Automatic segmentation of speech into syllabic units

Paul Mermelstein

Haskins Laboratories, New Haven, Connecticut 06510
(Received 14 April 1975; revised 19 June 1975)

As a first step toward automatic phonetic analysis of speech, one desires to segment the signal into syllable-sized units. Experiments were conducted in automatic segmentation techniques for continuous, reading-rate speech to derive such units. A new segmentation algorithm is described that allows assessment of the significance of a loudness minimum to be a potential syllabic boundary from the difference between the convex hull of the loudness function and the loudness function itself. Tested on roughly 400 syllables of continuous text, the algorithm results in 6.9% syllables missed and 2.6% extra syllables relative to a nominal, slow-speech syllable count. It is suggested that inclusion of alternative fluent-form syllabifications for multisyllabic words and the use of phonological rules for predicting syllabic contractions can further improve agreement between predicted and experimental syllable counts.

Subject Classification: 70.40, 70.60.

INTRODUCTION

Automatic phonetic analysis of speech, such as that carried out as part of a continuous speech understanding system, requires a mapping from acoustic signal to phonetic segments whose direct implementation has eluded speech researchers for many years. Liberman[1] reviews the case for considering the conversion between phone and sound to be a complex grammatical recoding that may prevent one from ever finding a direct replacement of sound segments by phones. In agreement with that point of view, we consider an alternative, indirect approach which segments the speech stream into syllable-sized units and decodes the phonetic segments of those units by considering the acoustic information contained in the entire syllable.[2] This paper presents results of experiments in automatic segmentation of continuous speech into such syllable-sized units.

The syllable has been defined linguistically as "a sequence of speech sounds having a maximum or peak of inherent sonority (that is, apart from factors such as stress and voice pitch) between two minima of sonority."[3] To arrive at an operational definition that is computationally implementable, one must define sonority in terms of physical measures on the speech signal. This requirement leads quickly to the realization that "inherent sonority" cannot be empirically defined because the same parameter, intensity, signals (in part) both sonority and stress, and the division between the two factors is rather arbitrary. The argument that stress values are assigned to entire syllables and sonority varies from phone to phone within the syllable cannot be applied to separate the two factors since it is precisely the operational determination of syllables that we are trying to achieve.

Stowe[4] attacked this problem by a hierarchic series of segmentation procedures, each operating on a different time function computed from the speech signal. Sargent *et al.*[5] also used two functions for syllable detection, one measuring peak-to-peak amplitude, the other root-mean-square (rms) intensity. In this work we explore a new approach. We attack the resolution of the above problem by defining a "loudness" measure for the speech signal, a time-smoothed and frequency-weighted summation of its energy content. Relative loudness maxima are interpreted as potential syllabic peaks and relative loudness minima as potential syllabic boundaries. To differentiate between syllables generally defined on the phonological level and the speech segments that may be located in the signal by phonetic criteria, we introduce the term "syllabic unit" for the syllable-sized speech segments that are to be found automatically. Boundaries located by loudness criteria do not necessarily segment the speech signal at points that can be identified as phone boundaries, or even word boundaries. The syllabic units are found to depend strongly on the phonetic performance of the speaker; in fact, they serve to describe that performance by grouping segments into larger units that generally form units of production as well.

In order to arrive at a segmentation of the signal into syllable-sized units, we find that one must define a measure of significance that permits classifying loudness minima as to whether they denote actual boundaries. Otherwise, the number of realized segments greatly exceeds the number of syllables one would count perceptually. Further, the measure of significance must be a function of the context of any particular loudness minimum. A local loudness minimum separated by less than 100 msec from another local minimum with lesser loudness may be insignificant, yet the same minimum with no other minima within 500 msec would generally signal a syllabic boundary.

The significance of loudness maxima must be similarly evaluated. In order to prevent segmentation into fragments that do not contain adequately strong syllabic peaks, we reject any segment whose loudness maximum is more than a given threshold below the overall loudness maximum, the syllabic peak of the loudest syllable of the utterance. Similarly, a minimum syllabic-unit duration of 80 msec is imposed, and segmentation that would result in shorter fragments is rejected.

