Stress and vowel duration effects on syllable recognition[a]

Charles W. Marshall[b] and Patrick W. Nye
Haskins Laboratories, New Haven, Connecticut 06510

(Received 13 December 1982; accepted for publication 15 April 1983)

Systems designed to recognize continuous speech must be able to adapt to many types of acoustic variation, including variations in stress. A speaker-dependent recognition study was conducted on a group of stressed and destressed syllables. These syllables, some containing the short vowel /ɪ/ and others the long vowel /æ/, were excised from continuous speech and transformed into arrays of cepstral coefficients at two levels of precision. From these data, four types of template dictionaries varying in size and stress composition were formed by a time-warping procedure. Recognition performance data were gathered from listeners and from a computer recognition algorithm that also employed warping. It was found that for a significant portion of the data base, stressed and destressed versions of the same syllable are sufficiently different from one another as to justify the use of separate dictionary templates. Second, destressed syllables exhibit roughly the same acoustic variance as their stressed counterparts. Third, long vowels tend to be involved in proportionally fewer cross-vowel errors but tend to diminish the warping algorithm's ability to discriminate consonantal information. Finally, the pattern of consonant errors that listeners make as a function of vowel length shows significant differences from that produced by the computer.

PACS numbers: 43.70.Gr, 43.70.Sc, 43.70.Dn

INTRODUCTION

To keep the analysis task within practical bounds, some form of segmentation of the acoustic signal into analyzable units is an intrinsic feature of all current computer-based speech recognition methods. The choice of segments actually employed in recognition algorithms and in recognition studies has encompassed a wide variation in duration. This has ranged, for example, from centisecond units (Bahl *et al.*, 1978) to phonemic segments (Klatt, 1978) to demisyllables (Dixon and Silverman, 1977; Rosenberg *et al.*, 1981) and beyond to syllables (Fujimura, 1975) and to words (Rabiner and Wilpon, 1979). Moreover, among these different choices, syllables and syllable-sized units have been lately receiving increasing attention.

There are several important features that qualify the syllable as a recognition unit. First, one must acknowledge the evidence that both speakers and listeners are aware of the existence of syllables and that they are usually in good agreement as to the number present in a given utterance. Second, syllables are the smallest units that can be uttered in isolation and for which, in many instances, it can be claimed that they are produced by completely executed articulatory gestures (roughly defined as maneuvers involving a single opening and closing of the vocal tract which, in turn, cause transient increases in the acoustic energy contour). Third, further merit stems from the fact that, especially for closed syllables (CVCs), the coarticulation effects between the phones within the syllable can be assumed (on average) to be stronger than they are across syllable boundaries. Hence, in principle, the selection of the syllable as a recognition unit should present a simpler segmentation task because the boundaries are located in the less strongly coarticulated regions of the signal (Fujimura, 1975).[1] Fourth, syllables may also be said to hold a strong claim to being the authentic building blocks of speech because they constitute many common words in their entirety and can be combined in appropriate sequences to form all the multisyllabic words as well. And finally, syllables provide the basis for an important feature of word and sentence patterning whereby, through the exercise of selective syllable emphasis (stressing) and lack of emphasis (destressing), information about the syntactic structure and semantic content of a sentence is encoded in the acoustic signal.

However, variations in syllable stress bring about significant changes in the acoustic duration and spectral composition of most syllables. The magnitude of these changes can vary considerably with speaking rate, syntactic role, and phonetic context. Thus the effects of stress variation are an inherent feature of speech acoustics—a feature that must be accommodated by all recognition systems. Included among these systems are, of course, those that seek to identify linguistically relevant entities such as syllables, usually by matching acoustic segments to a dictionary of templates. Proposals for countering acoustic variation have generally taken one of two extreme positions, which can be referred to as *collection* versus *computation*. These positions hold that the template dictionary should either include (1) a *collection* of all the allophonic variants of each syllable to be recognized, or (2) only canonical, or stressed, examples from which all the expected variants are *computed* by an algorithm. The former approach carries the requirement of a large memory capacity, while the latter one promises a significantly lower memory cost which has to be traded against

[a] This is a revised and expanded version of an oral paper presented at the 102nd Meeting of the Acoustical Society of America, 30 November–4 December 1981, Miami Beach, FL [J. Acoust. Soc. Am. Suppl. 1 70, S61 (1981)].

[b] Present address: Department of Computer Science, Yale University, New Haven, CT 06520.

 0001-4966/83/080433-11$00.80

a somewhat increased computation cost and is consequently of practical as well as theoretical interest.

In this paper, we report on a preliminary investigation into the problem of linguistic variation and dictionary composition and describe data that have a bearing on the collection versus computation issue. Using selected sets of syllable-sized segments—some stressed and some destressed—taken from continuously spoken speech, we examined the recognition performance of a computer algorithm and compared it with that of human listeners. For computer recognition purposes, we used a syllable recognition algorithm prepared by Mermelstein (1978). Because it was expected that the severity of stress effects might vary as a function of phonological vowel length, two groups of syllables were employed, one incorporating the short vowel /ɪ/ and the other the long vowel /æ/. The study obtained empirical estimates of the error rates that occur during the recognition of stressed and destressed syllables (1) as a function of vowel length and (2) for dictionaries containing different combinations of stressed and destressed syllables. A study of the cluster structures produced by stressed and destressed syllables in a cepstral distance space was also undertaken.

I. METHODS

A. Selection of syllables

Twenty-three pairs of vocabulary words were employed from a set of 24 pairs that had been originally selected. (The 24th pair was eliminated after a preliminary examination of the acoustic data.) Twelve pairs contained CVC syllables with an /ɪ/ vowel nucleus while the remainder contained similar syllables incorporating the vowel /æ/. One word of each pair (e.g., *tid*bit) contained the target syllable [tɪd] in stressed form while another word (e.g., wan*ted*) contained its destressed counterpart. When choosing the words containing destressed examples of each syllable, a deliberate attempt was made to select only those in which, in the judgment of our linguist colleagues, the color of the nuclear vowels, when spoken by eastern American speakers, would not be likely to go to shwa when destressed.[2] Table I contains the vocabulary items that were included in a total of 58 sentences. The sentences were structured in such a way that the contrast between stressed and destressed syllables was retained and the placement of any of the vocabulary words in sentence-final position was carefully avoided.[3] For example, one of the sentences was "Old Bag*dad* on the Tigris offered an array of *fan*tastic delights," which contained the syllables [dæd] and [fæn]. The sentences were composed in a variety of syntactic forms to induce the production of different speaking rhythms and to offset any reader tendency to adopt a sing-song or monotonous delivery. Each vocabulary word occupied at least two different contexts in the sentence set. However, four syllables were inadvertently included three times. They were the stressed syllables [læm], [tæd], [mæn], and the destressed syllable [dɪg].

TABLE I. Syllables employed in recognition study.

