AD-767 392

PROSODIC AIDS TO SPEECH RECOGNITION. III. RELATIONSHIPS BETWEEN STRESS AND PHONEMIC RECOGNITION RESULTS

Wayne A. Lea, et al

Sperry Rand Corporation

Prepared for:

Advanced Research Projects Agency

21 September 1973

DISTRIBUTED BY:

NTIS

National Technical Information Service
U. S. DEPARTMENT OF COMMERCE
5285 Port Royal Road, Springfield Va. 22151

SPERRY UNIVAC
COMPUTER SYSTEMS

PROSODIC AIDS TO SPEECH RECOGNITION:

III. RELATIONSHIPS BETWEEN STRESS AND PHONEMIC RECOGNITION RESULTS

by

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

Defense Systems Division
St. Paul, Minnesota
(612) 456-2430

Semiannual Technical Report Submitted To:

Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209

Attention: Director, IPT

21 September 1973

Report No. PX 10430

The research was supported by the Advanced Research Projects Agency of the Department of Defense under Contract No. DAHC 15-73-C-0310, ARPA Order No. 2010. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U.S. Government.

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* Univac Defense Systems Division P. O. Box 3525 St. Paul, Minnesota 55165	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

Prosodic Aids to Speech Recognition III: Relationships Between Stress and Phonemic Recognition Results

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Semiannual Technical Report; 1 March 1973 - 31 August 1973

5. AUTHOR(S) *(First name, middle initial, last name)*

1) Wayne A. Lea
2) Mark F. Medress
3) Toby E. Skinner

6. REPORT DATE 21 September 1973	7a. TOTAL NO. OF PAGES ~~34~~ 43	7b. NO. OF REFS 10
8a. CONTRACT OR GRANT NO DAHC15-73-C-~~310~~ 0310 b. PROJECT NO c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) Univac Report No. PX 10430 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)* None	

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Advanced Research Projects Agency 1400 Wilson Boulevard Arlington, Virginia 22209

13. ABSTRACT

A strategy is being implemented for acoustic aspects of speech recognition, whereby prosodic features are used to detect boundaries between phrases, then stressed syllables are located within each phrase, and a partial distinctive features analysis is done within stressed syllables. Programs for fundamental frequency tracking and detection of syntactic boundaries have been improved. Frequency-limited sonorant energy functions, spectral derivatives, and other parameters for segmental analysis have been developed. Several algorithms are being investigated for locating stressed syllables in continuous speech. Preliminary experiments have shown some success in locating sibilants and determining their places of articulation. Partial distinctive features analysis on stressed vowels has been attempted. Location of stop consonants and sibilants, and sibilant place of articulation determination have been more successful in stressed syllables. Studies are being conducted on the relative successes of vowel and obstruent categorizations in stressed, unstressed, and reduced syllables, for data reported by participants at the C-MU Segmentation Workshop. These studies in segmental analysis, and companion studies in stress perception and automatic location of stressed syllables, are being conducted on 31 ARPA Sentences, but later work will be based on new speech texts now being designed. Further work will involve continued applications of prosodic features to distinctive features estimation, plus prosodic aids to syntactic parsing.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Speech Perception						
Speech Recognition						
Speech Analysis						
Linguistic Stress						
Prosodies						
Prosodic Features Extraction						
Syntactic Boundary Detection						
Distinctive Features Estimation						
Phonemic Recognition						

PREFACE

This is the third in a series of reports or Prosodic Aids to Speech Recognition. The first report, subtitled "I. Basic Algorithms and Stress Studies", appeared 1 October 1972, as Univac Report No. PX 7940. (The subtitle did not appear on all copies of that report.) The second report, subtitled "II. Syntactic Segmentation and Stressed Syllable Location", appeared 15 April, 1973, as Univac Report No. PX 10232.

This research was supported by the Advanced Research Projects Agency of the Department of Defense, under Contract No. DAHC15-73-C-0310, ARPA Order No. 2010. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U. S. Government.

SUMMARY

Sperry Univac is continuing its implementation and testing of a strategy of speech recognition, whereby certain acoustic features (called "prosodic features") are used to segment the speech into grammatical phrases and to identify those syllables that are given prominence, or stress, in the sentence structure. Then, partial distinctive features analysis is to be done within each stressed syllable and wherever else reliable segmental analysis can be readily accomplished. An algorithm has previously been developed for marking phrase boundaries at the bottoms of fall-rise valleys in fundamental frequency (F_0) contours (cf. Lea, Medress, and Skinner, 1972b). A refinement in that computer program, as described in this report, eliminates one common source of false boundary detections.

An algorithm has also been devised for locating stressed syllables, based on local increases in F_0 and large integrals of energy within a syllable (Lea, Medress, and Skinner, 1973). Implementation of this algorithm as a FORTRAN program is now in progress. In addition, several alternative methods of stressed syllable location are being implemented, for comparison with this previously-described algorithm. (See Appendix B.)

These algorithms for syntactic segmentation and stressed syllable location require fundamental frequency and energy data as input information. The fundamental frequency tracker uses an autocorrelation technique, which has recently been revised to involve an absolute addition method of computation rather than multiplication, plus an autocorrelation of only the first half of the time window with the whole window. These revisions reduce computation time and are expected to be more efficiently implemented in real-time hardware. Some adjustments of thresholds in fundamental frequency tracking have also reduced the likelihood of erroneous F_0 values being obtained, but at the expense of occasionally not assigning an F_0 value in time segments that are apparently voiced.

Two frequency-delimited energy functions (60 to 3000 Hz and 650 to 3000 Hz) have been incorporated to provide means for segmenting speech into syllables. The 60-3000 Hz energy function has been used in conjunction with the refined F_0 data to provide improved results in locating the nuclei of stressed syllables. Other functions, such as a ratio of low-frequency to high-frequency energy, a very low frequency energy function, and a spectral derivative, have been incorporated to provide voicing decisions and means for sibilant and stop location.

In conjunction with the Carnegie-Mellon University Segmentation Workshop, 31 ARPA test sentences were subjected to these analysis tools, to provide data about voiced portions of speech, locations of stressed syllabic nuclei, and syntactic boundaries. Thirteen of these sentences had previously been processed (Lea, Medress, and Skinner, 1973; Lea, 1973a). Listeners were also asked to indicate, for each syllable in these sentences, whether they perceived that syllable as stressed, unstressed, or reduced. About 86% of the syllables perceived as stressed by the listeners were correctly located by a hand analysis with the stressed syllable location procedure. This agrees with previous location scores for other texts (Lea, 1973a). Studies of differences between algorithmic locations and stress perceptions, and of confusions between stress perceptions from time to time and listener to listener, are being conducted, and will be reported in a forthcoming paper (see Appendix A). To aid in such analyses, an automatic procedure is being developed for comparing times of algorithmically located "stressed syllables" with perceptions, and for providing confusion matrices and majority votes from various perception trials by several listeners.

A crucial assumption of the Sperry Univac speech recognition strategy has been that consonants and vowels should prove to be easier to accurately distinguish or categorize in stressed syllables than in unstressed or reduced syllables. Preliminary experiments in segmental analysis at Sperry Univac, plus extensive analyses of results from the Carnegie-Mellon University Segmentation Workshop, are permitting the testing of this hypothesis. Partial results from part of the Segmentation Workshop data suggest that vowels are, in fact, more reliably categorized (as front/central/back, high/mid/low, or rounded/unrounded, etc.) in stressed syllables than in unstressed or reduced syllables. Complete results for the relative success in categorization of vowels and obstruents in stressed, unstressed, and reduced syllables will be presented in a forthcoming paper (see Appendix C).

Preliminary studies in segmental analysis at Sperry Univac have shown that the front/back and high/low features of "steady state" regions of stressed vowels are accurately determined from simple spectral measurements. Sibilants (or coronal strident fricatives) were located for 91% of their occurrences in stressed syllables, 86% in unstressed syllables, and 66% in reduced syllables, for 31 ARPA test sentences. This was based on simple threshold conditions on the ratio of low to high frequency energy. Place of articulation for sibilants (for example, whether /s/ or /ʃ/ was spoken) was also correctly determined for 89% of the located sibilants, using a two-coefficient linear predictive analysis. Location of stop consonants from simple tests for low energy (silence) followed by a region of high spectral derivative (indicating a stop burst) yielded correct location of 46% of the stops in stressed syllables, 26% of the stops in unstressed syllables, and 22% of the stops in reduced syllables.

For these preliminary stop and sibilant location experiments, an analysis showed that higher percentages of prevocalic consonants were located than for postvocalic consonants. Higher percentages of single stops were located than for stops within consonant clusters. The highest percentage of stops locations was for prestressed single stops.

All these results suggest that phonemic categorizations are indeed most successful (at least with the preliminary techniques tested) in stressed syllables, and that sibilants may provide fairly robust phonemic information, even in the unstressed or reduced syllables of continuous speech.

These preliminary studies of segmental analysis, including the effects of stress, consonant clustering, and position within the syllable, will be continued, using increasingly more sophisticated algorithms and further segmental data. Voicing decisions, nasal detectors, formant tracks, and other analysis tools will be investigated. In addition to further studies with the 31 ARPA sentences, and some studies with other texts previously processed at Sperry Univac, studies will be done with the texts which are specifically being designed to isolate prosodic, syntactic, and phonetic effects.

The design of an extendable set of speech texts has begun. This set of texts will provide controlled environments in which specific effects of sentence type, syntactic constructions, intonation contours, stress patterns, and phonetic sequences may be studied. Sentences with only sonorant sounds in them are being devised, to eliminate local fundamental frequency variations, that result from voiced and unvoiced obstruents. Other sentences with unvoiced consonants in syllabic structures will provide easier syllabication than all-sonorant sentences do. Simple sentence structures (originally, without embeddings) are being selected, to study various effects of syntactic structures. These texts will be recorded by several talkers and processed through the available prosodic and segmental analysis routines.