One important application of syllabic-unit segmentation is as an aid to lexical analysis where one would like the same text spoken by different speakers to show at most a small number of alternative syllabic-unit representations. Fricatives are generally not tightly bound to the syllabic units with which they are associated, but

are frequently separated from them by a short interval of weak voicing or even silence. On the basis of loudness criteria alone, they form valid syllabic units. For the purposes of evaluating the results of our segmentation procedures and for accessing a lexicon of syllabic forms, we require that syllabic units have nonfricative nuclei. If subsequent analysis reveals that a syllabic unit manifests significant frication near the syllabic peak, it is labelled as a syllabic fragment and not counted as an independent syllabic unit.

I. SEGMENTATION USING A CONVEX-HULL ALGORITHM

In order that our empirically determined loudness function roughly approximate the subjective loudness function, loudness is obtained from the speech power spectrum by weighting frequencies below 500 Hz and above 4 kHz according to a function that drops off at 12 dB/octave outside these frequencies. To eliminate variations in loudness due to the phase of the fundamental frequency of excitation, the loudness function is low-pass filtered at 40 Hz. Our implementation computes loudness from the short-time power spectrum, but it could be equally well derived by directly filtering the speech wave.

Initially, a segment of speech between apparent pauses (silent interval exceeds 200 msec) is selected for analysis. The convex hull of the loudness function is defined as the minimal-magnitude function that is monotonically nondecreasing from the start of the segment to its point of maximum loudness, and is monotonically nonincreasing thereafter. Within the segment, the difference between the convex hull and the loudness function serves as a measure of significance of loudness minima. The point of maximal difference is a potential boundary. If the difference there exceeds a given threshold, the segment is divided into two subsegments.

Segmentation is carried out recursively. The convex hulls newly computed for the subsegments nowhere exceed the convex hull of the original segment. Hence, after any segmentation step, only less significant minima remain. If the maximal hull-loudness difference within the segment is below the threshold, no further segmentation of that segment is possible. The algorithm makes use of the loudness context implicitly by extracting minima in order of significance. A minimum may not be significant if another more significant minimum is located close by. Segmentation removes the more significant minimum and allows reconsideration of the significance of the less significant one.

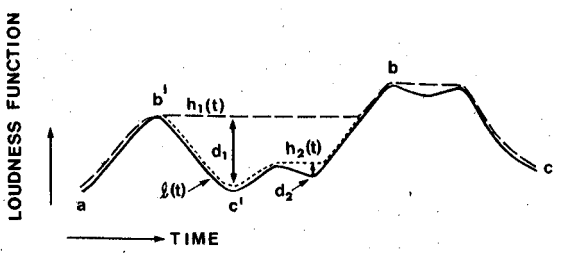

FIG. 1. Loudness function and convex hull for speech segment.

Figure 1 illustrates the implementation of the convex-hull algorithm. An original speech segment over the interval $(a-c)$ is found to possess a loudness function $l(t)$ with maximum at point b. The convex-hull computed for the segment $(a-b-c)$ is $h_1(t)$. Over the interval $(a-c)$, the maximum hull-loudness difference is d_1 at c'. If d_1 exceeds the threshold, segment $(a-b-c)$ is cut up into segment $(a-c')$ followed by segment $(c'-b-c)$. The hull for segment $(a-c')$, defined around the new maximum point b', follows the loudness curve. This results in a zero hull-loudness difference over that interval and hence that portion is not segmented further. The hull for segment $(c'-b-c)$, denoted by $h_2(t)$, is shown by the short dashed line where it differs from $h_1(t)$ over the segment interval. The new maximum hull-loudness difference is found to be d_2. If d_2 does not exceed the threshold then the segment $(c'-c)$ is not divided further.

The algorithm does not proceed from left to right in time. It assumes that the entire utterance is stored before processing commences, but requires only that a complete segment delimited by silent intervals be captured before segmentation starts. Where real-time operation is essential, the algorithm can be modified to operate from left to right with possible backtracking over an interval of no larger than the maximum syllabic unit interval, roughly 500 msec.

