PROSODIC AIDS TO SPEECH RECOGNITION: V. A SUMMARY OF RESULTS TO DATE

Wayne A. Lea

Sperry Univac

Prepared for:

Advanced Research Projects Agency

31 October 1974

Unclassified
Security Classification

AD/A-003931

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)	2a. REPORT SECURITY CLASSIFICATION
Univac Defense Systems Division P. O. Box 3525 St. Paul, Minnesota 55165	Unclassified 2b. GROUP

3. REPORT TITLE

Prosodic Aids to Speech Recognition V. A Summary of Results to Date

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

Final Technical Report; 1 March 1973 – 31 August, 1974

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

Wayne A. Lea

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
31 October, 1974	40	22

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
DAHC15-73-C-0310 b. PROJECT NO	Univac Report No. PX 11087
c d	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 summary of results to date is presented. 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. Experiments showed that listeners' perceptions of stressed syllables were quite consistent, and corresponded closely to the locations of stressed syllables obtained from prosodic features. Analysis of phonetic recognition results by several research groups showed that automatic phone categorization is much more accurate in stressed syllables. Studies showed that stressed vowels in several recorded texts tended to be roughly equally spaced in time, but the number of intervening unstressed syllables had a much more prominent effect on interstress interval than might have been expected from published hypotheses.

Prosodic features appear to be potentially useful for providing cues to sentence type, syntactic bracketing, occurrences of coordination and subordination, and specific semantic structures. Preliminary studies of some sentences that gave problems to speech understanding systems showed that prosodies do differ in yes/no questions versus commands, and that ambiguous syntactic structures can be disambiguated from prosodic patterns. A set of speech texts have been designed for careful analysis of the effects on prosodic patterns due to various contrasts in syntactic structure, semantics, stress patterns, and phonetic sequences.

DD FORM 1 NOV 68 1473

Unclassified
Security Classification

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

SPERRY UNIVAC

PROSODIC AIDS TO SPEECH RECOGNITION: V. A SUMMARY OF RESULTS TO DATE

by

Wayne A. Lea

Defense Systems Division
St. Paul, Minnesota
(612) 456-2434

Final Technical Report Submitted To:

Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209

Attention: Director, IPT

31 October 1974

Report No. PX 1087

This 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 author, 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.

PREFACE

This is the fifth in a series of reports on 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. The third report, subtitled "III. Relationships Between Stress and Phonemic Recognition Results", appeared 21 September 1973, as Univac Report No. PX 10430. The fourth report, subtitled "A General Strategy for Prosodically-Guided Speech Understanding", appeared 29 March 1974, as Univac Report No. 10791.

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 author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advance 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 distinguishing features analysis is to be done within each stressed syllable and wherever else reliable segmental analysis can be readily accomplished. Positions of the boundaries between grammatical phrases, stressed syllable locations, pauses and specific rhythmic patterns, and special intonational features are to be used to guide the selection of a candidate grammatical structure for the spoken sentence. This preliminary hypothesis about the grammatical structure of the sentence is thus made from prosodic features only, without prior determination of the words in the sentence. In essence, then, prosodic features are used to guide the efficient determination of the vowels and consonants making up the words in the sentence, and the determination of some aspects of the grammatical context in which those words are found.

From the partial distinguishing features analysis, and some knowledge of contextual constraints, words can be hypothesized as occurring at specific places in the utterance, with particular emphasis given to the hypothesizing of important words centered around the stressed syllables. These hypothesized words, the acceptable grammatical structures, the semantic constraints, and knowledge of the limited types of things that can be said in any specific task situation all provide the information needed to make a total hypothesis as to the identity of the spoken sentence. By an analysis-by-synthesis procedure, such an hypothesis can be submitted to sound structure rules which generate a comprehensive sound structure ("acoustic phonetic") pattern which can be compared to the sound structure pattern of the input. If the generated pattern is very similar to the input pattern, the hypothesized sentence structure is asserted to be the identity of the input sentence. If the patterns differ substantially, an error signal is produced, to guide the selection of another structural hypothesis to try.

Not all aspects of this general prosodically-guided analysis-by-synthesis strategy are being implemented for ARPA, but the strategy provides a framework within which substantial contributions to speech understanding can be provided by the judicious use

of prosodic features. This approach to speech understanding is motivated by several factors. For one thing, there is considerable evidence that human listeners make initial decisions about large linguistic units (phrases) before they attempt to decide upon the complete phonetic structures (that is, the sequence of vowels and consonants) in an utterance. They appear to use prosodic patterns such as intonation, stress, pauses, and rhythm to guide their decisions about large linguistic units.

In addition, some of the most reliable information about intended vowels and consonants is to be found in the stressed syllables, which are more carefully articulated. A study of five different methods for automatic classification of segments of speech into specific vowel and consonant categories showed that, with any of the available methods, the categorization of vowels, stop consonants (p, t, k, b, d, g), and fricatives (s, z, f, v, etc.) was far more reliable in the stressed syllables than elsewhere in the utterances. This demonstrated that stressed syllables provide "islands of reliability" in the sound structure of spoken English. These stressed syllables also occur in the most semantically important words of a sentence.

Since stressed syllables do provide some of the most important and reliably decoded information in the speech wave, a procedure for locating stressed syllables in connected speech has been developed. This procedure uses acoustic parameters of energy, syllabic duration, and voice fundamental frequency ("pitch") to locate the vowel and semi-vowel sounds forming the "nucleus" of a stressed syllable. Computer programs have been implemented for classifying the stressed vowel into one of five categories, depending upon the natural vocal tract resonances (formants) to determine whether the speaker's tongue is high or low, front or back, or retroflexed in his mouth.

Automatic detection of a few particular consonant categories is also attempted, independent of the locations of stressed syllables. These detected phone categories include sibilants [s, z, ʃ, ʒ, tʃ, dʒ], r-like sounds [r, ɝ], unvoiced stops [p, t, k], and nasals [m, n, ŋ]. Some, but not all, of the occurrences of these consonants will be within stressed syllables. Then, the remaining portions of stressed syllables, not found within the stressed vowel, or within one of the detected consonantal segments, are classified into gross left-over categories of unvoiced consonant, voiced consonant, or silence. The unstressed syllables are not totally segmented, and thus there are isolated islands of detected sound structure, some being the total stressed syllables and others being one of the four types of detected consonants.

Among the studies at Sperry Univac has been the comparison of three approaches to stressed syllable location. Methods based on only the duration of high energy chunks, or upon only the length of time that fundamental frequency (F_0) was not falling significantly, did not perform as well as the original algorithm based on archetype F_0 contours in phrases and local searches for high-energy chunks of speech. The archetype contour algorithm was also least sensitive to the type of sentence being processed, while the other algorithms showed quite different performance in yes/no questions.

Prosodic information can be used in several important components of speech understanding systems. Stressed syllables can form the anchors around which a search for occurrences of words can be attempted. We plan to investigate how detected positions of constituent boundaries, located stressed syllables, features of intonation, pauses, and rhythms can be used to determine: the type of sentence spoken (that is, whether or not it was a yes/no question); the correct grouping of portions of the utterance into phrases and specific linguistic units; the occurrence of coordinate structures; the subordination of one phrase under another; and the occurrence of specific semantic structures like co-reference, contrast, and emphasis.

We have undertaken a study of rhythm in our available speech texts. These studies suggest that, while stressed syllables tend to occur at about .4 to .5 seconds apart, the time intervals between stresses are very much affected by the number of unstressed syllables between stresses. In none of our texts did more than four unstressed (or reduced) syllables ever occur between two stressed syllables. There seems to be an almost equal increment in the size of the interval between onsets of stressed vowels with each increase in the number of intervening unstressed syllables. There is some suggestion in the rhythm data that the preferred rhythmic structure is an alternation of stress and unstress.

These rhythm studies also showed that the time intervals between detected phrase boundaries tended to be integral multiples of the mean time between stressed vowels. Pauses between embedded clauses tended to be of the same duration as the mean time interval between stressed vowels, while durations of pauses between sentences tended to be about twice that interstress interval.