A new speech research facility is being implemented to provide faster and more powerful speech analysis tools, including a hardware fast Fourier transform processor, speech synthesis facilities, and a Very Distant Host connection to the ARPA Network.

TABLE OF CONTENTS

1. INTRODUCTION

This is a report on work currently in progress in the Univac Speech Communications Group, under contract with the Advanced Research Projects Agency (ARPA). As a part of ARPA's total program in research on speech understanding systems, the research reported herein is concerned with extracting reliable prosodic and distinctive features information from the acoustic waveform of connected speech (sentences and discourses). Studies are being concentrated on problems of detecting stressed syllables and syntactic boundaries, then doing distinctive features analysis within stressed syllables.

At Univac, the viewpoint is that versatile speech recognition will proceed by making use of reliable information in the acoustic data, in combination with early use of linguistic regularities. As has been outlined in a previous report (Lea, Medress, and Skinner, 1972a), recognition is to be accomplished by using prosodically-detected stress patterns and syntactic structure in aiding a partial distinctive features estimation procedure. Prosodically-detected syntactic structure will also be used to aid syntactic parsers and semantic processors.

Prosodic cues to sentence structure, and prosodic aids to the location of reliable acoustic phonetic information, have been given little or no attention in previous speech recognition efforts. The strong motivations for the use of prosodic patterns in speech recognition procedures were thus presented in some detail in an earlier report (Lea, Medress, and Skinner, 1972a, section 2). Improvements in the Univac facilities for extracting prosodic features, spectral data, and formants, and a program for detecting boundaries between syntactic phrases (constituents), were described in a subsequent report (Lea, Medress, and Skinner, 1973). Extensive experiments were also described in that report, which were conducted to: (1) determine the success of detecting boundaries between major syntactic units from fall-rise patterns in fundamental frequency contours; (2) determine listeners' abilities to perceive stressed, unstressed, and reduced syllables in read texts and spontaneous utterances; and (3) determine the success of locating stressed syllables by an algorithm which used rising fundamental frequency and high energy integral as major acoustic correlates of stressed syllables in the constituents delimited by the boundary detector.

This previous work provided abilities to detect about 90% of all major syntactic boundaries from acoustic data, to locate 85% or more of the stressed syllables in connected speech, to provide reliable results about listeners' perceptions of stress levels, and to provide basic parameterization tools such as linear prediction, formant tracking, fundamental frequency tracking, and energy contours. It was assumed that stressed syllables would provide the most reliable information about phonemic content of an utterance and thus, when good distinctive features estimation procedures were developed (presumably based on the available parameterization techniques), they would work best in the stressed syllables. An essential remaining task was to implement the algorithm for stressed syllable location as a computer program, since the previous experiments had been based on hand analysis of energy and fundamental frequency contours. These new speech analysis tools were to be tested on extensive speech data, including new speech texts designed to specifically isolate effects of intonation, stress, lexical content, phonetic sequences, and syntactic structures.

The recent modifications and additions to prosodic and distinctive features extraction procedures, which will be described in section 2, provide improved fundamental frequency tracking, two new "sonorant energy" functions, voicing decisions independent of fundamental frequency tracking, and elimination of about half of the "false alarms" in syntactic boundary detection. With techniques similar to those presented at the Carnegie-Mellon University Segmentation Workshop, significant success in vowel classification and strident fricative location has been attained in some preliminary experiments.

Implementation of the stressed syllable location algorithm described in an earlier report (Lea, Medress, and Skinner, 1973) is in progress, along with several alternative ways of locating stressed syllables from energy and fundamental frequency contours, to be described in section 3. In addition, algorithms are being written for automatic comparison of stress perceptions from trial to trial, listener to listener, etc., plus comparison between perceptions and automatically-located "stressed syllables". Perception tests have been extended to include more ARPA test sentences.

A major new effort which dramatically justifies the Univac strategy (that is, speech recognition by early analysis of stressed syllables) is described in section 3.3. Segmentation and classification of vowels and consonants in continuous speech is shown to be more successful in stressed syllables, for each of five different segmentation and classification procedures reported at the Carnegie-Mellon University Speech Segmentation Workshop. This extensive study, when completed, should firmly demonstrate the validity of what has previously been a general assumption of more reliable decoding in stressed syllables.

The design of test sentences has begun, for isolating effects due to syntactic structures, stress patterns, lexical insertions, and phonetic content (see section 3.4).

Conclusions and references will be given in sections 4 and 5. Appendices are included which contain the abstracts of three papers to be presented to the Acoustical Society of America.

2. SYSTEMS FOR EXTRACTING PROSODIC AND DISTINCTIVE FEATURES

2.1 Parameter Extraction Procedures

Some modifications have been made to the fundamental frequency (F_0) processing technique (Lea, Medress, and Skinner, 1973, Appendix) to increase speed and accuracy of computation. The autocorrelation vector is now computed using absolute addition as opposed to multiplication, and contained (first half of the analysis window correlated with the entire window) versus circular autocorrelation. Thus, the AUTOCORRELATION EQUATION is now formulated as follows:

$$A_j = \sum_{i=1}^{N/2} | C_i + C_{i+j-1} | \; ; \; j = 0_L, 0_L + 1, \ldots, 0_M$$

Obviously, in the multiplication formulation, if either factor of a term in the product is zero, the term will be zero. This is also true in the logical implementation of the absolute addition formulation. Techniques which are more sophisticated (both in concept and implementation) might further enhance the absolute addition formulation (for example: if the two factors of a term differ in sign, assign a value of zero to the term); however, such enhancements do not appear to be necessary at this time.

Both formulations (circular multiplication and contained absolute addition) for the AUTOCORRELATION EQUATION produced very similar autocorrelation functions and resultant F_0 time functions when tested on some of the ARPA sentences. This is most likely due to the stability of the technique (i.e., the freedom permitted in the computational definition of autocorrelation) and the effect of the fifty millisecond analyzing time window (usually several fundamental periods per window) averaging out small variations in the different formulations. Absolute addition is naturally more attractive due to faster computation speed and ease of potential hardware implementation, and because the dynamic range of the numbers involved is reduced. The contained autocorrelation function has a flat slope (autocorrelation magnitude vs offset), unlike the circular autocorrelation function which has a variable slope dependent upon the alignment of the signal periodicity and the analyzing window. The fundamental frequency processing is about 10% faster using absolute addition as opposed to multiplication. Using contained autocorrelation, the processing is approximately 14% faster than with circular autocorrelation. The total savings in computation time for the contained absolute addition formulation as opposed to circular multiplication is about 22%.

Another change to the F_0 processing algorithm was to make the frequency search limits exclusive. That is, should the maximum autocorrelation offset be conincident with either offset corresponding to the bounds on F_0, the true maximum of the autocorrelation function may be outside the range of the F_0 offset limits. If this occurs, the time segment is declared unvoiced.

The initial energy thresholding technique has also changed from requiring the entire analyzing time window energy to exceed a threshold to necessitating that both the first and second halves of the time window be in excess of the threshold minus three decibels. This results in more precise F_0 onsets and offsets.

A valid maximum of the autocorrelation function within the offset search limits must now exceed 45% of the function at zero offset (previously this threshold was 28%). This threshold increase rules out some valid F_0 responses (expecially during rapidly changing F_0) and most invalid F_0 responses. A voicing function may be instituted to at least indicate the binary decision of voicing in these marginal areas.

The program for detecting syntactic boundaries from fundamental frequency contours has also been modified, to require that each new maximum or minimum in the F_0 contour must last for at least 20 ms (two time segments). This requirement that F_0 values be beyond each threshold of 7% rise or fall for at least 20 ms should eliminate about one-half of the false alarms in boundary detection (Lea, 1972, pp. 67-70).

In addition to these improvements in F_0 tracking and boundary detection, various frequency-delimited time functions have been incorporated for use in segmental analysis. Frequency spectra were computed every 10 ms for a 25.6 ms time segment using the technique of Linear Prediction (L-P). Prior to L-P analysis, each time segment was software preemphasized and Hanning windowed. Fourteen predictor coefficients were used in the L-P process, and Fourier transforms were performed on the jw-axis using a transform size of 256. The resultant spectra were then used in computing frequency-delimited energy measures. For each spectrum, dB were converted to power over the desired frequency limits, the power values were summed, and the sum was converted back to dB. This yielded a time function which reported a value of frequency-delimited energy every 10 milliseconds.

Total energy (60 to 5000 Hz), Sonorant energy (60 to 3000 Hz) and High Frequency Sonorant energy (650 to 3000 Hz) functions provide various degrees of syllabic segmentation of continuous speech. The Total energy function does not syllabicate effectively since it remains relatively high even during obstruents. The sonorant energy function performs best in isolating syllabic sonorant clusters; and the High Frequency Sonorant energy function further separates the vowel nucleus of a sonorant cluster from surrounding nasals, liquids and glides. Very Low Frequency energy (60 to 100 Hz) and the Ratio of Low to High Frequency energy (60 to 900 Hz/3000 to 5000 Hz) functions are being investigated for possible use as voicing determinants to augment the F_0 processing. A spectral derivative, which indicates the similarity of successive spectra, was computed over the broadband frequency range from 60 to 5000 Hz.

The 31 ARPA Sentences used in the Carnegie-Mellon University Segmentation Workshop were processed using the improved F_0 tracking algorithm, the new frequency-delimited energy functions, the revised algorithm for boundary detection, the alternative voicing detectors, and the spectral derivative. Analysis of these results is now in progress, as will be outlined in sections 2.2 and 3.3.

2.2 Studies On Distinctive Features Extraction

A survey of segmental analysis techniques (including those presented at the Carnegie-Mellon University Segmentation Workshop) has been conducted, and work has been initiated to conjoin segmental recognition with the philosophy that stressed syllables and other minimally coarticulated sounds, such as sibilants, are most reliably encoded. Some preliminary experiments have been conducted (on the 31 ARPA Sentences), including vowel classification, sibilant location and place of articulation determination, stop location, and nasal location.