II. EXPERIMENTAL RESULTS

The performance of the algorithm was evaluated by processing 11 sentences read by each of two male subjects at their comfortable reading rate. The first six sentences (text A) make up the well-known "Rainbow Passage," and contain both monosyllabic and multisyllabic words. The last five (text B) consisted of only monosyllabic words and were taken from material composed by Lea.[6] The differentiation in text material was utilized to explore the dependence of segmentation errors on the frequency of multisyllabic words in the text.

Figure 2 illustrates typical results for the text "...a boiling pot of gold at... ." The segmented loudness function is plotted above a computer-generated spectrographic representation of the utterance. The spectral data have been preemphasized at 6 dB/octave above 300 Hz. Use of a uniformly weighted intensity for the loudness function would miss the high-frequency energy discontinuity for [boj–liŋ]. By using loudness as defined, high-frequency energy variations are emphasized and the boundary is located.

By varying the segmentation threshold parameter d, we can control the relative frequency of extra syllabic units found and the frequency of syllables missed. A threshold $d=0$ will result in too many extra syllabic units due to segmentation even at points of minimal variation in the speech loudness. A high threshold, $d>3$ dB, will result in many significant segmentation points within voiced segments being missed. The segmentation results at d values of 2 dB as compared to 1 dB showed that 12 extra syllables in the corpus of 418 syllables had

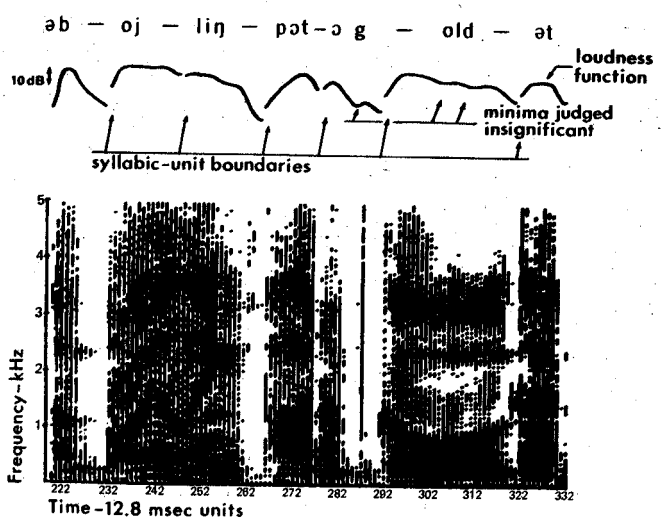

FIG. 2. Example of syllabic segmentation results for text "... a boiling pot of gold at...."

been eliminated and only two new missed syllables had been introduced. Further small increases in the value of d did not result in any appreciable difference in performance; therefore, all further results are given for the $d = 2$ dB condition.

Differences between the output of the segmentation algorithm and a nominal syllable count are given in Table I, classified by category and speaker. Since the syllable count is dependent on fricative detection, errors resulting from incorrect fricative detection are indicated separately, denoted by categories E2 and M2, respectively. The major source of extra syllabic units was in prepausal position (category E1) where significant release gestures were associated with final stops and liquids. The syllabic-unit loudness peaks for these cases were well above the −25-dB syllabic peak threshold, a value arrived at by empirical adjustments to eliminate most syllabic fragments. The frequency of prepausal extra syllabic units was highly speaker dependent, 1.2% for subject LL and only 0.5% for GK.

Syllabic units were missed primarily due to the contraction of an unstressed and stressed syllable-pair into one stressed syllabic unit (category M1). Most such junctures had a loudness minimum that was not less than 1 dB below the last convex hull computed; some in fact had no loudness minima associated with them at all. In the monosyllabic text B, such errors were encountered mostly in open syllables, e.g., /so aj/, /hɪ həd/, but their frequency was rather low (0.6%). Possible contractions across words that may result in a syllable count for a sequence of words that is smaller than the sum of the individual syllable counts may best be handled on the phonological level through a set of rules predicting such phenomena. In the multisyllabic text the frequency of syllables missed was significantly higher (1.4%). Many of these (10 of 22) were encountered in both speakers' productions, i.e., the syllable count for the same word or words for both speakers was consistent but different from a nominal syllable count that one would expect in slow speech. Typical examples are [hraj–zən] and [əp–pɛrn–lɪ] for "horizon" and "apparently." These forms must be considered to constitute acceptable productions alternative to those that would contain an additional syllabic unit for each word. Our results suggest a frequency of occurrence of problematic words whose syllabification cannot be adequately treated on the phonetic level. For speech recognition applications it seems advisable to handle these multisyllabic word problems on the lexical level by including alternative admissible syllabifications in any lexicon of syllabic forms.

