A Phonetic-Context Controlled Strategy for Segmentation and Phonetic Labeling of Speech

PAUL MERMELSTEIN, MEMBER, IEEE

Abstract—**This paper considers a sequential strategy for acoustic-phonetic speech analysis. Each analysis process is applied to an appropriately labeled speech segment and results in a possible subsegmentation of the original segment. The segments resulting from the analysis are labeled according to the analysis results.**

The advantages of the strategy are that no more segments are considered than those actually differentiated by the analysis steps. The extraction of acoustic cues pertinent to a phonetic feature can be tuned to classes of sounds separated on the basis of other cues, and this serves to increase the reliability of segment labeling. The analysis sequence yields a structure for the syllabic units of the speech signal that may be used to retrieve similar syllabic units for detailed comparison.

INTRODUCTION

WHAT IS the relationship between the acoustic cues of the speech signal and its phonetic features? Evidence available today appears to indicate that there is no simple transformation from the cues directly extractable from the signal by signal processing techniques to the phonetic features, the distinguishing characteristics of the individual phonetic elements. Rather, a complex encoding takes place so that information about a particular feature of a segment may in fact be carried by neighboring segments. A feature may be signaled by cues that differ, depending on other features present in the same segment, as well as on the contextual environment in which that segment is embedded. This paper outlines a strategy for drawing inferences about the phonetic features of segments from a sequence of acoustic processing steps, each of which characterizes in increasing detail the acoustic information present.

A syllabic unit is defined as a segment of the speech signal delimited by significant minima in a loudness function, a time-smoothed frequency-weighted summation of the signal spectrum. We focus our attention on characterizing segments within a syllabic unit and on the relationships between these segments. The relationships between the segments reveal a structure for the syllabic unit which may be used to select units of similar structure from a store of syllabic forms. At any point within the processing sequence, the cost of further characterization of the segments is weighed against the remaining ambiguities in the possible structural matches. The sequence of acoustic analysis steps set up as generally applicable may be modified, in light of the ambiguities remaining at any point, to derive the maximum useful information from any particular analysis step.

STRATEGY FOR SEGMENTAL ANALYSIS

Following Fant [1], I consider the speech signal to be composed of "a sequence of minimal sound segments, the boundaries of which are defined by relative distinct changes in the speech wave structure" [1, p. 7]. Consider a sequence of one or more such minimal sound segments that have some common acoustic property as an acoustic segment. By focusing in turn on different properties, we can isolate and appropriately label segments that exhibit these properties to differing extents. We may contrast two strategies for the segmentation and phonetic labeling of acoustic segments.

Manuscript received March 29, 1974; revised August 13, 1974. This work was supported in part by the Advanced Research Projects Agency of the Department of Defense under Contract N0014-67-A-0129-002, monitored by the Office of Naval Research. The views presented here do not necessarily represent the views of the Department of Defense. This paper was presented at the IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, Pa., April 15–19, 1974.

The author is with Haskins Laboratories, New Haven, Conn. 06510.

PRINTED IN THE U.S.A. Annals No. 502AS010

1) A set of acoustic cue detectors is constructed to operate on the speech signal in parallel and independently of each other. Whenever a change is noted in at least one acoustic cue, the signal is divided into separate segments. Call this the parallel cue detection strategy.

2) A number of acoustic cue detectors are applied to the speech stream sequentially. The selection of the detector to be applied next follows a decision tree. Call this the sequential cue detection strategy.

The parallel detection strategy is applicable to a model of speech analysis that considers the momentary speech signal to be a function, probably nonlinear, of independent acoustic features. Certain features manifest themselves quite independently. Voicing and frication can be considered independent features from this point of view. The amount of aperiodic energy needed to call a segment fricative in the presence of voicing is larger than that required to call it fricative in the absence of voicing. In fact, the outputs due to the separate excitation sources are known to combine nonlinearly. All sound segments are searched for all features. As a result, the sound segment is located in the hyperspace of acoustic features.

The sequential strategy makes use only of a minimal set of cues adequate to characterize the sound segment. Cues outside the minimal set are considered redundant. Since the phonological units are not always represented by the same acoustic cues, acoustic properties considered redundant in some cases may be used to aid the general transformation from acoustic cues to phonetic segments. There exists evidence today for independent human storage of phonetic features, for example, place and manner of production (Wickelgren [3]). The corresponding acoustic cues are, however, not generally independent. The perception of an acoustic cue underlying a particular phonetic feature may vary with changes in the acoustic cues underlying other phonetic features (Pisoni and Sawusch [2]). For example, place of production cues are functions of the voicing feature, although the voicing cues are generally independent of the place feature. Therefore, independent search for acoustic cues appears undesirable.

An important property of the sequential strategy is that segmentation and labeling are results of the same operations. A stretch of the speech signal is segmented if some analysis operation yields significantly different results over that stretch. Simultaneously, differing labels are attached to the newly derived segments. This procedure runs counter to the traditional pattern recognition strategy of complete segmentation followed by analysis of segments.