A time reference, within a stressed syllable nucleus, for performing vowel categorization may be defined as the instance of minimum spectral derivative, minimum second formant slope, minimum zero crossings, maximum total energy, or maximum sonorant energy. These acoustic parameters relate to the notions of steady-stateness and nearest approach to target (phonemic characteristic) attainment. Places of minimum second formant slope and maximum total energy have briefly been investigated (for the first 4 of the 31 ARPA Sentences) as areas to perform stressed vowel front/back, high/low classification. The results are encouraging, since most of the stressed vowels were correctly categorized.

Applying an algorithm which required the Ratio of Low to High Frequency energy (60 to 900 Hz/3000 to 5000 Hz) to be less than a threshold of minus 20 for at least 40 ms, 86% (74) of the 90 sibilants (/s, z, ʃ, ʒ , tʃ , dʒ/) were correctly detected in the 31 ARPA sentences, while only two false alarms (/t/'s in sentence RC8) were reported. Among the sibilants not located are those in sentences CV1300 and CV2300, in which the sibilant energy was observed on the spectrogram to be above 5 KHz.

Two separate techniques were used to determine place of articulation for the 74 of 90 sibilants correctly located in the 31 ARPA Sentences: (1) frequency of the maximum spectral peak (14 coefficient L-P), and (2) the single-pole (2 coefficient L-P) frequency. The categorization criteria were: (a) less than 3300 Hz is palatal, (b) greater than 3700 Hz is alveolar, and (c) between 3300 and 3700 Hz is undecided. The results were as follows. For the frequency of maximum amplitude spectral peak, place of articulation was correctly determined for 60 sibilants, while 12 were undecided, and place was incorrectly determined for 2 sibilants. For the single-pole frequency, place of articulation was correctly determined for 66 sibilants, while 7 were undecided and place was incorrectly determined for 1 sibilant. The incorrect place of articulation assignments occurred for the /ʃ/ portion of the affricate /tʃ/, which was bounded on both sides by the stop /t/ (e.g., "EACH TYPE"), which has the alveolar place of articulation and thus may have denied the palatal form for the /ʃ/.

The high percentage of sibilant location and accurate place of articulation determination for the 31 ARPA Sentences, despite their variety of speakers and recording conditions, suggests that sibilants are indeed robustly encoded in the speech signal.

Eighty-one of the 205 phonemic stops occurring in the 31 ARPA Sentences were correctly located by an algorithm requiring a spectral derivative in excess of a threshold of 600 (to represent the concept of 'stop burst') preceded by at least three 10 ms frames each having total energy less than 50 dB, thus indicating a stop closure). This technique also incorrectly reported 23 non-stops, of which 4 were phonetic oral stops and 5 were glottal stops. Other false alarms occurred at abrupt sonorant onsets and thus perhaps a modification to the algorithm requiring formant transitory movement during the time period immediately following the stop release will remove some of these false alarms in addition to eliminating the detection of the glottal stops. Phonetic stops which are not phonemic are probably best resolved at a non-segmental level of analysis.

Several parameters are being investigated as possible nasal detectors, including: significant differences between the Sonorant energy and High Frequency Sonorant energy functions, large Ratio of Low to High Frequency energy, low spectral derivative, low first formant frequency, and high value of Low Frequency energy.

Success in segmental analysis for these experiments can be correlated with perceived syllable stress. Sixty-six percent of all located stops were in stressed syllables. Also, 46% of all stops in stressed syllables were located, while 26% of all stops in unstressed syllables, and 22% of all stops in reduced syllables, were located. Thus, stop location is better in stressed syllables (at least with the present preliminary location scheme). Sibilants, on the other hand, show more reliable location even in stressed and reduced syllables. Sibilants in stressed syllables were correctly located in 91% of their phonemic occurrences in the 31 ARPA sentences, while sibilants were located in 86% and 66% of their occurrences in unstressed and reduced syllables, respectively.

Whether a consonant occurs in a prevocalic or a postvocalic position within a syllable, and whether it occurs as a single consonant or within a consonant cluster, might also be expected to affect phonetic location scores. An analysis was done on the separate effects on stop location of prevocalic versus postvocalic positions, single versus clustered consonants, and stress levels. A slightly higher percentage (5% higher) of prevocalic stops were located than for postvocalic stops. Higher percentages of single stops (by about 15%) were located than for stops within clusters. As noted before, stops in stressed syllables were located in about twice the percentages of the occurrences as stops in reduced or unstressed syllables were. The highest percentage of stop locations was 60%, in "prestressed" single stops (just before stressed vowels).

Similarly, preliminary studies of the interacting effects of prevocalic versus postvocalic position, clustering versus single consonant positions, and stress were also done for sibilant locations. Higher percentages (over 10% higher) of prevocalic sibilants were located than for postvocalic sibilants. There was no clear evidence of clusters yielding different sibilant location scores than single sibilants yielded. As noted before, location scores increased as stress level of the syllable increased, but were consistently higher than location scores for stops.

All these experimental results, while quite preliminary and likely to be affected by the exact procedures for segmental recognition, do suggest that stressed syllables are most reliably decoded, and that sibilants may provide fairly robust phonemic information, even in the unstressed or reduced syllables of continuous speech.

2.3 Improvements in the Interactive Speech Research Facility

The Univac speech research facility that is being used in this investigation has been described in an earlier report (Lea, Medress, and Skinner, 1972a). A new and enhanced research facility is now being implemented to provide a much faster and more powerful speech processing system, as shown in Figure 1. The heart of this system is a Univac 1616 computer with 48 kilowords of 16-bit memory, a 1.2 microsecond cycle time, and 16 I/O channels controlled by a separate input/output controller. In addition to improved versions of the kinds of peripherals found on the present research facility, the new system will have a hardware fast Fourier transform processor (HFFT), a digital speech synthesizer, and a graphical input tablet for synthesizer control.

The new research facility will have several important advantages over the old one. Of course, the HFFT will perform fast Fourier transform and similar operations very quickly. In addition, the memory will be contained in two separate memory banks, each of which will have multiple access ports. As a result, both the 1616's central processor unit and the HFFT will be able to operate simultaneously and independently. Other advantages come from the operating system for the new facility, which is being designed to permit efficient utilization of the facility's resources by overlapping processing and I/O whenever possible, and by providing file-structured storage on the disk storage subsystem.

In a separate, internally-funded project at Univac, a Very Distant Host interface is being implemented to connect the new speech research facility and other devices (initially, a teletype) to the ARPANET. An available Univac 1218 computer will serve much like the usual Terminal Interface Message Processor (TIP), but will not have packet forwarding and routine responsibilities, since it is at the end of a Very Distant Host circuit.

All of the 1218 software, including a Network Control Program (NCP), Reliable Transmission Package (RTP), and local terminal handlers, has been coded and partially debugged. The necessary interface hardware, which has as its main function the handling of the cyclic redundancy check and the

