained speakers. The comparison indicates that statistically the variations in pitch between successive voiced vocoder frames were zero a much larger percent of the time than found in natural speech, and also that large changes in pitch occurred more often than in the reference data. This is partly the result of using a quantization interval in the vocoder that is too large to permit encoding the small pitch variations which account for most of the pitch perturbations that occur in natural speech. The extension of this work using a larger number of speakers under different stress and background noise conditions would yield data of value on the adequacy of the pitch extractor in presently designed channel vocoders, the need for speaker training, and possible constraints on vocoder use.

PROBLEM STATUS

This is the final report on the analysis of the KY-537/U Vocoder; work on other phases of the problem is continuing.

AUTHORIZATION

NRL Problem Number 54R01-19
BUSHIPS Subproject SF 006-11-01 Task 7260

PITCH VARIATIONS IN VOCODED VOICE

INTRODUCTION

The U. S. Naval Research Laboratory (NRL) has been engaged in a general problem for the Bureau of Ships (BUSHIPS) aimed at improvement of the capability of handling digital data at rates up to 2400 bits/second for transmission on hf radio. One of the requirements is for the transmission of encrypted voice signals. Recent work (a) in this area included a Laboratory evaluation of a late model, 16 channel, pitch-excited vocoder (KY-537/U). In addition to the work reported on in reference (a), a brief study was performed of the pitch variations in the vocoded digital data signal.

This study was prompted by the work of P. Lieberman (b), (c) of the Air Force Cambridge Research Laboratory on pitch perturbations in normal speech, and by the work of Tierney and Gold (d), (e) of Lincoln Laboratory. This latter work showed that vocoded synthesized speech could be improved if an accurate measure were made of voiced/unvoiced condition, accurate pitch extractors were used, and a technique called spectral flattening were used in the synthesizer to reduce amplitude distortion at the channel modulators.

The study undertaken in this experiment was to obtain a statistical expression for the variation in pitch between successive voiced vocoder frames. The results were compared to published data of the pitch variations occurring in normal speech.

PITCH VARIATIONS IN NORMAL SPEECH

Some of the concepts used in the design of pitch-excited vocoders are (1) the voice pitch varies slowly enough that a sampling rate of 40 to 50 samples/second may be used, (2) the minimum frequency quantization interval may be three to five cps, and (3) speech is either voiced or unvoiced. Lately, these concepts have been changing as evidence has been obtained that performance of vocoders may be improved with better pitch indication. In reference (b), Lieberman studied the pitch of normal speech and concluded that the majority of the variations in pitch are small perturbations which are essential for the naturalness of speech, and that statistically the variations differ at the onset and end of voicing and for different emotional modes. In 86 percent of the periods examined, the pitch was not steady over a sample time equal to three cycles of the pitch frequency. The change in pitch period ($\Delta T = T_n - T_{n-1}$) was greater than 0.6 milliseconds 20 percent of the time and greater than 1.0 milliseconds 15 percent of the time. The durations of the pitch periods were measured with a resolution of 200 microseconds. Later work by Lieberman using 50 micro-

seconds resolution resulted in smaller perturbation values. For six male speakers with similar median pitch frequencies, the perturbations were ≥ 0.5 milliseconds 15 percent of the time. In this work one of the conclusions was that perturbations ≥ 0.5 milliseconds reflected variations in the vocal cord vibratory cycle. The term perturbation factor (PF) was given to the measurement of the percent of time that a speaker's pitch variations were ≥ 0.5 milliseconds.

The degree to which a vocoder can reflect the small perturbations less than 0.5 milliseconds, which account for 85 percent of the pitch variations, depends on the accuracy of the pitch extractor and the frequency quantization interval used. For example, in order to have a resolution of 0.5 milliseconds at a pitch frequency of 70 cps the frequency quantization interval would have to be not more than 2.5 cps

$$\left(\text{i.e. } \frac{1}{70} - \frac{1}{72.5} = 0.49 \times 10^{-3} \text{ sec}\right).$$

Some of the recordings used in reference (c) were used also by Shaffer (f) in his study of the information rate necessary to transmit pitch period durations for connected speech. One of the conclusions reached in this work was that the average range of variations of the pitch period duration within a voiced portion was 1.5 milliseconds. Also, that the voiced portions having from 6 to 20 pitch periods were all approximately equally likely and constituted about 50 percent of the sample population. The shortest voiced portions had two pitch periods and the largest 95 pitch periods. In this study by Shaffer all portions of speech were classed as either voiced or unvoiced. In the development of computer techniques for the extraction of speech parameters, Weiss and Harris (g) formulated rules to identify the portions of speech which contained both voiced and unvoiced energy. Information was not given as to the percent of time such a condition may be found to exist. As yet, no pitch-excited vocoders permit recognition or synthesis of speech containing both voiced and unvoiced energy. The pitch extraction and voicing circuitry of the KY-537/U vocoder is essentially the same as that of the older HY-2 vocoder, differing principally in the encoding method.

KY-537/U VOCODER VOICING CIRCUITRY AND PITCH EXTRACTOR

In the 2400 bits/second mode of operation, one frame of data in the KY-537/U vocoder consists of 54 bits repeated at 44.44 frames/second. Six bits of each frame are devoted to the transmission of a binary number of one to sixty-three representing the pitch frequency during a voiced condition. An unvoiced condition inhibits the pitch encoding function resulting in the transmission of six bits with logic values of zero. The voicing indicating circuitry determines that the speech sound is voiced if the energy in the frequency

range of 200 to 800 cps is greater than the energy above 4400 cps. A voiced condition must exist for at least 20 milliseconds before a voiced condition is recognized. This is equivalent to two pitch periods at 100 cps.

The pitch extractor circuitry consists of a 350 cps low pass filter and a variable frequency filter (tracking filter) followed by a waveform shaping circuit, a constant current generator, and a low pass filter. The output of the low pass filter is a slowly varying dc voltage proportional to the basic voice pitch of the talker. This dc signal is encoded into a binary number for transmission. Each number represents one of sixty-three frequencies spaced approximately 3.6 cps apart between 74 and 300 cps. (The accuracy of the pitch extractor and encoder was verified with a steady state signal.) The pitch analysis and encoding takes place nine times during each frame period, but transmission occurs only once per frame. In the 1200 bits/second mode, the frame rate is still 44.44 frames/second, but only four pitch sampling periods occur during a frame period.

EXPERIMENTAL PROCEDURE

The basic procedure used was to select, out of the serial stream of data from the transmitting vocoder, the six bits in every frame representing pitch frequency and to read these into a small digital computer (G-15D) 24 bits at a time. Each 24 bits form one computer word. The selection and reading was under the control of a computer program which permitted storage of pitch data from a maximum of 3200 successive frames. This is equivalent to 112 seconds of data. A simplified block diagram of the system is shown in Figure 1. A framing pulse from the vocoder receiver, preceding the first bit of pitch data in every frame, was used to recognize the pitch data in the serial stream. These pitch bits were clocked into a 24-bit shift register six bits at a time by the vocoder clock. After four frame periods, the data in the shift register were read into the computer by the computer's 100 kcps clock.

The audio input to the vocoder consisted of three male speakers reading 100 short questions such as follow:

What letter comes between A and C?

Are moths dangerous to clothing?

How many states are in the United States?

These questions were part of the speech material that was used to assess the voice communication efficiency of the operational Soft Talk System. The three speakers used in the present experiment were not trained talkers.

Computer programs were prepared to operate on the stored data to provide the following outputs:

1. A printed record of the stored data in decimal numbers (P) of zero to 63
2. A distribution of the percent of time the pitch number corresponded to each of the possible values of 1 through 63
3. A distribution of the percent of time the pitch number corresponded to each of the 63 possibilities at the onset of voicing
4. A distribution of the percent of time the pitch number corresponded to each of the 63 possibilities at the end of voicing; when a voiced period was equal to a single frame period, the pitch frequency was tabulated in both the onset of voicing and the end of voicing distributions
5. A distribution of the percent of time the change in pitch number ($\Delta P = P_n - P_{n-1}$) between two successive voiced frames corresponded to each of the possible values of zero through 62
6. A distribution of the percent of time the change in pitch number between the first two voiced frames at the onset of voicing corresponded to each of the possible values of zero through 62
7. A distribution of the percent of time the change in pitch number between the last two voiced frames at the end of voicing corresponded to each of the possible values of zero through 62, by definition, a voiced period equal to a single frame period was not included in any of the ΔP distributions
8. A distribution of the percent of time the change in period ($\Delta T = T_n - T_{n-1}$) of the pitch signal between two successive voiced frames occurred in each of 92 ranges between zero and 5.6 milliseconds.

RESULTS

Information regarding the pitch frequency of a voiced signal is encoded as a single frequency for each frame period of 22.5 milliseconds. Figure 2 is a plot of the percent of time the pitch frequency of three speakers was

encoded as the pitch number P. The encoded pitch in cps is approximately equal to $70 + 3.6 P$. (The lines connecting the points of data in all figures involving P and ΔP are for the convenience of the reader and do not imply a continuum of data.) The encoded pitch data for speakers ONE and TWO were very similar. In each case, approximately 50 percent of the data was concentrated within the range $P = 11$ to $P = 15$, or from approximately 110 cps to 124 cps. The peak at $P = 15$ was six to ten percentage points higher than either of the other two peaks within this range. The pitch frequency for speaker number THREE was much lower than the other speakers. Approximately 67 percent of the data was within the range of $P = 3$ to $P = 7$, or from approximately 81 cps to 96 cps. Again the data contained three peaks, but they were of nearly uniform amplitude. The reasons for the three peaks are not known.

It should be pointed out that when the audio input to the vocoder is from a magnetic tape loop, repeated encoding of the same recorded word or message does not always yield identical encoded pitch patterns. This could be because the vocoder pitch sampling periods may occur at different points within the message each time it is replayed. If the vocoder were available for further work, it would be of value to determine the degree of variation in the pitch patterns for repeated transmission of a recorded message. Too large a variation between pitch patterns could be an indication either that the sampling rate is too low, or that the pitch extractor has time constant problems. If the data is studied at the pitch encoder, the sampling rate is 400 samples/second rather than the 44.44 samples/second obtained when the encoded pitch is extracted from the serial data stream.

Figures 3 and 4 show the percent of time the pitch frequency of the three speakers was encoded as the pitch number P at the onset and end of voicing, respectively. The peak occurring at $P = 15$ for either speaker ONE or speaker TWO was very prominent at the onset of voicing (Figure 3) and much reduced at the end of voicing (Figure 4). This indicates the tendency to begin voicing at a particular pitch with less constraint at the end of voicing. Speaker TWO used a larger range of pitch at the onset and end of voicing more often than did the other speakers.

Figure 5 shows the pitch patterns for three single syllable words spoken by talker number TWO. For the sample words chosen, the word SCOUT resulted in a pitch pattern where the pitch was gradually decreasing from $P = 17$ to $P = 4$ in eleven frame periods. The pitch pattern for the word ACT was U-shaped lasting for 12 frame periods and having pitch numbers between 10 and 15. The third word,

DWARF, had a pattern where the pitch fluctuated at the beginning of the word and was steady for a large part of the time within the word. A gradual increase in pitch from $P = 6$ to $P = 11$ in 18 frames was the general trend in the pattern. These three pitch patterns illustrate the different shaped pitch patterns obtained. Also, they show the range over which the pitch varied within a given word. The median ΔT for the three words was 0.3, 0.3, and 0.6 milliseconds, respectively.

The three talkers used in this experiment were not trained speakers. Use of recordings of typical users of vocoders under different stress and background noise conditions would yield data of value on the adequacy of the pitch extractor, need for speaker training, and possible constraints on vocoder use.

The data in Figures 2, 3, and 4 have been replotted in Figures 6, 7, and 8 to consolidate the data for each speaker.

