Toward a model for lexical access based on acoustic landmarks and distinctive features

Kenneth N. Stevens[a)]
Research Laboratory of Electronics and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachussetts 02139-4307

(Received 16 January 2001; accepted for publication 10 January 2002)

This article describes a model in which the acoustic speech signal is processed to yield a discrete representation of the speech stream in terms of a sequence of segments, each of which is described by a set (or bundle) of binary distinctive features. These distinctive features specify the phonemic contrasts that are used in the language, such that a change in the value of a feature can potentially generate a new word. This model is a part of a more general model that derives a word sequence from this feature representation, the words being represented in a lexicon by sequences of feature bundles. The processing of the signal proceeds in three steps: (1) Detection of peaks, valleys, and discontinuities in particular frequency ranges of the signal leads to identification of acoustic landmarks. The type of landmark provides evidence for a subset of distinctive features called articulator-free features (e.g., [vowel], [consonant], [continuant]). (2) Acoustic parameters are derived from the signal near the landmarks to provide evidence for the actions of particular articulators, and acoustic cues are extracted by sampling selected attributes of these parameters in these regions. The selection of cues that are extracted depends on the type of landmark and on the environment in which it occurs. (3) The cues obtained in step (2) are combined, taking context into account, to provide estimates of ''articulator-bound'' features associated with each landmark (e.g., [lips], [high], [nasal]). These articulator-bound features, combined with the articulator-free features in (1), constitute the sequence of feature bundles that forms the output of the model. Examples of cues that are used, and justification for this selection, are given, as well as examples of the process of inferring the underlying features for a segment when there is variability in the signal due to enhancement gestures (recruited by a speaker to make a contrast more salient) or due to overlap of gestures from neighboring segments.
[DOI: 10.1121/1.1458026]

PACS numbers: 43.71.An, 43.72.Ar, 43.72.Ne [KRK]

I. INTRODUCTION

This article describes a proposed model of the process whereby a listener derives from the speech signal the sequence of words intended by the speaker. The proposed model contains a lexicon in which the words are stored as sequences of segments,[1] each of which is described in terms of an inventory of distinctive features. Acoustic cues are extracted from the signal, and from these cues a sequence of feature bundles is derived. This pattern of estimated feature bundles is matched against the items in the lexicon, and a cohort of one or more sequences of words is hypothesized. A final feedback stage synthesizes certain aspects of the sound pattern that could result from each member of the cohort, and selects the word sequence that provides the best match to the measured acoustic pattern.

This article is concerned primarily with the part of the model that leads to a description of the signal in terms of a sequence of discrete phonological segments, i.e., in terms of bundles of distinctive features. That is, we are attempting to model the speech perception process up to the point where the analog acoustic signal has been interpreted as a sequence of discrete phonological units.

There is ample evidence that words are stored in memory in terms of sequences of segmental units, and that these segmental units are further represented in terms of the values of a set of binary features. That is, the lexical representation is discrete in at least two ways: each word is an ordered sequence of discrete segments, each of which is represented by a discrete set of categories. Some of this evidence comes from acoustic studies of sounds produced by various manipulations of the vocal tract, showing certain distinctive and stable acoustic patterns when the vocal tract is in particular configurations or performs particular maneuvers. These combinations of acoustic and articulatory patterns are based on the physics of sound generation in the vocal tract, including theories of coupled resonators, the influence of vocal-tract walls on sound generation, and discontinuities or stabilities in the behavior of sound sources (Stevens, 1972, 1989, 2001). Evidence for features also comes from quantal aspects of auditory responses to sound, such as responses to acoustic discontinuities and to closely spaced spectral prominences (Chistovich and Lublinskaya, 1979; Delgutte and Kiang, 1984). There is further evidence that these features are grouped together in a hierarchical structure (Clements, 1985; McCarthy, 1988; Halle, 1992; Halle and Stevens, 1991). The aim of the acoustic processing of the speech signal, then, is to uncover the features intended by the speaker, so that these

a)Electronic mail: stevens@speech.mit.edu

 0001-4966/2002/111(4)/1872/20/$19.00

features can be matched against the lexicon, which is also specified in terms of segments and features.