Syllables containing /ɪ/		Syllables containing /æ/	
Stressed	Destressed	Stressed	Destressed
*Rig*marole	Out*rigger*	*Cat*alog	*Cat*astrophic
Dignification	*Indignation*	*Tactics*	Tictactoe
Indigenous	*Indigestion*	*Lamb*faced	*Lamb*aste
*Fil*trate	In*fil*trate	*Fat*uous	Ara*fat*
Simple	*Simplicity*	*Tangent*	*Tang*erine
Per*miss*able	Pre*mise*	*Fantail*	*Fantastic*
Distant	*Distinguish*	*Dad*dy	Bag*dad*
*Tid*bit	Wan*ted*	Auto*mat*ic	Auto*mat*
*Lit*mus	Star*lit*	*Hap*less	Mis*hap*
Bin	Coal-*bin*	*Manic*	Bag*man*
History	*Historic*	*Bag*man	Grab-*bag*
Sister	Cathar*sis*		

B. Speaker characteristics

Two male speakers were employed (DZ and LL) to allow speaker-dependent effects to emerge. Both were natives of the eastern United States who had accents typical of that region. Each speaker read the list of sentences under instructions to imagine himself in circumstances in which each of the sentences might have been spoken and to reproduce them in an extemporaneous manner. During a preliminary examination of their speech data, it was found that one of the originally selected syllables failed to retain its vowel color when destressed and, therefore, it was eliminated from the study, leaving a total of 23 syllables. Four recording sessions were scheduled for each speaker at minimum intervals of about two weeks. Two recordings of the sentences were made at each recording session. Thus the speakers provided eight different readings of each sentence and at least 16 examples of each syllable pair (the four syllables noted above each yielded 24 examples).

C. Parametric conversion procedures

After low-pass filtering at 4.9 kHz, the speech material was digitized at a 10-kHz rate and stored. A phonetician then isolated the target syllables by examining a display of the digitized waveform, adjusting a pair of cursors to mark the head and tail of each syllable at a zero crossing point in the waveform, and verifying the identity of the segment by listening to its output reproduced through a digital-to-analog converter and loudspeaker. The phonetician also made vowel duration measurements on a portion of the speech data from both speakers. Segmentation by visual inspection was preferred over automatic segmentation in order to keep the number of segmentation errors to an absolute minimum. Earlier work with an automatic segmentation algorithm (Mermelstein, 1975) has revealed the types of segmentation errors that automatic processing tends to introduce.[4]

Having been isolated by hand, the sampled representations of the syllables were converted into sequences of cepstral coefficient vectors at two levels of precision. For the first precision level (PL1) spectral values were obtained by FFT analysis of the digitized segments at a frame interval of 128 samples; for the second precision level (PL2) the interval was set at the higher resolution level of 64 samples per frame interval. In both cases, a frame consisted of 256 samples weighted by a Hamming window. Then, to shape the spec-

tral energy content of the data so that it more closely resembled the frequency response of the human ear, the logarithms of the spectral amplitudes were weighted by a group of 20 triangular filters located at equal intervals along the mel-scale of frequency. This was done to gain the enhanced performance achieved previously with this transform (Davis, 1979; Davis and Mermelstein, 1980). Next, vector arrays of six cepstral coefficients were computed at PL1 and 10 coefficients at PL2 for successive time-frame intervals (the gain-dependent zeroth coefficient was omitted from these arrays). Therefore, for any given syllable, the number of PL2 coefficients exceeded the number of PL1 coefficients by a factor of 3.3.

D. Template construction and distance measurement

The procedure for creating syllable templates from the available tokens employed a dynamic programming algorithm described by Mermelstein (1976, 1978). This algorithm was based on principles employed in earlier work (Velichko and Zagoruyko, 1970; Bridle and Brown, 1974; Itakura, 1975) but differed from that work in some important details.

Each syllable was represented by a temporal sequence of mel-scale cepstral coefficient vectors. These vectors formed a matrix with the nth row representing the feature vector for the nth time frame. The nonlinear warping consisted of selectively repeating or deleting rows in pairs of matrices.

Before warping any pair of syllables together to form a template, an initial optimum alignment was found by adding to each end of the shorter syllable an amount of silence equivalent to the difference in duration. Then this syllable, plus its silent attachments, was shifted with respect to the longer syllable until an interim minimum in the distance between the syllables (i.e., a minimum in the summed squares of the cepstral differences of corresponding time frames) had been found. At this point, the excess silence at the edges of the shorter syllable was pruned away so that the two matrices contained the same number of rows.

Following this length equalization and alignment, the nonlinear warping algorithm was used to dynamically form the pattern of repetitions and deletions of rows from each matrix that gave the best match between them. The procedure involved the warping of both matrices onto a third time sequence (Sakoe and Chiba, 1978) and the derivation of a symmetric distance measure based on the sum of the squares of corresponding vector elements. The possible warps were constrained in such a way that the ends of the matrices always remained aligned together. Out of the warping procedure, the optimum path and its associated minimum distance were obtained. The optimum path was used to specify the corresponding time frames that were subsequently averaged together during template construction and, during recognition, the inverse of the computed distance was employed as a measure of the likelihood that a token represented the same syllable as a given template.

Having averaged two tokens together to form the first interim version of a template, this template was then warped together with a new token and the average of the resulting pair of matrices was computed by a procedure that weighted the matrix representing the interim template in proportion to the number of tokens it already contained. This process was repeated until the supply of tokens was exhausted—usually after the fourth or eighth warp.

TABLE II. The speaking sessions that served as tokens and templates.

Run No.	Tokens		Templates
1	Session 1	tested against	Session 2
2	Session 1	tested against	Session 3
3	Session 2	tested against	Session 4
4	Session 2	tested against	Session 3
5	Session 3	tested against	Session 4
6	Session 4	tested against	Session 1

The tokens used to construct templates were warped together in a fixed order but, to minimize possible order effects, four groups of dictionaries (one from each of the four speaking sessions) were formed and distance measurements were computed between each of these dictionaries and tokens drawn from one or more of the other three sessions. Thus tokens to be recognized were never components of the template sets (dictionaries) against which they were matched; they were, however, drawn from the same words and sentence contexts as the templates, and they were spoken by the same speaker but at a different session. The pattern of comparisons is shown in Table II.

E. Composition of the dictionaries

The four groups of syllable tokens produced by each of the two speakers (one group per speaking session) were converted into parametric form at both levels of precision. Following conversion, the tokens of each group were warped together by the dynamic programming technique (Mermelstein, 1978; Rabiner *et al.*, 1978) to give three classes of templates from which four dictionaries per speaker were derived (see the flowchart shown in Fig. 1).

The "stressed" (S) dictionaries contained templates formed by warping together only stressed tokens, while the "destressed" (D) dictionaries contained templates formed exclusively from destressed tokens. Consequently, each of these dictionaries contained 23 entries. The "combined" (C) dictionaries were formed by warping the stressed and destressed occurrences of each syllable token together and, therefore, also numbered 23 entries. The "both" (B) dictionaries contained the union of the stressed and destressed templates formed from a given speaking session (i.e., dictionaries S plus D); hence, they were twice the size of the other dictionaries and contained a total of 46 templates. As already noted, one dictionary was formed from each speaking session. Therefore, the total number of dictionaries produced amounted to 32 (four sessions × two speakers × four dictionary types).

