DOCUMENT RESUME

ED 088 288 FL 005 220

AUTHOR Lea, Wayne A.; And Others
TITLE Use of Syntactic Segmentation and Stressed Syllable Location in Phonemic Recognition.
PUB DATE Nov 72
NOTE 12p.; Paper presented at the Meeting of the Acoustical Society of America (84th, Miami, Florida, November 1972)

EDRS PRICE MF-$0.75 HC-$1.50
DESCRIPTORS Articulation (Speech); Computers; Consonants; *Distinctive Features; Graphs; Language Patterns; *Language Research; Oral Expression; Pattern Recognition; Phonemes; *Phonemics; *Phonological Units; Phonology; *Phrase Structure; Sentence Structure; Syllables; Syntax; Vowels; Word Recognition

ABSTRACT

Automatic speech recognition is expected to be more successful when syntactically-related information is incorporated into early stages of recognition. Phonemic decisions, in particular, are expected to be more accurate and less ambiguous when contextual information is considered. A computer program detected about 90% of all boundaries between major syntactic constituents from fall-rise fundamental frequency (Fo) contours in the Rainbow Script as read by two talkers. A procedure was devised for locating stressed syllables, in constituents, from (1) high-energy and increasing-Fo portions near the peaks of Fo contours and (2) local increases in Fo from an archetype falling contour. The procedure succeeded in locating most syllables which had been perceived by listeners to be stressed. A recognition strategy is outlined for using detected syntactic boundaries and stressed syllable locations in estimating distinctive features of some phonemes in connected speech. Vowel and consonanat recognition would be attempted first in the stressed syllables. Other readily-detected segments, such as coronal strident fricatives, would be found. A pilot study showed that front/back decisions for vowels are more reliable in stressed than unstressed syllables. (Author/DD)

USE OF SYNTACTIC SEGMENTATION AND STRESSED SYLLABLE LOCATION IN PHONEMIC RECOGNITION

Wayne A. Lea
Mark F. Medress
Toby E. Skinner

Univac DSD
P.O. Box 3525
St. Paul, Minn. 55165

ABSTRACT

Automatic speech recognition is expected to be more successful when syntactically-related information is incorporated into early stages of recognition. Phonemic decisions, in particular, are expected to be more accurate and less ambiguous when contextual information is considered. A computer program detected about 90% of all boundaries between major syntactic constituents, from fall-rise fundamental frequency (F_0) contours in the Rainbow Script, as read by two talkers. A procedure was devised for locating stressed syllables, in constituents, from (a) high-energy and increasing-F_0 portions near the peaks of F_0 contours and (b) local increases in F_0 from an archetype falling contour. The procedure succeeded in locating most syllables (83% for one talker, 98% for another) which had been perceived by listeners to be stressed. A recognition strategy is outlined for using detected syntactic boundaries and stressed syllable locations in estimating distinctive features of some phonemes in connected speech. Vowel and consonant recognition would be attempted first in the stressed syllables. Other readily-detected segments, such as coronal strident fricatives, would be found. A pilot study showed that front/back decisions for vowels are, indeed, more reliable in stressed than in unstressed syllables.

Presented at the 84th meeting of the Acoustical Society of America, Miami, Florida, November, 1972.

This research was supported in part by the Advanced Research Projects Agency of the Department of Defense under Contract DAHC15-72-C-0138, and in part by the Univac Independent Research and Development Program.

USE OF SYNTACTIC SEGMENTATION AND STRESSED SYLLABLE LOCATION IN PHONEMIC RECOGNITION

Wayne A. Lea, Mark F. Medress, and Toby E. Skinner

1. Structural Aids to Phonemic Recognition

Linguistic and perceptual arguments (Lea, 1972) suggest that devices which recognize speech will have to make use of grammatical structure in early stages of the recognition procedures. This can be accomplished, in part, by using the prosodic features to segment the speech into grammatical phrases, and to identify those syllables that are given prominence, or stress, in the sentence structure.