We plan to study interstress time intervals, intervals between detected phrase boundaries, and time intervals between all syllables (or, equivalently, number of

syllables per unit time) as acoustic measures of rhythm and rate of speech. Information about rate of speech may be used in selecting the appropriate phonological rules to apply in determining underlying phonemic structure from the slurred, coarticulated phonetic sequences. "Fast speech" rules show more slurring, coarticulating, and dropping of speech sounds. In addition to such phonological use of rate of speech, specific rhythmic effects such as interruptions of rhythm (pauses, "disjunctures", etc.) could be useful in hypothesizing the grammatical structure of a sentence.

Several "problem sentences", which were quite similar in the sequence of words they were composed from, but which had different syntactic structures, were submitted to our prosodic analysis procedures. These sentences, obtained from Bolt Beranek and Newman, included a yes/no question ("Have any people done chemical analyses on this rock?") and the command that resulted when the first word was incorrectly recognized ("Give any people done chemical analyses on this rock."). The command is structurally ambiguous, with one interpretation referring to the process of giving to any people those chemical analyses that have been done, and another interpretation referring to any chemical analyses that were done by people (that is, "people-done" analyses). An example of a pronunciation with each intended interpretation was provided, along with an apparently "neutral" pronunciation which was presumably intended to not indicate which of the two interpretations was correct. Another pronunciation of the wording of the yes/no question, with more like the intonation of a declarative or command, was also provided.

A study of the prosodic patterns in these sentences gave encouraging indications that prosodies can provide important cues to syntactic structure. The auxiliary verb "have" in the yes/no question was unstressed while the command verb "give" was stressed. The general slope of the F_0 contour in the yes/no question was flatter than the more-falling contour of the command, although the F_0 rise expected at the end of the yes/no question did not occur. The word "done" was stressed except when it was in the compound construction "people-done", in which case the first syllable of the following word "chemical" was stressed. There was a phrase boundary between "people" and "done" in each command except the "people-done" interpretation, in which case the boundary occurred after "done", thus marking "people-done" as a unit. The time intervals between vowel onsets also showed distinctions between the two interpretations. The "neutral" version of the command turned out to be identified with

the first ("[any people][done chemical analyses]") interpretation by every available prosodic cue. This is a harmony with an expectation that the first interpretation is the most likely, and the most like a neutral, "unmarked" interpretation, since other possible structures, like the yes/no question, would make "any people" a unit, and not "people-done" a likely unit.

Studies with these few sentences are only suggestive of possible prosodic cues to syntactic structures. Further studies with many more utterances are needed. We have designed a set of 902 sentences which provide "minimal pairs" of sentences with nearly identical word sequences but contrasting structures. These sentences include explicit tests of the prosodic effects of sentence type, contrastive syntactic bracketing, subordination, coordination, syntactic categories (such as pronouns, verbals, compound nouns, etc.), movement of stress within phrases, coreference, etc. Prosodic patterns to be studied for these sentences include: performance of the program for detecting phrase boundaries from valleys in F_0 contours; acoustic correlates of stressed syllables, and performance in automatic stressed syllable location; acoustic measures of rhythm and rate of speech; overall F_0 contour shapes; and local variations in prosodic features due to phonetic sequences. Also, we have designed a set of 178 sentences which include all word-initial consonant-vowel (CV) sequences, and all word-final vowel-consonant (VC) sequences. These "phonetic-sequence sentences" provide the speech data needed for efficiently testing automatic procedures for vowel and consonant classification. For example, five sentences provide instances of all distinguishable stressed vowels of American English, coupled with the sibilants [s, ʃ], in initial CV and final VC positions.

From extensive studies with such designed sentences, we hope to develop experimentally-validated intonation rules and other prosodic rules. These rules will then be used to guide parsing, semantic analysis, phonological analyses, and word matching procedures in ARPA speech understanding systems.

TABLE OF CONTENTS

1. INTRODUCTION

This is a Final Report for the Advanced Research Projects Agency (ARPA), under ARPA Contract No. DAHC 45-73-C-0310, and it consequently summarizes progress to date. Yet, in a technical sense, it is another in a series of reports on work currently in progress in the Univac Speech Communications Group. 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, doing distinctive features analysis within stressed syllables, and using prosodic features to guide syntactic parsing and phonological and semantic analyses.

Under a previous contract, Sperry Univac developed some basic tools for prosodic and distinctive-features analysis, and conducted some initial experiments dealing with prosodic patterns in connected speech. That work is reviewed briefly in Section 2 of this report. In Section 3, the many areas of progress during the current contract are briefly reviewed. Section 4 provides plans for further studies to be conducted under new contracts.

Section 5 gives a listing of the publications and reports resulting from the ARPA program at Sperry Univac. References are listed in Section 6.

2. RESULTS FROM PREVIOUS CONTRACTS

2.1 Reasons for Prosodic Analysis

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). Linguistic arguments were given, showing that there is not always a distinguishable segment in the speech wave that corresponds to each abstract phonemic segment, and some of the cues to the presence of a particular phoneme may appear in nearby or distant acoustic segments. In addition to the fact that the encoding of phonemic and prosodic information into the acoustic waveform is a complex one involving overlapping in time and environmental dependence, the encoding itself is often performed incompletely and with considerable variability. Indeed, in some utterances, whole phonemes or syllables may be "missing" from the pronunciation. A speech recognition system based on acoustic manifestation of all phonemes or all distinctive features would thus frequently fail.

Linguists argue that "in general, the perceiver of speech should utilize syntactic cues in determining the phonemic representation of an utterance" (Chomsky and Miller, 1963, p. 314, emphasis added). Perception theorists (e.g., Miller, 1962, p. 81) have also argued that large units, on the order of phrases and clauses, are used in early stages of human perception of speech, and that detailed phonemic decisions are made later, and then only where they are needed to fill in information about the large units. Experiments have shown that clicks superimposed on speech were perceived as occurring near certain major syntactic boundaries, regardless of the actual timing of the clicks within the speech continuum (cf. review by Gleitman and Gleitman, 1970). The perceiver appears to wait until the ends of such syntactic units before making decisions about detailed sound structure. Speech perception thus appears to involve making use of certain expectations and received cues to determine the syntactic structure of a sentence, and then using such syntactic information in guiding phonemic decisions.

Among the cues encoded in the acoustic signal that may be used in making the preliminary syntactic hypotheses, the prosodic features of juncture, intonation, stress, and rhythm seem pre-eminent. These are acoustically manifested by silences, lengthened time intervals between vowel onsets, variations in fundamental frequency (F_0), relative energy levels, and durations of specific units like syllable nuclei, vowels, and consonants. Linguists have claimed for decades that intonation marks the syntactic structure of English sentences (see reviews in Lea, 1972, and Lea, Medress, and Skinner, 1972a). Chomsky and Halle (1968) provided rules for generating English stress patterns from syntactic structures. Time intervals between vowel onsets have been shown to be one cue to boundaries between syntactic units. It is known that interruptions of rhythm mark major syntactic and semantic boundaries.

These earlier arguments about prosodic cues to syntactic structure have been reinforced by studies to be reported in subsequent sections of this report.

2.2 Prosodic and Segmental Analysis Tools

At Sperry Univac, we have implemented highly flexible and interactive systems for processing and studying speech. The first version of the Speech Communications Laboratory included a sound isolation room for making high quality audio tapes, plus tape recorders, microphones, analog-to-digital and digital-to-analog converters, and a computer system with various graphical and alphanumeric displays and mass storage devices, toggle switches, pushbuttons, potentionmeters, keyboard, and interactive system software. Linear predictive analysis was implemented on the speech research facility, and a formant tracking algorithm similar to one by Shafer and Rabiner (1970) was implemented to use the smoothed linear prediction spectra. An autocorrelation method of fundamental frequency tracking (cf. Sondhi, 1968; Lea, Medress, and Skinner, 1973b, Appendix) was implemented, and total speech energy was computed from the sum of the squares of the time waveform values. Band-limited energy functions were computed from summing the squares of the smoothed spectral magnitudes within a frequency band, and then converting the sum to dB. Plots of formant values versus time, and F_0 and energy functions versus time, were obtained both from the interactive graphical displays and computer listings.