During the recognition procedure, a warping was performed for each token with each of the templates in the appropriate dictionary (see Table II) and the "recognized" syllable was identified as the top member of the list of

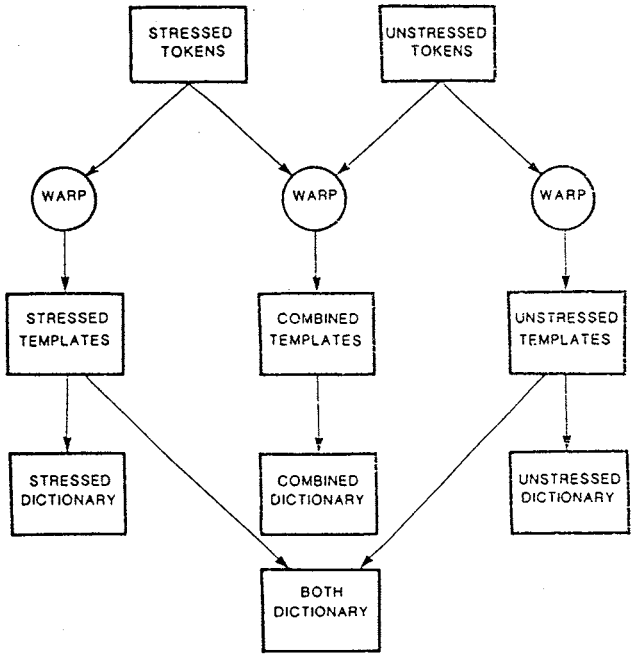

FIG. 1. Flowchart illustrates the production of four types of dictionaries labeled B, C, S, and D. For each such dictionary, the source data were stressed and destressed tokens extracted from a single speaking session.

hypothesized candidate syllables ranked in order of increasing token-template distance. These lists were employed in later studies that examined, in cases where the top candidate was in error, the frequency with which the correct choice appeared later in the list.

F. Collection of data from listeners

To establish a baseline from which to assess and, perhaps, to gain further insights into the performance of the computer recognition algorithm, a recognition test using the same isolated speech segments was presented to a group of ten listeners. These listeners consisted of colleagues and their graduate students. All had taken part in many previous experiments of a similar nature and were fully familiar with the phonetic alphabet. They were given a list of the 23 syllables in phonetic transcription, informed that each presentation would be drawn from that list, and instructed to record each identification (or guess if necessary) by placing a check in a column below the appropriate entry in the list. The listeners were not asked to record stress levels. The syllables were delivered to the listeners at 5-s intervals via TDH-39 earphones from a tape recording of the computer output. Five seconds between each stimulus provided sufficient time for the listeners to make their responses. However, to ensure the detection and avoidance of missed responses, an 8-s interval was inserted after each group of five syllables and a 10-s interval after every 20th syllable. The listeners heard (in random order) all of the target syllables produced by both speakers. Each one was repeated four times. Four of the syllables, as noted earlier, were inadvertently repeated six times. Hence, each subject heard 192 syllable presentations from each speaker. The subjects' identification data were then entered into the computer and stimulus/response matrices for both the stressed and destressed syllables of each speaker were constructed.

II. RESULTS

A. Introduction

The results were examined from several points of view. To verify that our speech data did actually contain the expected durational variations, vowel duration and syllable duration measurements were examined. Then, the computer recognition errors were sorted and analyzed by precision level, vowel type, dictionary type, and stress. The data gathered from human listeners were, where possible, sorted and analyzed in similar fashion and compared with the computer results. Finally, the acoustic parameters were examined by means of a multidimensional scaling technique to reveal the clustering structures of stressed and destressed syllables.

B. Phonological versus physical vowel durations

Phoneticians have long believed that the vowel /æ/ has a longer duration in American English speech than the vowel /ɪ/. The classic experimental support for this assertion was provided by Peterson and Lehiste (1960), who showed that the intrinsic durations of /æ/ and /ɪ/ as syllabic nuclei in American English averaged 330 and 180 ms. However, they also observed that the length of a syllabic nucleus varied according to whether it was followed by a voiced or voiceless consonant. Since the final consonants of the CVC syllables employed in this study were drawn from both voiced and voiceless classes without regard to ensuring equal representation, it was necessary to verify empirically that a significant difference in duration was retained for the syllables we had chosen. To do this, it was deemed sufficient to perform vowel duration measurements on a representative portion of the data base and, for this purpose, data from one session by each speaker were selected. In contrast with the measurement procedure adopted by Peterson and Lehiste, which tended to include a large portion of the consonantal transition as a part of the vowel, the vowel durations measured in this study were confined to so-called steady-state regions of the syllables. These regions were defined as those portions of the syllables in which the cepstral frequencies did not deviate by more than 10% from their central values. Average overall durations of the syllables containing /æ/ and /ɪ/ were computed from the total numbers of samples stored per syllable.

The results of the vowel duration measurements are shown in Table III and reveal that, on average, the durations obtained from speaker DZ were just a few percent shorter than those obtained from speaker LL. (The difference between the speakers in overall *syllable* duration was, however, considerably larger—about 35%.) The difference in median duration between stressed and destressed productions of the vowel /ɪ/ are shown in the table to be 9 ms in the case of LL and 11 ms for DZ. Smaller reductions are apparent for the vowel /æ/. (A difference of the same sign was also evident in the overall *syllable* durations.) Thus the syllables incorporating long vowels tended to retain the property of vowel length, while those incorporating short vowels were

TABLE III. Vowel duration measurements in milliseconds.

	Subject DZ				
	Stressed		Destressed		Overall
	/æ/	/ɪ/	/æ/	/ɪ/	averages
Median	70	45	64	34	53.2
Std. Dev.	14	9	11	10	11.2
Average	69	43	66	39	54.2
Maximum	102	64	99	77	85.5
Minimum	49	25	49	26	37.2

	Subject LL				
	Stressed		Destressed		Overall
	/æ/	/ɪ/	/æ/	/ɪ/	averages
Median	67	49	64	40	55.0
Std. Dev.	12	8	12	10	10.6
Average	71	45	64	42	55.5
Maximum	97	61	88	78	81.0
Minimum	50	30	40	27	36.8

found to exhibit even further shortening in their destressed forms. In addition, it was found that destressing caused the consonantal regions of the syllables to be reduced in amplitude and overall spectral definition.

C. Overall errors in computer syllable recognition

The overall effects of stress on the performance of the recognition algorithm are best summarized in terms of the average error per syllable. Figure 2 shows the percentages of recognition errors made per syllable on the speech of LL and DZ as a function of the dictionary type and precision level. The data were obtained by averaging over six recognition runs. Each run was "open" and speaker dependent and compared all 192 tokens from one session with each of the four dictionary types (containing 23 or 46 templates). The syllables obtained from each recording session were employed once as the raw material for a group of dictionaries and one or more times as the unknowns (see Table II). The error data for dictionary B in Fig. 2 neglected errors in stress assignment.