One way in which prosodic information, and resulting syntactic segmentation and stress pattern analyses, may be used to aid distinctive features estimation is as follows. At an early stage in recognition, one detects boundaries between major syntactic constituents from prosodic features. Then, the highest-stress syllables within each constituent are located, using reliable prosodic cues to stress. Some distinctive features are then estimated within these stressed syllables, since the consonants and vowels are expected to be more reliably encoded in stressed syllables than in weakly stressed or reduced syllables (Hughes, Li, and Snow, 1972). Next, the partial distinctive features description is matched with generated or stored patterns for possible stressed syllables or words in the lexicon. Then a guess as to the word content of the constituent is made, based on the reliable feature information from the stressed syllables, plus other reliable data within the constituent (such as presence of coronal strident fricatives, etc.; cf. Medress, 1972). If reliable decisions cannot be made based on such minimal feature information within the constituent, analyses are then applied to other words or syllables at lower stress values, and a guess based on the two or more moderately-stressed syllables is made. Iteration would continue until all syllables are analyzed, if necessary. Each iterative guess as to constituent identity would be combined with those for other constituents in the sentence until a satisfactory set of hypotheses for all constituents

yielded the grammatical, meaningful sentence.

A computer program was previously implemented for detecting boundaries between major grammatical constituents from fall-rise "valleys" in fundamental frequency (F_o) contours (Lea, 1972). It has successfully detected 80 to 90% of all predicted boundaries, for weather reports, newscasts, and stories composed of monosyllabic words, as read by six talkers.

On the other hand, no algorithm for locating stressed syllables in connected speech has previously been developed. The use of boundary detections and stressed syllable locations in aiding distinctive features estimation has also not been tested. One problem in testing a stressed syllable locator is to first establish what are properly considered as the actual stressed syllables in the speech.

Experiments have been designed to study syntactic boundaries and stress patterns in the first paragraph of the well-known "Rainbow Passage" (Fairbanks, 1940; Lea, Medress, and Skinner, 1972). Perception tests were conducted to provide the standard stress decisions whereby acoustic cues to stress could be tested.

2. Perceived Stress Patterns in the Rainbow Script

Three listeners (WAL, MFM, and TES) individually heard the Rainbow Script as recorded by two male talkers (ASH and GWH). Following a modification (cf. Lea, Medress, and Skinner, 1972) of the procedure used by Hughes, Li, and Snow (1972), each listener heard clauses or sentences in the Rainbow Script repeated at will (by rewinding and replaying a tape). The listener was instructed to mark (in whatever way he chose), for each syllable, whether he heard that syllable as stressed, unstressed, or reduced. To facilitate marking for each syllable, the script was typed on a sheet of paper with vertical slashes between syllables. A mark was required for each syllable (between two slash marks). The listener received one such sheet for each talker.

Each listener repeated the experiment at least twice, to establish listener consistency from one time to another. Figure 1 shows resulting

"confusion" matrices of perceptions from one trial to the next. Differences from talker to talker were slight, so that data for both are pooled in Figure 1. In general, the preponderance of judgments were on the diagonal, showing that perceived stress levels were consistent from repetition to repetition. A third trial by listener WAL again showed similar results. Listeners WAL and MFM reported that they categorized as follows: "Is the syllable stressed? If not, is it reduced? If not, (that is, for the left-overs) it is unstressed". Listener TES used an alternate strategy whereby reduced syllables were the left-over category. For all listeners, confusions between reduced and unstressed syllables were more frequent than those between stressed and unstressed syllables.

Comparisons of the perceptions of one listener (WAL) with the others are shown in Figure 2, for the first trials by each listener. Listener TES, with his different strategy, perceived far more reduced syllables, and far fewer stressed syllables than the other two listeners.