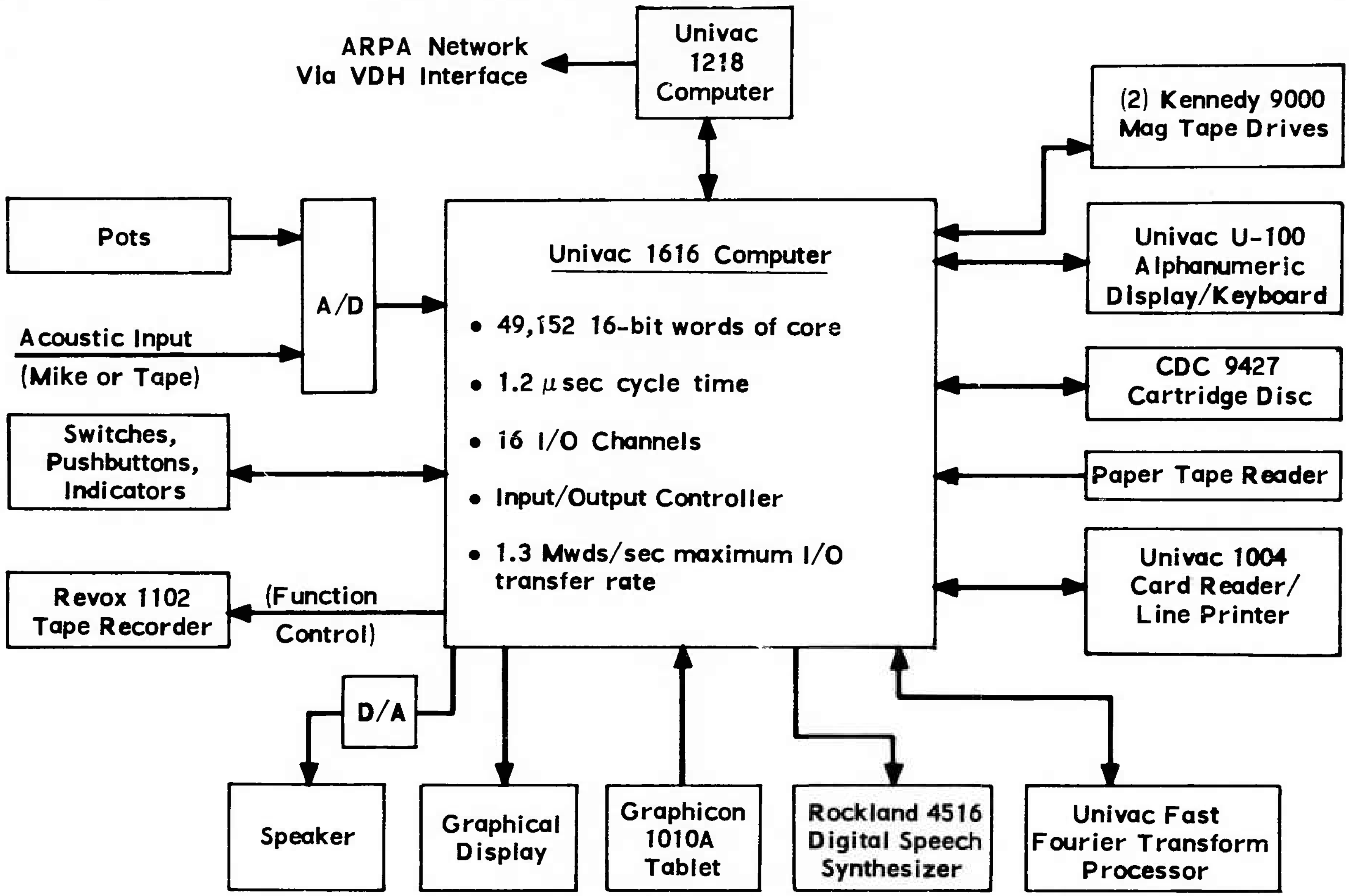

ARPA Network
Via VDH Interface
Univac 1218 Computer
(2) Kennedy 9000 Mag Tape Drives
Pots
A/D
Acoustic Input
(Mike or Tape)
Univac 1616 Computer
• 49,152 16-bit words of core
• 1.2 μsec cycle time
• 16 I/O Channels
• Input/Output Controller
• 1.3 Mwds/sec maximum I/O transfer rate
Univac U-100 Alphanumeric Display/Keyboard
CDC 9427 Cartridge Disc
Switches, Pushbuttons, Indicators
Paper Tape Reader
Revox 1102 Tape Recorder
(Function Control)
Univac 1004 Card Reader/ Line Printer
D/A
Speaker
Graphical Display
Graphicon 1010A Tablet
Rockland 4516 Digital Speech Synthesizer
Univac Fast Fourier Transform Processor

Figure 1. Block Diagram of the New and Enhanced Speech Research Facility

transparent transmission conventions, has been checked out with both the 50 kilobit modem and 1218 computer. On-line network testing will begin shortly, and the entire network connection should be available for use by November of this year.

After some initial experience is gained with the ARPANET, additional local ports may be added, such as a modem for any local dial-up terminal, and connections for other local computers. The software may also be expanded to allow such higher-level protocols as the File Transfer Protocol.

With the ARPANET connection, the new speech research facility will be able to access the Lincoln Laboratories' speech data base and other contractors' programs and hardware for speech understanding research. The teletype can then be used simultaneously for interactive communication, including message sending and receiving.

3. EXPERIMENTS ON STRESSED SYLLABLE LOCATION AND PHONETIC CLASSIFICATION

3.1 Implementation of Stressed Syllable Location Algorithms

The Sperry Univac strategy for speech recognition requires demarcating constituents, finding stressed syllables, and doing a partial distinctive features analysis on the presumably reliable data within the stressed syllables. A method for demarcating constituents has been implemented (Lea, 1972, 1973b) and tested with extensive speech data (Lea, 1973a). A recent improvement was outlined in section 2.1 of this report. Investigation of methods for partial distinctive features estimation has begun, as described in section 2.2. The strategy for stressed syllable location was outlined in previous reports (Lea, Medress, and Skinner, 1972a and b; 1973), and a hand analysis showed that the algorithm successfully located about 85% of the syllables perceived as stressed by a panel of listeners (Lea, 1973a). Here we discuss work on the implementation of the algorithm and its evaluation in comparison to alternative ways of locating stressed syllables. In addition, methods will be described for automatically determining percentages of correctly located stressed syllables, misses, and false alarms, and for providing confusion matrices for comparing perception and automatic location results.

As outlined in previous reports (Lea, 1973a; Lea, Medress, and Skinner, 1972b, 1973), the algorithm used for stressed syllable location assumes that local increases in F_0 and high energy integral are the most reliable correlates of stressed syllables. The increasing F_0 near the beginning of each constituent detected by the boundary detector is assumed to be attributable to the first stressed syllable in the constituent (Lea, 1973a, section 5). A stressed "HEAD" to the constituent is thus associated with a portion of the speech which is high in energy with rising F_0, and bounded by substantial (5 db or more) dips in energy. Other stressed syllables in the constituent are expected to be accompanied by local increases in F_0. Since the usual ("archetype") shape of the F_0 contour in a constituent is a rapid rise followed by a gradual fall in F_0, we expect that local 'increases' in F_0 due to later stressed syllables will show local rises above the gradually falling F_0 contour, even if F_0 does not rise absolutely near the stressed syllable. The stressed syllable is located within a high-energy-integral region near this local rise above the archetype F_0 contour. A flowchart of this complete algorithm was presented by Lea (1973a, p. 96).

Implementation of this algorithm as a FORTRAN program began by first developing a subroutine ("CHUNK") which finds all peaks and dips in the sonorant energy function and delimits syllabic nuclei as all contiguous points within 5 db of the maximum intensity value in that "chunk" or syllable. Preliminary tests with a few files of speech data show that this subroutine finds almost all syllables, with very few "extra" chunks. Thus, good syllabication of the speech is accomplished. The only extra chunks obtained are unvoiced stop bursts or fricatives, which may be ruled out as syllabic nuclei by simple voicing and frication tests. The few occasions when more than one syllable are included in a single chunk result from lack of sufficient energy dips in intersyllabic sonorants.

The overall stress location algorithm ("STRESS") calls CHUNK to obtain syllabication results. Input data, read from cards or mass storage, include F_0 contours in eighth tones, the sonorant energy contours in dB, and the output from the syntactic boundary detector (a function which is zero except where it takes on one nonzero value at each syntactic boundary, another nonzero value at each position of maximum F_0 in a constituent, and a third nonzero value at each sentential pause).

After reading the data and obtaining the syllabication results from the subroutine CHUNK, the STRESS program then calls on subroutine INTGRL to determine the duration and energy integral of each high-intensity chunk. This energy integral information will be used later in STRESS to locate the highest-energy syllable near F_0 increases, for stressed syllable location. However, since it is available and we know from past studies (Medress, Skinner, and Anderson, 1971) that energy integral is among the best cues for stressed syllables, a study has been undertaken to determine whether stressed syllables can be accurately located using energy integral alone. Preliminary results to date, with only about 20 seconds of speech, showed that a threshold (minimum) duration of 100 ms for the chunk, or a threshold on the energy 'integral' (sum of dB values in the time segments within the chunk) of about 600 dB, located about 21 of the 23 syllables perceived as stressed by listeners, while falsely locating 5 chunks from among the 22 syllables that were not perceived as stressed.

These good preliminary results with a simple energy-integral method of stressed syllable location suggest the need for evaluating alternative simple methods for stressed syllable location, before one firmly adopts the complex archetype-contour-based algorithm which has previously been described. Consequently, the implementation of the total complex algorithm is being accompanied by studies of how well several alternative strategies work for stressed syllable location. In addition to the method which simply says that all syllabic nuclei (or chunks) with duration greater than a threshold, or energy integral greater than a threshold, are considered stressed, several methods are considered which only use F_0 increases or inflections to mark stress, and others are considered which use simple combinations of F_0 and energy cues.

A subroutine "ONLYFO" has been implemented to locate all portions of speech with rising or non-falling F_0, and to locate all portions where the slope of F_0 is increasing positively. Both such features are expected to be associated with stressed syllables (Bolinger, 1958), but the increasing slope feature allows such regions as a flat F_0 contour in the midst of a general fall to be a candidate for a stressed syllable, while excluding cases where F_0 is rising merely due to continuations of trends in surrounding stressed syllables. Subroutine ONLYFO thus provides information about the potential of stressed syllable detection from F_0 contours alone. (Another F_0 parameter, the peak F_0 in the vowel or nucleus, has been shown to be a useful stress cue in isolated words (cf. e.g. Lea, 1972, Ch. 5), but obviously is not suitable in complete sentences, where the later portions almost always have lower F_0 than earlier portions. A simple threshold on peak F_0 values could thus not work. On the other hand, a search for local F_0 maxima, surrounded by F_0 valleys, is exactly what is involved in the syntactic boundary detections used as inputs for the location of HEAD stressed syllables in the archetype-contour-based algorithm.) In general, it is probably much more difficult to accurately define the limits (beginning and ending) of a stressed syllable using F_0 alone than with the natural chunking accomplished by energy contours.

Having considered some simple techniques of stressed syllable location from F_0 contours alone and energy contours alone, we may consider possible combinations of the two types of cues. There are several possibilities short of the total complex algorithm previously used in stressed syllable location. One may select all chunks whose duration or energy integral is above a certain threshold, and whose associated F_0 contour is rising (or not falling). This constitutes location by energy contours, and subsequent selection by F_0 contours. Alternatively, one may detect possible candidates from regions of rising F_0 or increasing F_0 slope, and locate the syllables as within nearby chunks of large energy integral. If an algorithm simply detects regions of substantial rise in F_0, and locates the earliest high-energy integral chunk within that rising F_0 portion, that would be equivalent to finding all HEADs of constituents, as is to be done by subroutine HEADER of the complete archetype-contour-based algorithm. An alternative to the use of the archetype line for locating other (non-HEAD) stressed syllables in the constituent would be to look for any other chunks (between HEADs) whose durations or energy integrals are larger than some large threshold value.

All of these combinations are being investigated. In addition, subroutine HEADER has been implemented to find the HEAD of each constituent, as described in the detailed description of the original algorithm for stressed syllable location (Lea, 1973a). Subroutine OTHERS is being implemented to establish the archetype line of falling F_0, to search for local rises above the archetype, and to locate nearby high-energy-integral chunks. (See Appendix B.)