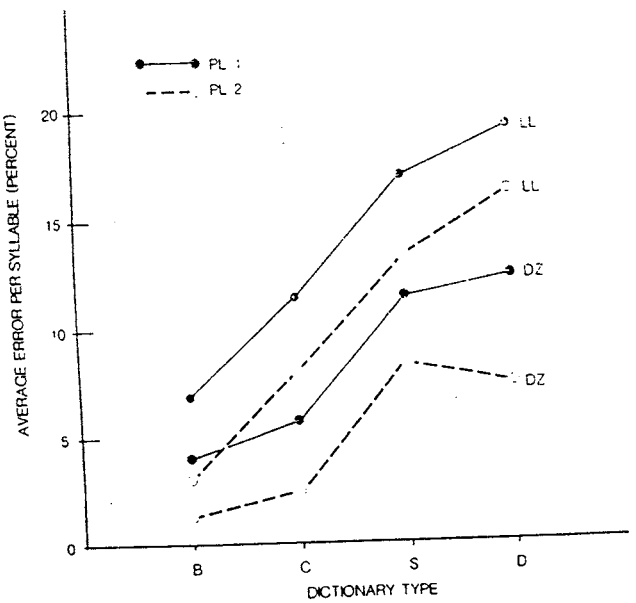

FIG. 2. The average error per syllable plotted against dictionary type for two speakers (LL and DZ) and at two precision levels (PL1 and PL2). At PL1 spectral values were computed at a frame interval of 128 samples and at PL2 the frame interval was set at 64 samples. Window size remained fixed at 256 samples.

The unknown tokens comprised equal numbers of stressed and destressed syllables whereas the dictionaries, except for B, contained only one template per syllable. Hence, recognition by the algorithm was considered correct when the syllable identity of the token (without regard to its stress) agreed with that of the template. Only for dictionary B was it possible to get separate estimates for errors of identity and of stress level. Confusion matrices for each of the individual recognition runs were formed and these were later summed together to create a single matrix from which were calculated the average error for each dictionary type, precision level, and speaker.

Four principal findings emerge from these data. The first is that the B dictionary gives the best overall performance. Second, the C dictionary is superior to both the S and D dictionaries. Third, the performance for the higher precision level (PL2) is significantly better than that for the lower precision level (PL1). Finally, all these features are apparent in the data of both speakers.

The results clearly show that the degree to which stress variation is included in syllable template formation is reflected in subsequent performance. For both speakers, the best recognition performance occurred when using the B dictionaries that contained *both* stressed and destressed templates and employed the higher precision spectral coefficients.

The next best performance emerged when the C dictionaries were used. Here the results show that, although occupying half of the storage space employed by the B dictionaries and the same space as the S and D dictionaries, the C dictionaries successfully embodied a high proportion of the variation due to stress—sufficient indeed to easily outperform the S and D dictionaries. Moreover, since the average error rate obtained with the C dictionaries was less than twice that of the B dictionaries this suggests that, in principle, it should be possible to replace the least reliable C templates by separate stressed and destressed templates and thereby create hybrid dictionaries that perform as well as B dictionaries but occupy less storage space than B dictionaries demand.

Figure 2 also shows a systematic speaker difference, with the speech of DZ yielding lower error rates than the speech of LL under the same conditions. This difference is comparable to the difference introduced by variations in dictionary type and is larger than the difference brought about by a change in precision level. It is of interest to note that the same speakers were employed in an earlier study that compared the effects on recognition performance arising from the use of different types of acoustic coefficients (Davis and Mermelstein, 1980). In that study, a similar speaker difference was found with each type of coefficient.

Furthermore, Fig. 2 indicates that between dictionaries B and C and for a given error rate, there exists the opportuni-

ty to trade dictionary type (structure) against coefficient resolution. However, since computational complexity varies as the square of the number of coefficients involved, it is apparent that if the coefficient resolution were doubled for dictionary C, twice as many computational operations would be necessary to recognize a token using C as would be necessary to perform a recognition using dictionary B following a doubling of the number of templates in that dictionary. Hence, a greater increase in recognition accuracy per datum (bit) can be achieved by carefully increasing the number of templates than by using a larger number of higher-resolution coefficients per template. Also, once a *lower bound* has been reached for errors through improvements achieved by increasing coefficient resolution, it is apparent that further improvements may still be achieved by increasing the number of allophonic variants represented in template form to a point where a balance is found between the benefits of error reduction and an increasing computational cost.

D. Errors classified by vowel identity

The computer recognition errors classified as a function of dictionary type and vowel identity are shown in the upper half of Table IV. In all four types of dictionary, more recognition errors occurred between syllable tokens and templates incorporating the same vowel nucleus than occurred between syllables having different vowel nuclei. Moreover, a larger number of *syllable identity* errors was associated with the longer of the two vowels. This evidence strongly suggests that the errors arose because the vowel /æ/, constituting a substantial portion of the syllable, made a larger contribution to the distance measurement than did the flanking consonants. In other words, the presence of long vowels tended to "dilute" the consonant discriminability.

Table IV also shows that if the cross-vowel errors involving /æ/ are expressed as a proportion (P %) of all errors involving /æ/, this proportion is smaller than the corresponding proportion for the vowel /ɪ/. This is true for both speakers and all dictionaries with the exception of B where, against the background of a small total number of cross vowel errors involving /ɪ/, the proportions (P %) exhibit the opposite relationship because this total is exceeded by an isolated set of confusions peculiar to the speech of LL. Thus taken as a whole, the number of errors involving long vowels tends not to include a substantial proportion of cross-vowel errors. Since long vowels constitute a prominent proportion of the syllables they occupy, they offer more information about their spectral structure and, hence, provide greater inherent protection against cross-vowel error.

Finally, Table IV prompts the observation that if cross-vowel errors from dictionaries B, C, and S only are considered in that order, the number of those errors involving the vowels /æ/ and /ɪ/ increases at a roughly equal rate despite the differences in vowel duration. The major reason for this result probably stems from the properties of the dynamic warping algorithm whose nonlinear adjustment of the time axis has a tendency to provide some compensation for differences in vowel duration.

E. Comparison with human listeners

The lower section of Table IV shows the listeners' data classified by vowel and stress level. An examination of the *cross-vowel* errors shows agreement with the bulk of the computer error data (upper section) inasmuch as the largest proportion of the human errors also involved /ɪ/ as compared with /æ/. The result suggests, of course, that the listeners were also able to make good use of the greater amount of vowel information available in the stimuli containing long vowels. The closest agreement with the listeners' overall performance is offered by dictionary C; here, both the proportion of cross-vowel errors (P %) and the total number of errors are of similar magnitude (listeners, 283; dictionary C, 310). However, the listeners' data differ from the computer results by posting a higher *total* of errors involving the vowel /ɪ/ (i.e., listeners, 180 versus dictionary C, 119). Hence, the data provide evidence that the listeners' abilities to recognize

TABLE IV. Syllable errors classified by dictionary type and vowel. Symbols /æ/ and /ɪ/ at the left of the table refer to vowel nuclei of misidentified syllable tokens while the same symbols located at column heads refer to the nuclei of syllable templates that were mistakenly selected. P % refers to the proportion of cross-vowel errors expressed as a percentage of all errors involving that vowel.