Shown in Figures 3 (for talker ASH) and 4 (for talker GWH) are the syllable-by-syllable perceptions of stress levels. The majority decisions of all three trials by listener WAL were combined with those of the first trials by listeners MFM and TES, to yield overall impressions of the relative stressedness of each syllable. Plotted for each of the syllables in the Rainbow Script are the number of <u>stressed</u> judgments minus the number of <u>reduced</u> judgments, for the three listeners. <u>Unstressed</u> judgments were assigned values of zero. Cases where the <u>reduced</u> judgment of listener TES cancelled a <u>stressed</u> judgment by another listener are shown by double-ended arrows (↕). For example, in the syllable <u>sun</u>, all three listeners heard it as <u>stressed</u>, so that a value of +3 was plotted; if two perceived a syllable as <u>reduced</u>, and the other perceived it as <u>unstressed</u>, (such as with -<u>sion</u> in the second sentence), a value of -2 resulted. The syllables which were most definitely <u>stressed</u> (i.e., perceived by all listeners as stressed) thus were at the top of the scale; those definitely perceived as <u>reduced</u> were at the bottom of the scale. (Also shown on Figures 3 and 4 are lines separating grammatical constituents, and boxes and circles surrounding various stressed syllables, to be explained in sections 3 and 4.)

3. Boundary Detection Results

A hand analysis was done on the F_o contours of the Rainbow Script, strictly following the algorithm for detecting boundaries between major syntactic constituents, but incorporating a few slight refinements for eliminating some false boundary detections. The detected boundaries are marked in Figures 3 and 4 by lines between the printed syllables they separate. As might be expected from previous studies (Lea, 1972), most (86% for ASH, 92% for GWH*) of all boundaries between major syntactic constituents, as predicted by an independent syntactic analysis, were correctly detected (but not located). Six boundaries between minor syntactic constituents were also detected in the data for talker GWH. One "false" (syntactically unrelated) boundary detection was obtained at a consonant-vowel transition for ASH, and two for GWH.

Figure 5 shows how many stressed syllables occurred in the detected constituents, for each talker. In only two of the detected constituents, for each talker, did no stressed syllable occur. These resulted from improperly placed or "false" boundaries. On the other hand, well over half of the constituents had exactly one stressed syllable in them. In a sense, the constituent boundary program may then be said to be detecting some of the stressed syllables (but not all). If each constituent had exactly one lexical word with a major-stressed syllable within it (Chomsky, 1965; Emonds, 1970), we might expect constituent detections to be thus closely associated with the presence of stressed syllables.

It would appear that the boundary detection program could be used to detect many stressed syllables. To locate the stressed syllable or syllables within each constituent, further techniques were needed.

4. Stressed Syllable Location

Many studies have shown that peak F_o, local increases in F_o (Bolinger, 1958), and energy integrals (Medress, Skinner, and Anderson, 1971) are among the best acoustic correlates of stress, at least in isolated words. Intonation studies (Armstrong and Ward, 1926) have shown that, in connected texts, F_o peaks near the first stressed syllable (the so-called "HEAD") of each breath group, and falls gradually until the last stressed syllable,

*These scores neglect Noun Phrase - Verbal boundaries, which previous studies (Lea, 1971; 1972) had shown are not reliably manifested.

after which may occur the rapid fall of an utterance-final "TUNE I" contour or the rise in F_o at the end of "TUNE II" contours (which mark "incompletion").

These concepts were incorporated in an algorithm for stressed syllable location. In the algorithm, it was assumed that each major constituent would have a TUNE I or II contour, that the peak F_o would be near the first stressed syllable in the constituent, that the first stressed syllable ("HEAD") could be located by high-energy and rising-F_o portions (bounded by dips in speech intensity at syllable limits). Other stressed syllables in the constituent are assumed to be manifested by deviations in F_o above an archetype line (a straight line on a semi-log plot) from the F_o peak to the F_o value at the end of the constituent. Again, such stressed syllables are delimited by decreases in energy.