Later work with simple peak-picking with the LPC spectra was so successful that the complex Shafer and Rabiner algorithm for formant tracking was discarded.

The refined spectral analysis procedures then involved:

- low-pass filtering at 4782 Hz (using a seventh order elliptic function Cauer low-pass filter provided by Lincoln Laboratories);
- sampling at a rate of ten thousand samples per second;
- software pre-emphasizing by first order differencing;
- applying a Hanning weighted analyzing time window of width 25.6 milliseconds (ms) with a 10 ms advance; and
- performing a 256-point Fast Fourier transform.

The LPC analysis was done with fourteen predictor coefficients, and evaluation of the spectrum at -75 Hz off the jw-axis. Peak picking was done to estimate formants, and these spectral peaks were smoothed by bringing back into line any one or two formant values that were out of line with preceding and following formant values.

2.3 Syntactic Boundary Detection

Lea's earlier research (1971, 1972, 1973b) showed that a decrease (of about 7% or more) in F_0 usually occurred at the end of each major syntactic constituent, and an increase (of about 7% or more) in F_0 occurred near the beginning of the following constituent. A computer program, based on the regular occurrence of F_0 valleys at constituent boundaries, correctly detected over 80% of all syntactically predicted boundaries. Over half of the "missing" boundaries were between noun phrases and auxiliary or main verbs. About 10% of all detected boundaries were "false", in that they were not associated with syntactic boundaries. These false boundaries were primarily caused by F_0 variations near voiced and unvoiced obstruents.

Sentence boundaries were always accompanied by fall-rise F_0 valleys. In fact, the rise in F_0 (around 30% change) after a sentence boundary was substantially larger than the usual rises (about 40% or less) after non-sentential constituent boundaries. In addition, sentence boundaries were usually (in over 90% of all cases) accompanied by long (35 centisecond or longer) stretches of unvoicing. Here "sentence boundaries" refer to both boundaries between matrix (unembedded) sentences and boundaries between embedded full-clausal sentences.

These initial studies also showed that syntactic categories had some effects on boundary detection. Coordinate noun phrases or coordinate adjectives were always

accompanied by F_0 valleys between the conjuncts. Around 95% of all boundaries before prepositional phrases were detected by F_0 fall-rise valleys. Less than half of the boundaries between noun phrases and following verbals were accompanied by F_0 valleys.

Lea's program was implemented as a FORTRAN program on the Sperry Univac speech research facility, with output syntactic boundary decisions marked at appropriate times on plots of F_0 and sonorant energy. This implementation was tested with six talkers reading the Rainbow Script, two talkers reading a paragraph composed of only monosyllabic words, and a collection of 13 sentences involving eight talkers. The 13 sentences were taken from a set of 31 "ARPA Sentences", twenty-seven of which had been selected by Wayne Lea on the basis of interesting syntactic variation, from a set of about 250 sentences spoken by workers at five ARPA contractors. Some of these 13 sentences were read from written texts, but others were obtained from simulated man-computer interactions (Lea, Medress and Skinner, 1973a). The boundary detection algorithm was found to correctly detect 79% of all linguistically predicted boundaries in the readings of the Rainbow Script, 83% of all predicted boundaries in the "Monosyllabic" Script, and 74% of all predicted boundaries in the 13 ARPA sentences. The 'spontaneous' speech of the ARPA sentences thus showed the lowest detection score, partly due to the monotonic intonation with which some ARPA sentences were spoken, and partly due to unpredictable F_0 inflections and hesitation pauses that occurred in the simulations of man-machine interactions.

2.4 Listener's Perceptions of Stress Patterns

We anticipated that stress patterns in spoken sentences would provide us with some cues to syntactic structure and some guidelines as to where the most reliably-encoded phonetic information would be. An algorithm for locating stressed syllables from acoustic patterns thus was needed. However, any algorithm for locating stressed syllables can only be evaluated in comparison with some "standard" that specifies which syllables are actually stressed. We thus conducted experiments to determine the effectiveness and stability of listener's perceptions of stress patterns for six talkers reading the Rainbow Script, two talkers reading the Monosyllabic Script, and eight talkers involved in the 13 ARPA sentences (Lea, 1973a).

Individual listeners were allowed to repeatedly listen to portions of the tapes, until they could mark each syllable as either stressed, unstressed, or reduced. These perception tests were repeated three times (separated by at least a few days), by each of three listeners. For two listeners, an individual listener's perceptions of which syllables were stressed agreed from one trial to the other trials, for 95% of all syllables perceived as stressed. Thus, a listener confused about 5% of the syllables between levels of stressed and unstressed from one trial to another. His confusions between unstressed and reduced levels were much more frequent. Also, two of the listeners agreed with each other on 95% of their judgments as to which were the stressed syllables.

We concluded that it was possible to determine, within about a 5% tolerance, which syllables are actually stressed in connected speech, by obtaining listeners' perceptions in the manner that was used in these experiments. Then, if a stressed syllable location algorithm could locate 95% of all syllables perceived as stressed by majority votes of two or more listeners, it would be doing as well as one repetition of the perception tests would do for predicting the perceptions from another repetition of the experiment. It would also be doing as well as one listener would do in comparison to another listener. We can demand no better than 5% precision in stressed syllable location from acoustic data.

We also found that listeners tended to be more confused as to which were the stressed syllables in questions than in declaratives or commands, with yes/no questions yielding the most confusion, and declaratives yielding the fewest confusions. Such effects of sentence type on perceived stress patterns required further study, since only a few questions had been included in the texts studied.

2.5 Algorithmic Location of Stressed Syllables

The extensive previous studies of acoustic correlates of stress showed that local increases in F_0 and large values of energy integral in a syllabic nucleus were among the most reliable correlates of stressed syllables. An algorithm for locating stressed syllables was developed. The increasing F_0 near the beginning of each constituent detected by the boundary detector was assumed to be attributable to the first stressed syllable in the constituent. A stressed "HEAD" to the constituent was thus associated with a portion of the speech which was high in energy with rising F_0, and bounded

by substantial (5 dB or more) dips in energy. Other stressed syllables in the constituent were 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 expected that local 'increases' in F_0 due to later stressed syllables would be evident as local rises above the gradually falling F_0 contour, even if F_0 did not rise absolutely near the stressed syllable. The stressed syllable was then located within a high-energy-integral region near this local rise above the archetype F_0 contour.

A hand analysis using this algorithm (cf. Lea, 1973a) succeeded in locating about 85% of all syllables perceived as stressed by the majority votes of the panel of three listeners. The Monosyllabic Script, with its more prominent stresses on monosyllabic content words, yielded quite high location scores (90% and 95%). The spontaneous ARPA sentences, which were more monotone and which gave more difficulties to the boundary detection algorithm, showed the lowest stressed syllable location scores. About 7% to 28% of all algorithmically located stretches of speech did not enclose any syllable perceived as stressed by the majority vote of the listeners. Some of these "false alarms" pointed to syllables that at least one listener did perceive as stressed. It appeared that false alarm rates could be reduced by improvements in the boundary detector, and by a refinement of not demanding stressed HEADS in short constituents (such as those less than 200 ms in duration). Further studies were needed to reduce false alarm rates and simultaneously improve the scores for correct locations. There thus appeared to be room for improvement in the performance of the stressed syllable location algorithm, and it seemed necessary to investigate whether simpler procedures could work as well. Also, the algorithm needed to be implemented as a computer program.

3. PROGRESS DURING THIS CONTRACT

3.1 The Need for Further Work

We thus completed our first ARPA contract, satisfied that good performance in constituent boundary location and stressed syllable location had been demonstrated. An adequate method had been devised for obtaining listeners' judgments of which syllables in connected speech are stressed, unstressed, or reduced. However, several forms of further work appeared to be needed. The program for constituent boundary detection could be refined to produce fewer false alarms, by requiring each new F_0 maximum or minimum to remain beyond the 7% thresholds for at least two time segments (20 ms). Thus, single F_0 values that were out of line with surrounding F_0 values would not trigger the detection of substantial F_0 rises and falls. Also, it appeared desirable to incorporate an overall confidence measure for each boundary, based on the percentage decrease in F_0 before the apparent boundary, the percentage increase after the boundary, the shape of the contour near the boundary, and the time between that boundary and the immediately preceding or following ones. Thus, cusp-like changes at boundaries between unvoiced consonants and sonorants (of the form – ⌣) and very brief F_0 dips or jumps (of such a form as ⌣ –) might be assigned very low likelihood of being boundaries, while major gradual changes (of the form ⌣) would be assigned higher confidence ratings. One or both of two boundaries separated by short times (in the order of 200 ms or less) might be considered suspect, and assigned a low confidence rating.