Differences categorized E2 and M2 denote extra and missed syllabic units due to incorrect categorization of the unit as nonfricative and fricative, respectively. Missed units result if a short vowel-like interval is missed and the unit is interpreted as completely fricative, e.g., [t^h ʊ], due to a previously discussed decision not to count fricative-like syllabic units as independent. Extra units result if a voiced fricative shows voicing sufficiently strong that it is interpreted as vowel-like. Presumably, these errors could be eliminated as a result of improvements in the fricative detection algorithm.

In summary, the overall frequency of syllable-count differences with respect to nominal, slow-speech syllable count was 9.5%, consisting of 6.9% missed and 2.6% extra. We have previously reported 2.7% boundaries missed and 9% extra syllabic units found by essentially the same algorithm for a different text of 430 syllables.[7] There the algorithm was not optimized for the value of threshold d and errors were counted relative to a perceptual syllable count on the same spoken material. The difference in missed boundaries arises from the difference in the standard of comparison. For lexical applications, a maximal syllabic form appears as the most useful standard.

The two sets of results, those reported here and those previously reported,[7] carried out on data from different speakers and collected under different recording conditions, yield roughly comparable difference rates. The previous study used data on a total of 31 sentences recorded by some five speakers, three male and two female. No large differences in overall syllable-count difference rates are observed as long as the speech is spoken carefully and at a moderate rate.

III. LOCATION OF SYLLABIC-UNIT BOUNDARIES

The boundaries located by the algorithm do not bear a simple relationship to general syllabification rules followed in "phonological" syllabic boundary assignment where the main criterion appears to be whether words occur in the language with that particular initial or final

TABLE I. Differences between algorithm-derived syllable count and manual slow-speech count.

		Different category									
		E1		E2		M1		M2		Total syllables	
Speaker	Text	A	B	A	B	A	B	A	B	A	B
LL		1	1	1	-	12	2	2	2	123	86
CK		1	4	2	1	5	3	2	1	123	86

cluster. Based on these criteria, the syllabification of words containing intervocalic nasals and liquids is generally ambiguous, the sonorant may be assigned to either syllable-initial or syllable-final position. Linguists generally assign the maximum initial consonant sequence to the stressed syllable.[8] The algorithm locates a boundary within the consonant roughly at the point of minimal first-formant frequency. The major part of the consonantal segment is generally found assigned to the syllable carrying heavier stress due to its greater loudness. Where allophonic variations are associated with syllabic position, e.g., [ɹ'ɑʊnd] vs [ə'rɑʊnd], the syllabification resulting from use of the algorithm is generally consistent with our phonetic expectations.

The change in loudness or intensity at the onset of a syllable is generally more abrupt than at its end; thus there is less uncertainty about the onset time of a syllable than about its termination. Therefore, silent segments or those whose loudness is below the noise threshold are arbitrarily assigned to syllable-final position. This results in inclusion of nonreleased final stops in the previous syllable, but released stops straddle the syllabic boundary.

Intervocalic clusters are generally divided up. Compounds such as /sʌnlajt/ are segmented in accordance with morphemic criteria as loudness is found to decrease over the nasal and to increase over the liquid. Initial or final clusters may, however, be frequently broken up by the syllabification when unstressed syllables precede or follow them. For example, /t ʊ/+/grit/ may map to [tʊg–rit], or /pajlz/+/ɔf/ to [pajl–zɔf], where – indicates the position of the boundary within the phonetic segment stream. Generally, the effect is to couple an initial or final cluster constituent with the preceding or following syllable if that ends or starts with a vowel. These effects occur sufficiently consistently, at least in our limited data, so that syllable reorganization may be predictable by rules.