The results of such algorithmic locations of stressed syllables were compared with the perceptions of stress. For talker ASH, 29 of the 33 detected HEADS had been perceived as stressed by two or three listeners (that is, had a Stress Score SS = +2 or +3), and two other detected HEADS were perceived as <u>stressed</u> by two listeners, yet <u>reduced</u> by listener TES. One false HEAD, due to a false boundary detection, was not perceived as stressed by any listener. Ten other syllables with SS = +2 or +3 were also found by the stressed syllable locator program, while eight were missed. Thus, for talker ASH, 83% of all syllables with SS = +2 or +3 were located, while three syllables (7% of all located syllables) perceived as stressed by two listeners, but reduced by the other, were detected, and five located syllables (10% of all located syllables) were <u>not</u> perceived as stressed by at least two listeners (SS $\leq$+1).

For talker GWH, 98% of all stressed (SS = +2 or +3) syllables were located by the algorithm, with one miss on a syllable of SS = +2, plus six false alarms (with no more than one listener perceiving the syllable as stressed). Three syllables perceived as stressed by two listeners, but reduced by TES, were also located.

Figures 3 and 4 illustrate these results in more detail. The stress score for each HEAD which was correctly detected has been boxed in (⊡); while every other (non-HEAD) syllable located by the algorithm is marked by a triangle △ around the stress score for that syllable. It is important to

realize that the stressed syllables enclosed by the squares and triangles in Figure 3 and 4 were not always the only data included within the high-energy speech portions as located by the algorithm. Extended voiced sequences, and especially sonorant sequences, may have no significant energy dips, so that sequences such as -orizon and boiling may be included in the located "stressed syllables". In fact, three cases (long round, one end, and man looks) are shown in Figure 4 where, because of no substantial energy dips between syllables, more than one stressed syllable was included in the HEAD. Such syllables could be considered located, even though they are not separated from a neighboring stressed syllable, since they still would be included in the data processed by the strategy which estimates distinctive features within the located stressed "syllables".

Thus, the large majority of stressed syllables and major syntactic boundaries were correctly established by the two algorithms based on prosodic patterns.

5. Applications to Distinctive Features Estimation

A pilot study of the differences in the effectiveness of distinctive features estimation in stressed (SS = +2 or +3) versus unstressed (SS = -1, 0 or +1) syllables was undertaken, for the data of talker GWH. For each of the 40 stressed and 23 unstressed syllables (excluding ones with the diphthongs aI or ɔI), the center position of the energy maximum was determined. Then a program for identifying whether a vowel was front or back was applied to the three time segments centered at this position, and the majority vote of the three frame decisions was taken as the front/back identity of the vowel. We may define the error rate as the percentage of such majority decisions that were front for phonemically back vowels, or vice versa. The error rate was 22% for the unstressed vowels and 8% for the stressed ones, indicating that front/back decisions were more likely to be in error in unstressed than in stressed vowels.

This is only one indication of the differences expected in distinctive features estimation within stressed versus unstressed syllables. Success in identifying obstruents is expected to be better within stressed syllables, and other vowel features (such as the distinction between high and low vowels) will undoubtedly be affected. The strategy of distinctive features estima-

tion is thus expected to be different, depending upon whether the syllable is stressed or unstressed.

The better strategies for distinctive features estimation are expected to be substantially aided by the automatic detection of syntactic boundaries and the location of stressed syllables, as described in this paper. In addition, boundaries and stressed syllable information may be useful in syntactic parsing and other aspects of speech recognition.

ACKNOWLEDGEMENT

The authors are indebted to George W. Hughes and Kung-Pu Li of Purdue University for providing the speech data analyzed in this research. The procedure for perceptual judgments of stress is a modification of that used by Hughes, Li, and Snow (1972).

REFERENCES

ARMSTRONG, L. E. and WARD, I. C. (1926), Handbook of English Intonation. Cambridge: Heffer (2nd Edit.).

BOLINGER, D. (1958), A Theory of Pitch Accent in English. Word, vol. 14, p. 109.