Figure 9 is a plot of the percent of time that the change in pitch frequency between two adjacent voiced frames resulted in a shift in the encoded pitch numbers equal to $\Delta P = P_n - P_{n-1}$. The shift in cps is approximately equal to 3.6 (ΔP). For speakers ONE, TWO, and THREE respectively, ΔP was equal to zero for 16.5 percent, 31.5 percent, and 45 percent of the time. The range of 16.5 percent to 45 percent for the measured percent of time that ΔP was equal to zero for three speakers is a wide spread and is in conflict with Lieberman's data for normal speech. Lieberman found that in only 14 percent of the periods examined that the pitch was steady over a sample time equal to three cycles of the pitch frequency. This was not dependent on the speaker's fundamental pitch frequency. The quantization interval was 200 microseconds. The use of a constant frequency quantization interval in the vocoder makes a direct comparison with Lieberman's data difficult. Table I shows the results of a comparison based on the median pitch period. Speaker ONE had a median pitch frequency of approximately 117 cps. Thus, the measurement period of one frame time was equivalent to approximately 2.7 pitch periods at the median frequency. This is compared to a measurement period of three pitch periods in the reference work for natural speech. The quantization interval in the present work was 3.6 cps or approximately 260 microseconds at the median frequency. This is compared to 200 microseconds used by Lieberman. In the vocoder data for speaker ONE, no pitch variations occurred 16.5 percent of the time compared to 14 percent for normal speech. Such a close agreement as this was not obtained for the other two speakers. Although speaker TWO had a median pitch

frequency near that of speaker ONE, the data for speaker TWO shows no pitch variations between successive frames twice as often as for speaker ONE. For speaker THREE the median pitch frequency was approximately 88 cps, which gave approximately 2.0 pitch periods in the frame time. The quantization interval was equivalent to approximately 460 microseconds at the median frequency, or 2.3 times that used by Lieberman. This could be one of the reasons why the results show no pitch variations for 45 percent of the time.

Figures 10, 11, and 12 show the results of converting pitch frequencies to period (milliseconds) enabling changes in pitch to be expressed as changes in period (ΔT). The data for speakers TWO and THREE for $\Delta T > 0.6$ milliseconds are in fair agreement with the reference data obtained with a 50 microsecond quantization interval. Speaker ONE'S data for $\Delta T > 0.6$ milliseconds corresponds closer to the reference data obtained with 200 microseconds quantization intervals. The ΔT data for the vocoded speech are quantized in intervals of 200 microseconds. The quantizing of data can result in a graph that may be misleading when the data being quantized are not continuous. In this case, the ΔT data are not continuous; this is the result of shifting the pitch frequency in equal frequency increments. Therefore, a second curve is plotted in Figure 12 showing the ΔT data quantized in 100 microseconds increments for $\Delta T > 1.0$ milliseconds. This is a truer picture of the condition that existed. It shows fair agreement with the reference data at $\Delta T = 0.4$ milliseconds and clearly shows the inability to express small pitch period variations with a frequency quantization interval as large as 3.6 cps when the median pitch frequency is as low as 88 cps.

Figures 13 and 14 show the percent of time that the change in pitch number between two adjacent voiced frames equaled ΔP at the onset and end of voicing, respectively. For these conditions, variations between speakers maintained the same general relationships as were discussed in reference to Figure 9 when all ΔP data were used. The data in Figures 9, 13, and 14 are replotted for the individual speakers in Figures 15, 16, and 17. The data for speaker ONE, Figure 15, shows a larger pitch modulation than that of the other speakers. Also, it shows that there was an extremely small difference in ΔP at the onset and end of voicing from the ΔP for all events. For speaker THREE, the tendency to modulate more at the onset of voicing than at other times is clearly evident. This is not to imply that large changes in pitch between successive frames are to be desired. On the contrary, the reference data shows that while large, sudden changes in pitch do occur in natural voice signals, it is the small perturbations, less than 0.5 milliseconds, which

account for 85 percent of the variations in pitch. The ΔP data for the three conditions (all events, onset of voicing, and end of voicing) are plotted in cumulative distribution curves in Figures 18, 19, and 20. From Figure 18 it may be seen that ΔP was ≤ 5 approximately 16 percent of the time for speaker number ONE. In comparison, for the same percent of the time ΔP was ≤ 2 for speaker number THREE. Stated another way, ΔP was ≤ 2 , 16 percent of the time for speaker number THREE and 58.5 percent of the time for speaker number ONE.

SUMMARY

According to references (b) and (c), pitch variations in natural speech occur almost continuously. The instantaneous changes in pitch periods are predominantly less than 500 microseconds in magnitude and are considered essential for maintaining the naturalness of speech. The results of analyzing the pitch variations in the digital vocoded speech of three untrained male speakers at this Laboratory show that:

1. The pitch variations between successive voiced vocoder frames (22.5 millisecond intervals) were equal to or greater than 500 microseconds for 31 percent to 54 percent of the time, depending on the speaker. This was approximately 2 to 3.6 times as often as the average given in reference (c) for natural speech for trained speakers.
2. No pitch variations occurred between successive, voiced vocoder frames for 16 percent to 45 percent of the time depending on the speaker. This, also, was a larger percentage of the time than found in natural speech for trained talkers.

The extension of this work using recordings of typical users of vocoders under different stress and background noise conditions would yield data of value on the adequacy of the pitch extraction in presently designed vocoders, need for speaker training, and possible constraints on vocoder use.

ACKNOWLEDGEMENTS

The author is indebted to Mr. F. C. Kahler, and Mr. M. Y. McGown of this Laboratory for their assistance in the conduction of the experiment and the programming of the computer.

REFERENCES

1. NRL Memorandum 5410-688A:WMJ:dm of 9 Nov 1964, "KY-537/U Vocoder; Laboratory Analysis of and Comparison with TSEC/HY-2 Vocoder"
2. Phillip Lieberman, "Perturbations in Vocal Pitch", Journal of the Acoustical Society of America, Vol 33, No. 5, PP 597-603, May 1961
3. Phillip Lieberman, "Some Acoustic Measures of the Fundamental Periodicity of Normal and Pathologic Larynges", Journal of the Acoustical Society of America, Vol 35, No. 3, PP 344-353, March 1963
4. Bernard Gold, "Computer Program for Pitch Extraction", Journal of the Acoustical Society of America, Vol 34, No. 7, PP 916-921, July 1962
5. Joseph Tierney, Bernard Gold, Vincent Sferrino, J. A. Dumanian, and Everett Aho, "Channel Vocoder with Digital Pitch Extractor", Journal of the Acoustical Society of America, Vol 36, No. 10, PP 1901-1904, Oct 1964
6. Harry L. Shaffer, "Information Rate Necessary to Transmit Pitch-Period Durations for Connected Speech", Journal of the Acoustical Society of America, Vol 36, No. 10, PP 1895-1900, Oct 1964
7. Mark R. Weiss and Cyril M. Harris, "Computer Technique for High-Speed Extraction of Speech Parameters", Journal of the Acoustical Society of America, Vol 35, No. 2, PP 207-214, Feb 1963

Table I
Comparison of the Measurement Conditions and Results Obtained
for Percent of Time No Variation Occurred in Vocoder
Extracted Pitch for Three Speakers and for Reference Data for Natural Speech

Speaker	Median Pitch Number	Median Pitch Freq. (cps)	Measurement Period for No Pitch Variation	Quantization Interval at Median Pitch Frequency (msec)	Percent of Time No Pitch Variations Occurred
1	13	117	22.5 msec, or ≈2.7 Pitch Periods	.26	16.5
2	14	121	22.5 msec, or ≈2.7 Pitch Periods	.24	31.5
3	5	88	22.5 msec, or ≈2.0 Pitch Periods	.46	45.0
Reference	---	---	3 Pitch Periods	.20	14.0