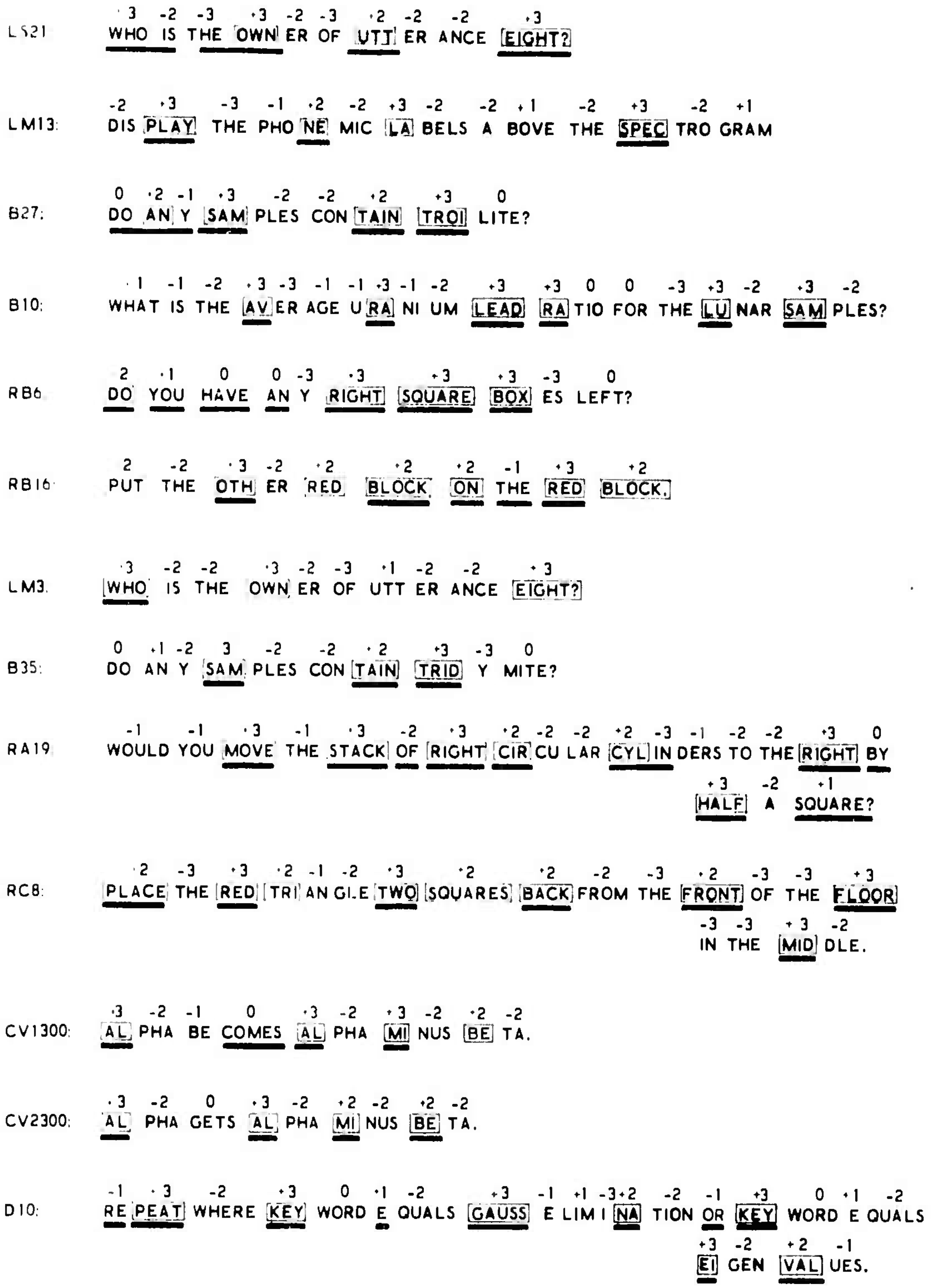
These automatic locations of stressed syllables must be evaluated in comparison with perceived stress patterns. Subroutine COMPAR is being implemented to automatically compare the times of perceived stressed syllables with the times of located "stressed syllables". Scores showing the number of instances where a location overlaps with the perceived stressed syllables will be provided, as will 'false' locations and any failures to locate syllables perceived as stressed. A subroutine CONFUS will provide tabulations of such successes and confusions, and will allow the display of confusion matrices for perception results (for repetition-to-repetition confusions, confusions from listener to listener, etc.). A related subroutine MAJORT will give majority perception results from several trials, and provide the type of stress score plots shown in previous reports (Lea, Medress and Skinner, 1972a, 1972b; Lea, 1973a).

These algorithms will be applied to the Monosyllabic and Rainbow Scripts spoken by ASH and GWH, and to the 31 ARPA Sentences analyzed at the Carnegie-Mellon University Segmentation Workshop. If results substantially agree with previous hand analyses, the next applications will be on the new designed texts.

3.2 Extensions of Stress Perception Tests

A method has previously been described for presenting recorded scripts to individual listeners, to obtain their personal judgments as to which syllables are stressed, unstressed, or reduced (Lea, Medress, and Skinner, 1972b, 1973; Lea, 1973a). These stress perception tests have been extended to include the 31 ARPA Sentences. Each listener repeated the perception test on the 31 ARPA sentences three times (with at least one week between trials). Confusions from trial to trial and from listener to listener will be described in a future report, using the automated confusion analysis techniques. Here we shall consider the overall majority decisions about the stress level in each syllable. As discussed before (Lea, 1973a, p. 22), this overall stress score is obtained by first determining, for each listener, his majority vote, from the three trials, as to the stress level of each syllable. Then the results for all three listeners were pooled, to obtain scores between +3 (for all listener's majority votes saying the syllable is stressed) to -3 (for all listener's majority votes saying the syllable is reduced).

Figures 2, 3, and 4 show the resulting stress score above each syllable in the sentences. Also shown are boxes around each syllable perceived as stressed by two or more listeners (stress score equal to +2 or +3) with the recent series of perception tests. Dark lines are shown under each portion which



```
         +3  -2  -3    +3 -2  -3    +2  -2   -2      +3
LS21     WHO IS  THE  OWN ER  OF   UTT  ER  ANCE   EIGHT?

         -2   +3     -3   -1  +2  -2  +3  -2    -2  +1    -2   +3    -2   +1
LM13:    DIS PLAY   THE  PHO  NE  MIC LA  BELS  A  BOVE  THE  SPEC  TRO GRAM

         0  +2 -1  +3    -2   -2    +2     +3     0
B27:     DO AN  Y  SAM  PLES  CON  TAIN   TROI  LITE?

         +1   -1  -2  +3 -3 -1  -1 +3 -1  -2    +3    +3  0   0   -3  +3 -2   +3  -2
B10:     WHAT IS  THE AV ER AGE U  RA NI  UM   LEAD   RA TIO FOR THE LU NAR  SAM PLES?

         2   +1    0    0  -3   +3     +3      +3  -3   0
RB6      DO  YOU  HAVE  AN  Y  RIGHT  SQUARE  BOX  ES  LEFT?

         2    -2    +3  -2  +2    +2     +2  -1   +3    +2
RB16     PUT  THE  OTH  ER  RED  BLOCK   ON  THE  RED  BLOCK.

         +3  -2  -2    +3  -2  -3   +1  -2   -2      +3
LM3.     WHO  IS  THE  OWN ER  OF  UTT  ER  ANCE   EIGHT?

         0  +1 -2  3    -2    -2    +2    +3  -3   0
B35:     DO AN  Y  SAM PLES   CON  TAIN  TRID  Y  MITE?

          -1    -1   +3   -1   +3   -2   +3   +2  -2 -2  +2  -3  -1 -2  -2   +3   0
RA19     WOULD YOU MOVE THE STACK OF RIGHT CIR CU LAR CYL IN DERS TO THE RIGHT BY
                                                          +3   -2   +1
                                                         HALF   A  SQUARE?

          +2    -3  +3  +2 -1 -2  +3    +2      +2    -2   -3   +2   -3  -3    +3
RC8:     PLACE THE RED TRI AN GLE TWO SQUARES BACK FROM THE FRONT OF THE FLOOR
                                                          -3  -3   +3  -2
                                                          IN THE  MID DLE.

         +3  -2   -1   0     +3  -2  +3 -2   +2 -2
CV1300:  AL PHA  BE  COMES  AL  PHA  MI NUS  BE TA.

         +3  -2   0    +3  -2  +2 -2   +2 -2
CV2300:  AL PHA GETS  AL  PHA  MI NUS  BE TA.

         -1  +3    -2     +3   0   +1  -2      +3   -1 +1 -3+2 -2   -1  +3    0  +1  -2
D10:     RE PEAT WHERE   KEY  WORD  E QUALS  GAUSS  E LIM I NA TION OR  KEY  WORD E QUALS
                                                                  +3  -2  +2  -1
                                                                  EI GEN  VAL UES.
```

Figure 2. Comparison of Algorithmically Located Stressed "Syllables" with Perceived Stress Patterns, for the 13 ARPA Sentences.