CHOMSKY, N. (1965), Aspects of the Theory of Syntax. Cambridge, Mass.: M.I.T. Press, Chapter 2.

EMONDS, J. E. (1970), Root and Structure Preserving Transformations, Ph.D. Thesis, Linguistics Dept., M.I.T.

FAIRBANKS, G. (1940), Voice and Articulation Drillbook. New York: Harper and Row.

HUGHES, G. W., LI, K.-P, and SNOW, T. B. (1972), An Approach to Research on Word Spotting in Continuous Speech, Proc. 1972 Conf. on Speech Communication and Processing, Newton, Mass., pp. 109-112.

LEA, W. A. (1971), Automatic Detection of Constituent Boundaries in Spoken English, J. Acoust. Soc. Amer., vol. 50, p. 116(A).

LEA, W. A. (1972), Intonational Cues to the Constituent Structure and Phonemics of Spoken English, Ph.D. Thesis, School of E.E., Purdue University. Portions of that research appeared in "An Approach to Syntactic Recognition without Phonemics", Proc. 1972 Conf. on Speech Communication and Processing. Newton, Mass.: pp. 198-201.

LEA, W. A., MEDRESS, M. F., and SKINNER, T. E. (1972) Prosodic Aids to Speech Recognition. Semiannual Technical Report, ARPA Contract DAHC15-72-C-0138. Univac Report No. PX 7940, October, 1972.

MEDRESS, M. F. (1972) A Procedure for the Machine Recognition of Speech, Proc. 1972 Conf. on Speech Communication and Processing, Newton, Mass., pp. 113-116.

MEDRESS, M. F., SKINNER, T. E., and ANDERSON, D. E. (1971), Acoustic Correlates of Word Stress, Presented to 82nd Meeting, Acoustical Society of America, Denver, Colorado, October 20, (Paper K3).

TRIAL 1 BY LISTENER WAL

TRIAL 2 BY LISTENER WAL	STRESSED	UNSTRESSED	REDUCED
STRESSED	110	4	0
UNSTRESSED	7	18	6
REDUCED	0	12	97

(a)

TRIAL 1 BY LISTENER MFM

TRIAL 2 BY LISTENER MFM	STRESSED	UNSTRESSED	REDUCED
STRESSED	104	5	0
UNSTRESSED	2	70	32
REDUCED	0	5	35

(b)

TRIAL 1 BY LISTENER TES

TRIAL 2 BY LISTENER TES	STRESSED	UNSTRESSED	REDUCED
STRESSED	29	8	0
UNSTRESSED	9	44	11
REDUCED	0	12	141

(c)

Figure 1. Perceived stress levels from repetition to repetition, with results for both talkers pooled. (a) Listener WAL; (b) Listener MFM; (c) Listener TES.

JUDGMENTS BY LISTENER WAL

JUDGMENTS BY LISTENER MFM	STRESSED	UNSTRESSED	REDUCED
STRESSED	101	5	0
UNSTRESSED	13	22	45
REDUCED	0	4	63

(a)

JUDGMENTS BY LISTENER WAL

JUDGMENTS BY LISTENER TES	STRESSED	UNSTRESSED	REDUCED
STRESSED	38	0	0
UNSTRESSED	58	5	1
REDUCED	18	26	108

(b)

Figure 2. Perceived stress levels for one listener versus the other listeners, with results for both talkers pooled. (a) Listener WAL versus listener MFM; (b) listener WAL versus listener TES.