The boundary predictions also needed to be improved, by defining and applying a strict set of rules for syntactic bracketing and prediction of intonation contours. Intonation rules were needed, along with the selection of an adequate grammar to define the syntactic structure that would be part of the input to such intonation rules. To develop and test such rules, careful controlled studies of intonation in various syntactic structures needed to be undertaken.

The algorithm for locating stressed syllables still had to be implemented as a computer program and tested carefully to see that it performed at the level of success attained in the previous hand analyses. Also, several improvements were needed. Among those apparently worthy of further investigation were better procedures for defining the slope of the archetype F_0 contour in a constituent, a careful "tuning" of

all the parameters and detailed steps for selecting HEAD's and other stressed syllabies, use of a low-frequency "sonorant" energy function rather than the broadband energy function (so that better syllabication might be attained), and the incorporation of procedures for locating other possible stressed syllables before the HEAD (or peak F_0 position) when the peak F_0 occurs late in a constituent (say more than 400 or 500 ms after the preceding boundary).

It also seemed reasonable to compare the results from the archetype stressed syllable algorithm (either before or after it was implemented as a computer program) with results in stressed syllable location by other possible procedures. For example, if one called all long-duration portions where energy was above a threshold value as stressed syllabies, how many of the perceived stressed syllables would be detected and how many false alarms would result? Alternatively, could one get comparable success by looking for all F_0 rises or upward inflections and choosing the high energy portion nearest such places, without use of boundaries and archetype contours in his procedures?

More extensive experiments were needed wherein the various variables of sentence type, talker, lexical forms, phonetic content, position in sentence and intonation contour, and such could be independently controlled. In particular, such studies could test further the apparent difficulty in listeners' assignments of stress within yes-no questions, and the relative successes in boundary detection and stressed syllabie location within questions versus declaratives or commands.

The application of boundary detections and stressed syllabie locations to guiding a partial distinctive features analysis was also yet to be done. Until such details of the distinctive features analysis were better defined, the question couldn't be resolved as to whether higher "hit" rates or lower "false alarm" rates were more important to attain in the boundary detection or stressed syllable location algorithms. Also, we saw that techniques must be explored for applying boundary and stressed syllable information to the aid of syntactic parsers. Such efforts were obviously critical to implementing a prosodically-guided speech recognition strategy at Univac.

3.2 Improved Facilities and Analysis Tools

A new and enhanced sppech research facility has been implemented, using a Univac 1616 computer, improved peripherals, a hardware fast Fourier transform

processor, and more efficient memory usage with a file-structured system. A Very Distant Host Interface has been implemented on a Univac 1219 computer, to the point where it now provides teletype and line printer usage of the ARPANET.

Several modifications and additions were made to the prosodic and distinctive features extraction procedures. In fundamental frequency (F_0) tracking, the autocorrelation vector is now computed using absolute addition rather than multiplication, and contained autocorrelation (the first half of the analysis window correlated with the entire window). This yielded about a 22% savings in computation time. Also, an energy thresholding technique was introduced which required both the first and second halves of the time window to have energy in excess of a threshold amount. This resulted in more precise F_0 onsets and offsets.

Some detailed improvements in our spectral analysis techniques were made, and functions of 'total energy' (60 to 5000 Hz), 'sonorant energy' (60 to 300 Hz), and 'high frequency sonorant energy' (550 to 3000 Hz) were obtained from the spectra. The sonorant energy function performs best in isolating syllabic sonorant clusters, while the high frequency sonorant energy function permits separating the vowel nucleus of a sonorant cluster from surrounding nasals, liquids, and glides. Very low frequency energy (60 to 400 Hz) provides an independent voicing function. A ratio of low to high frequency energy (60 to 900 Hz/3000 to 5000 Hz) provides a cue to presence of sibilants. A spectral derivative, which indicates the similarity of successive spectra, provides a cue to the presence of sudden spectral changes at stop releases.

3.3 Stress Patterns and Boundaries in 31 ARPA Sentences

Sperry Univac participated in the comparison of speech segmentation and classification procedures conducted at the Carnegie-Mellon University Segmentation Workshop. The 31 ARPA Sentences used in the Segmentation Workshop were processed using the improved F_0 tracking algorithm, the new frequency-delimited energy functions, the algorithm for boundary detection, alternative voicing detectors, and the spectral derivative. Sperry Univac provided data on the beginning and ending of each voiced portion in the speech, the beginning and ending of stressed nuclei (obtained as 5-dB-down points in a hand analysis of the sonorant energy function), and locations of syntactic boundaries as determined by the constituent boundary detection program.

This Workshop effort was a major step in our attempts to define how to use prosodic information in aiding segmental analysis of speech. Also, it involved extending our previous prosodic analyses from 13 ARPA Sentences to the complete set of 31 ARPA Sentences. Three trials of the stress perception tests had to be conducted on the 31 ARPA Sentences. F_0 and energy functions were obtained and the boundary detection program applied to get all detected syntactic boundaries. Then a hand analysis of the F_0, energy, and boundary information was done, following the "archetype-contour" algorithm for stressed syllable location. The algorithm correctly located 86% of all syllables perceived as stressed, with 23% of all locations being false. (This was an improvement of 6% in correct locations, and 8% reduction in false alarms, for the 13 sentences that had previously been processed. The improvements resulted from the new conditions on F_0 tracking, the refinement of the boundary detector which required F_0 maxima and minima to be of 20 ms minimum duration, and the use of the sonorant energy function rather than the total energy function.)

3.4 Evidence that Algorithms for Phonetic Categorization Work Best in Stressed Syllables

Some preliminary experiments were conducted (on the 31 ARPA Sentences) to classify vowels, locate sibilants and determine their place of articulation, and locate stops. An algorithm for locating sibilants from low values of the ratio of low to high frequency energy found 86% of all sibilants, with only two false alarms. Place of articulation was correctly determined for 89% of the located sibilants. These results suggest that sibilants are robustly encoded in the speech signal. In contrast, a first attempt at stop location succeeded in locating less than half of all phonemic stops.

Of most interest in these initial studies was the effect of perceived syllable stress on the reliability of segmental locations and categorizations. Sixty-six percent of all located stops were in stressed syllables (even though only about one-third of all syllables were stressed). 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 was better in stressed syllables (at least with that preliminary location scheme). Sibilants, on the other hand, showed 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.

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. 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.

All these experimental results were quite preliminary and could be expected to be affected by the exact procedures for segmental recognition. Would other more-sophisticated techniques for categorizing phonetic segments from acoustic features better match phonetic transcriptions in stressed syllables than in unstressed or reduced syllables? To answer this, we studied the results of five groups participating in the Carnegie-Mellon University Speech Segmentation Workshop, and compared their segmentations to the transcriptions provided by a linguist (Linda Shockey).

The sophistication and detailed methods of segment categorization used by these groups varied considerably, but for each group a chart was compiled giving a score for the level of correspondence ("phonetic similarity") between each of the acoustically-derived segment labels and the phonetician's labels (Lea, 1973c). If the score was below a threshold, the categorization was considered unacceptable (an "error"). Error rates were compiled separately for syllables perceived (by a majority of listeners) as reduced, unstressed, or stressed. The results showed that vowels and obstruents were least likely to be inadequately categorized in stressed syllables, for each group's segmentation scheme.

We may conclude that, while ideal methods might be devised to phonetically categorize as well in unstressed and reduced syllables as in stressed syllables, several (available and practical) methods for phonetic segmentation correspond most closely

with phonetic transcriptions in stressed syllables. Combining this with the closer correspondence expected between phonetics and underlying phonemic structure in stressed syllables, and the semantic importance assigned to stressed syllables, one can see the value of early attention to stressed syllables in procedures for recognition of continuous speech.