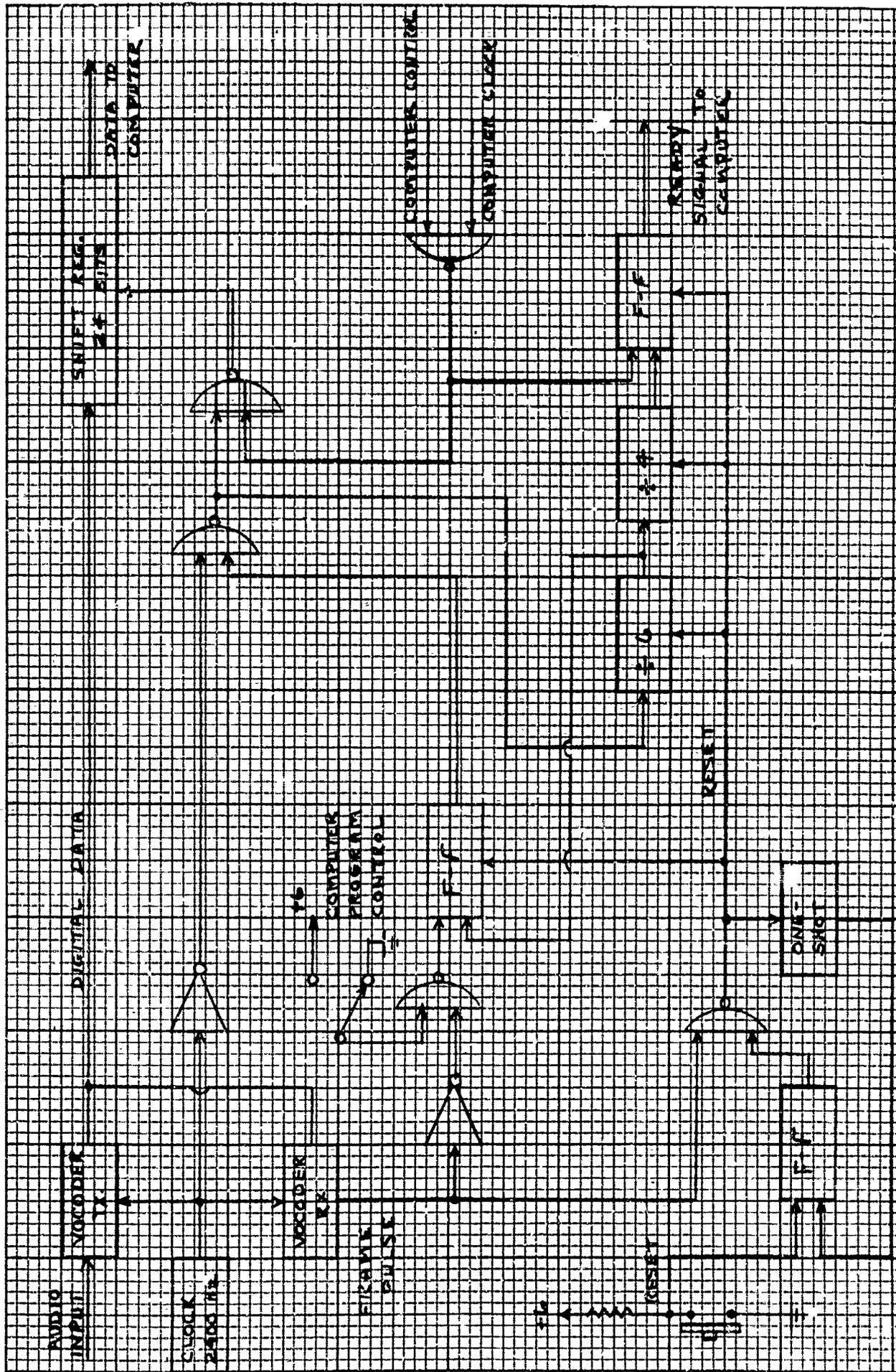


Figure 1 - Block diagram of pitch data selector

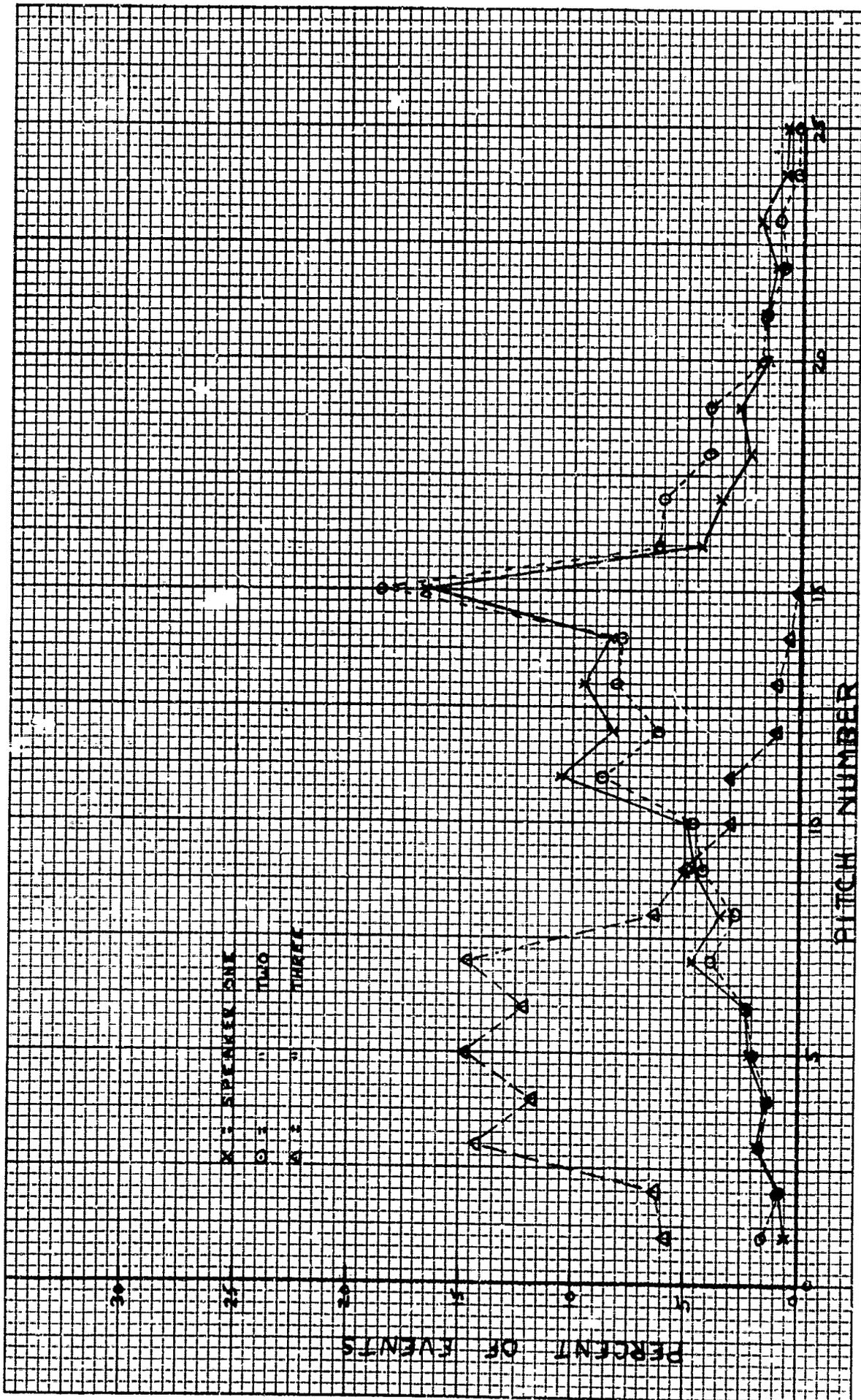


Figure 2 - Frequency distribution of vocoder extracted pitch for three speakers

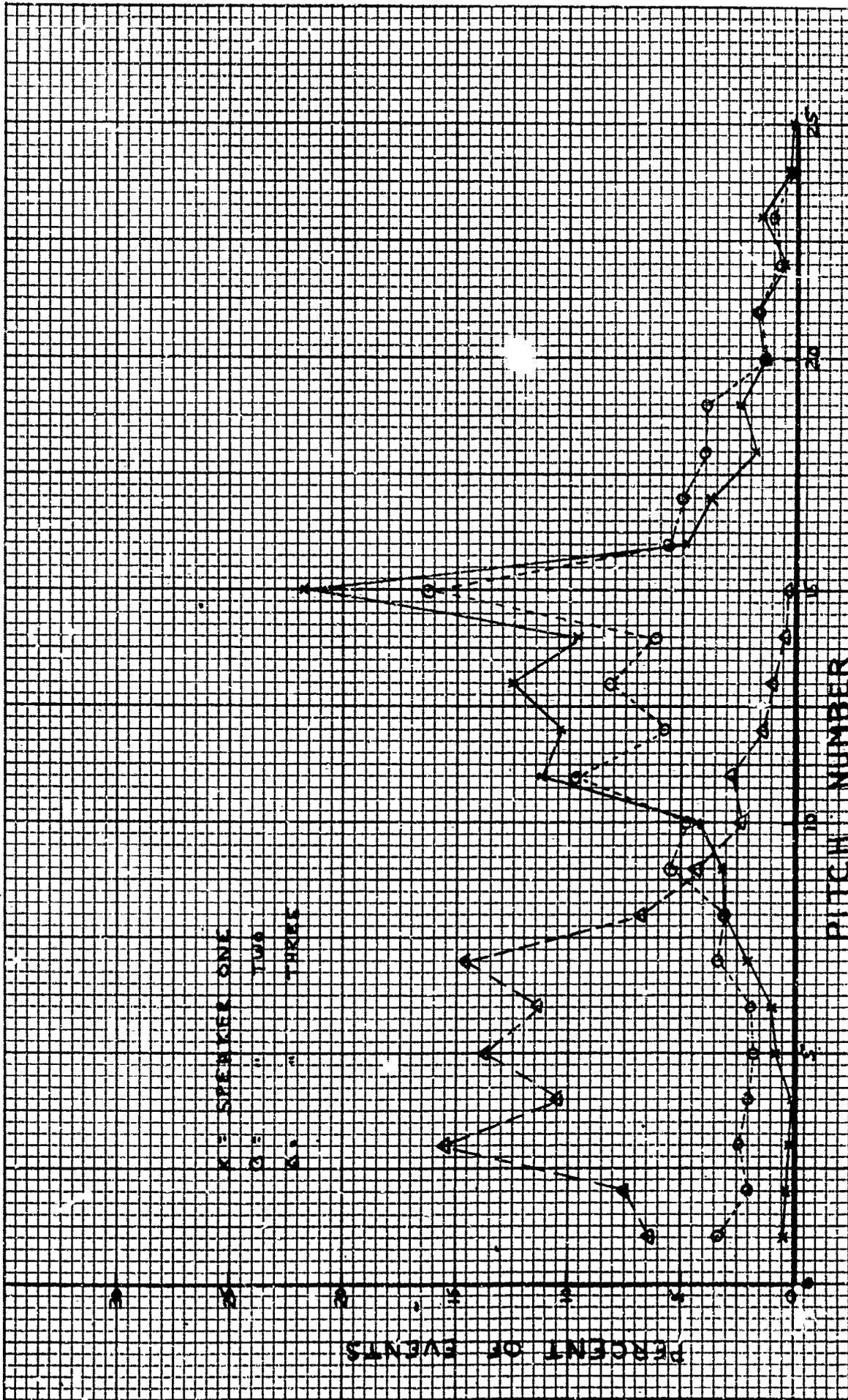


Figure 3 - Frequency distribution of vocoder extracted pitch at the onset of voicing for three speakers

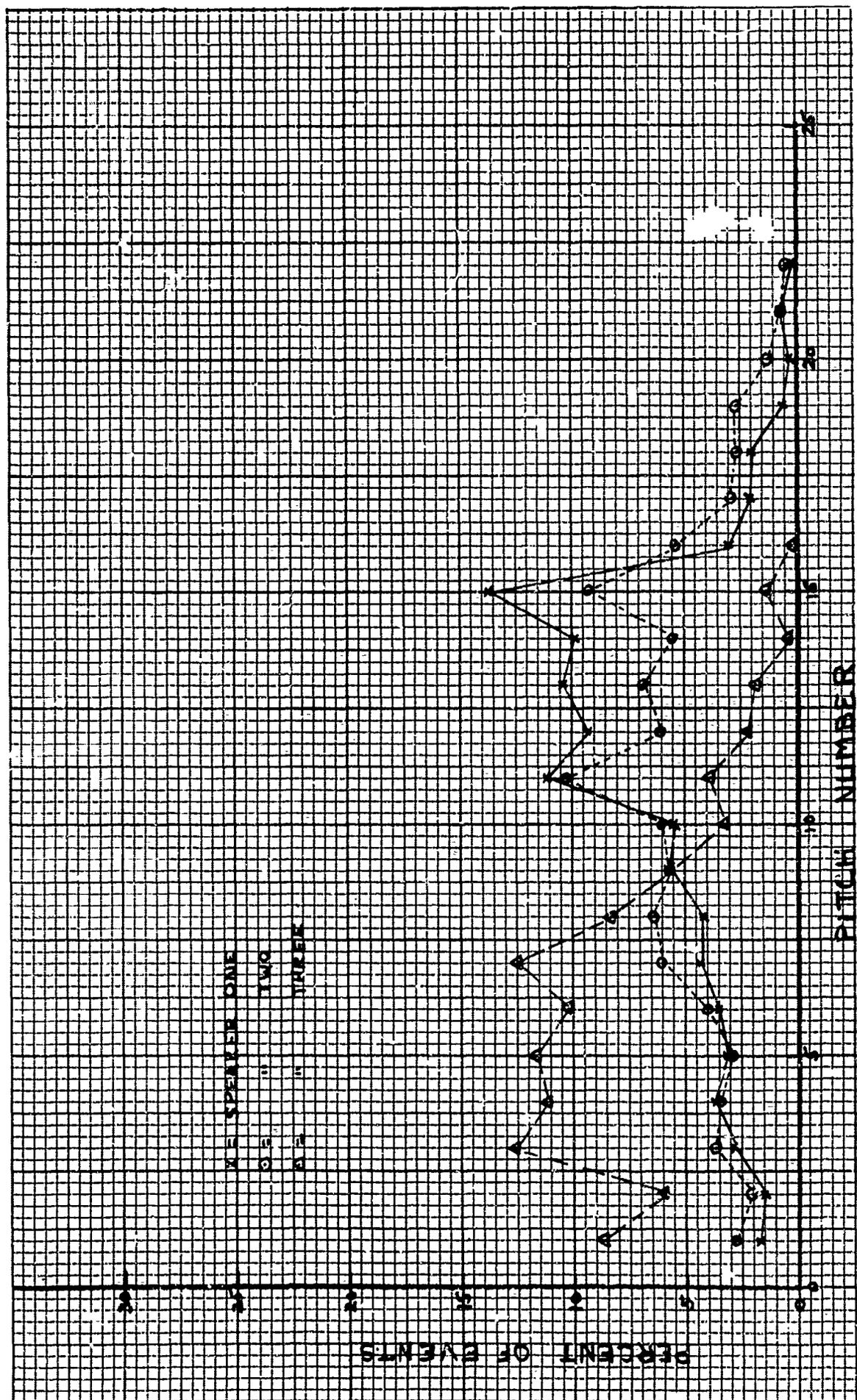


Figure 4 - Frequency distribution of vocoder extracted pitch at the end of voicing for three speakers