In contrast to the discretely specified phonological representation of an utterance, the acoustic signal that is radiated from the mouth of a speaker is continuous. It is an analog signal that is generated by continuous movements of a set of articulatory and respiratory structures. As has been observed, however, the relations between the articulatory and the acoustic representations of speech have certain quasi-discrete or quantal characteristics that are exploited in speech production (Stevens, 1972, 1989). These quantal attributes help to simplify the process of uncovering the discretely specified segmental and categorical representations that are building blocks of words.

In this article we first examine how the acoustic signal provides evidence for the presence of the discrete segmental units, and we review the inventory of segmental features and their defining or primary acoustic and articulatory correlates (Sec. II). A formal structure for storing words in the lexicon is then described (Sec. III). We then observe (in Sec. IV) that in running speech there are several factors that combine to create variability in the acoustic representation of the segments and features. These factors introduce additional cues for the features depending on the context in which they occur. It is argued that this variability can be reduced by selecting acoustic cues that are closely tied to articulation, since the variability can be traced to modifications in articulation that are governed by well-defined principles. The process of estimating the discretely specified underlying segments and their features from analysis of the continuous speech signal is then outlined (Sec. V).

II. SEGMENTS AND THEIR FEATURE REPRESENTATIONS

The distinctive features describe the contrasts between words in language. That is, a change of the binary value of a feature for a segment in a word has the potential of generating a different word. For example, the word pairs seat and sheet or bat and pat differ by a single feature in the first segment. As discussed below, each distinctive feature is assumed to have a defining articulatory action and a corresponding defining acoustic correlate.

A. Segments and articulator-free features

There is a set of features that classify segments into broad classes—classes that can be described roughly as vowels and some general classes of consonants. These are called articulator-free features since they do not specify a particular articulator that is activated to produce the segment (Halle, 1992). Rather, they refer to general characteristics of constrictions within the vocal tract and the acoustic consequences of producing these constrictions.

Probably the most fundamental distinction in speech is between vowels and consonants. Vowels are produced with a relatively open vocal tract, so that air can flow freely through the tract without obstruction and without significant pressure drop across any narrowing that may be present in the supraglottal airway. True consonants, on the other hand, are produced with a significant narrowing of the airway in the oral cavity.

Acoustically, vowels have greater intensity than consonants. The first formant frequency is generally higher in vowels than in consonants because the oral cavity is relatively open for vowels and there is a narrow constriction in the oral region for consonants. The generation of true consonants involves a sequence of two steps: the formation of a narrowing in the oral cavity, and the subsequent release of that narrowing. By definition, true consonants cause a particular type of discontinuity in the acoustic signal at one or both of these times of closing and releasing. The acoustic discontinuity is a consequence of a rapid change in the pressures and flows within the vocal tract—either a rapid increase or decrease in pressure behind the constriction in the oral cavity (to produce obstruent consonants) or an abrupt switching of the airflow to a different path within the oral and nasal cavities without an increase in intraoral pressure (to produce sonorant consonants). In the case of sonorant consonants, it is possible to produce the requisite acoustic discontinuity only by creating a complete closure of an oral articulator in the midline of the oral cavity. For obstruent consonants, however, an acoustic discontinuity can be formed either by making a complete closure in the midline or by creating a partial closure that is narrow enough to cause a pressure increase and to create continuous flow through the constriction.

One of the ways that true consonants differ from vowels (in addition to the presence of the acoustic discontinuity noted above) is that the spectrum amplitude in the low and midfrequency regions in the consonant region is weaker than the corresponding spectrum amplitude in the adjacent vowels. A reduced spectrum amplitude of this kind can also be produced without generating a constriction that is sufficient to cause an acoustic discontinuity. Segments produced in this way are glides. In English, they include /w/ and /j/, which are produced by raising the tongue dorsum to produce a narrowing between the dorsum and the palate, and, in the case of /w/, a rounding of the lips. The consonant /h/ is also a glide, and the reduced amplitude is produced by spreading the glottis without creating a significant narrowing in the vocal tract above the glottis.