Recognition by computer
(summed over speaker, stress and precision level)

	Dictionary B			Dictionary C			Dictionary S			Dictionary D		
	/æ/	/ɪ/	P %	/æ/	/ɪ/	P %	/æ/	/ɪ/	P %	/æ/	/ɪ/	P %
/æ/	86	14	14.0	170	21	11.0	295	33	10.1	308	53	14.7
/ɪ/	9	67	11.8	19	100	16.0	40	220	15.3	122	166	42.7
	Total 176			Total 310			Total 588			Total 649		

Recognition by listeners
(Summed over speaker)

	Stressed			Destressed			Totals		
	/æ/	/ɪ/	P %	/æ/	/ɪ/	P %	/æ/	/ɪ/	P %
/æ/	42	6	12.5	50	5	9.1	92	11	10.7
/ɪ/	8	15	34.8	21	136	13.3	29	151	16.1
	Total 71			Total 212			Total 283		

TABLE V. Syllable identification errors classified by speaker. Computer data obtained using parameters at PL2 and dictionary B.

Comparison of listener recognition and computer recognition		
Recognition method	Speaker	Percent error
Listening	DZ	10.0
Listening	LL	7.5
Computer	DZ	1.4
Computer	LL	3.0

the *consonants* of a syllable were not impaired by the presence of a long vowel, suggesting that the recognition processes in the two cases are quite different. This conclusion is further supported by a comparison of listener and computer data in respect to the ten most frequently made consonant errors. These data reveal that virtually no consonant confusions were shared in common. Furthermore, a classification of these errors in terms of voicing, manner, and place of articulation (occurring either alone or in combination) showed no systematic differences—they appeared in both groups of data with roughly equal frequency.

Further results from the listening experiments are given in Table V, classified by speaker. The table shows that the syllables produced by LL were more accurately recognized by listeners than those produced by DZ—a result that is again at variance with that obtained by computer. In addition, for both speakers, and contrary to our expectations, the error percentages indicate that the overall human recognition performance was somewhat worse than the best computer performance (i.e., at PL2).

F. Errors classified by stress

A more revealing comparison of the listeners' recognition results with the computer results, and of the effects of dictionary type on computer performance, can be obtained if the errors are separately calculated for stressed and destressed tokens. Turning first to the computer data, Fig. 3 indicates that the difference between stressed and destressed error rates was smallest when the B and C dictionaries were in use—notwithstanding the relatively larger difference that emerged from the speech of LL. A comparison of the listeners' recognition data with the computer data also reveals some marked speaker-dependent effects. While the listeners' error rate for stressed-token recognition of LL's speech is closely comparable to the error rate turned in by the computer, their corresponding error rate on DZ's stressed speech shows a threefold increase over the computer error rate. A reason for this difference was revealed by a detailed examination of the listeners' errors on stressed tokens. This showed that 38% of the errors could be accounted for by two confusions, namely, those between DZ's articulation of [mæn] versus [mæt] and [hɪs] versus [dɪs]. In the destressed syllable data, however, no similar pair of confusions accounted for a comparably large proportion of the errors and the listeners' overall error rate consistently exceeded that delivered by the computer. Thus in summary, there was evidence that on the stressed tokens, the listeners tended to

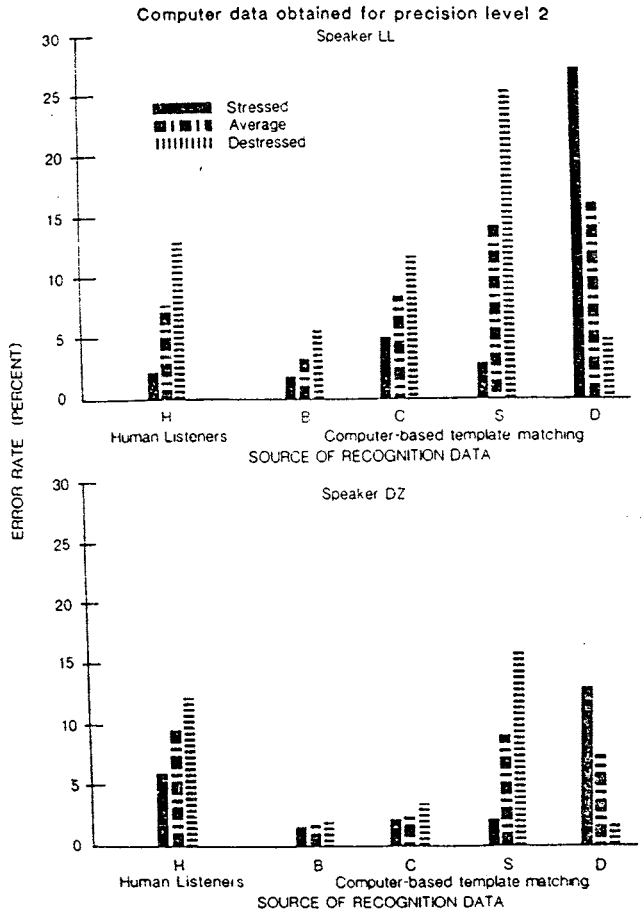

FIG. 3. Comparison of error rates for human and computer recognition of syllables supplied by speakers LL and DZ. Results labeled H were obtained from listeners. Labels B, C, S, and D refer to the four types of computer dictionary formed from coefficient data computed at PL2 (see text for explanation). The computer employed a dynamic warping and recognition algorithm with each dictionary in turn to recognize a closed set of unknown tokens.

perform only slightly worse than the computer, while on destressed tokens their performance was considerably below the computer using dictionary B.

A review of the composition of the four dictionaries can assist in explaining a substantial proportion of the error-rate differences appearing in Fig. 3. In the case of the B and C dictionaries, the computer error rates for stressed and destressed tokens differed from one another by small amounts relative to the corresponding differences for dictionaries S and D, with the B dictionary evidencing a lower error rate on both stress types. Since only the B and C dictionaries contained both stressed and destressed information, their overall superiority was certainly to be expected. Meanwhile, using the S dictionary, the error rate for *stressed* tokens emerged as being nearly identical with that obtained when using the B dictionary. Destressed tokens, on the other hand, fared about four times worse when using dictionary S than when using dictionary B, a direct consequence of the lack of destressed information in S dictionaries. Conversely, when dictionary D was in use, errors involving *destressed* tokens occurred at roughly the same frequency as they did when using dictionary B, while the stressed tokens submitted to

TABLE VI. Recognition scores using dictionary B. (A = Correct syllable identity and stress. B = Correct syllable identity but incorrect stress. C = Incorrect syllable identity.)

Speaker		Classified by speaker, precision level, and stress of token	A	B	C	Totals
DZ	PL1	Stressed	495	68	25	588
		Destressed	439	105	21	564
	PL2	Stressed	521	61	7	588
		Destressed	473	82	9	564
LL	PL1	Stressed	445	110	33	588
		Destressed	442	75	47	564
	PL2	Stressed	499	79	10	588
		Destressed	470	69	25	564
	Totals		3784	649	176	4608

dictionary D yielded, as expected, an extremely high error rate.