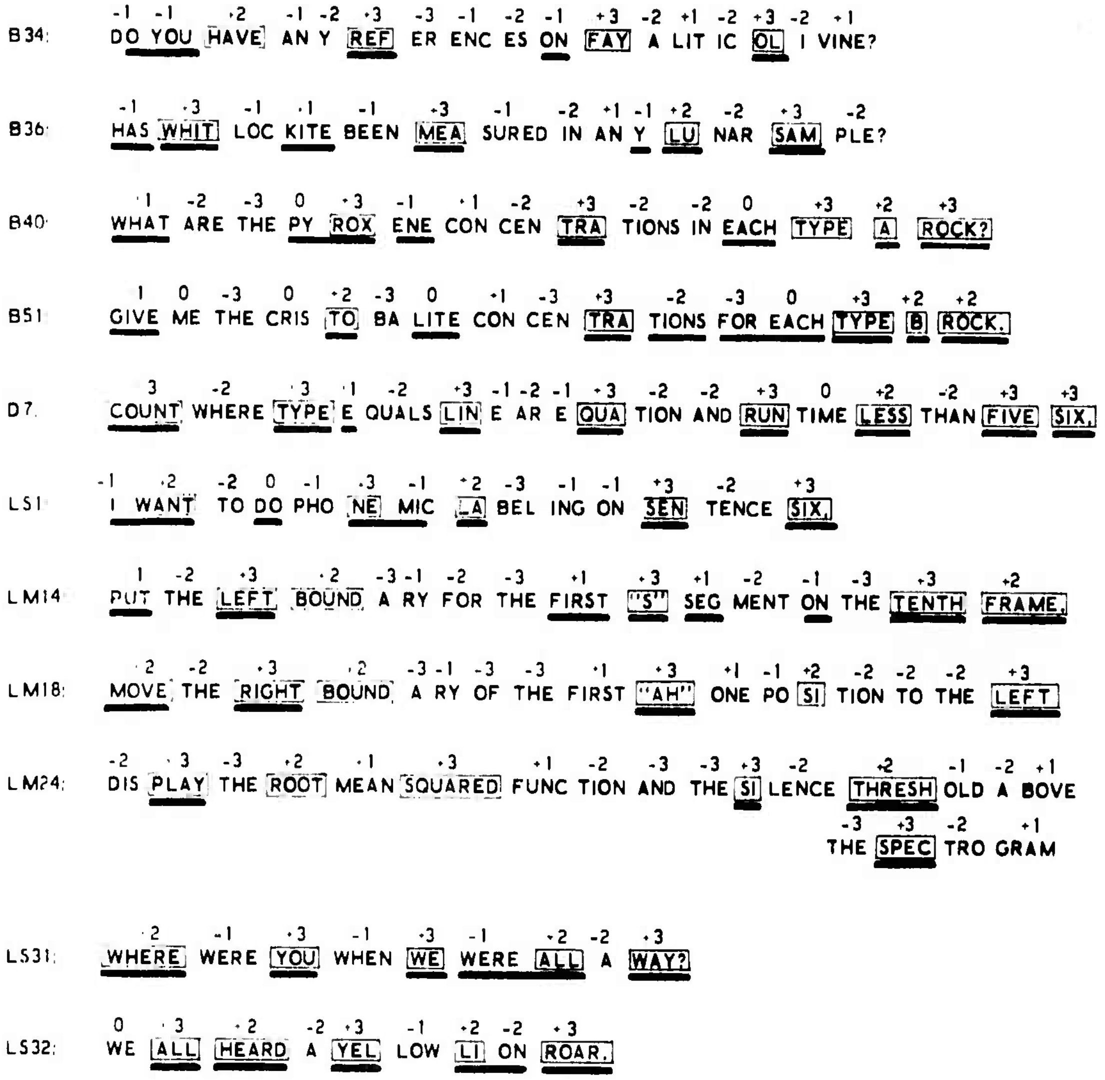

B34:
-1 -1 +2 -1 -2 +3 -3 -1 -2 -1 +3 -2 +1 -2 +3 -2 +1
DO YOU HAVE AN Y REF ER ENC ES ON FAY A LIT IC OL I VINE?
B36:
-1 +3 -1 +1 -1 +3 -1 -2 +1 -1 +2 -2 +3 -2
HAS WHIT LOC KITE BEEN MEA SURED IN AN Y LU NAR SAM PLE?
B40:
+1 -2 -3 0 +3 -1 +1 -2 +3 -2 -2 0 +3 +2 +3
WHAT ARE THE PY ROX ENE CON CEN TRA TIONS IN EACH TYPE A ROCK?
B51:
1 0 -3 0 +2 -3 0 +1 -3 +3 -2 -3 0 +3 +2 +2
GIVE ME THE CRIS TO BA LITE CON CEN TRA TIONS FOR EACH TYPE B ROCK.
D7:
3 -2 +3 +1 -2 +3 -1 -2 -1 +3 -2 -2 +3 0 +2 -2 +3 +3
COUNT WHERE TYPE E QUALS LIN E AR E QUA TION AND RUN TIME LESS THAN FIVE SIX.
LS1:
-1 +2 -2 0 -1 +3 -1 +2 -3 -1 -1 +3 -2 +3
I WANT TO DO PHO NE MIC LA BEL ING ON SEN TENCE SIX.
LM14:
1 -2 +3 +2 -3 -1 -2 -3 +1 +3 +1 -2 -1 -3 +3 +2
PUT THE LEFT BOUND A RY FOR THE FIRST "S" SEG MENT ON THE TENTH FRAME.
LM18:
+2 -2 +3 +2 -3 -1 -3 -3 +1 +3 +1 -1 +2 -2 -2 -2 +3
MOVE THE RIGHT BOUND A RY OF THE FIRST "AH" ONE PO SI TION TO THE LEFT
LM24:
-2 +3 -3 +2 +1 +3 +1 -2 -3 -3 +3 -2 +2 -1 -2 +1
DIS PLAY THE ROOT MEAN SQUARED FUNC TION AND THE SI LENCE THRESH OLD A BOVE
-3 +3 -2 +1
THE SPEC TRO GRAM
LS31:
+2 -1 +3 -1 +3 -1 +2 -2 +3
WHERE WERE YOU WHEN WE WERE ALL A WAY?
LS32:
0 +3 +2 -2 +3 -1 +2 -2 +3
WE ALL HEARD A YEL LOW LI ON ROAR.

Figure 3. Comparison of Algorithmically Located Stressed "Syllables" with Perceived Stress Patterns, for Additional ARPA Sentences Recorded by BBN, SDC, and Lincoln Laboratory.

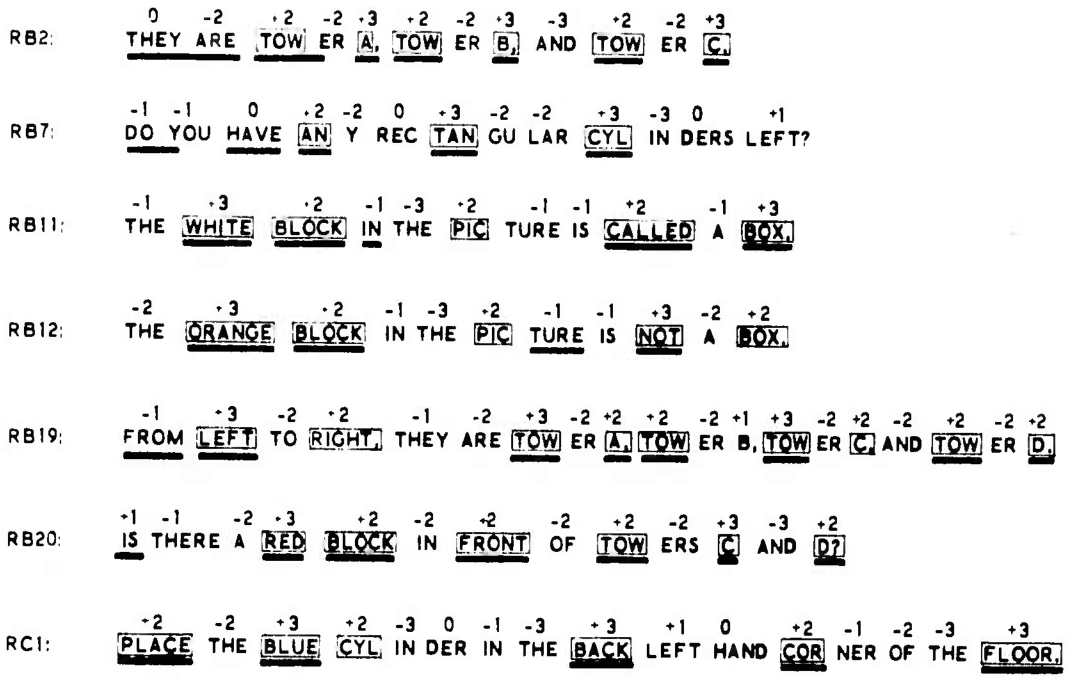

RB2:
0 -2 +2 -2 +3 +2 -2 +3 -3 +2 -2 +3
THEY ARE TOW ER A, TOW ER B, AND TOW ER C.
RB7:
-1 -1 0 +2 -2 0 +3 -2 -2 +3 -3 0 +1
DO YOU HAVE AN Y REC TAN GU LAR CYL IN DERS LEFT?
RB11:
-1 +3 +2 -1 -3 +2 -1 -1 +2 -1 +3
THE WHITE BLOCK IN THE PIC TURE IS CALLED A BOX.
RB12:
-2 +3 +2 -1 -3 +2 -1 -1 +3 -2 +2
THE ORANGE BLOCK IN THE PIC TURE IS NOT A BOX.
RB19:
-1 +3 -2 +2 -1 -2 +3 -2 +2 +2 -2 +1 +3 -2 +2 -2 +2 -2 +2
FROM LEFT TO RIGHT, THEY ARE TOW ER A, TOW ER B, TOW ER C, AND TOW ER D.
RB20:
+1 -1 -2 +3 +2 -2 +2 -2 +2 -2 +3 -3 +2
IS THERE A RED BLOCK IN FRONT OF TOW ERS C AND D?
RC1:
+2 -2 +3 +2 -3 0 -1 -3 +3 +1 0 +2 -1 -2 -3 +3
PLACE THE BLUE CYL IN DER IN THE BACK LEFT HAND COR NER OF THE FLOOR.

Figure 4. Comparison of Algorithmically Located Stressed "Syllables" with Perceived Stress Patterns, for Additional ARPA Sentences Recorded by SRI.

was located by a hand analysis with the algorithm for stressed syllable location. These algorithmic results are those determined for the Carnegie-Mellon University Segmentation Workshop. Whenever an underlined portion includes a boxed-in stressed syllable, a correct location has been obtained, A boxed-in syllable which is not underlined is a "miss" for the algorithm. Cases where an underlined portion did not include a boxed-in syllable (that is, no part was perceived as stressed by two or more listeners) are false locations of stressed syllables.