In summary, then, there are three broad classes of segments: vowels, glides, and consonantal segments. Production of a vowel causes a maximum in the low- and midfrequency spectrum amplitude. An acoustic discontinuity occurs at the times when a consonantal constriction is formed or released. And the acoustic correlate of a glide is a reduction in low- and midfrequency spectrum amplitude but without acoustic discontinuities. Evidence for these three kinds of segments within an utterance is provided by landmarks in the signal: a peak in low-frequency amplitude for a vowel, a minimum in low-frequency amplitude, without acoustic discontinuities, for a glide, and two acoustic discontinuities for a consonant, one of which occurs at the consonant closure and one at the consonant release. Figure 1(a) displays a spectrogram of a sentence, and shows the placement of vowel and consonant landmarks; examples of glide landmarks are shown in Fig. 1(b). In Fig. 1(a), the arrows at the top of the spectrogram

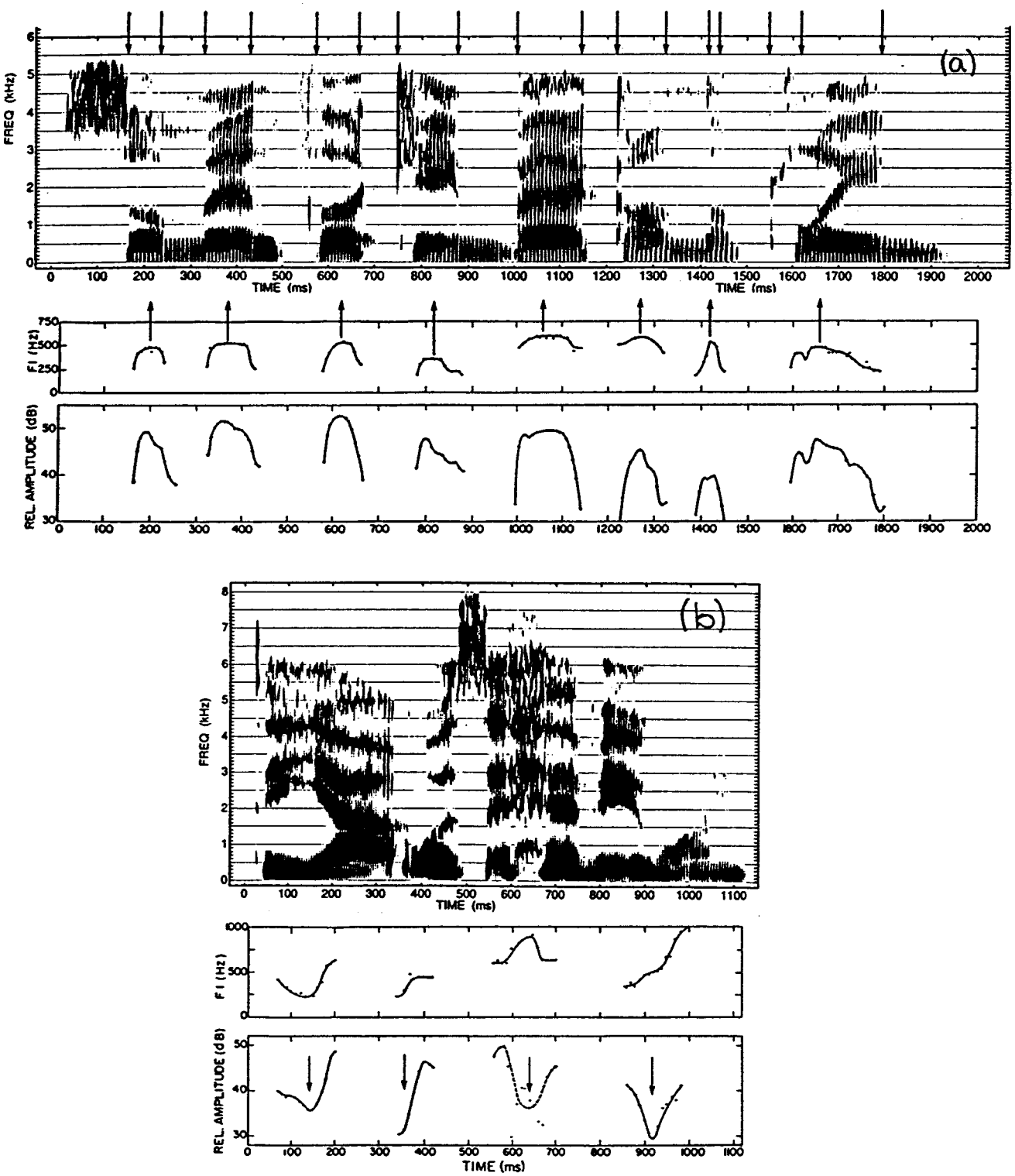


FIG. 1. (a) Shown at the top is a spectrogram of the sentence Samantha came back on the plane, produced by an adult male speaker. The plot immediately below the spectrogram gives the frequency of *F*1 versus time during the vocalic regions, measured at 7.5-ms intervals. At the bottom is a plot of the relative amplitude of the first-formant prominence during the vocalic regions. Each point on this amplitude plot and on the plot of frequency of *F*1 is measured from a spectrum which is obtained by averaging a 10-ms sequence of spectra, each of which is calculated using a 6.4-ms Hamming window. The arrows below the spectrogram indicate vocalic landmarks and the arrows at the top identify consonantal landmarks. See text (from Stevens, 1998). (b) The spectrogram is the sentence The yacht was a heavy one, produced by a female speaker, illustrating acoustic attributes of glides. The first-formant frequency in the vicinity of the glides is shown immediately below the spectrogram. The plot at the bottom is the