The foregoing analysis ignored stress assignment as long as a syllable's identity was found correctly. Dictionary B provides the only opportunity to analyze stress-only-errors and Table VI presents these data. The results show that, summed across both speakers and precision levels, errors in stress assignment occurred with 3.7 times greater frequency than did errors in syllable identity (cf., column B = 649 and column C = 176).

G. Examination of recognition rank

An analysis was made of the number of times that the correct syllable appeared in second, third, fourth, and fifth positions in the rank of ordered distance measures obtained during the recognition computations. The results showed that about 70% of the syllables that failed to occupy the first rank (and, therefore, be "recognized") appeared in the second rank. Overall, the third rank captured about 18% of the unrecognized syllables and the fourth rank accounted for a further 5%. Speaker differences were another major feature of these data. In the case of LL, the proportions of syllables appearing in the various ranks did not vary significantly as a function of precision level. Speech data from DZ, on the other hand, showed higher proportions of unrecognized syllables entering the second rank in runs employing PL2. The magnitude of this shift was particularly prominent in the data for dictionary B, which indicates that this effect was related to the lower number of errors arising under PL2 conditions.

H. Geometry of the stress distance space

The more significant features of the results just described can be explained by reference to the concept of a syllable distance space. Within this space, five possible configurations of the stressed and destressed tokens can be intuitively expected. Four of these are shown in Fig. 4. The fifth configuration (asymmetric clusters; equal discriminability) shares features illustrated by configuration types (II) and (IV) and has been omitted. In each case, Fig. 4 shows the theoretical relationship of two phonetically close syllables A and B occurring in both stressed (A′) (B′) and destressed (A) (B) forms. The heavy vertical bar that bisects an imaginary line linking the midpoints of the A and B distributions marks the position of the decision boundary between distributions A and B, which are assumed to be of similar size and conformation. (X) represents an unknown token. The first case, type (I), assumes that destressed syllables have the same central tendency as stressed syllables and form a large (noisy) cluster surrounding a smaller, more dense, cluster of stressed

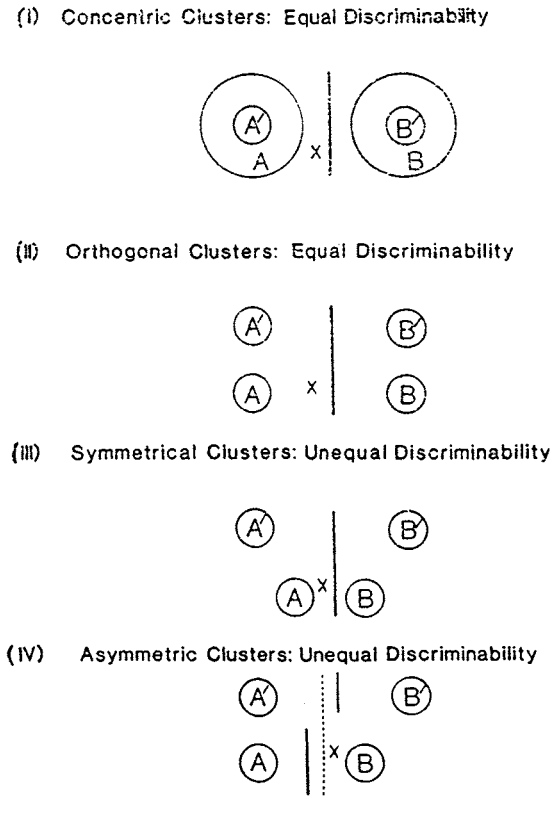

FIG. 4. The symbol (X) represents the spatial location of an unknown token. Four types of cluster patterns for A and B are shown. Types (I), (II), and (III) are so distributed that a single decision boundary would serve for recognition of both stressed and destressed syllables and would lead to the classification of (X) as a member of the class "destressed A." For type (IV), different boundaries are required for unbiased decisions between stressed and destressed A and B. Hence, the token (X) is potentially classifiable as a "destressed B" or "stressed A."

tokens. This pattern would predict that a dictionary of stressed syllables (S) should serve well with both syllable types and, therefore, outperform all the other dictionaries. However, the data we have reported do not fit this prediction. Types (II) through (IV) postulate different formations of separate clusters for stressed and destressed syllables. Type (II), consisting of four symmetric and orthogonal clusters, would suggest that stressed and destressed syllables should be found to be equally discriminable. Type (III) might arise when the discriminability of destressed pairs is less than that of stressed pairs but a single decision boundary can still serve to determine whether token (X) belongs to A or B. The fourth cluster configuration, type (IV), also gives rise to unequal discrimination but additionally requires the adoption of a second decision boundary to ensure the proper classification of the unknown (X).

To determine which of these theoretical models best fits the data, the distances obtained during recognition calculations were assembled in matrix form and input to the multidimensional scaling program KYST (Kruskal *et al.*, undated). This program enabled us to generate graphic displays of the actual cluster structures of stressed and destressed syllables under a variety of dimensional constraints. The first observation to note is that, viewed overall, the clusters of *destressed* tokens consistently appeared to be only slightly less compact than the clusters of stressed tokens and, therefore, to possess a different but almost equally distinct acoustic form. In the two-dimensional case, the results contained examples of clusters that fitted each of the last three cases shown in Fig. 4. For example, Fig. 5 shows some actual distributions for both speakers obtained from data accumulated over all their speaking sessions. The spatial distributions are for the syllables [dɪg], [dɪj], and [dɪs], chosen because they represent minimal pairs (i.e., pairs of syllables that differ by a single phoneme). For the speaker LL, the upper half of the figure provides an example of orthogonal clusters resembling type (II) of Fig. 4, while below is shown an equivalent group of clusters for the speaker DZ. In the latter case, the clusters tend to be asymmetrical and to resemble type (IV). In fact, by far the largest proportion of examples studied could be classified as type (IV). Thus overall, the fourth case emerged as the best general model for the recognition data.