The algorithm correctly located 86% of all syllables perceived as stressed by two or more listeners. Twenty-three percent of all locations were false (that is, did not include a syllable perceived as stressed). These results were comparable to those obtained in previous hand analyses. In particular, the 13 ARPA Sentences shown in Figure 2, which yielded 86% correct locations and twelve percent false alarms in this recent hand analysis, were found to yield 80% correct location and 20% false alarms in the earlier study (Lea, 1973a, p. 62).[1] The improvements resulted from several changes in parameterization: the new conditions on F_0 tracking as described in section 2; the refinement of the boundary detector which requires F_0 maxima and minima to be of 20 ms minimum duration; and the use of a sonorant energy function, rather than the total (0-5000 Hz) energy function used in previous studies. A comparison of the perceptual and algorithmic results of Figure 2 with those previously shown for the same sentences (in figures C-10 and C-11 of Lea, 1973a, pp. 105 and 106) also shows that the sonorant energy function more precisely brackets the stressed syllable, so that underlined portions now do not as frequently include both the stressed syllable and one or more of its surrounding unstressed or reduced syllables.

Comparison of Figure 2 with the earlier ones for the 13 ARPA Sentences also shows that the majority perceptions of stress levels from the recent three trials differ somewhat from those for the earlier trials. While some

[1] Sentence B10 in the C-MU Segmentation Workshop data is actually not the same utterance as that used in the previous studies of the 13 ARPA Sentences. It apparently was a second recording (by another talker) of the same written text.

difference may have been introduced by the re-recording, digitizing, and digital-to-analog conversions involved in obtaining the second tape, most differences are presumably due to the instability of listener's perceptions from trial to trial. An analysis of confusions between the majority decisions (specifically, the stress scores) from the first three trials and those from the recent three trials showed that less than 8% of the syllables were confused between stressed (ss = +2 or +3) and unstressed (ss = +1, 0, or -1), or between unstressed and reduced (ss = -2 or -3). This compares with 13% to 19% for trial-to-trial confusions for the individual listeners in the three earlier trials, and 22% to 52% confusions from listener-to-listener on those earlier trials (Lea, 1973a, pp. 26 and 31). Obviously, the pooling of listeners and trials does reduce overall confusions, and provides more stable stress perception results.

In the preliminary study of effects of sentence type on stress level confusions, reported by Lea (1973a, pp. 40-42), it appeared from the 13 ARPA Sentences that questions tended to give more confusions than declaratives or commands. With the larger set of 31 sentences, this tendency can be tested more completely. This will be done when confusion matrices are obtained from the automated analysis now being implemented.

All these stress perception results will be reported on in the ASA paper abstracted in Appendix A.

3.3 Reliability of Phonemic Classification Results in Stressed Syllables

The availability of speech segmentation and classification results from the Carnegie-Mellon University Segmentation Workshop makes possible the determination of whether stressed syllables are more readily decoded than unstressed or reduced syllables. During the Workshop, a preliminary study was conducted on the correctness of vowel segment classifications for two sentences (LM3 and LS21) for which segmentation data, algorithmic stress locations, and stress perceptions were all available. In that preliminary study, all of the vowels in syllables located as stressed by the algorithm were correctly categorized (essentially as front/central/back, high/mid/low, and rounded/unrounded) by four of the five groups that had provided vowel identifications. Only one (10%) of all the unstressed and reduced syllables were correctly categorized by at least four of the five groups. Pooling all the results for all five groups (which is best not done in a more thorough analysis, but which suggests general trends), 90% of all categorizations were correct for vowels either perceived as stressed or located by the stressed syllable location algorithm, while only 60% of all categorizations were correct in unstressed vowels, and only 38% were correct in reduced vowels.

These results suggest that vowels are more correctly categorized, by available automatic segmentation and labelling schemes, when they are stressed. With stress perceptions now available for the 31 ARPA Sentences, and with the complete segmentation results soon to be available for those sentences, this study can be completed for all 31 sentences. In addition, some of the participants at the Workshop have agreed to provide similar segmentation data for Univac's Monosyllabic Script and Rainbow Script, recorded by two talkers (ASH and GWH). This will provide substantial evidence about the ability of a stress-location algorithm to lead one to the most readily decoded portions of speech. Effects of stress on consonant recognition will also be studied. Previous studies, such as Klatt and Stevens' (1972) studies of spectrogram reading, have shown that <u>consonants</u> are much more readily categorized in pre-stressed positions.

To make more precise the previous subjective judgments of "correctness" of segment categoriza ion results, a scoring procedure is being devised based on the number of major distinctive (or "distinguishing") features that are correctly assigned for each phone. Thus, a vowel should be located as a vowel, then assigned a positive point for each major feature correctly determined (say +1 each for determining high/mid/low and front/central/back, and an extra point for each additional clear categorization such as rounded, retroflex, etc.). A consonant should be located as a non-vowel portion, and points assigned for stop/fricative/sonorant determination, place of articulation, and such restricted features as strident/mellow, liquid-glide/nasal, etc. Points may be subtracted for each erroneous feature, such as labelling a fricative as a sonorant.

This study of segment categorizations will not involve careful study of segment boundary positions. Only the presence of a reasonably labelled segment in the region of a phone will be demanded. Other studies of segmentation accuracy could be attempted if one wanted to assess performance in placing segment boundaries.

The results of the careful analysis of segment categorizations will be summarized in a forthcoming paper to be presented at a meeting of the Acoustical Society of America. The Abstract appears in Appendix C.

3.1 Design of Extendable Texts

In previous reports (Lea, Medress, and Skinner, 1972a, 1973), we have proposed the design of an extendable set of speech texts which can isolate the effects of intonation contours, sentence types, syntactic constructions, phonetic content, and semantic structure on speech recognition facilities. Design of such texts has begun, with an expansion of goals to relate to three major purposes:

(1) Isolation of ways in which various factors (sentence type, phonetic sequence, constituent structure, stress patterns, and position in intonation contours) affect F_0 contours, syntactic boundary detection, stressed syllable location, and distinctive features estimation;

(2) On-line demonstration of specific capabilities in parameterization, syntactic boundary detection, stressed syllable location, distinctive features estimation, lexical hypothesizing, parsing, and sentence recognition; and

(3) Preliminary definition of necessary, desirable, and expendable features of "natural" languages for restricted man-computer communication with speech.

The primary objective remains that of developing a succession of sets of sentences, each set being extended from the previous set to allow more and more versatile and natural sentences for addressing a computer, while carefully controlling various features so that, by minimal contrasts between two or more sentences, one can establish exactly what it is about a sentence that causes it to yield specific prosodic patterns, phonetic recognition successes and difficulties, etc.

To date, several decisions have been made about the design of sentences which isolate one prosodic, phonetic, or syntactic factor from another. To begin with, a subset of sentences will be recorded which are entirely sonorant; that is, no fricatives, oral stops, or affricates occur anywhere in any of the sentences. This is being done to eliminate the confusing effects that obstruents have on F_0 contours. In stressed syllables, fundamental frequency will often start high after unvoiced consonants, and rapidly fall for a few centiseconds, while during voiced obstruents F_0 dips about 10%, and rises in the first part of following vowels or sonorants (Lea, 1972; 1973c). Such phonetic effects on F_0 contours interact with stress effects, so that, for example, unstressed syllables following stressed syllables may have falling contours, even if the consonant which precedes the unstressed vowel is voiced (Lea, 1972, Chapters 4 and 5, 1973c).

If one were to determine stress by rising F_0 contours such as Bolinger (1958) suggests, such phonetic influences on F_0 values and slopes would thus interfere with stressed syllable location. Similarly, such phonetic effects on F_0 contours have repeatedly caused false detections of syntactic boundaries (Lea, 1972a, p. 67-70. Lea, 1973a, p. 9 and 16).

All-sonorant utterances also are substantially constrained in terms of possible syntactic structures and lexical insertions. Articles and determiners are confined to be a, an, all, any, no, none. The only modal auxiliaries possible are will and may (not shall, must, can, would, etc.); WH-words are confined to why and when; no perfect constructions are possible (since they require have been); almost all past-tense verbs are excluded, as are passives with is or was; prepositions are confined to along, among, in, on; and the subvocabularies for adverbs, adjectives, nouns, verbs, possessives, conjunctions, pronouns, and the like are also highly constrained. A preliminary study of several technical dictionaries for aeronautical discussions, for example, showed at most a few hundred possible words in the total vocabulary. The use of all-sonorant sentences is thus one way to dramatically reduce the alternatives in lexical insertion and sentence structure, while eliminating a most troublesome interaction between phonetic and prosodic patterns.

On the other hand, syllabication from energy contours is considerably more difficult when non-vowel sonorants are the only intervocalic consonants. Consequently, for easy syllabication (and subsequent stressed syllable location), sentences are best designed to have only unvoiced consonants (such as only unvoiced fricatives) between vowels. A subset of sentences is being designed with only such vowel-unvoiced fricative alternations in all positions or certain positions in the phonetic structure. With one sentence whose structure is all sonorant, and a second sentence which has one sonorant word of the other sentence replaced by a fricative-vowel word, one can study effects of phonetic contrasts on prosodic patterns.

Also possible with such subsets of sentences with controlled phonetic structure is the determination of phonetic recognition success in various phonetic environments. Stressed /i,a,u/, which have been found to be more reliably identified than other vowels (Klatt and Stevens, 1972), will be contrasted with other vowels. Single nasals, which were found to be more readily identified than clusters or other single sonorants, will be given early attention.

The designed subsets of sentences will also include minimal pairs (or near-minimal pairs) of sentences with similar syntactic structure and phonetic content, but alternative positions of the stressed syllable within a constituent (such as stress immediately after a syntactic boundary, or one, two, or more syllables later). Such controlled contrasts may determine under what stress pattern conditions the constituent boundaries are "delayed" in their F_0 manifestation. With the same syntactic structure but alternative words whose stressed syllables are in different positions within the word, one may study lexical stress effects, in contrast to phrasal stress effects. With the same word in different positions in a sentence, one can study effects of position in the overall intonation contour on syllable duration, F_0 contours, etc.