amplitude of the *F*1 prominence. The arrows at the amplitude minima for the glides identify the glide landmarks. The irregularities in the amplitude for /h/ are smoothed with the dashed line. Measurements are made at 10-ms intervals using the procedures described in the legend for part (a) (from Stevens, 1998).

TABLE I. Articulator-free features for some consonants in English.

	t,d	s,z	θ,ð	n
Continuant	−	+	+	−
Sonorant	−			+
Strident		+	−	

identify acoustic discontinuities due to consonant closures or releases. The two panels below the spectrogram are plots of the first-formant frequency versus time and the amplitude of the F1 spectral prominence versus time during the vowels. Vowel landmarks are indicated by the arrows near the times when F1 is a maximum, which is close to the time when the low-frequency amplitude is a maximum. For the four glides in Fig. 1(b), the arrows identify times when the amplitude of the F1 prominence is a minimum.

In the case of consonantal segments, we have observed that two further subclassifications can be assigned. A consonant can be formed with a complete closure in the oral cavity in the midline (designated by the articulator-free feature [−continuant]) or with a partial closure that permits a fricative consonant to be produced with continuous turbulence noise (designated as [+continuant]). For [−continuant] segments, there is a further subdivision into [+sonorant] (no increase in pressure behind the consonant closure) and [−sonorant] (increased intraoral pressure).

Still another articulator-free feature is used to implement a contrast for [+continuant] consonants produced with the tongue blade. With appropriate shaping of the cavity anterior to the constriction, and by directing the air stream against the lower teeth, the resulting sound spectrum of the fricative consonant at high frequencies has a much greater spectrum amplitude than the adjacent vowel in the same frequency range. Such a consonant is designated as [+strident], whereas a fricative with weak high-frequency noise is [−strident].