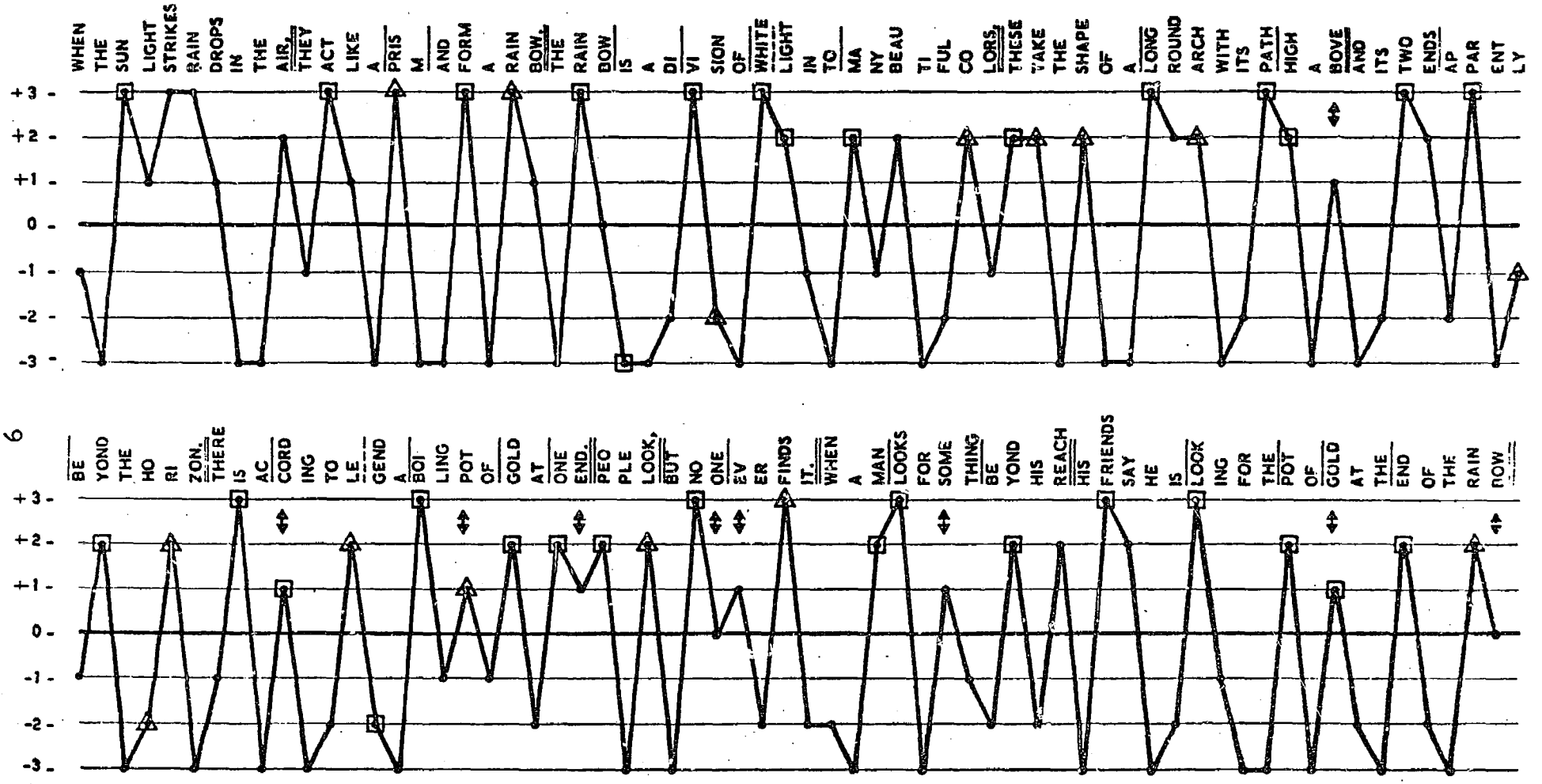


Figure 3. Summary of structural results for talker ASH reading the Rainbow Script. Plotted for each syllable is the number of judgments of the syllable as stressed minus the number of judgments of the syllable as reduced. Unanimous judgment as stressed thus yields the top value of +3, whereas judgments as reduced pull the plotted value down toward -3. Cases where the reduced judgment of listener TES cancelled a stressed judgment of another listener are shown by double-ended arrows (⇕). Solid lines between the printed syllables mark positions of correct computer-detected syntactic boundaries, double lines mark detected sentence boundaries, and dotted lines mark false (syntactically-unrelated) boundary detections. Stressed syllables located as HEADS of constituents are shown within boxes (⊡). Other "stressed" syllables located by the algorithm are enclosed by triangles (△). See text.