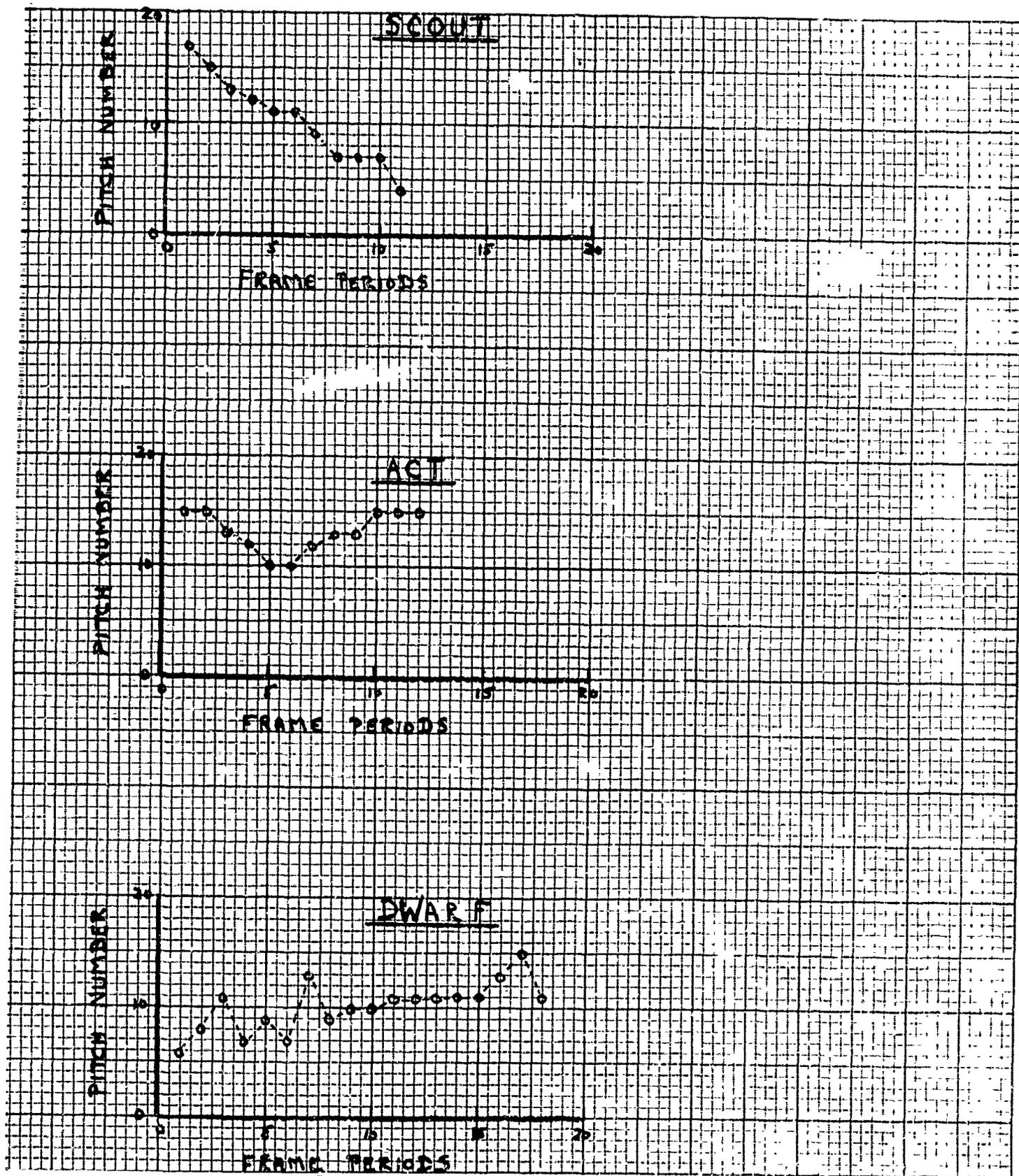


Figure 5 - Vocoder extracted pitch patterns for three single syllable words by speaker number two

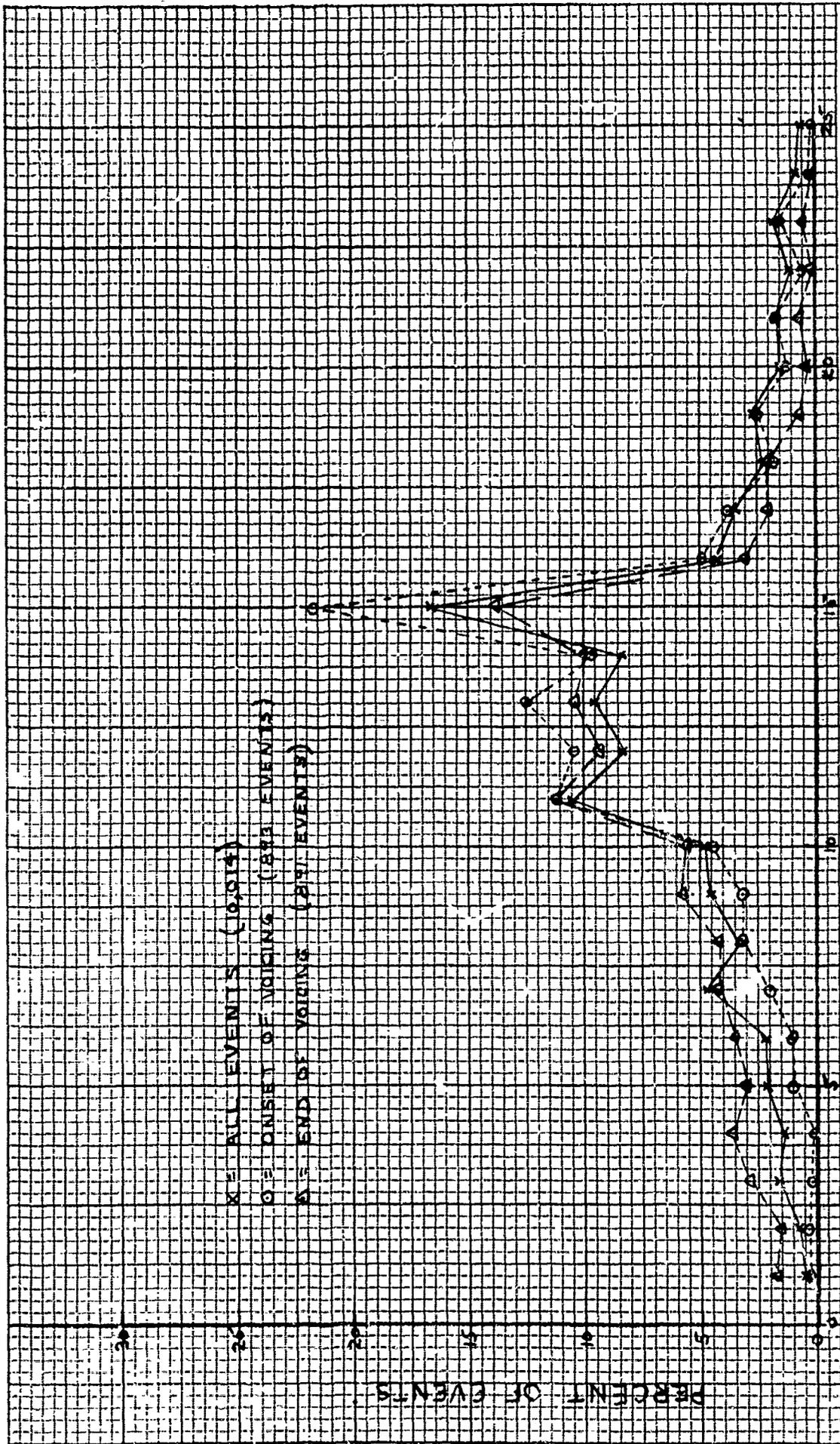


Figure 6 - Frequency distribution of vocoder extracted pitch of speaker number one for all voiced events, for the onset of voicing, and for the end of voicing

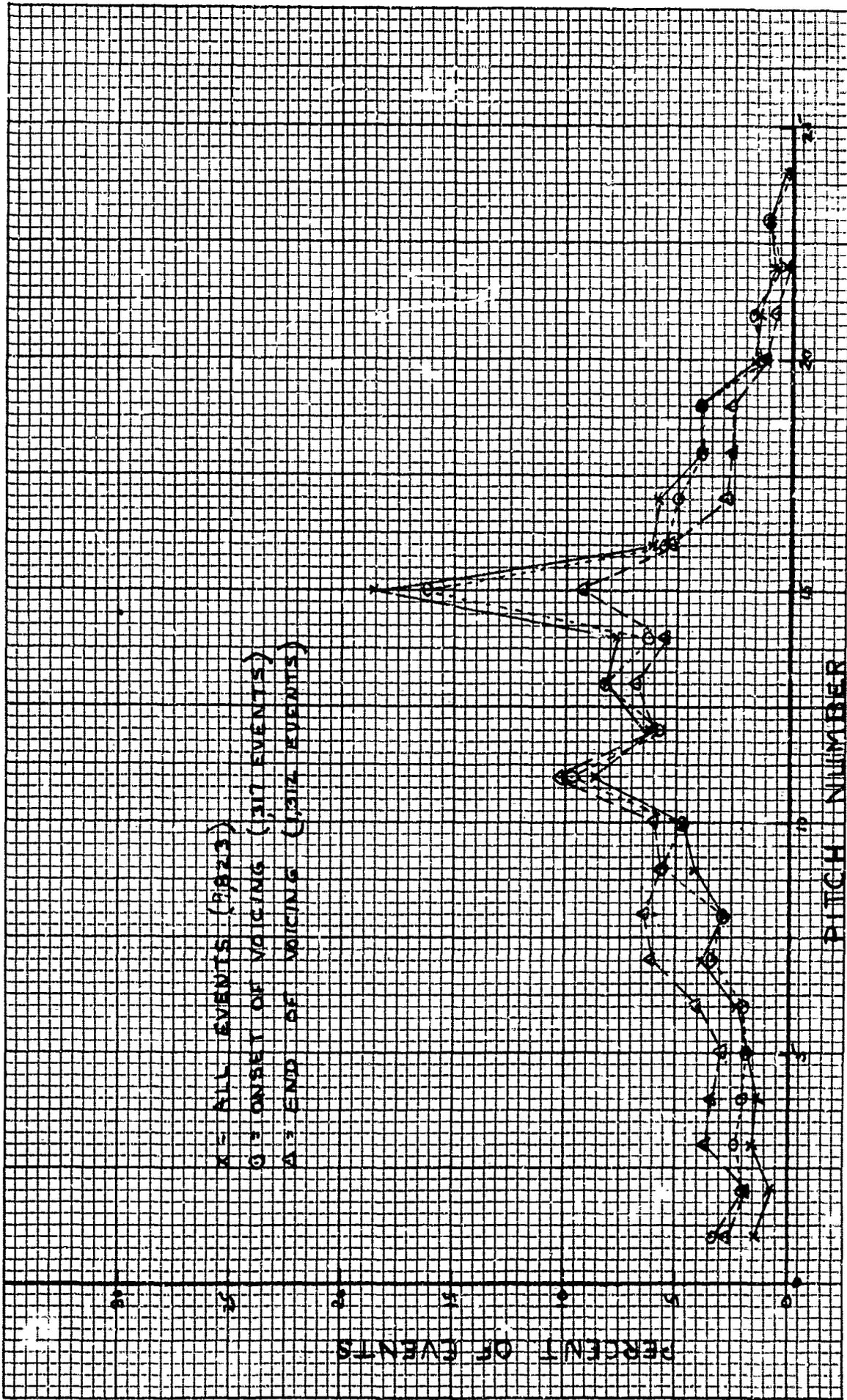


Figure 7 - Frequency distribution of vocoder extracted pitch for speaker number two for all voiced events, for the onset of voicing, and for the end of voicing

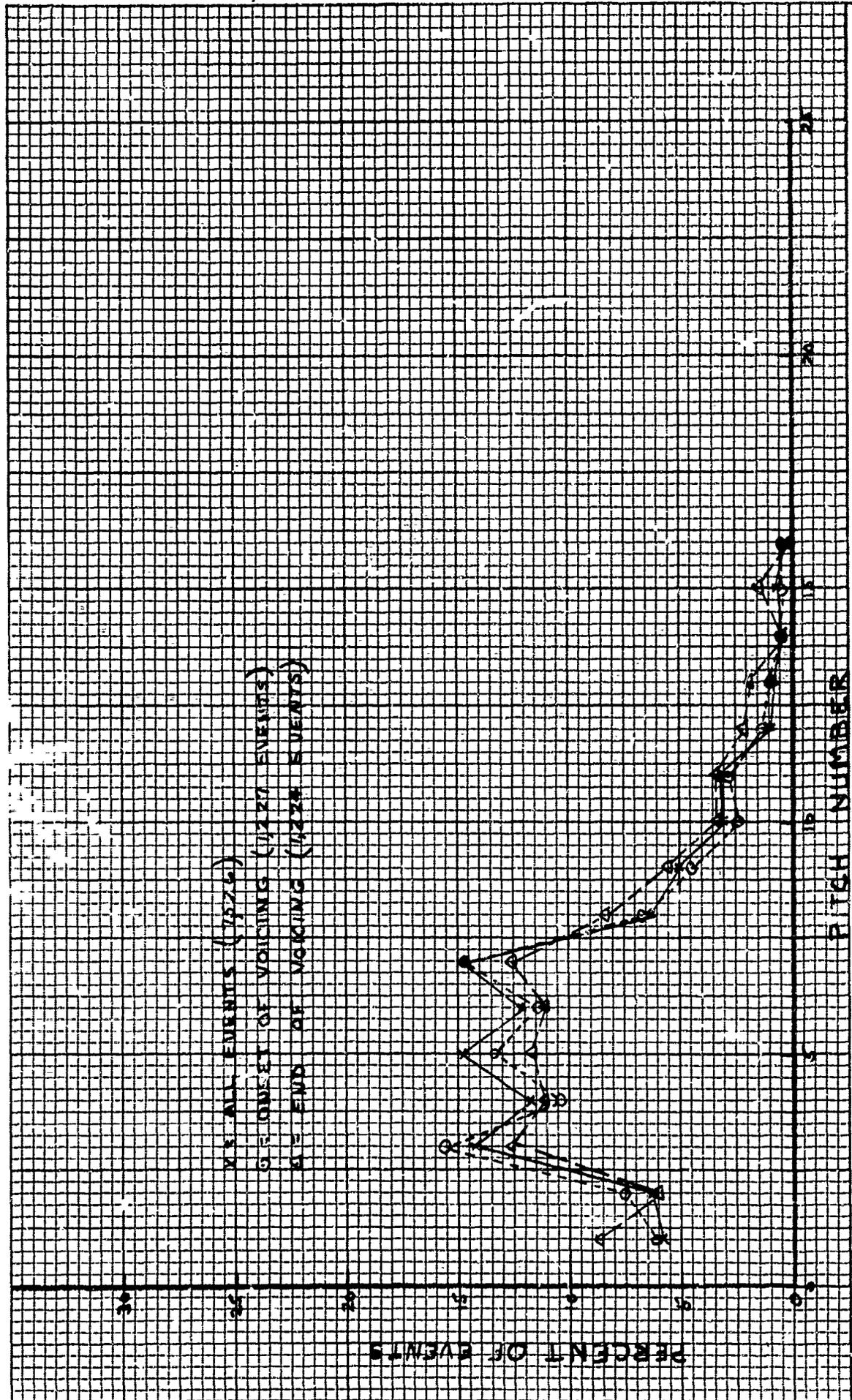


Figure 8 - Frequency distribution of vocoder extracted pitch for speaker number three for all voiced events, for the onset of voicing, and for the end of voicing