The type (IV) configuration (Fig. 4) illustrates that, if a destressed token (X) is submitted to an S dictionary, the difference in location of the stressed decision boundary (upper vertical bar) will result in (X) being recognized as belonging to the class A. This makes it clear why poor recognition performances were obtained when the tokens were of different stress than the available templates. The diagram can also offer an explanation as to why a dictionary containing the combined templates was found to give better results and why better performance will always be achieved by using both stressed and destressed templates. To follow the explanation offered in this case, we must assume that the clusters representing the combined templates for A and B will lie midway along the axis joining the centers of the stressed and destressed distributions. Therefore, the decision boundary (dotted vertical line) will now move to a point midway between the original stressed and destressed boundaries and

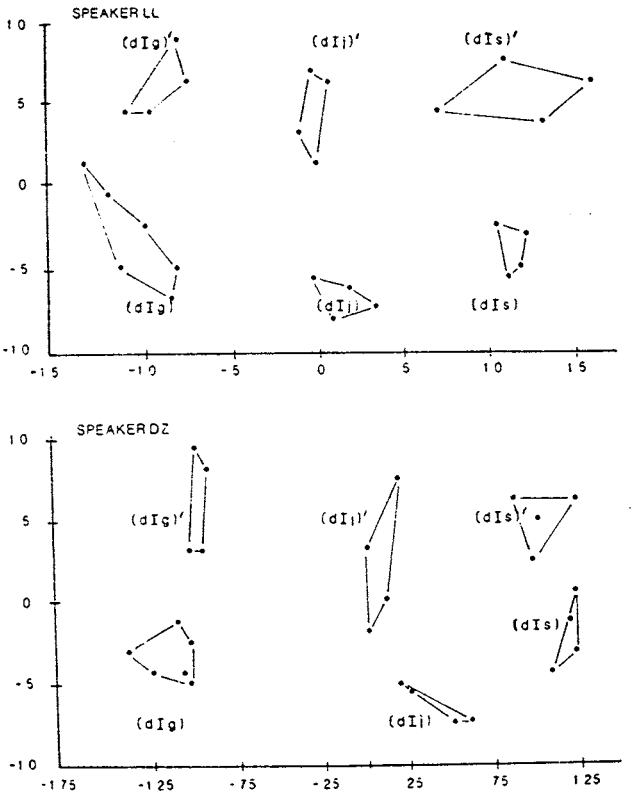

FIG. 5. Cluster configurations of the syllables [dɪg], [dɪj], and [dɪs] obtained by analysis performed by the multidimensional plotting program KYST. Primes indicate stressed syllables. Syllable data were extracted from single sessions delivered by speakers LL and DZ.

(X) referred to this new boundary would now be correctly classified as belonging to class B.

III. CONCLUSIONS

A. Summary and comments

We have conducted an investigation into the effects of stress and vowel duration on the performance of a recognition algorithm and we have compared some aspects of this performance with data gathered from listeners. In an effort to gain better control over our speech data, we chose to examine a form of stress variation that, while present in continuous speech, was sufficiently constrained that it could not be claimed to be representative of the more extreme forms that stress reduction can take. We deliberately omitted those types of stress reduction that result in (1) the syllabic vowel being pronounced as a shwa and, (2) the consonantal features being severely attenuated. Nevertheless, despite the relatively modest amount of stress variation present, its effects on recognition performance were quite large.

Our results showed that recognition accuracy for stressed and destressed syllables can be improved in three ways; these are, in increasing order of effect: (1) by increasing the resolution of the acoustic parameters as exemplified by exchanging PL1 for PL2, (2) by *combining* the acoustic features of stressed and destressed syllables into a single template dictionary or, (3) by doubling the size of the dictionary to include templates for *both* stressed and destressed sylla-

bles. Moreover, the results indicate that when computational economy is at issue, the nature of the trade-off between parameter resolution and dictionary size promises greater gains in recognition accuracy per bit of information from dictionary enlargement (the inclusion of individual stressed and destressed templates) than from increases in parameter precision.

We also examined the cluster structure adopted by pairs of linguistically different stressed and destressed syllables and found that the bulk of them can be classified as asymmetric distributions offering unequal discriminability for stressed and destressed forms. Moreover, we found that destressed tokens form clusters that are only marginally less compact than their stressed counterparts. This observation was confirmed by the fact that the overall recognition rate for destressed tokens submitted to a dictionary of destressed templates was very similar to the rate observed for stressed tokens matched against a dictionary of stressed templates.

The reason for the unusual compactness of the destressed tokens must almost certainly be sought in the environments in which these syllables were produced. The restrictions that were placed on the amount of stress reduction we wished to permit imposed strict limitations on the number of syllables and lexical environments that were available. Thus the fact that any given word containing a target syllable appeared in only two different sentence environments provided little opportunity for a variety of coarticulation effects to extend from neighboring phones to the target syllables. Moreover, the experimental conditions fostered the likelihood that the magnitude of any coarticulatory interaction would vary according to a target syllable's position within a word. For example, as seen in Table I, destressed syllables occupied word-initial, mid-word, and word-final positions on a roughly equal basis whereas stressed syllables appeared prominently in word-initial position. Hence, to the extent that the strongest coarticulatory influence was likely to occur between target syllables and immediately adjacent phones, it may be assumed that, by virtue of the constancy of their immediate environment, approximately one-third of the destressed syllables were produced with substantially the same coarticulation.

In addition, we confirmed that the phonologically short vowels were, according to our measurement criteria, shorter than phonologically long vowels. We also found that the shortening of vowel length and syllable length that accompanies stress reduction is greater in the case of the shorter vowel. Since some degree of time normalization is an intrinsic feature of the warping algorithm, one might expect that any bias in favor of longer vowels would be offset. Certainly this is suggested by the fact that cross-vowel error rates for short and long vowels increase at an approximately equal rate across dictionaries B, C, and S. However, the study also indicates that long vowels have two important advantages and suffer one disadvantage when subjected to warping and recognition procedures. First, among the advantages is the fact that identification errors involving long vowels tend to include a smaller proportion of cross-vowel errors than is found to be included among the identification errors involving short vowels. Second, long vowels tend to be associated with lower vowel-error rates than short vowels. The disadvantage that long vowels face is due to the preponderant contribution they make to the distance measure. This contribution is so large that it masks or "dilutes" consonant information to such a degree that syllable identity errors increase. We must therefore conclude that in future attempts to develop improved distance metrics, an effort directed at enhancing the contribution made by consonants should be given priority.

Another group of observations made in this study centered on the similarities and differences between recognition performances delivered by listeners and those produced by the computer. Evidence indicated that listeners could achieve a recognition accuracy on stressed tokens that is roughly comparable with that achieved by computer. On the other hand, computer recognition rates for *destressed* syllables under the most favorable conditions were found to be superior to the rate achieved by listeners. One tentative explanation for this possibly surprising observation rests on the notion that the listeners tend to be biased (or pre-primed) for stressed-item recognition by the phonetically spelled syllable transcriptions displayed on their response forms. Yet another explanation acknowledges the fact that listeners must carry in their heads many more syllable templates than were listed on the response form. Given this fact, an unknown destressed token X may not be directly identified with the nearest syllable (A') listed on the response form but can be identified instead with template (C), not included in the response list, because distance $D(X,A') > D(X,C)$. Subsequently, X having lost its own acoustic identity (by decay of short-term memory) and assumed that of C, a search for the nearest template identified in the response list leads to the incorrect selection of template (B') because $D(C,B') < D(C,A')$. Finally, it might be noted that the recognition of stressed syllables is a highly practiced task whereas the recognition of destressed syllables is not because, in continuous speech, destressed syllables are normally recognized with the aid of their context. To recognize them in isolation is a relatively unfamiliar task and consequently poorer performance is to be expected. Of course, the present data provide no opportunities to properly examine these alternative hypotheses. In the final analysis, it has to be conceded that the behavior of listeners and the behavior of the computer algorithm are so different as to make it obvious that the recognition principles employed by both are quite different.