Application of any particular analysis function to an appropriate acoustic segment can yield only a small number of alternative productions, as suggested by the phonological rules of the language. By thus limiting the number of segments produced, we avoid the requirement for independent analysis of time-synchronous chunks of speech, we limit the total number of decisions made, and we reduce the possibilities for phonologically inconsistent labeling of segments.

ANALYSIS RULES

The segmentation and analysis rewrite rules given below are formally context independent. The decision whether a segment is to be further subdivided is based only on acoustic information contained in that segment. These rules govern mainly the number of segments to be produced and their labeling as to voicing and manner of production. Further place of production analysis rules can be expected to be context dependent, and they will be considered in greater detail below. The entire strategy consists of two stages of analysis, one context independent, determining the number of subsegments, and one context dependent, deriving further information about the individual segments.

In constructing an appropriate acoustic analysis sequence, we may profitably utilize the phonological rules of the language that restrict the segmental makeup of syllabic units. The rules on the manner of production of the sequence of units are particularly strong. For example, according to our definition, a syllabic unit must be completely voiced, be composed of a voiced segment bounded by voiceless segments on one or both sides, or be a syllabic fragment that is completely unvoiced. Segments differing in voicing and manner of production can be ordered according to sonority so that in going from the initial boundary through the syllabic peak to the final boundary, sonority is first monotonically increasing, then monotonically decreasing. This suggests that a strategy for manner of production cue analysis should proceed from the edges of the syllabic unit to the center and look for manner of production cues that accompany increasing degrees of sonority.

The place of production rules, not yet implemented, allow the analysis of segments to be carried out according to a sequence dependent on the previously derived manner of production information. The selection of vowels appears to be least dependent on the neighboring consonantal segments; therefore, a preliminary vowel decision can be made first. This preliminary decision, classifying the vowels only into three groups, /i/-, /a/-, and /u/-like, can be followed by a subsequent analysis taking coarticulation rules into account once the neighboring consonants have been identified in greater detail. Determination of the place of production of consonant classes is context dependent in the sense that the syllabic vowel color is taken into consideration when making that decision. Consonants are considered in order of decreasing sonority moving outward from the syllabic peak.

IMPLEMENTATION OF REWRITE RULES

The following rewrite rules for segmentation and labeling have been implemented and are undergoing evaluation. Each rule transforms the given segment into the indicated subsegments if the criteria for the results of the acoustic analysis are satisfied. The subsegments correspond to the nodes on the segment-structure tree that are descendants of the original segment.

TABLE I
REWRITE RULES FOR SYLLABLE ANALYSIS

1.	[Sentence] → [Syllabic unit] (Sentence)	//Syllabication//
2.	$\text{[Syllabic unit]} \rightarrow \left\{\begin{matrix}\text{[Voiceless segment]}\\ \text{(Voiceless segment) } [V_1] \text{ (Voiceless segment)}\end{matrix}\right\} \text{ (Silent segment)}$	//Voicing decision//
3.	$\text{[Voiceless segment]} \rightarrow \left\{\begin{matrix}\text{(Voiceless burst) } \left\{\begin{matrix}\text{[Voiceless fricative]}\\ \text{[Aspiration]}\end{matrix}\right\}\\ \text{[Voiceless burst]}\end{matrix}\right\}$	//Voiceless subsegments//
4.	$[V_1] \rightarrow \text{(Voice bar) (Voiced burst) } [V_2] \text{ (Voice bar)}$	//Voiced stops//
5.	$[V_2] \rightarrow \text{(Voiced fricative) } [V_3] \text{ (Voiced fricative)}$	//Voiced fricatives//
6.	$[V_3] \rightarrow \left\{\begin{matrix}\text{(Nasal consonant) } [V_4] \text{ (Nasal consonant)}\\ \text{[Syllabic nasal]}\end{matrix}\right\}$	//Nasal//
7.	$[V_4] \rightarrow \text{(Liquid) } [V_5]$	//Prevocalic liquid//
8.	$[V_5] \rightarrow \left\{\begin{matrix}[V_6] \text{ (Liquid)}\\ \text{[Syllabic liquid]}\end{matrix}\right\}$	//Postvocalic liquid//
9.	$[V_6] \rightarrow \left\{\begin{matrix}\left\{\begin{matrix}\text{(Short vowel)}\\ \text{(Semivowel)}\end{matrix}\right\} \text{ [Long vowel]}\\ \text{[Long vowel] } \left\{\begin{matrix}\text{(Semivowel)}\\ \text{(Short vowel)}\end{matrix}\right\}\\ \text{(Short vowel)}\end{matrix}\right\}$	//Vowel-like segments//

Legend: [ ] - mandatory segment
( ) - optional segment
{ } - ordered disjunction
// // - comments

Note that some intervocalic segments will be cut apart by the syllabic division rule. Thereafter, the two subsegments will be processed individually as parts of each syllabic unit. When the final results are analyzed, identically labeled segments that follow each other in time may be combined into one segment. The nine rules shown in Table I give the voicing and manner of production analysis. Further rules are to be added for analysis of vowels and place of production of consonants.