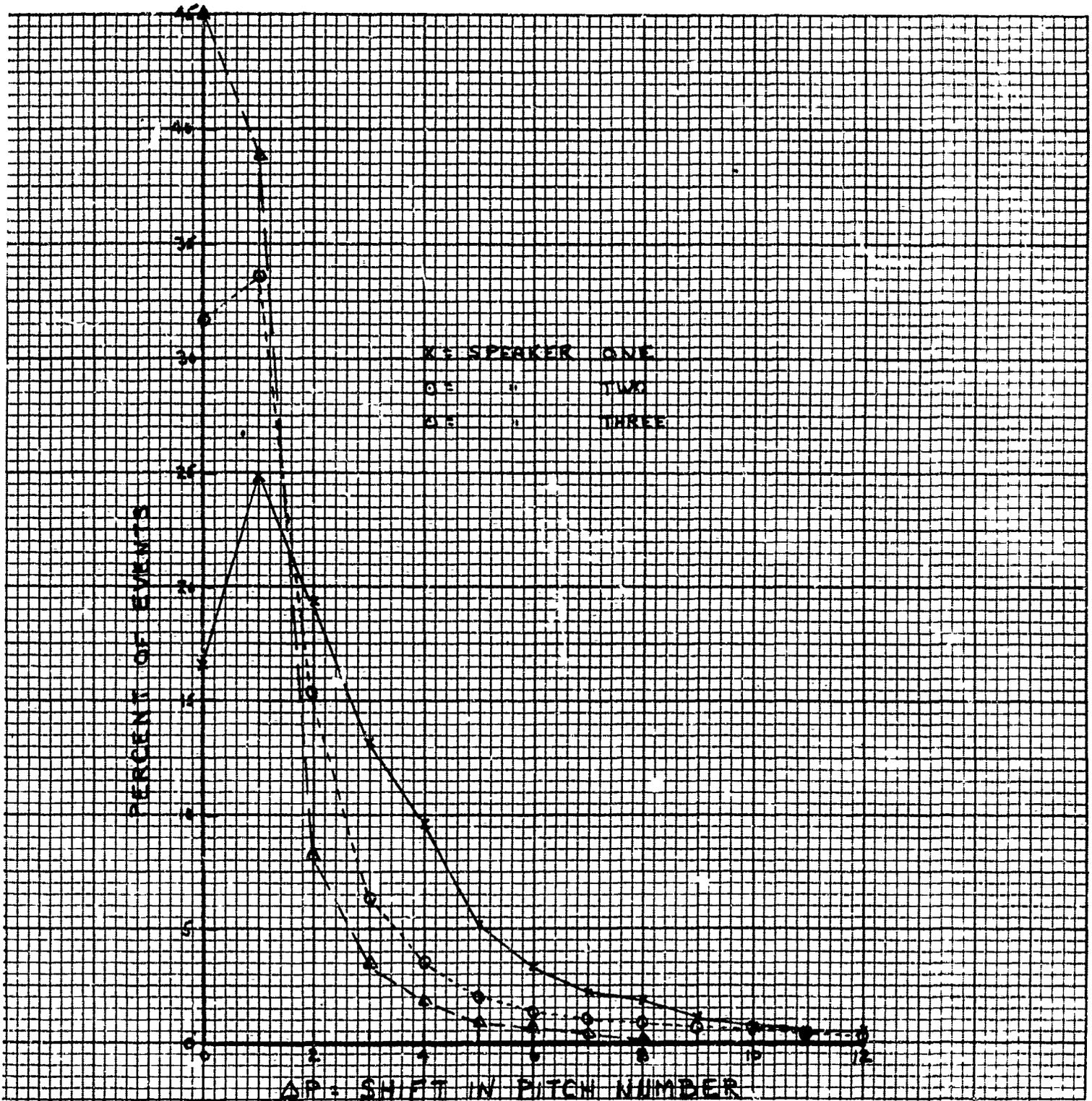


Figure 9 - Frequency distribution of the shift in the encoded pitch number between successive voiced vocoder frames, for three speakers

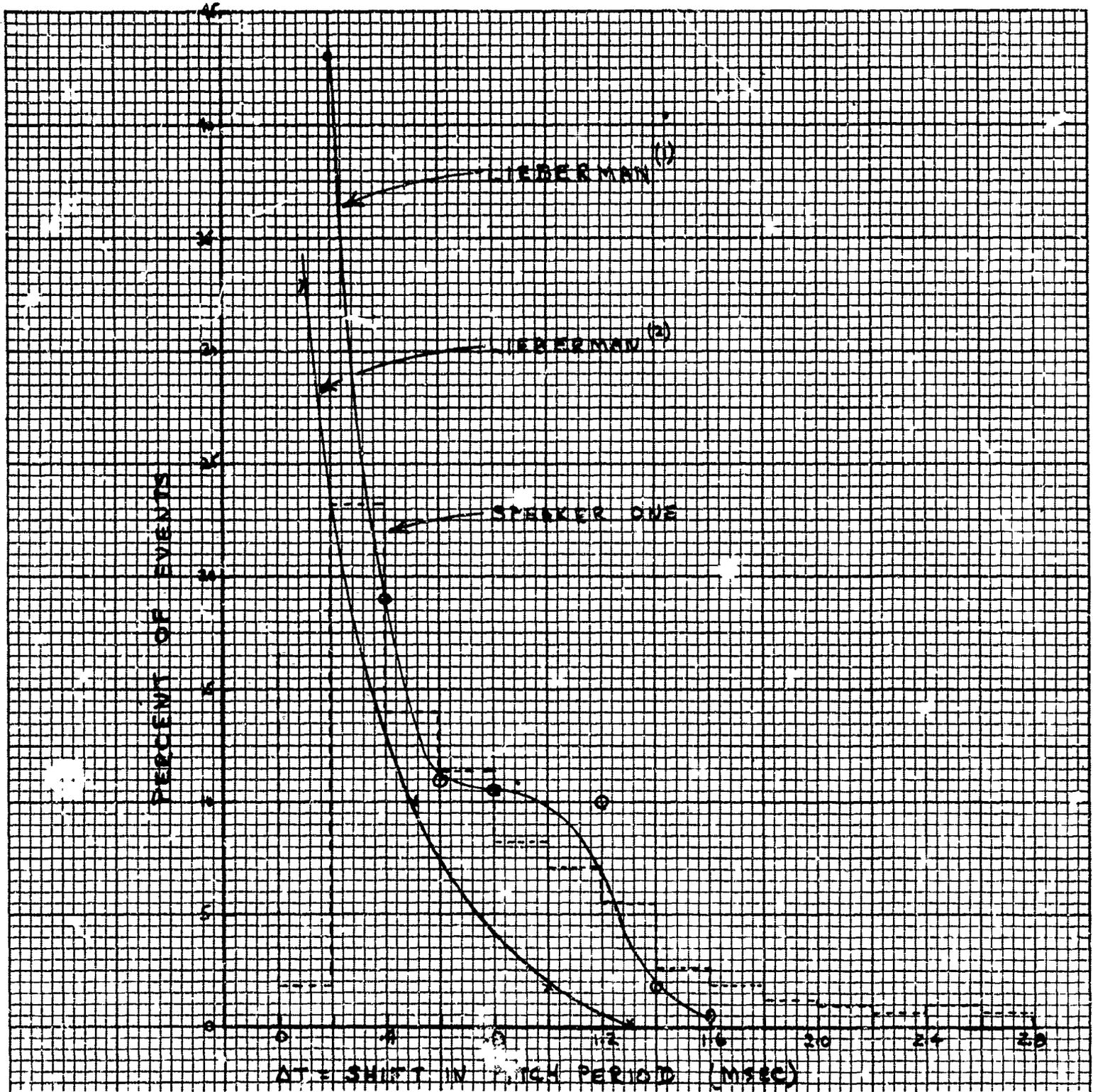


Figure 10 - Frequency distribution of the shift in the encoded pitch period between successive voiced vocoder frames quantized in 0.2 msec intervals for speaker number one, with reference data for natural speech

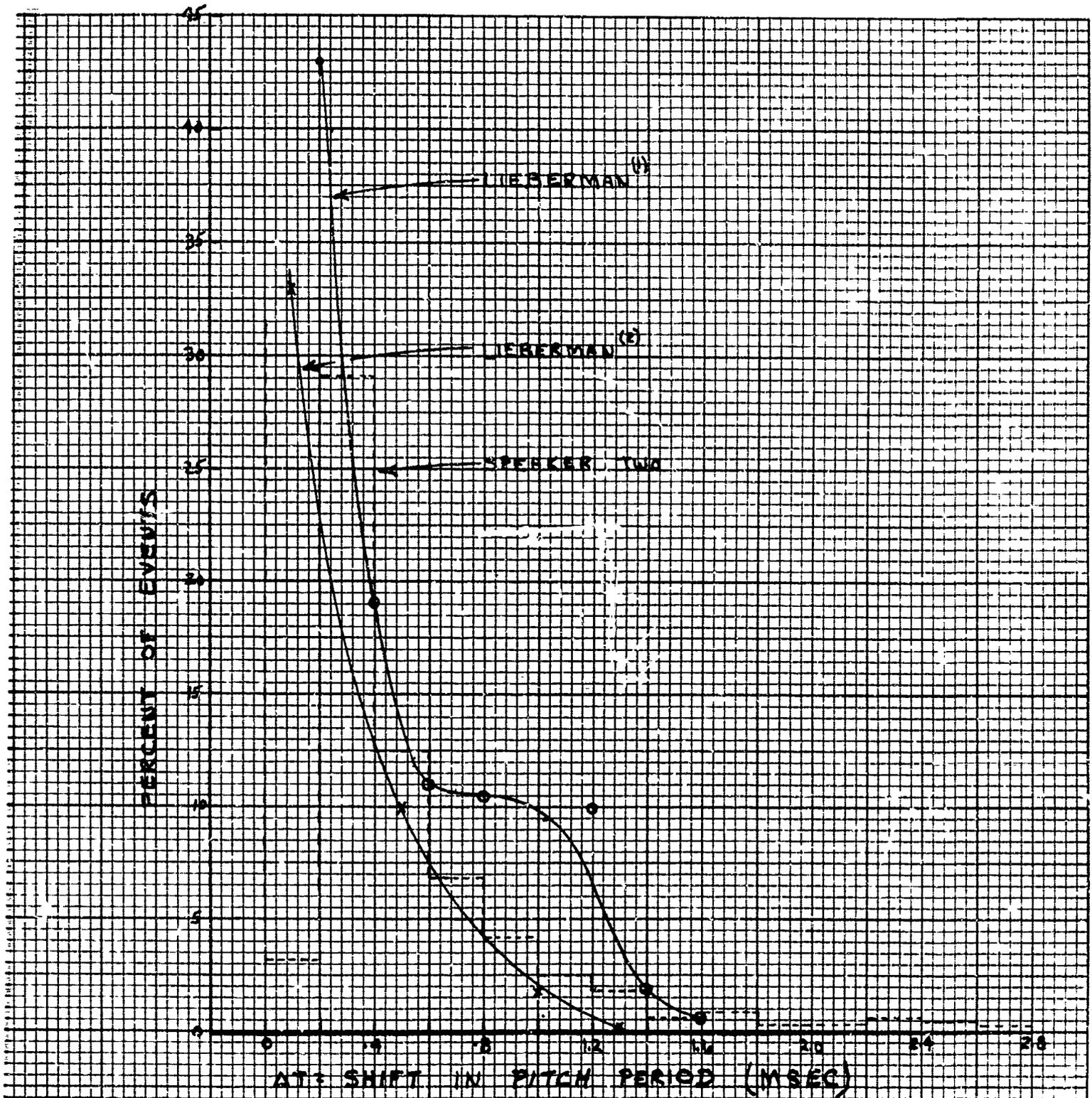


Figure 11 - Frequency distribution of the shift in the encoded pitch period between successive voiced vocoder frames quantized in 0.2 msec intervals for speaker number two, with reference data for natural speech