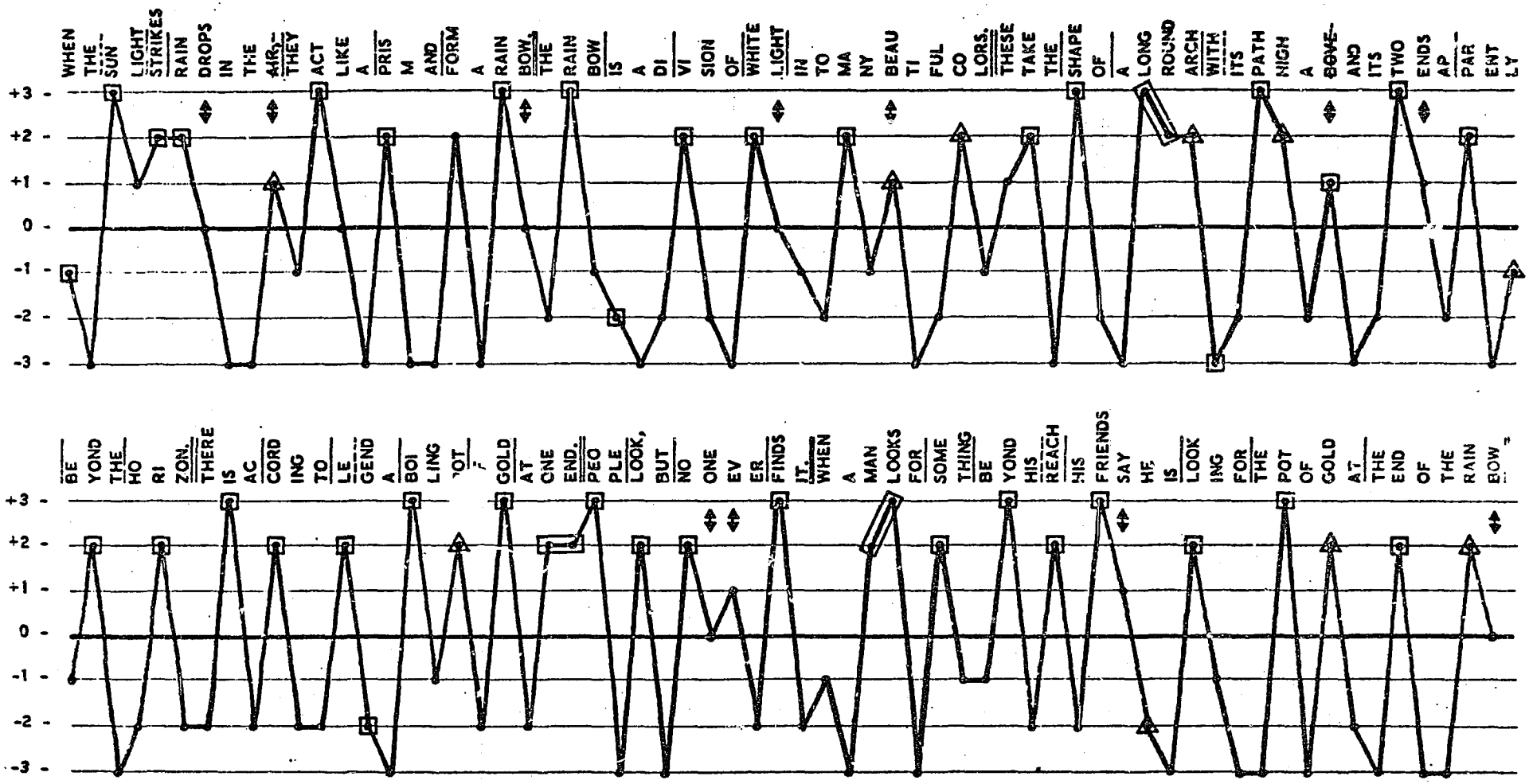
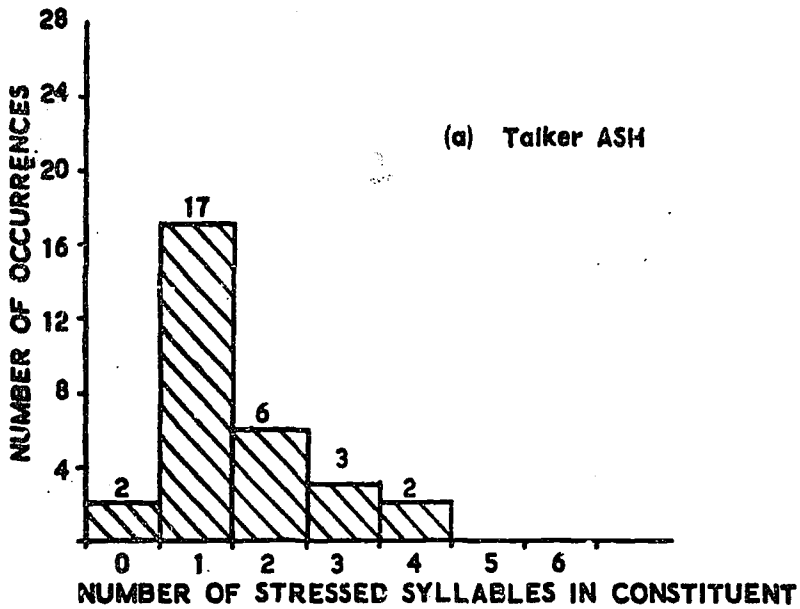