Our syllabic units are acoustic segments. They are derived from the actual production and thus do not correspond precisely to linguistic (phonological) syllables. In particular, two words may form one syllabic unit if the first ends in an open vowel, the second starts with an open vowel, and no glottal stop intervenes. For example, "the old" generally forms one syllabic unit /ðIold/ in which Rule 9 will attempt to find two vowels. Rules 7 and 8 are separated to indicate explicitly that the cues for prevocalic liquids may be different from those for postvocalic liquids.

The rewrite rules cited are not meant to be complete. Rather, they indicate the kinds of rules required to implement the strategy outlined here.

DATA STRUCTURES

The question of an appropriate data structure to express the results of the analysis operations is rather important. The structure most appropriate for hierarchic analysis is that of a tree whose root node corresponds to the complete utterance and whose subtrees correspond to each syllabic unit. Where an acoustic property is found present or absent for the entire segment, that segment is relabeled but not cut. Where a significant change is found for that property over the span of the segment, the segment is cut into two parts at the point of change. Branches equal in number to the number of subsegments are grown from the node corresponding to the original segment. The nodes branching from any higher node are ordered in time according to the time ordering of the segments corresponding to the nodes. Since the root nodes corresponding to the syllabic units are similarly ordered, a temporal chain of segments is maintained at all times in the processing sequence. This chain allows efficient reference to preceding and succeeding segments (nodes) even where these do not branch from the same parent node. By ordering the analysis operations so that those operations which can be expected to be more reliable are carried out first, we can construct a metric of differences for syllabic units that decreases with the level of the node where a difference is encountered. Highly similar units have the same structure and differ only in the label assigned to the terminal nodes. Less similar units may have a common set of topmost nodes to which differing subtrees appear to be attached.

We may assign a similarly structured map to each entry in the lexicon of admissible syllabic forms. Consider the situation where a phonologically acceptable derivation is found, but an error is suspected because no entry matching the derived structure can be located in the lexicon. Likely substitutions can be found by examining forms to which only a partial match exists. To this end, the lexical entries are assigned a code appropriate to the structural map. Retrieval is thus made possible on the basis of the complete code word or any initial subsequence thereof.

Fig. 1 illustrates the segmentation-derived structure for the word "spinster." Three syllabic units are found, the first a fragment containing the initial fricative [s] and the stop gap of [p], the second roughly corresponding

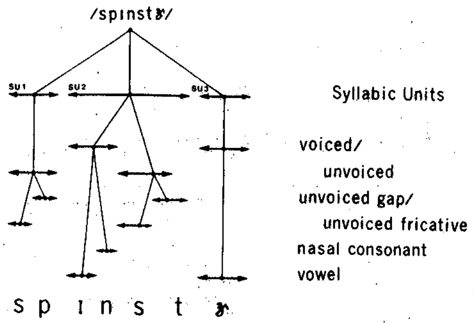

Fig. 1. Segment tree for word "spinster."

to the sequence [Inst], and the third the final vowel [ɝ ϑ]. The voicing decision separates the second syllabic unit into two parts, finds the first all voiced and the last all unvoiced. The unvoiced fricative detector segments the first syllabic unit and the unvoiced segment of the second. The nasal detector is applied only to voiced segments and it segments that portion of the second syllabic unit. The fact that seven segments are found for this word, equal in number to the constituent phonemes, is purely coincidental. The total number of segments may be larger than the number of phonemes, due to segmentation of stop releases, or smaller, due to incomplete segmentation of vowel-like sequences.

To date, the voicing, syllabic unit, frication, and nasal consonant indicators have been implemented. These are the distinctly different segments of the speech stream. The detection of liquids and semivowels lies immediately ahead. Thereafter, attention will focus on the information supplied by segments within the syllabic unit regarding the place of production of those segments themselves as well as of neighboring segments. Although to date our experiments have been restricted to short utterances, there appear few serious problems in extending the procedure to longer utterances. Perhaps the major problem apparent at the moment is the increased difficulty of segmenting unstressed syllabic units and the decrease in detail recoverable from them. This may require the treatment of the stressed and unstressed counterparts of the same syllabic units as different lexical forms for comparison purposes.

ACKNOWLEDGMENT

The author appreciates the assistance of F. S. Cooper, G. Kuhn, and L. Lisker, who participated in discussions leading to the formulation of the ideas presented here.

REFERENCES

[1] C. G. M. Fant, "Descriptive analysis of the acoustics of speech," *Logos*, vol. 5, pp. 3–17, 1962.

[2] D. B. Pisoni and J. R. Sawusch, "On the identification of place and voicing features in synthetic stop consonants," Haskins Lab., New Haven, Conn., Status Rep. on Speech Res., SR-35/36, pp. 65–80, 1973.

[3] W. A. Wickelgren, "Distinctive features and errors in short-term memory for English consonants," *J. Acoust. Soc. Amer.*, vol. 39, pp. 388–398, 1966.