3.5 Locating Stressed Nuclei

Because of this demonstrated importance of stressed syllables in speech recognition, an algorithm for automatically locating stressed syllables could be a valuable component of a speech understanding system. We have already reviewed the general success in stressed syllable location using our "archetype contour" algorithm.

To further evaluate the effectiveness of this archetype contour algorithm, the results with that algorithm were compared with results in stressed syllable location by other procedures. One alternative simple procedure finds all dips and peaks in the sonorant energy function, and delimits syllabic nuclei as all contiguous points within 5 dB of the maximum intensity value in each high-intensity "chunk" or syllable. Then, those chunks (or syllabic nuclei) that have a minimum duration of 100 ms are declared to be stressed.

Another simple routine locates all portions of speech where, for 100 ms or longer, fundamental frequency does not decrease more than one eighth tone per ten milliseconds (this is a relaxed form of a process of finding regions where fundamental frequency is steadily rising, or at least not falling rapidly).

Regions of increasing fundamental frequency were not as reliably related to stressed syllables as were the durations of high-energy "chunks", with poorest performance in the man-machine interactions of the ARPA Sentences. The archetype-contour algorithm was shown to perform better than either of these two simpler algorithms, particularly for the spontaneous ARPA speech (Lea, 1973d, 1974a).

Stressed syllable location by the algorithms was found to be affected by the type of sentence spoken. For each algorithm, false alarms are most frequent in yes/no questions. The lowest correct location score from chunk durations occurred in yes/no questions, while the highest correct location score from increases in fundamental frequency occurred in yes/no questions. This suggests the value of combining the two types of cues to improve success in stressed syllable location, such as is done in the archetype-contour algorithm.

Based on these results, the archetype-contour algorithm was seen to be good for stressed syllable location, and its implementation as a FORTRAN program was undertaken.

3.6 A General Strategy for Prosodically-Guided Speech Understanding

A major accomplishment in the current contract was to define a specific strategy for using prosodic information to guide both phonological and syntactic aspects of speech understanding systems. A total analysis-by-synthesis framework for recognition was outlined, in which the preliminary acoustic analyses include: extracting phonetic and prosodic features; obtaining phrase boundaries, stress patterns, and rhythms; analyzing the phonetic structures centered around reliable "anchor points" like stressed syllables, sibilants, r-like sounds, and nasals; and hypothesizing syntactic structures that might have produced the detected prosodic patterns. Following such prosodically-derived guesses as to large-unit structure of the utterance, and the prosodically-guided phonetic analysis, the system strategy uses lexical, syntactic, semantic, and pragmatic components to hypothesize total sentence structures, which are applied through phonological rules to generate expected patterns for the hypothesized structure. If the generated patterns agree closely with the input patterns, the hypothesis is accepted; otherwise, an error signal is fed back, to control the selection of the next sentence hypothesis to try.

In this overall strategy, prosodic features reduce the areas where one must perform costly spectral analyses, they provide pointers to some of the most reliable phonetic information, they provide information about rate of speech (which may be useful in selecting appropriate fast-speech phonological rules, etc.), and they provide cues to syntactic bracketing and categories (which will be used in aiding syntactic parsing).

3.7 Design of an Acoustic Front End

A block diagram of the preliminary acoustic analysis components of the speech understanding strategy was devised (Lea, Medress, and Skinner, 1974; Lea, 1974a, p.8). An initial implementation of many aspects of that system structure was used in an isolated-digit recognition demonstration, under funding of a separate internal research program at Sperry Univac (Skinner, 1974). The significance of this separate isolated-word recognition effort is that it has provided an actual implementation of many of the

preliminary analysis components to be included in the "acoustic front end" of a total prosodically-guided sentence-understanding strategy. The remaining components (including the stressed syllable algorithm, the procedures for analyzing phonetic structures in the region of the stressed syllable, the preliminary syntactic hypothesizer, and the prosodic and phonological rules) can be added as they are completed.

3.8. Compiling Phonological and Prosodic Rules

Phonological rules will form a major component in our prosodically-guided speech understanding strategy, as they are expected to in the strategies of other ARPA contractors. Wayne Lea has participated in the ARPA Rules Workshops, and presented talks on prosodic hypotheses and rules. We plan to compile hypotheses and experimentally-verified rules about prosodic regularities, such as rules relating stress patterns to syntactic structures, and rules showing how clause structures, phrases, stress patterns, and phonetic sequences superimpose effects on F_0 contours.

An unpublished ARPA "SUR Note" and a Sperry Univac technical report (Lea, 1974c) have provided some initial attempts at defining prosodic hypotheses. Further studies are closely tied to plans for prosodic analysis of our large new speech data base, which will be discussed later in this report (Sections 3.11 and 3.12).

3.9. Studies of Rhythm

The time intervals between the beginnings of the nuclei of stressed syllables were studied, for perceived stresses for two talkers reading the Rainbow and Monosyllabic Scripts, and for the 31 ARPA Sentences (Lea, 1974a). Stressed syllables did tend to be spaced about 400 ms apart, but the variation in interval sizes was quite large, even for a single talker within a single text. We concluded that the concept of English being a stress-timed language is not simply exhibited by exact equality of interstress intervals, or even by an unquestionable "tendency toward equality" of interstress intervals regardless of other factors. We found that, contrary to several published hypotheses, the average interstress interval increases about linearly with the number of unstressed syllables between the stresses. A tendency toward stressed-unstressed alternation was exhibited, and it is probably this tendency, plus the somewhat uniform durations expected for unstressed syllables, that yields the tendency for interstress intervals to cluster somewhat near an average of 400 ms or so.

This study also showed that pauses between clauses of a sentence tended to be about the same duration as interstress intervals, while pauses between sentences tended to be twice that duration. A pause is thus like an integer multiple of an inserted silent interstress interval. We also found that time intervals between detected syntactic boundaries tended to cluster in a multimodal distribution centered around multiples of the average interstress interval.

These results indicate that interstress intervals, pause durations, and intervals between detected boundaries all seem to relate to speech rhythm. Each of these, plus a measure like the number of syllables per second, may be useful as a measure of speech rate. Knowing speech rate may be useful in selecting what phonological rules to assume apply to a given utterance.

3.10 Prosodic Analysis of BBN Problem Sentences

We have conducted experiments on constituent boundary detection, stressed syllable location, reliability of phonetic analysis in stressed syllables, and rhythmic regularities like interstress intervals, pause durations, and intervals between constituent boundaries. How might ARPA contractors begin to use such prosodically-based information in speech understanding systems? To suggest answers to this question, we processed six "problem sentences" provided by Bolt Beranek and Newman, to see if prosodies could distinguish between yes-no questions and commands, and if they could disambiguate potentially ambiguous syntactic structures. We found that, contrary to a popular belief, a yes-no question is not simply marked by increasing F_0 values within or following the last stressed syllable of the sentence, even if the sentence sounds like a question to the casual listener. However, there was some indication that the yes-no question may have less fall in F_0 in the total F_0 contour after the first stressed syllable. Also, the auxiliary verb ("have", "do", etc.) that often begins a yes-no question is unstressed, while the verb of a command is stressed. Some disjunctures, or time intervals between onsets of syllabic nuclei, were also found to be different for the yes-no question versus command.

The stress positions, disjunctures, and positions of detected syntactic boundaries were also found to distinguish between structurally different structures like [any people] [done] [chemical analyses] versus [any [people-done] chemical analysis]. (NP V V NP NP NP Adj Adj NP NP)

Thus several cues (overall F_0 contours, positions of constituent boundaries, stress patterns, and disjunctures) do show promise of distinguishing yes/no questions from commands and of disambiguating structurally ambiguous sentences.

No strong claims could be made from these preliminary studies, but the results do reinforce our opinions that some aspects of linguistic structure can be usefully detected from prosodic features, and possibly used to guide syntactic hypothesizing. Further studies with extensive sets of utterances with controlled contrasts must be undertaken, to determine exactly what can be readily determined from prosodic patterns. As each prosodic cue to linguistic structure is firmly established from such experimental studies, that cue may be applied to actual tests with available ARPA speech understanding systems.