For consonantal segments, then, there are three articulator-free features: continuant, sonorant, and strident. Table I gives a classification of representative consonants in terms of these features. It is noted that the feature [sonorant] is distinctive only for [−continuant] consonants, whereas [strident] is distinctive only for [+continuant] consonants. We adopt here a convention of not designating a feature value when that value is redundant, i.e., can be predicted from the other features.

B. Articulator-bound features and their acoustic correlates

We turn now to a description of features that specify which articulators are active when the vowel, glide, or consonant landmarks are created, and how these articulators are shaped or positioned. Descriptions of the primary articulatory and acoustic correlates of these features have been given in a number of publications (Chomsky and Halle, 1968; Keyser and Stevens, 1994), and we will simply summarize those descriptions here.

There are seven articulators that can potentially be active in producing phonetic distinctions in language. These

TABLE II. Seven articulators and the features that specify phonetic distinctions that can be made with each articulator. The articulators are divided into three groups: those that can form a constriction in the oral cavity, those that control the shape of the airway in the laryngeal and pharyngeal regions, and the aspect of vocal-fold adjustment relating to stiffness.

lips	[round]
tongue blade	[anterior]
	[distributed]
	[lateral]
	[rhotic]
tongue body	[high]
	[low]
	[back]
soft palate	[nasal]
pharynx	[advanced tongue root]
glottis	[spread glottis]
	[constricted glottis]
vocal folds	[stiff vocal folds]

are (1) the lips, (2) the tongue blade, (3) the tongue body, (4) the soft palate, (5) the pharynx, (6) the glottis, and (7) adjustments of the tension of the vocal folds. These articulators can be manipulated in one or more ways to generate segments that are distinctive (i.e., that could potentially distinguish between words). Each of these ways of manipulating an articulator is described by a feature, which can have a plus or minus value. The articulators and the features associated with them are listed in Table II.

In Table II, the articulators are classified into three groups. The first group, consisting of the lips, the tongue blade, and the tongue body, involve adjustments of the configuration of the oral cavity. The soft palate, the pharynx, and the glottis form the second group, and manipulations of these articulators control the vocal-tract shape in the laryngeal and pharyngeal regions. The third class in Table II describes adjustments of the stiffness or slackness of the vocal folds, and these adjustments are not involved directly in changing the shape of the airway. Not all of the features in Table II are distinctive in English, and this is not an exhaustive list of the universal distinctive features that are used across languages.

Vowels and glides are specified completely by a listing of articulator-bound features. In general, the major features for vowels are the features specifying the position of the tongue body—the features [high], [low], and [back]. In English, additional features that describe contrasts for vowels are [round] and [advanced tongue root].[2] The features that distinguish between glides can involve any of the top six articulators in Table II. In English (and in most, if not all, languages), the inventory of glides is considerably smaller than the inventory of vowels or consonants.

A listing of the features for some vowels and glides is given in Table III. Note that the feature [round] is unspecified for front vowels, and the feature [spread glottis] is the only one that is specified for the glide /h/. Some vowels in English are diphthongs, such as the vowels in hide and boy. We represent these diphthongs as sequences of two segments, the second of which is a glide. An example of the feature representation for the diphthong /ɑj/ is given in Table III. Our convention for the offglides of such diphthongs is to leave as

TABLE III. Listing of the features for some vowels and glides in English.

	i	ɛ	æ	ɑ	ʌ	u	w	j	h	ɑj
High	+	−	−	−	−	+	+	+		−+
Low	−	−	+	+	−	−	−	−		+−
Back	−	−	−	+	+	+	+	−		+−
Round				−	−	+	+			
Advanced tongue root	+	−	−	−	−	+	+	+		−
Spread glottis									+	

unspecified the value of the feature [advanced tongue root] but to mark this offglide as a glide.