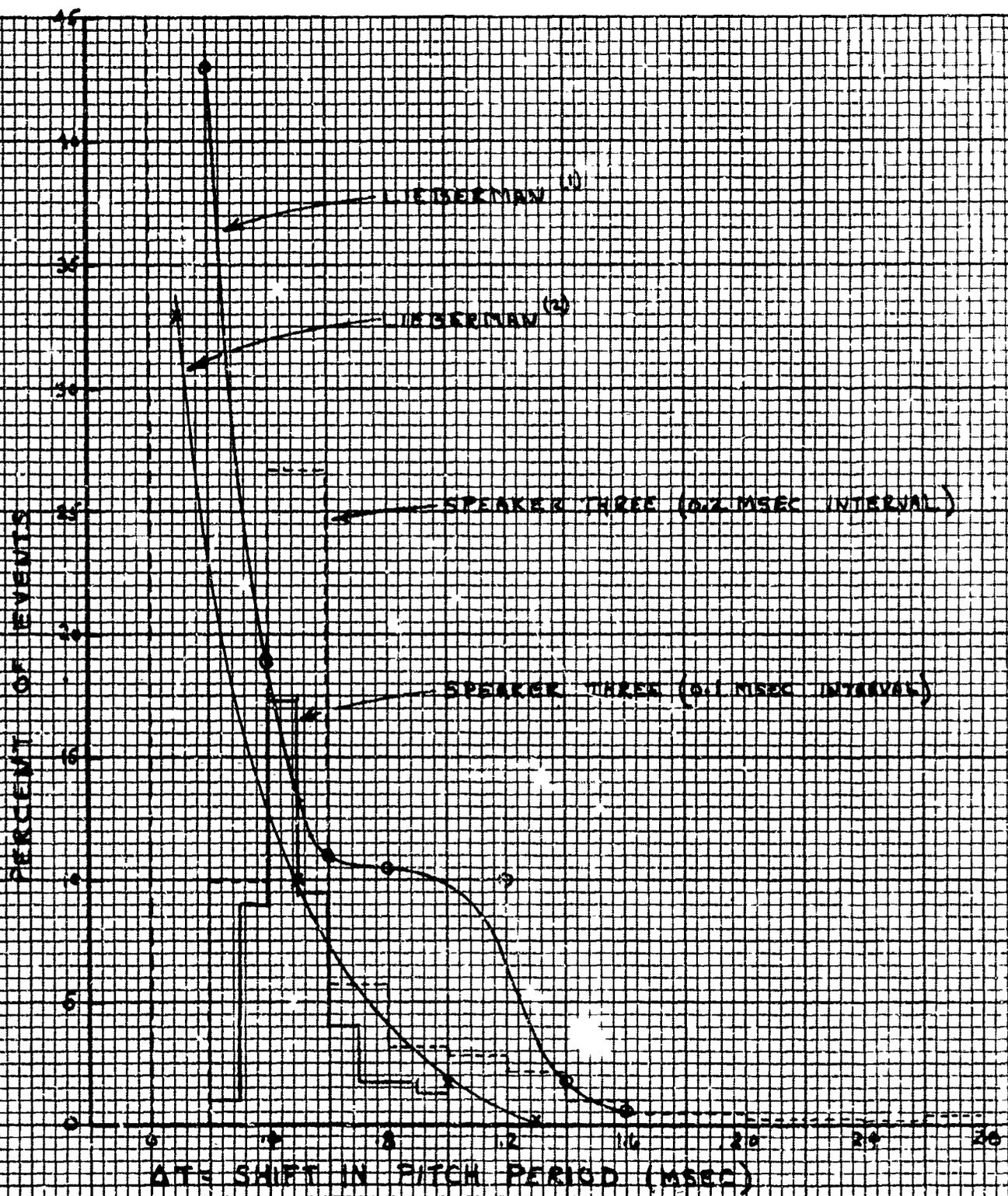


Figure 12 - Frequency distribution of the shift in the encoded pitch period between successive voiced vocoder frames quantized in 0.2 msec and 0.1 msec intervals for speaker number three, with reference data for natural speech

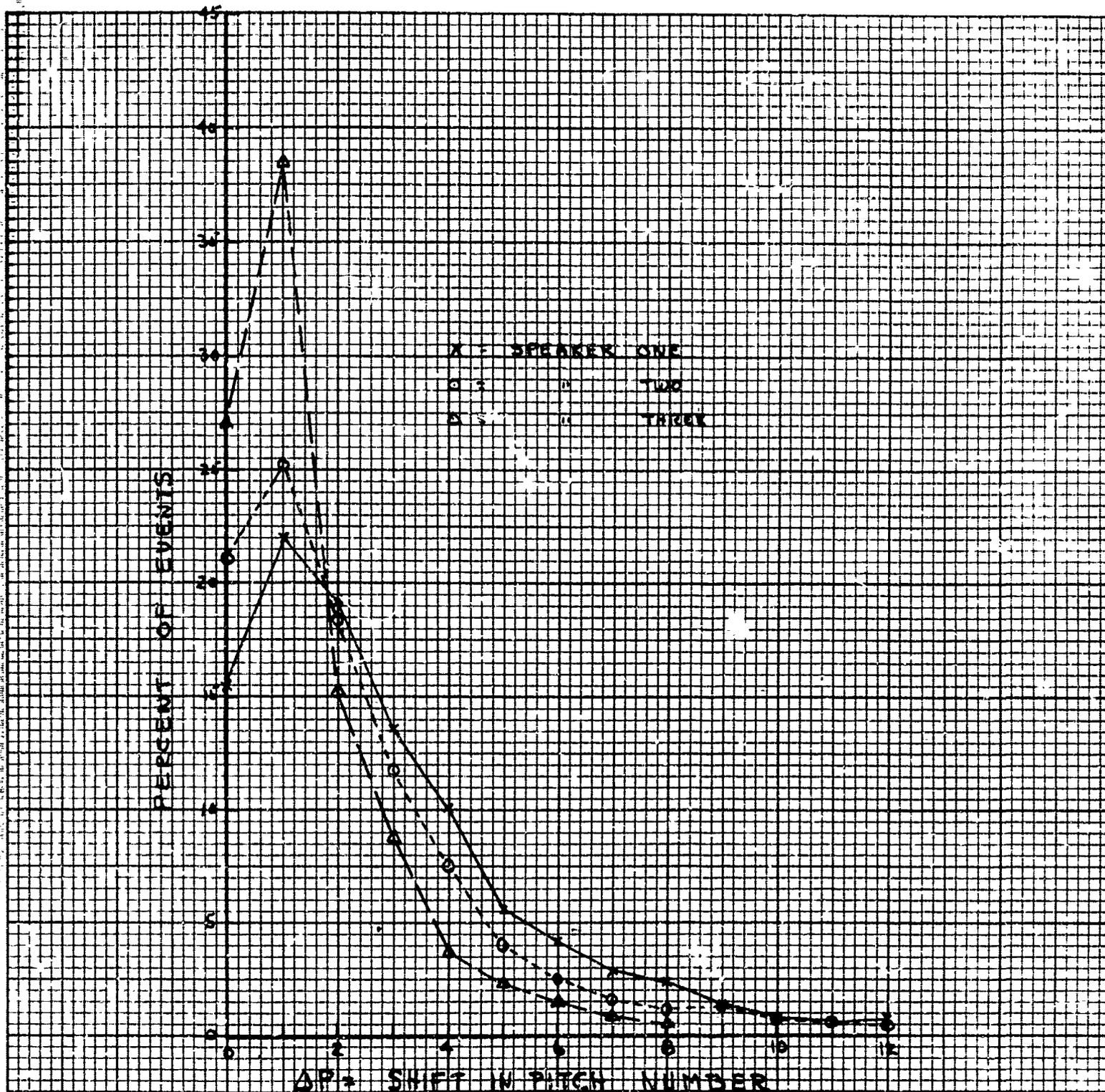


Figure 13 - Frequency distribution of the shift in the encoded pitch number between successive voiced vocoder frames at the onset of voicing, for three speakers

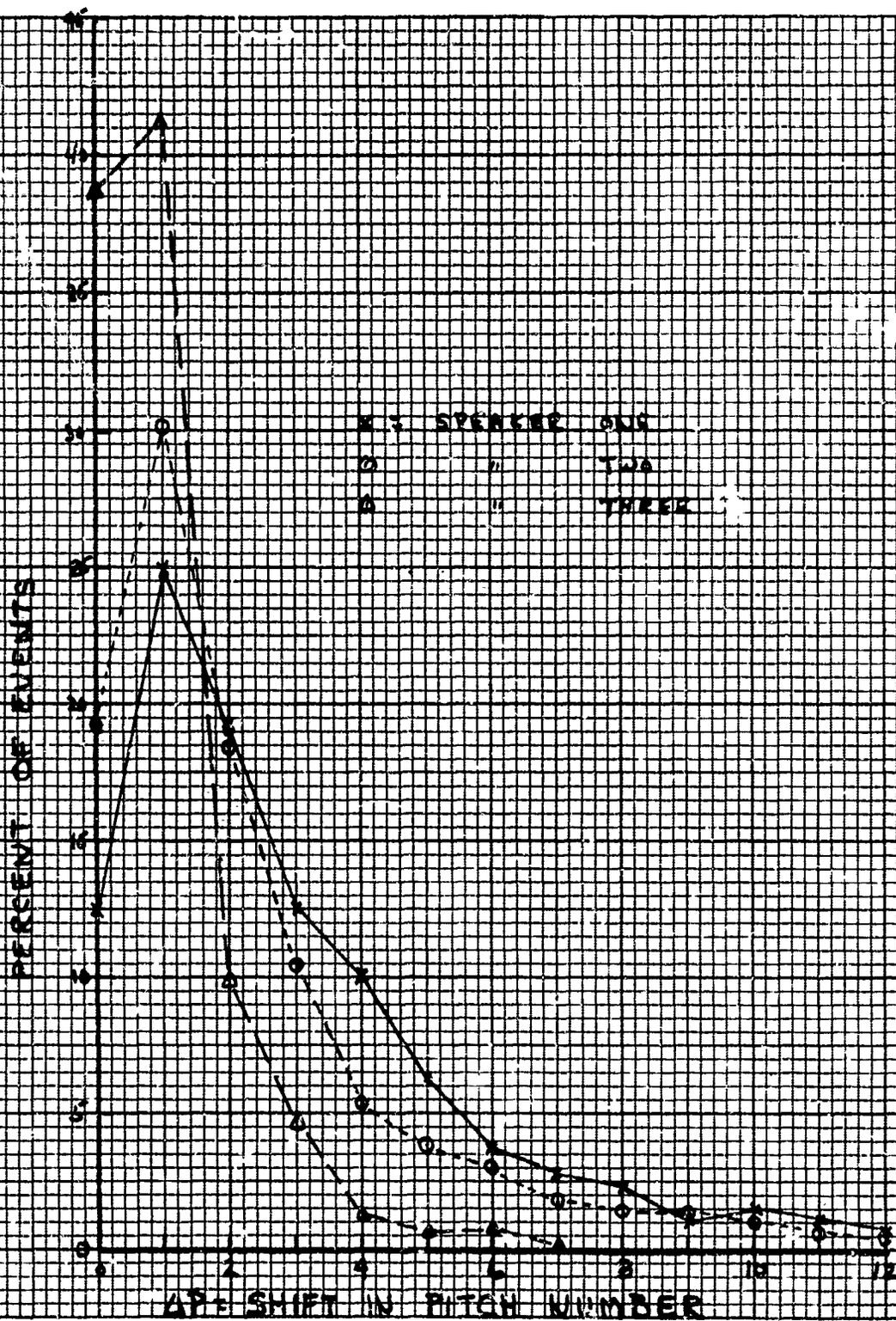


Figure 14 - Frequency distribution of the shift in the encoded pitch number between successive voiced vocoder frames at the end of voicing, for three speakers