3.11 Sentences for Controlled Testing of Acoustic Phonetic Components of Speech Understanding Systems

We have repeatedly noted the need for more controlled testing of prosodic and phonetic analysis procedures, to determine what are the underlying linguistic causes for various changes in acoustic features. This contrasts with the usual way in which researchers select speech texts to test the performance of system components. Frequently, researchers merely record and process a few arbitrary sentences which are selected from the set of possible inputs to the machine. These may be obtained (1) by writing down a few sentences and having prospective talkers read the sentences, or (2) by setting up a mock protocol, in which a talker pretends to be talking to the machine while performing the task for which he hopes to use the speech understanding system. The spontaneous utterances from a protocol are expected to be more representative of the kind of speech that will ultimately be handled by the machine, but the speech read from written texts is more easily controlled. The importance of controlled sentence designs cannot be overemphasized. For ease of processing, for efficiency, and to maximize the number of questions answered by the fewest number of sentences, one would like to pack as many occurrences of relevant phonetic sequences into as few sentences as possible.

A total of 178 "Phonetic Sentences" have been designed (Lea, 1974b) to test the acoustic phonetic components of speech understanding systems, such as acoustic parameterization (e.g., formant tracking and fundamental frequency tracking), phonetic segmentation and categorization (locating each /s/ in the speech,

or finding all unvoiced stops, or determining place of articulation for a voiced stop, etc.), and acoustic phonetic and phonological rules (how consonants are affected by position in a cluster, how pre-stressed and unstressed consonants differ, how vowels are affected by consonantal context, what happens at word boundaries, etc.). One subset of sentences includes occurrences of most English word-initial consonants or consonant clusters, combined with each of eleven stressed vowels, and the word-final consonantal sequences following the stressed vowels. Word-initial and word-final vowels are also included, as are diphthongs coupled with certain consonants. Another subset includes very simple sentence structures with contrasting phonetic categories, such as all-sonorant sentences, sentences where all consonants are unvoiced stops, etc. Finally, another subset includes vowel-vowel sequences at word boundaries, to study English anti-hiatic mechanisms like glottal stops.

A complete description of the Phonetic Sentences, including the total list of sentences, has been published (Lea, 1974b). The sentences will be recorded later, following a modification of procedures described in Sperry Univac Report PX 10952 (Lea, 1974b).

3.12 Sentences for Controlled Testing of Prosodic and Syntactic Components of Speech Understanding Systems

In previous reports (Lea, 1974a, c) we have described a number of specific issues about how prosodic patterns may indicate syntactic structures. A set of 902 sentences have been designed to test such issues (Lea, 1974c). One of the first issues to be systematically addressed with such sentences concerns the effects of the position of the first stressed syllable within a constituent. As stress is moved within a constituent, how do the associated positions of syntactic boundaries move, and how are the acoustic correlates of stress affected? Other questions to be considered include how prosodic patterns may give cues to sentence type, presence of coordinate structures, subordination, and contrastive syntactic bracketings. Various subsets of the sentences test prosodic effects of: adverbs, prenominal and predicate adjectives; possessives and quantifiers; pronouns; boundaries between noun phrases and verbals; negation; there-insertion; noun-verb stress pairs; restrictive and apositive relative clauses; etc.

These "Prososyntactic Sentences" will be recorded along with the Phonetic Sentences. They will permit careful study of the prosodic effects of various isolated differences in linguistic structure, and will provide the necessary information for development of experimentally-verified rules of prosodic structure.

3.13 Usage of ARPANET

As a separate, internally-funded project at Univac, a Very Distant Host (VDH) connection to the ARPANET has been implemented on a Univac 1219 computer (instead of on the Univac 1218 as previously planned). As a single-ended circuit, the VDH functions much like the Terminal Interface Message Processors (TIP's) on the network, without the responsibilities of message packet forwarding, but with the capabilities of efficient large-size packet transmissions.

Completely operational software on the 1219 includes a Network Control Program (NCP), Reliable Transmission Package (RTP), and handlers for a teletype and high-speed line printer. The hardware and software to connect the VDH to the Speech Research Facility 1616 computer is at the debugging stage. This will allow additional input/output devices to be used in network transmissions. The 1616 software might then be expanded to include the higher-level File Transfer Protocol (FTP),to allow high speed program and data exchange across the network.

The 1219 VDH facility has been used extensively for communications with other ARPA contractors, using both the mailbox and direct link capabilities of the network. This has proved a useful mode of getting and sending memos, getting questions answered, and exchanging technical information, including, for example, FORTRAN programs. We have begun to use program facilities on the network as our own work has suggested specific needs. In particular, we have learned how to access phonological rule testers at BBN and SDC. This has allowed us to discover what phonological rules are being formalized by other groups, and how these rules apply to our own vocabularies. We anticipate that such access will allow us to keep abreast of rules as they are incorporated, and aid our own thinking about how a phonological rules component might be incorporated in systems.

4. PLANS FOR FURTHER STUDIES

4.1 Summary

Our work on prosodic guidelines to speech understanding has progressed to a point of considerable success and encouraging results. We have presented theoretical arguments about the need for extracting from the acoustic speech signal some prosodic cues to the large-unit linguistic structure, without dependence upon the prior determination of phonemic structure and recognition of the words in the sentence (Lea, Medress, and Skinner, 1972a). Vital assumptions of a prosodically-guided approach to speech understanding have been verified from a variety of experiments. In particular, stressed syllables have been shown to be of prime importance in speech recognition, because of: (a) the occurrence of stressed syllables in semantically important words; (b) the close correspondence between detected phonetic structure and underlying phonemic structures in stressed syllables; (c) the much higher reliability of phonetic classification possible in stressed syllables (as evidenced by the analysis of results from the CMU Speech Segmentation Workshop); and (d) the vital cues to syntactic structure that stressed syllables provide (as evidenced by the different patterns of stress locations for the alternative interpretations of the BBN problem sentences).

A series of "natural experiments" (cf. Anderson, 1966) have been conducted to determine specific relationships between various acoustic prosodic features, on the one hand, and linguistic structures, perceptions, and abstract notions (such as rhythm, etc.), on the other hand. In such "natural experiments", one does not directly control an independent variable (such as syntactic bracketing) and study resultant changes in a dependent variable (such as valleys in F_0 contours); rather, he simply looks at the data obtained from naturally-occurring phenomena (such as the speech previously recorded and identified as the Rainbow Script, Monosyllabic Script, and the 31 ARPA Sentences). From studies of such speech texts, we have demonstrated that over 90% of all intuitively predicted syntactic boundaries are detected from substantial fall-rise valleys in F_0 contours. We have shown that over 85% of all syllables perceived as stressed are located by a particular combination of energy duration and F_0 cues, assuming archetype F_0 contours for constituents. Available methods for automatic phonetic classifications have been shown to be most reliable within the syllables perceived as stressed, for the 31 ARPA Sentences. Stressed syllabic nuclei have been

shown to have a rough tendency toward equal spacing in time, though the dependence upon the number of intervening unstressed syllables is very prominent. Pause durations also appear to correlate with average time intervals between stressed nuclei. Finally, a look at a few similar sentence structures in the BBN problem sentences has shown considerable hope for using prosodic cues to select correct syntactic structures. Those sentences showed distinctive stress patterns, positions of detected constituent boundaries, and F_0 contour parameters (such as the F_0 fall from the peak value in the sentence to the maximum value in the last stressed syllable), which can distinguish yes/no questions and commands, and determine the correct bracketing of the ambiguous word sequences.

In such natural experiments, one cannot be certain that some unknown third variable is not the source of any apparent relationships between the acoustic variable and the underlying abstract variable. Controlled experiments, with all variables except one fixed in the comparison of two utterances, provide the proper extension from the encouraging results of the natural experiments. The designed speech texts provide the necessary controls and sufficient data to extend these encouraging tendencies into well-defined rules relating prosodic variables and linguistic structure. Now that the controlled speech texts have been designed, and separate reports have been published describing the Phonetic Sentences (Lea, 1974b) and the Prosyntactic Sentences (Lea, 1974c), these controlled experiments can begin.