Figure 4. Same as Figure 3, but for talker GWH. Three cases are shown where two syllables (e.g., long round) were included within a single HEAD by the stressed syllable locator. See text.



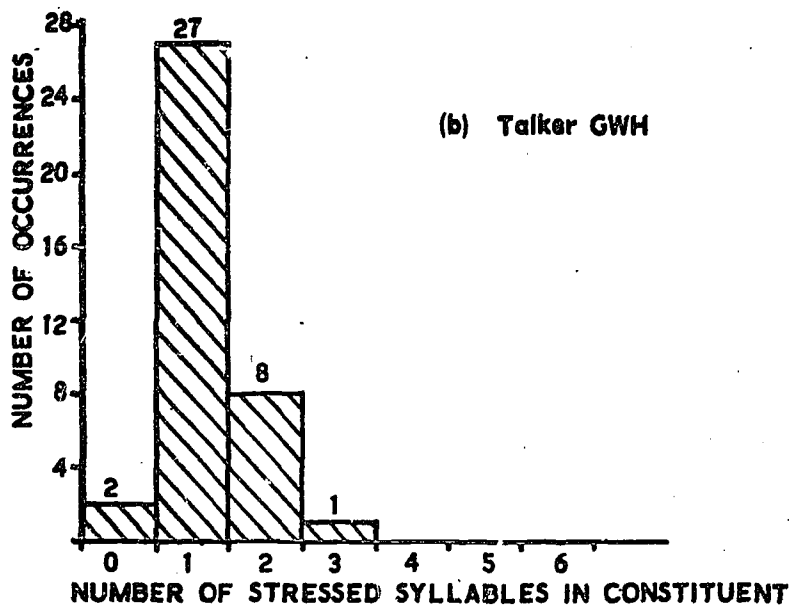


Figure 5. Frequencies of Occurrences of Perceived Stressed (SS = +2 or +3) Syllables within Detected Constituents of the Rainbow Script.