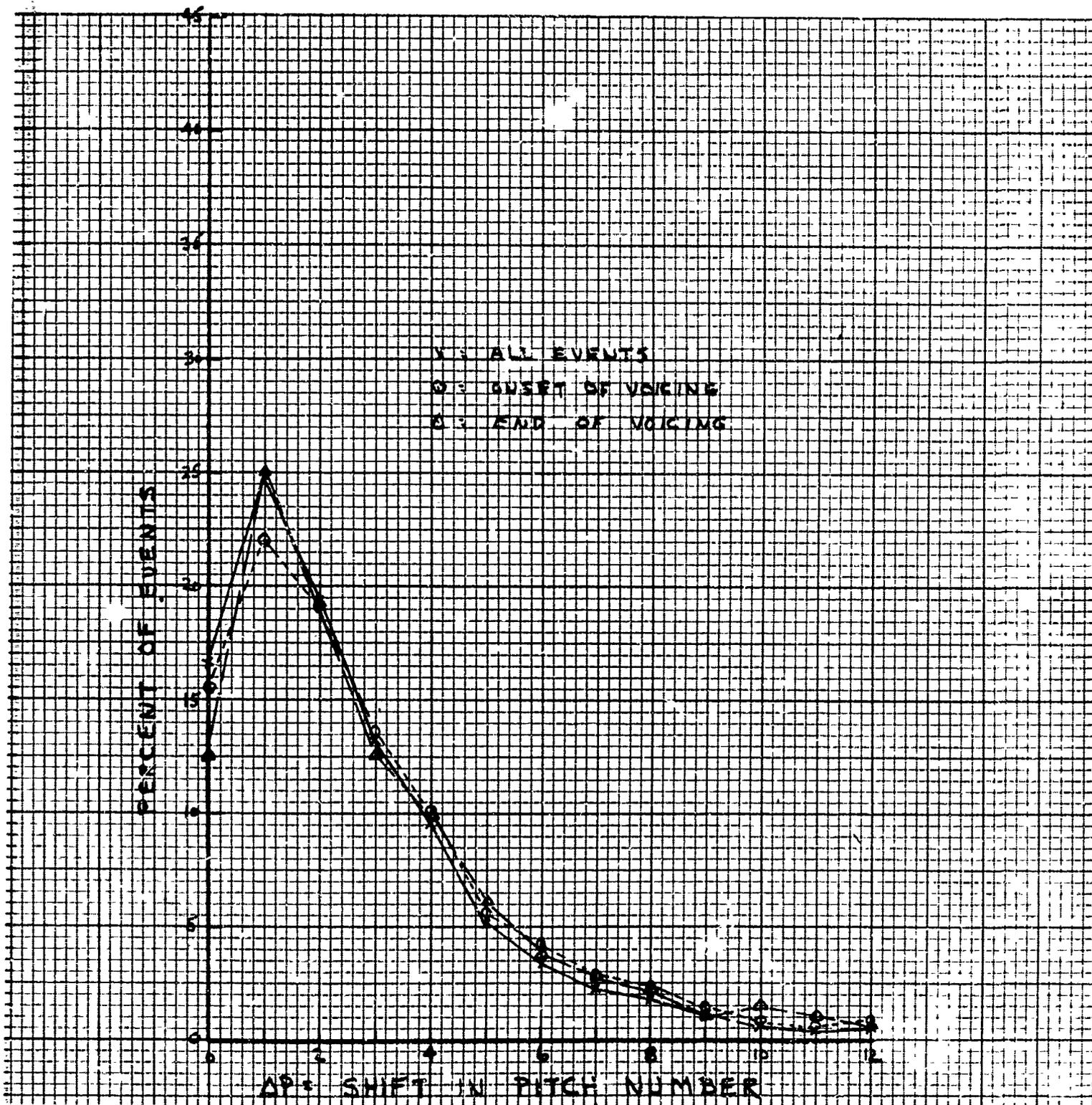


Figure 15 - Frequency distribution of the shift in the encoded pitch number between successive voiced vocoder frames for all events, at the onset of voicing, and at the end of voicing, for speaker number one

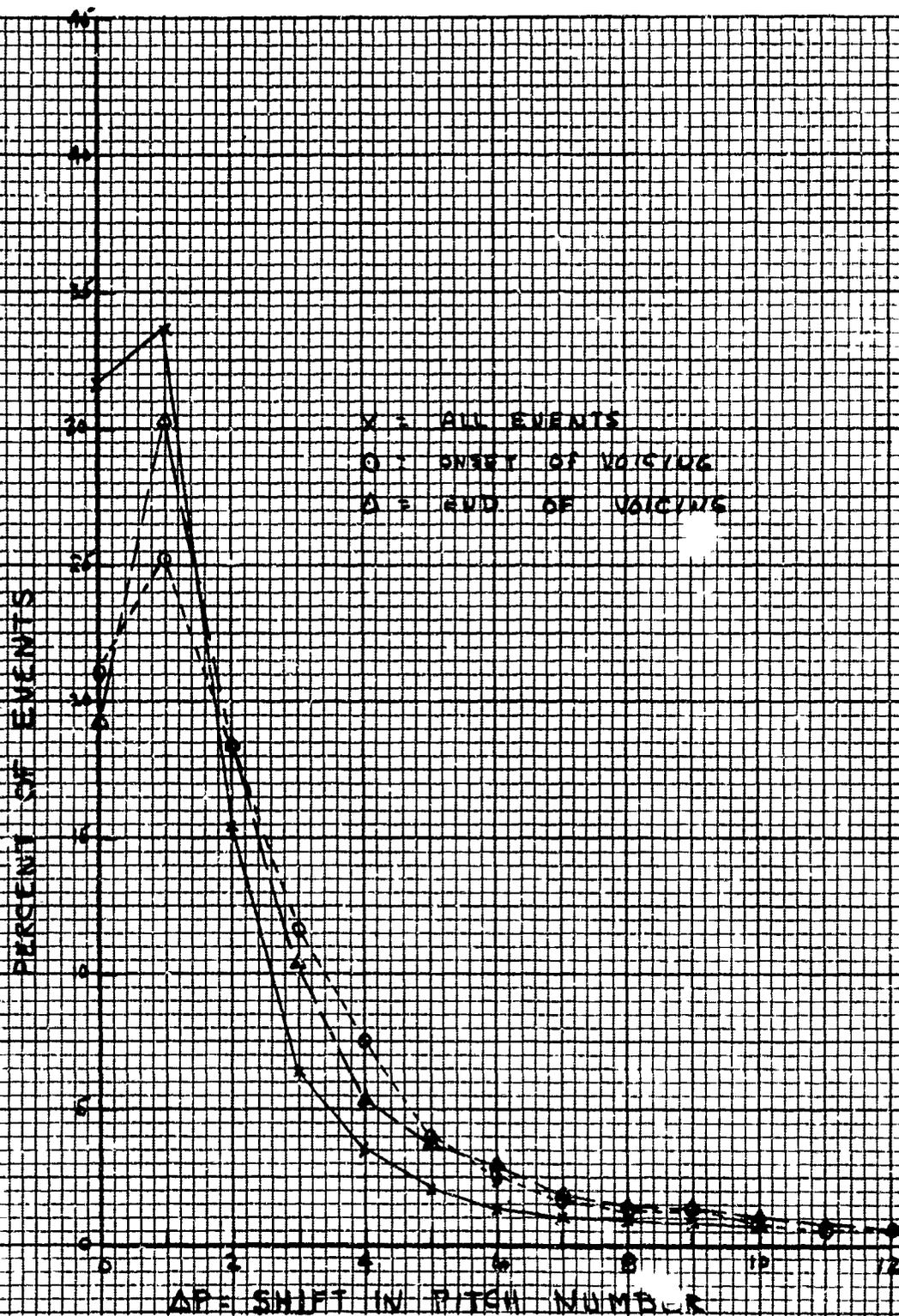


Figure 16 - Frequency distribution of the shift in the encoded pitch number between successive voiced vocoder frames for all events, at the onset of voicing, and at the end of voicing, for speaker number two

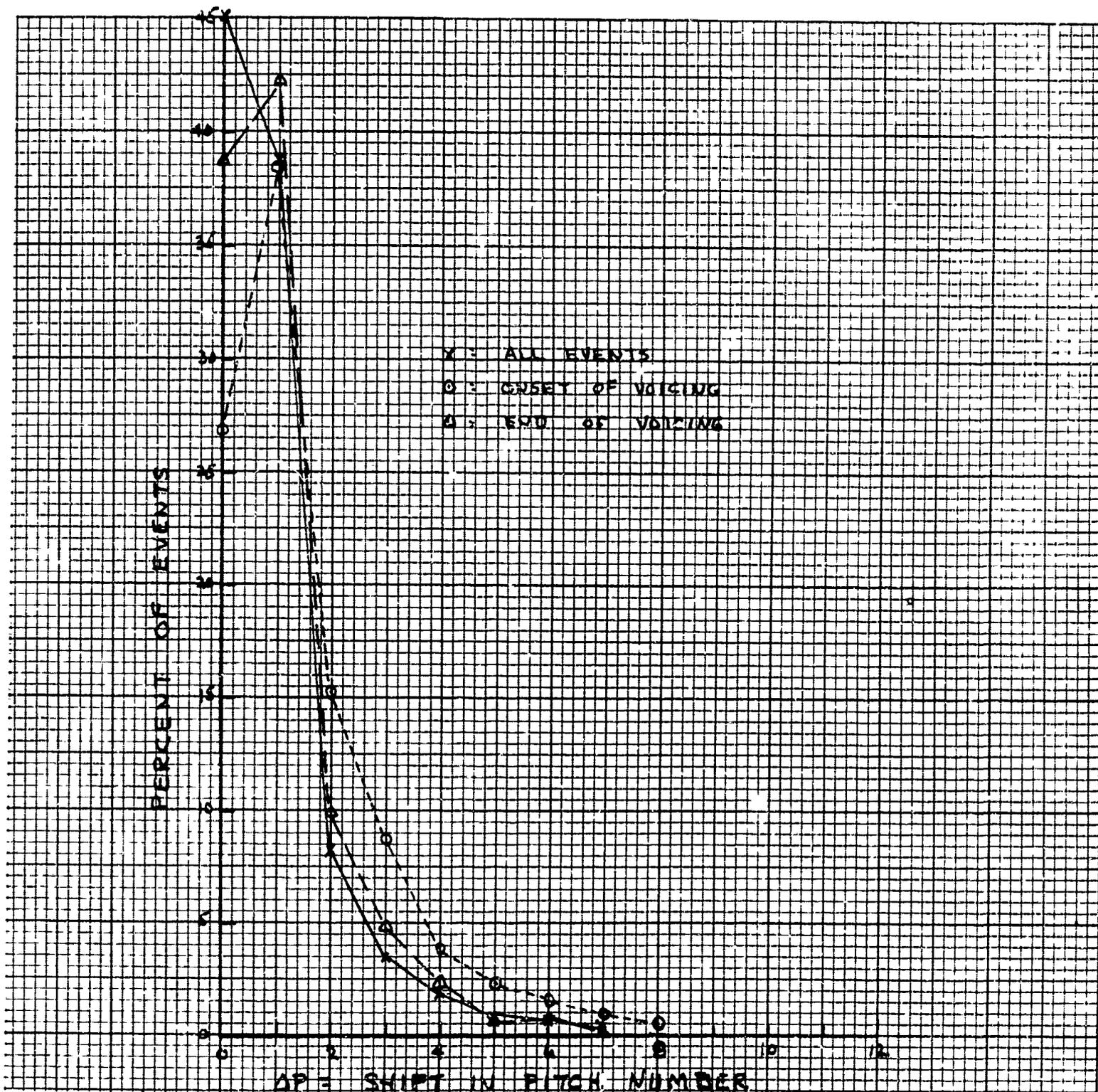


Figure 17 - Frequency distribution of the shift in the encoded pitch number between successive voiced vocoder frames for all events, at the onset of voicing, and at the end of voicing, for speaker number three