Besides such interactions between phonetic structure, syntactic boundaries, stress patterns, and positions in the sentence intonation contour, studies can be done on the effects of sentence type and phrase structure. Approximately 60 simple syntactic structures (without sentence embeddings such as relative clauses, complement structures, or conjunction) have been selected for consideration in early analysis. These include 12 declaratives, with a subject, optional auxiliary, verb, up to two noun phrases (direct and indirect object) in the predicate, and optional adverbial phrase. Also included are six simple command structures, twelve yes/no question structures (six with and six without DO-support), and thirty WH-questions (one for each of the twelve declarative structures with the first noun phrase questioned, one for each with the second noun phrase questioned, and one for each of the six structures which have a third noun phrase which can be questioned). These structures may not all be different enough to warrant inclusion in the final selection of the designed texts. Also, adjectives, passive structures, agent deletion, adverb preposing, reflexives, anaphoric pronouns, compound nouns, conjoined noun phrases and verb phrases, relative clauses, and complement structures will be considered in the original design and later extensions of such speech texts. Negatives will also be given particular attention.

These texts will be recorded several times by several talkers, but initial tests will be confined to one repetition by two or three talkers reading the first subset of selected sentences.

If the designed sentences are to have any applicability to specific tasks of man-computer interaction, they must be indicative of the types of sentences expected in an operational speech understanding system. For this reason, questions and commands suitable for querying or commanding a machine are being given particular attention in the design of texts. For graceful extension from very restricted subsets of possible sentences to more and more versatile communications, one must consider those features which are necessary, or at least desirable, in natural man-machine interaction.

These studies should provide a series of subsets of English sentences which are increasingly more versatile while providing the controlled environments in which specific effects of phonetic, prosodic, and syntactic structure may be determined.

4. CONCLUSIONS AND FURTHER STUDIES

This report has summarized work in progress. Most studies described herein are far from completed. The improved methods for fundamental frequency tracking, sonorant energy extraction, and syntactic boundary detection are not expected to change significantly. However, studies of distinctive features estimation techniques have just begun. The preliminary studies to date have indicated that stressed syllables are the most reliably decoded portions of continuous speech, but further studies are needed. Specifically, methods of vowel categorization will be investigated further, as will methods for sibilant and stop location and categorization. New studies will be conducted on voicing decisions and nasal location.

The complete set of segmentation results for the 31 ARPA Sentences, as obtained from several participants at the Carnegie-Mellon University Segmentation Workshop, will be studied, to determine the effects of stress on the accuracy of segment categorizations. These studies will also include some studies of segment categorization in the Monosyllabic Script and Rainbow Script.

The stressed syllable location algorithm will be implemented, and integrated into the Sperry Univac speech research facility. Alternative methods for stressed syllable location will also be investigated. In addition, routines will be implemented for automatically comparing stress perceptions with algorithmic stressed syllable locations, and for comparing perception results from time to time and listener to listener.

Further stress perception tests, syntactic boundary detections, algorithmic locations of stressed syllables, and other prosodic and segmental studies will be performed on the test sentences now being designed. These studies should permit developing more specific theories about prosodic patterns and their relationships to phonetic and syntactic structures. They also should yield refinements in methods for syntactic boundary detection, stressed syllable location, and segmental recognition.

With the new research facility now being developed, many of these additional studies should proceed more rapidly. The ARPANET connection will also permit access to other researchers' algorithms, such as parsers.

In summary, work now in progress should soon yield successful computer programs for: syntactic boundary detection; stressed syllable location; evaluation of stress perception and location results; partial distinctive features analysis in stressed syllables and in sibilants (and perhaps stops) of unstressed or reduced syllables; and access to other researchers' algorithms by way of the ARPANET. To date, basic prosodic analysis algorithms have been implemented, and extensive steps have been taken to use such prosodic aids in partial distinctive features estimation. Further work will more precisely explain previous successes and limitations of prosodic and phonetic analysis tools, by isolating effects in the designed texts. The next major effort to be undertaken will be in prosodic aids to syntactic parsing.

5. REFERENCES

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

KLATT, D. H. and STEVENS, K. N. (1972), Sentence Recognition from Visual Examination of Spectrograms and Machine-Aided Lexical Searching, Proc. 1972 Conference on Speech Communication and Processing. IEEE and AFCRL: Bedford, Mass., pp. 315-318.

LEA, W. A. (1972), Intonational Cues to the Constituent Structure and Phonemics of Spoken English, Ph.D. Thesis, School of E. E., Purdue University.

LEA, W. A. (1973a), Syntactic Boundaries and Stress Patterns in Spoken English Texts, Univac Report No. PX 10146, Univac Park, St. Paul, Minnesota.

LEA, W. A. (1973b), An Approach to Syntactic Recognition without Phonemics, IEEE Trans. on Audio and Electroacoustics, vol. AU-21, 249-258.

LEA, W. A. (1973c), Segmental and Suprasegmental Influences on Fundamental Frequency Contours. In Consonant Types and Tone (Proceedings of the First Annual Southern California Round Table in Linguistics, Ed. by L. Hyman), University of Southern California Press.

LEA, W. A., MEDRESS, M. F., and SKINNER, T. E. (1972a), Prosodic Aids to Speech Recognition: I. Basic Algorithms and Stress Studies, Univac Report No. PX 7940, Univac Park, St. Paul, Minnesota.

LEA, W. A., MEDRESS, M. F., and SKINNER, T. E. (1972b), Use of Syntactic Segmentation and Stressed Syllable Location in Phonemic Recognition. Presented at the 84th Meeting, Acoustical Society of America, Miami Beach, Florida, Nov. 27-30, 1972.

LEA, W. A., MEDRESS, M. F., and SKINNER, T. E. (1973), Prosodic Aids to Speech Recognition: II. Syntactic Segmentation and Stressed Syllable Location, Univac Report No. PX 10232, Univac Park, St. Paul, Minnesota.

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, 1971.

APPENDICES:

ABSTRACTS OF ASA PAPERS

APPENDIX A: Perceived Stress as the "Standard" for Judging Acoustical Correlates of Stress

ABSTRACT

Acoustical correlates of stress can only be evaluated in comparison with some "standard" specifying which syllables are actually stressed. The standard should be consistent from time to time, and largely independent of talker and listener idiosyncrasies. Three phonetically-trained subjects listened repeatedly to spoken texts and spontaneous sentences, until they could categorize each syllable as either stressed, unstressed, or reduced. This procedure was repeated three times for each speech text and listener. Two listeners differed from each other on only about 5% of all syllables as to whether they were perceived as stressed or not. Each also showed only about 5% confusions in decisions about stressed syllables from one trial to another. Unstressed and reduced levels were much more frequently confused. The third listener gave less consistent results. Subjects' judgments of stress when given only the written text were of comparable consistency, but did not correspond well with perceptions with speech, if the speech was spontaneous rather than spoken texts. Stress perceptions consequently may be suitable for evaluating acoustical correlates to within a 5% tolerance in overall location scores. Pooling the perceptions from several trials and several listeners may improve the stability of this "standard" for stress assignment.

Paper to be presented by Wayne A. Lea at the 86th Meeting of the Acoustical Society of America, Oct. 27-Nov. 2, 1973, Los Angeles, California.

APPENDIX B: An Algorithm for Locating Stressed Syllables in Continuous Speech

ABSTRACT

Local increases in fundamental frequency (F_0) and large integrals of energy in the syllabic nucleus are known to be among the best acoustical correlates of stress. Major syntactic constituents have been shown to have archetype rapid-rise-then-gradual-fall F_0 contours, with the rise into the maximum F_0 often associated with the first stressed syllable in the constituent. An automatic procedure for detecting constituent boundaries and maximum F_0 positions in constituents (Lea, W. A. (1973), An Approach to Syntactic Recognition without Phonemics, IEEE Trans. Audio and Electroacoustics, AU-21, No. 3), and sonorant energy and F_0 functions, provided input data for an algorithm for locating stressed syllables. The first stressed syllable of a constituent was associated with a high-energy-integral portion near the rising F_0 into maximum F_0 position. Other stressed syllables were associated with high-energy-integral portions near local increases in F_0 above a steadily-falling "archetype line" from the maximum F_0 position to the end of the constituent. For over 400 seconds of speech, including written texts, and questions, commands, and declarations for man-machine interaction (involving sixteen talkers), over 85% of all syllables perceived as stressed by a panel of listeners were correctly located.

Paper to be presented by Wayne A. Lea at the 86th Meeting of the Acoustical Society of America, Oct. 29-Nov. 2, 1973, Los Angeles, California.

APPENDIX C: Evidence that Stressed Syllables are the Most Readily Decoded Portions of Continuous Speech

ABSTRACT

Stressed syllables are presumed to be the most carefully articulated portions of speech, and thus the most likely to provide the reliably encoded information needed for automatic recognition of continuous speech. In conjunction with the Carnegie-Mellon Speech Segmentation Workshop, nine research groups used different automatic techniques to segment continuous speech (31 sentences) and identify the phonetic categories or phonemes. These segmentation and classification results were evaluated according to whether major distinguishing features of each of the phones (such as high/mid/low, front/central/back, and rounded/unrounded for vowels, and manner of articulation for consonants) were correctly determined. Listeners were asked to classify all syllables in the speech as stressed, unstressed, or reduced, and an algorithm for automatic location of stressed syllables also was used to delimit stressed nuclei. Vowels that were perceived as stressed and/or located by the algorithm were more accurately classified than unstressed or reduced vowels. Similarly, pre-stressed obstruents were more reliably categorized than other consonants.

Paper to be presented by Wayne A. Lea at the 86th Meeting of the Acoustical Society of America, Oct. 29-Nov. 2, 1973, Los Angeles, California.