The algorithm forms a useful tool for phonetic analysis. The word pair /rezd/ vs /redz/ forms an interesting example where attempts to use phonological criteria such as the measure of "vowel affinity" proposed by Fujimura[9] to constrain the admissible syllable structures in English break down. Here, /z/ and /d/ are phonemes that may occur in either order in syllable-final position, an exception to a general ordering of phonemes by increasing "vowel affinity" in syllable initial and decreasing "vowel affinity" in syllable-final position. The convex-hull algorithm invariably classifies /rezd/ as one syllabic unit and /redz/ as two, a proper unit followed by a syllabic fragment. The "vowel affinity" of the fricative is different in the two cases, as manifested by a difference in intensity of voicing. The fricative in /redz/ is but weakly voiced, frequently devoiced. The postvocalic /z/ preceding a voiced stop carries stronger voicing. When followed by an unstressed vowel-initial syllable, this difference manifests itself by an assignment of the fricative to the first syllabic unit in the case where /rezd/+/ɪn/ gives [rezd–din] (syllabic-unit boundary within the closure of the /d/), but to the second syllabic unit in /redz/+/ɪn/ mapping to [red–zɪn]. We conclude that for the purposes of phonetic analysis, information derived regarding syllabic units and fragments is in fact useful even though for syllable-counting purposes one may desire to minimize the number of such fragments.

IV. CONCLUSIONS

Syllabic units can be counted in continuous speech by simple automatic techniques. The number of syllabic units found will agree relatively reliably with a text-derived syllable count under the following conditions:

(1) The algorithm is tuned to minimize extra syllabic units and missed units by adjusting the significance threshold d.

(2) A moderate amount of postprocessing is performed to weed out fricative-like syllabic fragments because they do not constitute independent syllabic units.

(3) Phonological rules are employed to predict where separate words may be contracted to reduce the syllabic count of the total to less than the sum of the individual counts.

(4) Alternative fluent-production forms are recognized for many multisyllabic words.

Segmentation into syllabic units appears to be sufficiently consistent so that the units so delimited constitute appropriate units of the speech signal on which further analyses may be carried out to extract additional phonetic information.

ACKNOWLEDGMENTS

I wish to acknowledge with thanks discussions of the material presented here with my colleagues at Haskins Laboratories, F. S. Cooper, J. Gaitenby, G. Kuhn, and P. Nye. This research was supported in part by the Advanced Projects Agency of the Department of Defense under contract No. N00014-67-A-029-002 monitored by the Office of Naval Research. The views presented here do not necessarily represent the views of the Department of Defense.

[1]A. M. Liberman, "The Grammars of Speech and Language," Cogn. Psychol. 1, 301–323 (1970).
[2]P. Mermelstein, "A Phonetic-Context Controlled Strategy for Segmentation and Phonetic Labeling of Speech," IEEE Trans. Acoust. Speech Signal Process. ASSP-23, 79–82 (1975).
[3]R. H. Robins, *General Linguistics. An Introductory Survey* (Indiana U.P., Bloomginton, IN, and London, 1966).
[4]A. N. Stowe, "Segmentation of Speech into Syllables," J. Acoust. Soc. Am. 25, 806(A) (1963).
[5]D. C. Sargent, K. P. Li, and K. S. Fu, "Syllable Detection in Continuous Speech," J. Acoust. Soc. Am. 45, 410(A) (1974).
[6]W. A. Lea, "Sentences for Controlled Testing of Acoustic Phonetic Components of Speech Understanding Systems," Rep. PX 10952, Sperry Univac Defense Sys., St. Paul., MN (1974).
[7]P. Mermelstein and G. M. Kuhn, "Segmentation of Speech into Syllabic Units," J. Acoust. Soc. Am. 25, 806(A) (1973).
[8]J. E. Hoard, "Aspiration, Tenseness and Syllabication in English," Language 47, 133–139 (1971).
[9]O. Fujimura, "Syllable as a Unit of Speech Recognition," IEEE Trans. Acoust. Speech Signal Process. ASSP-23, 79–82 (1975).