4.2 Controlled Experiments with Designed Sentences

The designed sentences will soon be recorded by five male and three female talkers, with three repetitions spaced a week or more apart. Initially, a subset of those sentences (which subset investigates the effects of stress movement within constituents) will be analyzed. Such analysis includes: obtaining F_0 and energy functions; applying the boundary detection algorithm; automatically locating stressed syllables; obtaining listener's stress perceptions for the subset of sentences; conducting a thorough evaluation of the results of these prosodic analyses; and determining whether rules can be written relating stress positions within constituents to expected positions of detected constituent boundaries, acoustic correlates of stress, etc.

In subsequent controlled experiments with the designed texts, we plan to study prosodic cues to sentence type, syntactic bracketing, subordination, and coordination.

We are particularly interested in investigating the combined effects of syntactic, lexical, and phonetic structures on intonation contours, and developing intonation rules that relate fundamental frequency contours to linguistic structures. Only by such a systematic attack on the task of compiling experimentally-verified rules and developing consequent tools for acoustic prosodic analysis can one hope to provide the kind of reliability needed to make such prosodic data of major value in speech understanding systems.

4.3 Use of Prosodics in Speech Understanding Systems

The development of useful rules for relating prosodic patterns to linguistic structure also demands the direct application of those rules to working systems, to evaluate their accuracy and utility. Sperry Univac plans to be involved in several forms of interaction with ARPA contractors, to cooperatively evaluate various prosodic tools and rules. In particular, Sperry Univac is proposing to: analyze selected problem sentences for systems contractors, to see how prosodic information may help select correct structures; and to investigate prosodic aids to parsing (with the BBN system, for example). Sperry Univac will also be participating in other interactions such as phonological rules workshops, ARPA Speech Understanding Research (SUR) Steering Committee activities, and data base selections.

To effectively impact the ARPA program of building speech understanding systems that meet the original specifications by 1976, prosodic features should be introduced into the developing systems as early as possible. Sperry Univac believes that prosodies can provide crucial contributions to word matching, parsing, semantic analysis, and phonological analysis. It is noteworthy that the Steering Committee of the ARPA Speech Understanding Research (SUR) program has also suggested that a major attack should be mounted on the area of prosodics, since this source of knowledge has not been used in any previous system, though it offers the possibility of a unique contribution to sentence disambiguation and overall system control strategies.

From previous discussions with systems contractors, we already anticipate several ways in which prosodic information may be directly useful in current speech understanding systems. One way is in guiding procedures for locating and matching words in sentences. In long phonetic sequences (especially segment lattices which allow several alternative boundaries for segments, and which allow several segment

labels to be assigned to various segments), it is difficult to tell where a candidate word might begin and where it might end. The location and phonetic specification of stressed syllables may provide reliable anchor points around which the search for an adequate word match might be attempted. Particular weight might thus be given to the most reliable information in the word. Also, words that might appear to be possible candidates for insertion at various points in the phonetic sequence may be ruled out if they have the wrong stress patterns.

Because of this potential for stressed syllables aiding word matching, all three ARPA systems contractors have asked Sperry Univac to provide the stressed syllable location program for inclusion in their systems. We plan to supply the program and related documentation as soon as it is totally operational. Since that program works with syntactic boundary detections as input data, the boundary detection program will also be supplied, after it has been refined to eliminate some of the false detections, and to provide confidence measures for each detection.

In addition to providing systems contractors our improved procedures for syntactic boundary detection and stressed syllable location, we plan to investigate specific ways in which prosodic information may aid structural analyses. For example, we may use typed input to BBN TENEX, through the ARPANET, to access the BBN parser. After learning how to use the parser to parse some BBN sentences, we may study its use with some of the designed test sentences, to learn where ambiguities and problems occur that might be resolved with prosodic data. We plan to cooperate with BBN in developing ways of introducing useful prosodic data into their system, such as by using prosodic patterns to specify predicate functions attached to transition arcs used in parsing (cf. Bates, 1974). Prosodic patterns may be used to assign priorities or likelihood information to various transition arcs, and to provide further qualifying information which must be satisfied before pop-up procedures are undergone. Similar introductions of prosodic information into other parsers could be attempted.

As each hypothesis about prosodic cues to syntactic structure is tested with the designed sentences and found to provide useful information, we may investigate its "a posteriori" use in disambiguating, or selecting among competing structures for an utterance. We may also consider how that prosodic information can be introduced early in parsing procedures, to avoid fruitless paths of search in structural analysis, thus giving "a priori" guidelines to efficient parsing.

4.4 Compiling Useful Prosodic Rules

There is a definite need to develop precise rules for systematically relating prosodic patterns to underlying structures. The previous predictions of where phrase boundaries should occur in F_0 contours have been based on intuitive analyses of syntactic structures. Where expected boundaries do not occur, or wherever false or unexpected boundaries do occur, there has been no recourse indicating the source of the error. This is in part due to the intuitive predictions used, and in part due to the uncontrolled syntactic structures involved in the texts studied. Experiments with the designed sentences of known syntactic structure will indicate exactly what structural boundaries are marked, and will permit the writing of precise rules predicting where boundaries will occur in new sentences of similar structures. These rules for predicting detectable boundaries may then be useful in computer determination of possible underlying structure given the detected boundaries.

Similarly, precise rules for relating stress patterns to underlying structures are needed. Published (theoretical) hypotheses about sentence stress patterns can best be tested with experimental data such as that to be analyzed in the proposed studies. Such systematic studies are needed to ultimately write analytical procedures for predicting underlying structures from stress patterns.

Of primary importance in such prosodic studies is the development of English intonation rules. Intonation rules cut across the whole gamut of problems involved in speech understanding, including the explanation of why constituent boundaries are detectable in F_0 contours, what are the acoustic correlates of stressed syllables, how syntactic and semantic structures might be manifested acoustically, what are useful phonological rules and morphological rules in various contexts, and how stress rules might be inferred from acoustic data. If one can understand how the interacting effects of semantics, syntax, lexical structures, stress patterns, and phonetic sequences are superimposed in the F_0 and energy contours of controlled English sentences, he has some of the most essential tools for using acoustic prosodic data to guide speech understanding strategies.

We plan to investigate such rules of English intonation, plus other rules about stress patterns, rhythm, and rate of speech, as time and resources permit. We also plan to continue supporting and participating in the ARPA workshops on acoustic phonetic and phonological rules. Besides general participation in the compilation and evaluation of phonological rules, Sperry Univac will endeavor to provide and test any useful prosodic rules.

5. PUBLICATIONS AND REPORTS

The following is a complete list of Sperry Univac publications, reports, presentations, and unpublished ARPA SUR Notes, resulting from ARPA funding to data:

Publications and Reports

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

LEA, W. A., Influences of Phonetic Sequences and Stress on Fundamental Frequency Contours of Isolated Words, J. Acoust. Soc. of America, Vol. 53, January, 1973, 346(A).

LEA, W. A., MEDRESS, M. F., and SKINNER, T. E. Use of Syntactic Segmentation and Stressed Syllable Location in Phonemic Recognition, J. Acoust. Soc. of America, Vol. 53, January, 1973, 356(A).

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

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

LEA, W. A., MEDRESS, M. F., and SKINNER, T. E., Prosodic Aids to Speech Recognition: III. Relationships between Stress and Phonemic Recognition Results, Univac Report No. PX 10430, Univac Park, St. Paul, Minnesota, September, 1973.

LEA, W.A., MEDRESS, M. F., and SKINNER, T.E., A Prosodically-Guided Speech Understanding Strategy, Proc. IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, Penn., April, 1974, 38-44.

LEA, W. A., Prosodic Aids to Speech Recognition IV: A General Strategy for Prosodically-Guided Speech Understanding, Univac Report No. PX 10791, Univac Park, St. Paul, Minnesota, March, 1974.