Our finding that the spatial distributions of our stressed and destressed syllables do not greatly differ in size suggests that it might be possible to derive the acoustic properties of each destressed syllable by applying a warp in both the time and frequency domains to its appropriate stressed counterpart. Moreover, if warps of this kind proved to have properties that were common to a large class of syllables, say all CVCs of a given vowel type, this would be of considerable help in controlling the rate of dictionary growth. One way of applying such a warp would be by means of a matrix that would provide the opportunity to compute a composite or standard warp for a given syllable class by averaging together the warps obtained from many CVCs.

Stress effects are among the most difficult of the many

obstacles that lie in the path of achieving a practical continuous speech recognition capability. In this study, we have begun a systematic approach to this problem by attempting to generate controlled, yet realistic, data and to observe their interaction with recognition variables such as dictionary composition, parameter precision and widely used recognition techniques such as dynamic pattern matching. We have succeeded in identifying many of the interactions that take place and in several cases have been able to point out their boundary conditions. Future work on the problem of stress variation should involve the gradual relaxation of some of the input constraints adopted here, the collection of additional observations and the development of new and better algorithms.

ACKNOWLEDGMENTS

This research was supported by NSF grant MCS 79-16177 and BRS grant RR-05596 to Haskins Laboratories. The authors wish to thank Franklin Cooper and Paul Mermelstein for much useful advice given during the planning of these experiments. We also wish to express our appreciation to Franklin Cooper, Louis Goldstein, Ignatius Mattingly, Paul Mermelstein, and Bruno Repp for the helpful comments they offered on various drafts of this paper.

[1]It is the complex nature of the coarticulatory interaction between phones (particularly within syllables) that has proved to make segmentation strategies based on phonemic units so difficult to develop.

[2]Many linguists (Pike, 1945; Trager and Smith, 1951) have drawn attention to the fact that English speech has more than two levels of stress. Furthermore, the comments of our colleagues and reviewers have made it obvious that there is insufficient agreement on a terminology for stress designation to permit us to use the words "stressed" and "destressed" without the following explanatory remarks: The syllables employed in this study were obtained from words in which they customarily receive contrasting degrees of lexical stress. These stress contrasts were potentially subject to enhancement or reduction by the sentential context although the most obvious syntactic influences such as word-final lengthening were avoided. Therefore, a syllable labeled as "stressed" did not necessarily bear the primary or highest sentential stress. Syllables labeled as "destressed," on the other hand, always bore less stress than their stressed counterparts but were never so severely reduced as to cause the nuclear vowel to be produced as a shwa. In general, experience leads us to expect that the stress reduction exhibited by syllables incorporating /ɪ/ to be greater than the reduction for syllables incorporating /æ/.

[3]Because syllables in phrase-final position tend to undergo lengthening and because syllable lengthening is one of the principal correlates of stress (Fry, 1955), it was particularly necessary to avoid the interaction of such position effects with the syllables chosen for this study.

[4]Errors occurred primarily in the syllable-duration category and were due to a failure of the segmentation algorithm to include released bursts in final position as an integral part of the preceding syllable. A secondary problem was the occasional omission of destressed syllables. Such errors were not acceptable for the purposes of the present study.

Bahl, L. R., Baker, J. K., Cohen, P. S., Cole, A. G., Jelinek, F., Lewis, B. L., and Mercer, R. L. (1978). "Automatic recognition of continuously spoken sentences from a finite state grammar," IEEE Record. Int. Conf. Acoust. Speech Signal Process. (IEEE, New York), 418–421.

Bridle, J. S., and Brown, M. D. (1974). "An experimental automatic word recognition system," JSRU Rep. No. 1003, Joint Speech Research Unit, Ruislip, England.

Davis, S. B. (1979). "Order dependence in templates for monosyllabic word identification," IEEE Record. Int. Conf. Acoust. Speech Signal Process. (IEEE, New York), 570–573.

Davis, S. B., and Mermelstein, P. (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust. Speech Signal Process. ASSP-28, 357–366.

Dixon, N. R., and Silverman, H. F. (1977). "The 1976 Modular Acoustic Processor (MAP)," IEEE Trans. Acoust. Speech Signal Process. ASSP-25, 367–379.

Fry, D. B. (1955). "Duration and intensity as physical correlates of linguistic stress," J. Acoust. Soc. Am. 27, 765–768.

Fujimura, O. (1975). "The syllable as a unit of speech recognition," IEEE Trans. Acoust. Speech Signal Process. ASSP-23, 79-82.

Itakura, F. (1975). "Minimum prediction residual principle applied to speech recognition," IEEE Trans. Acoust. Speech Signal Process. ASSP-23, 67–72.

Klatt, D. H. (1978). "SCRIBER and LAFS: Two new approaches to speech analysis," in *Trends in Speech Recognition*, edited by W. A. Lea (Prentice-Hall, Englewood Cliffs, NJ).

Kruskal, J. B., Young, F. W., and Seery, J. W. (undated). "How to use KYST-II; A very flexible program for multidimensional scaling and unfolding," Computing Information Service, Bell Laboratories.

Mermelstein, P. (1975). "Automatic segmentation of speech into syllabic units," J. Acoust. Soc. Am. 58, 880–883.

Mermelstein, P. (1976). "Distance metrics for speech recognition—psychological and instrumental," in *Pattern Recognition and Artificial Intelligence*, edited by C. H. Chen (Academic, New York), pp. 374–388.

Mermelstein, P. (1978). "Recognition of monosyllabic words in continuous sentences using composite word templates," IEEE Record. Int. Conf. Acoust. Speech Signal Process. 708–711.

Peterson, G. E., and Lehiste, I. (1960). "Duration of syllable nuclei in English," J. Acoust. Soc. Am. 32, 693–703.

Pike, K. (1945). *Intonation of American English* (University of Michigan, Ann Arbor, MI).

Rabiner, L. R., Rosenberg, A. E., and Levinson, S. E. (1978). "Considerations in dynamic time warping algorithms for discrete word recognition," J. Acoust. Soc. Am. Suppl. 1 63, S79.

Rabiner, L. R., and Wilpon, J. G. (1979). "Speaker-independent isolated word recognition for a moderate size (54 word) vocabulary," IEEE Trans. Acoust. Speech Signal Process. ASSP-27, 583–585.

Rosenberg, A. E., Rabiner, L. R., Levinson, S. E., and Wilpon, J. G. (1981). "A preliminary study on the use of demisyllables in automatic speech recognition," IEEE Record. Int. Conf. Acoust. Speech Signal Process. (IEEE, New York), 967–970.

Sakoe, H., and Chiba, S. (1978). "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoust. Speech Signal Process. ASSP-26, 583–586.

Trager, G., and Smith, H. (1951). *An Outline of English Structure. Studies in Linguistics: Occasional Papers 3* (Battenburg, Norman, OK).

Velichko, V. M., and Zagoruyko, N. G. (1970). "Automatic recognition of 200 words," Int. J. Man Mach. Stud. 2, 223–234.