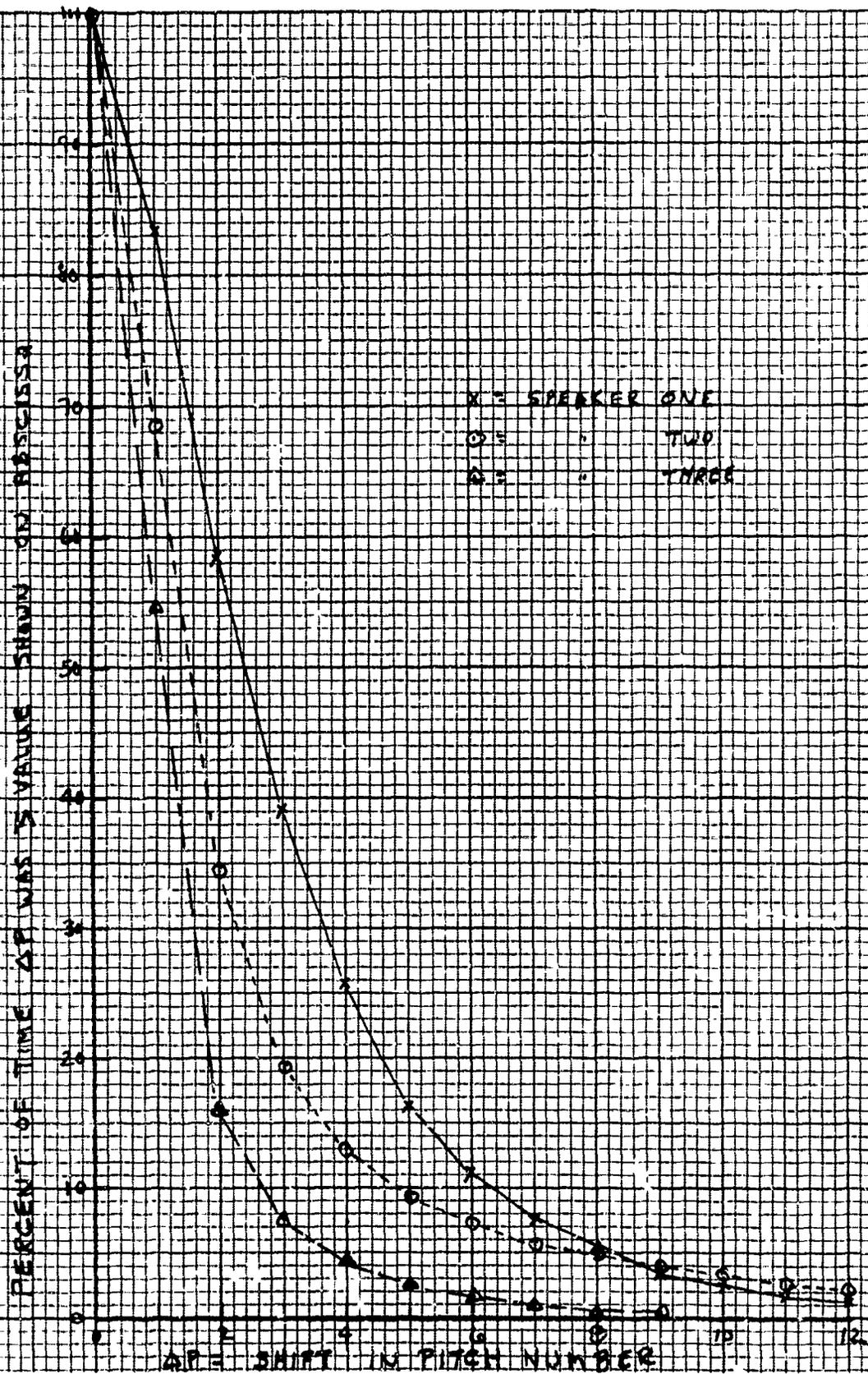


Figure 18 - Cumulative frequency distribution of the shift in the encoded pitch number between successive voiced vocoder frames for all events, for three speakers

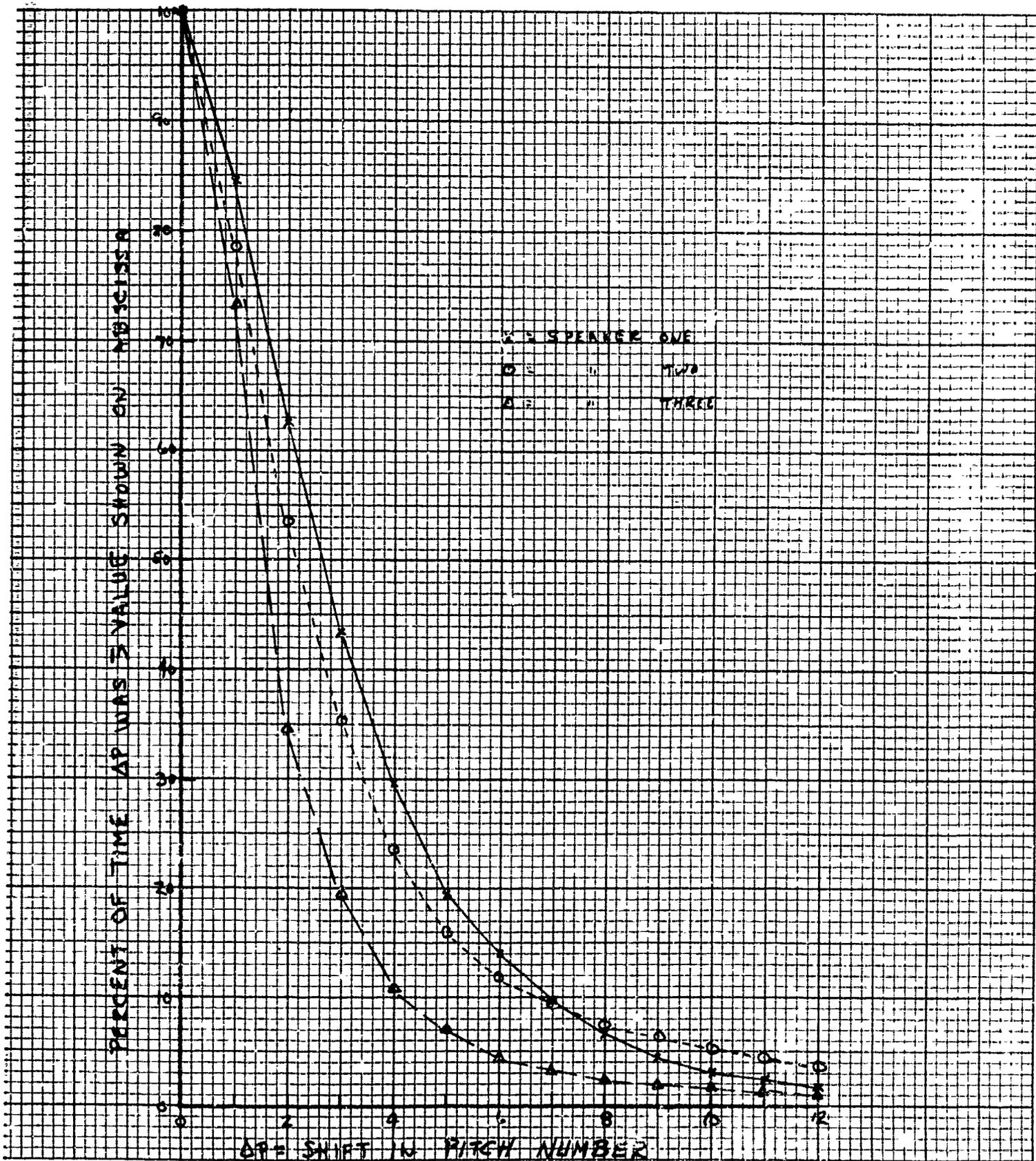


Figure 19 - Cumulative frequency distribution of the shift in the encoded pitch number between successive voiced vocoder frames at the onset of voicing, for three speakers

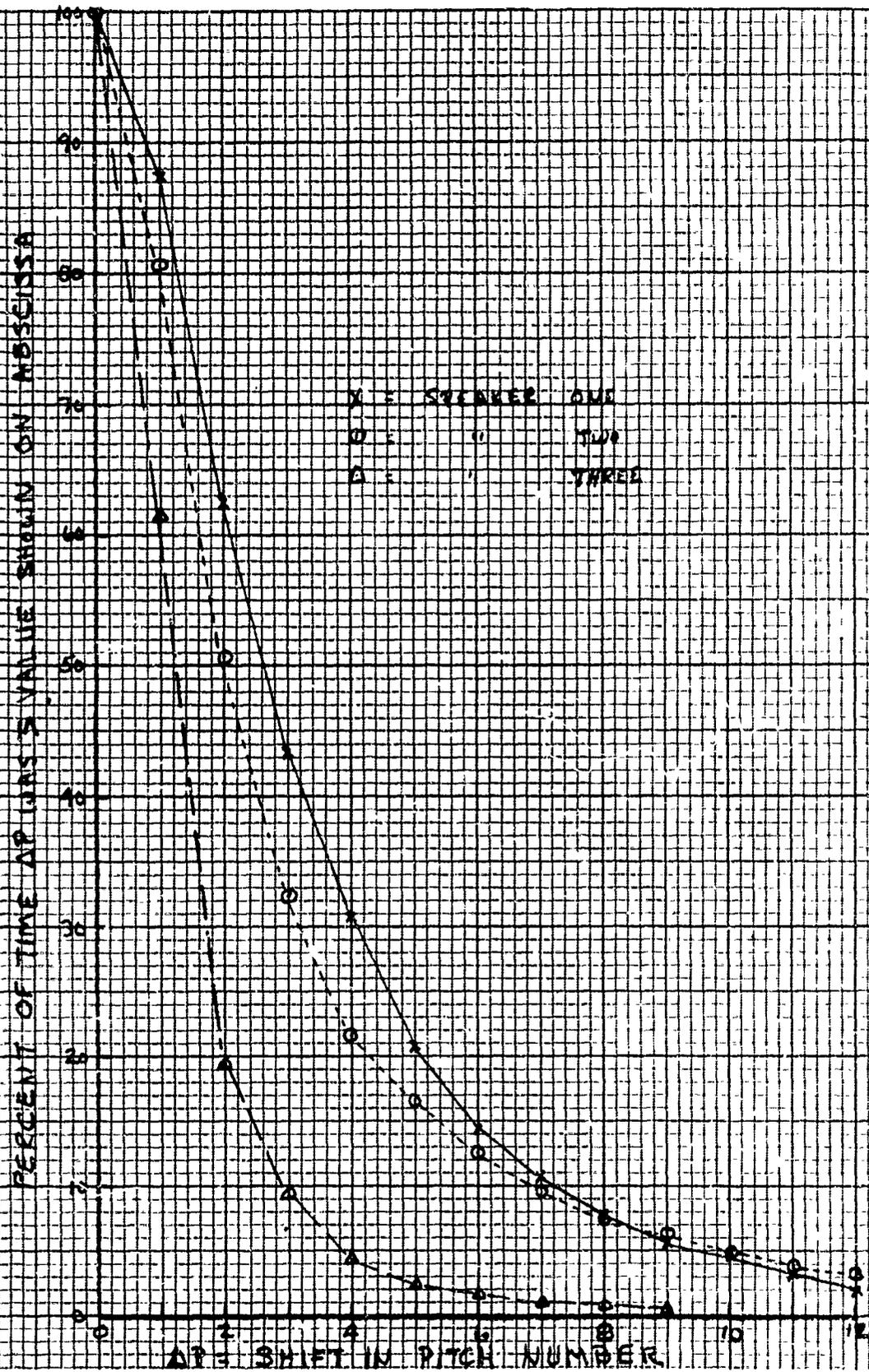


Figure 20 - Cumulative frequency distribution of the shift in the encoded pitch number between successive voiced vocoder frames at the end of voicing, for three speakers