LEA, W. A., Sentences for Controlled Testing of Acoustic Phonetic Components of Speech Understanding Systems, Univac Report PX 10952, Univac Park, St. Paul, Minnesota, September, 1074.

LEA, W. A., Sentences for Controlled Testing of Prosodic and Syntactic Components of Speech Understanding Systems, Univac Report PX 10953, Univac Park, St. Paul, Minnesota, October, 1974.

Oral Presentations

LEA, W. A., Influences of Phonetic Sequences and Stress on Fundamental Frequency Contours of Isolated Words, presented at the 84th Meeting of the Acoustical Society of America, Miami Beach, Florida, November, 1972.

LEA, W. A., Use of Syntactic Segmentation and Stressed Syllable Location in Phonemic Recognition, presented at the 84th Meeting of the Acoustical Society of America, Miami Beach, Florida, November, 1972 (coauthored with Mark F. Medress and Toby E. Skinner).

LEA, W. A., Prosodic Features and Linguistic Structure, presented at the ARPA Tutorial Lectures on Acoustic-Phonetic Characteristics of English Sentences, Cambridge, Massachusetts, December, 1972.

LEA, W, A., MEDRESS, M. F., and SKINNER, T. E., A Prosodically Guided Speech Understanding Strategy, presented at IEEE Symposium on Speech Recognition, Carnegie Mellon University, Pittsburgh, Pa., April, 1974.

LEA, W. A., "Perceived Stress as the 'Standard' for Judging Acoustical Correlates of Stress", presented at the 86th Meeting of the Acoustical Society of America, Los Angeles, California, November, 1973.

LEA, W. A., "Evidence that Stressed Syllables Are the Most Readily Decoded Portions of Continuous Speech", presented at the 86th Meeting of the Acoustical Society of America, Los Angeles, California, November, 1973.

LEA, W. A., "An Algorithm for Locating Stressed Syllables in Continuous Speech", presented at the 86th Meeting of the Acoustical Society of America, Los Angeles, California, November, 1973.

LEA, W. A , Prosodic Phenomena, session chaired at ARPA Phonological Rules Workshop, Systems Development Corporation, Santa Monica, California, June, 1974.

LEA, W. A., A Speech Data Base for Testing Components of Speech Understanding Systems, presented at 88th Meeting of the Acoustical Society of America, St. Louis, Missouri, November, 1974.

LEA, W. A., Prosodic Hypotheses and Rules, to be presented at the ARPA Workshop on Acoustic Phonetic and Phonological Rules, Bolt Beranek and Newman, Cambridge, Massachusetts, November, 1974.

Unpublished ARPA SUR Notes

2. MEDRESS, M. F., The Univac Speech Recognition Study (5 pages), December, 1971.

16. MEDRESS, M. F., Proposed Computer Phonetic Transcriptions (2 pages), February, 1972.

17. MEDRESS, M. F., Univac Speech Bibliography (1 page), February, 1972.

32. MEDRESS, M.F., Revised Computer Phonetic Representations (2 pages), April 1972.

36. MEDRESS, M.F., et al., "Acoustic Correlates of Linguistic Stress" (Literature Survey, 22 pages), June, 1972.

39. LEA, W. A., Acoustic Cues for Boundaries between Syntactic Units (6 pages), August, 1972.

45. LEA, W. A., Considerations in the Design of Good Speech Texts (7 pages), September, 1972.

48. MEDRESS, M. F., Plans for the First Segment of the Speech Data Base (2 pages), October, 1972.

49. MEDRESS, M. F., et al., Prosodic Aids to Speech Recognition (68 pages), October, 1972.

53. LEA, W. A., MEDRESS, M. F., and SKINNER, T.E., Use of Syntactic Segmentation and Stressed Syllable Location in Phonemic Recognition (11 pages), December, 1972.

54. LEA, W. A., Syntactic Factors in the Initial Selection of Sentences for the Data Base (7 pages), December, 1972.

63. LEA, W. A., Some Factors in the Selection of Utterances for Speaker Normalization (3 pages), February, 1973.

67. LEA, W. A., Acoustic Analysis of 13 ARPA Sentences (4 pages), February, 1973.

82. LEA, W. A., MEDRESS, M. F., and SKINNER, T.E., Univac Final Technical Report: Prosodic Aids to Speech Recognition II: Syntactic Segmentation and Stressed Syllable Location (34 pages), May, 1973.

108. LEA, W. A., MEDRESS, M. F., and SKINNER, T. E., Prosodic Aids to Speech Recognition: III. Relationships between Stress and Phonemic Recognition Results (6 pages), October, 1973.

139. MEDRESS, M. F., Prosodic Aids to Recognition: IV, A General Strategy for Prosodically-Guided Speech Understanding (65 pages), May, 1974.

141. LEA, W.A., Sentences for Testing Acoustic Phonetic Components of Systems (18 pages), July, 1974.

6. REFERENCES

ANDERSON, B. F. (1966). The Psychology Experiment, Belmont, California; Brooks/Cole Publishing Co.

BATES, M. (1974), The Use of Syntax in a Speech Understanding System, Proc. IEEE Symposium on Speech Recognition, Carnegie-Mellon Unuversity, Pittsburgh, Pa., 226-233.

CHOMSKY, N. and HALLE, M. (1968), The Sound Pattern of English, New York; Harper and Row.

CHOMSKY, N. and MILLER, G. A. (1963), "Introduction of the Formal Analysis of Natural Languages", in Handbook of Mathematical Psychology, pp. 269-321; Ed. R. D. Luce, R. R. Bush and E. Galanter, New York: John Wiley and Sons, Inc.

GLEITMAN, L. R., and GLEITMAN, H. (1970). Phrase and Paraphrase, New York, W. W. Norton and Co.

LEA, W. A. (1971), Automatic Detection of Constituent Boundaries in Spoken English, J. Acoust. Soc. of America, Vol. 50, 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.

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), Evidence that Stressed Syllables are the Most Readily Decoded Portions of Continuous Speech, presented at the 86th Meeting of the Acoustical Society of America, Los Angeles, October-November, 1973 (Paper Y14).

LEA, W. A. (1973d), An Algorithm for Finding Stressed Syllables in Continuous Speech, presented at the 86th Meeting of the Acoustical Society of America, Los Angeles, October-November, 1973 (Paper M9).

LEA, W. A. (1974a), Prosodic Aids to Speech Recognition IV. A General Strategy for Prosodically-Guided Speech Understanding, Univac Report No. PX 10791, Univac Park, St. Paul, Minnesota.

LEA, W. A. (1974b), Sentences for Controlled Testing of Acoustic Phonetic Components of Speech Understanding Systems, Univac Report No. PX 10952, Univac Park, St. Paul, Minnesota.

LEA, W. A (1974b), Sentences for Controlled Testing of Prosodic and Syntactic Components of Speech Understanding Systems, Univac Report No. PX 10953, Univac Park, St. Paul, Minnesota.

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

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

LEA, W. A., MEDRESS, M. F., and SKINNER, T. E. (1973b), "Prosodic Aids to Speech Recognition: III. Relationships between Stress and Phonemic Recognition Results", Univac DSD, Univac Report No. PX 10430, Univac Park, St. Paul, Minnesota.

LEA, W. A., MEDRESS, M. F., and SKINNER, T. E. (1974), A Prosodically-Guided Speech Understanding Strategy, Proc. IEEE Symposium on Speech Recognition, Carnegie-Mellon University. Pittsburgh, Penn., 38-44.

MILLER, G. A. (1962), Decision Units in the Perception of Speech, IRE Trans. on Info. Th., Vol. IT-8, pp. 81-3.

SHAFER, R. W. and RABINER, L. R. (1970), System for Automatic Formant Analysis of Voiced Speech, J. Acoust. Soc. Amer., Vol. 47, pp. 634-648.

SKINNER, T. E. (1974), A Speech Recognition System Using Reliably Encoded Segmental Information as Applied to Telephone Bandwidth Isolated Digits, Univac Report No. PX 10701, Univac Park, St. Paul, Minnesota.

SONDHI, M. M. (1968), New Methods of Pitch Extraction, IEEE Trans. on Audio and Electroacoustics, Vol. AU-16, pp. 262-266.