To fully describe a consonant, it is necessary to designate which primary articulator is used to form the constriction that produces the acoustic discontinuities or discontinuity that constitutes the landmark. Only three articulators can produce these consonantal landmarks: the lips, the tongue blade, and the tongue body. These articulators produce constrictions in the oral cavity. Associated with each articulator there is a set of one or more articulator-bound features. For a given primary articulator, some of these features must be specified. In addition to designating a primary articulator and its features, it is usually necessary to specify the features associated with secondary articulators. In English, these features are [nasal] and [stiff vocal folds]. (In some other languages, the features [spread glottis] and [constricted glottis] are used distinctively.)

A listing of articulators and articulator-bound features for some consonants in English, together with articulator-free features already given, is displayed in Table IV. Features that are redundant or do not contribute to a distinction in English for a given segment are left blank.

The articulator-bound features in Tables II–IV are identified with particular articulators and with the shaping and positioning of those articulators. However, these features also have well-defined acoustic and perceptual correlates. Linguists have argued that the universal set of features that are used to define contrasts in the languages of the world (of which those discussed above for English represent a subset) have the property that for each feature there is a coincidence between the articulatory description for the contrast and a distinctive acoustic correlate for that contrast. We illustrate this premise with three examples.

When a consonant is produced with a tongue-blade constriction, the constriction can be positioned forward of the alveolar ridge (in which case the defining feature is [+anterior]) or posterior to the alveolar ridge (i.e., [−anterior]). The strident fricatives /s/ and /š/ in English are distinguished by this feature, as Table IV shows. Acoustically, the tongue blade position for /s/ creates a resonance of the cavity anterior to the constriction that is in the range of the fourth or fifth formant for the speaker, giving rise to a spectral prominence in the sound in this frequency range. In the case of /š/, the tongue blade is moved to a slightly more posterior position, and is shaped somewhat differently, so that there is a spectral prominence in the third formant range. Thus the + or − values of the feature [anterior] define the position of the tongue-blade constriction with respect to a landmark on the hard palate. Coincident with this manipulation of the constriction, there are well-defined acoustic (and perceptual) consequences that specify which natural frequency of the vocal tract receives principal excitation from the frication noise source near the constriction.

Another example is the feature [back] for vowels. For [+back] vowels, the tongue body is displaced back to form a narrowing in the pharyngeal or posterior oral cavity. The acoustic consequence is a second-formant frequency that is low and close to the first-formant frequency. Vowels classified as [−back], on the other hand, are produced with the tongue body forward and a high second-formant frequency. A natural dividing line for [+back] and [−back] vowels is the frequency of the second subglottal resonance, which is in

TABLE IV. Listing of articulator-free features, articulators, and articulator-bound features for some consonants in English.

	b	d	g	p	f	s	z	š	m	l
Continuant	−	−	−	−	+	+	+	+	−	−
Sonorant	−	−	−	−					+	+
Strident						+	+	+		
Lips	+			+	+				+	
Tongue blade		+				+	+	+		+
Tongue body			+							
Round	−			−	−				−	
Anterior		+				+	+	−		+
Lateral									−	+
High			+							
Low			−							
Back			+							
Nasal									+	−
Stiff vocal folds	−	−	−	+	+	+	−	+		

the vicinity of 1500–1700 Hz for adults (Cranen and Boves, 1987). When *F*2 for a vowel comes close to this resonance, an acoustic perturbation can be created. This subglottal resonance forms a kind of ''acoustic berm.'' Vowels that are [−back] tend to maintain *F*2 above this perturbation; for [+back] vowels, speakers tend to place *F*2 below this frequency. There is also evidence that the auditory response to a front vowel, with *F*2 high and close to *F*3, is qualitatively different from that for a back vowel (Syrdal and Gopal, 1986). Again we see a coincidence of an articulatory positioning and an acoustic and perceptual consequence of that articulatory manipulation.

A third example is concerned with the features [stiff vocal folds][3] (Halle and Stevens, 1971). When this feature is used to characterize an obstruent consonant, there are two acoustic consequences in the vicinity of the consonant release. For a consonant classified as [−stiff vocal folds] (commonly referred to as a voiced consonant), the vocal folds are slackened, and glottal vibration during the time when there is a reduced transglottal pressure in the consonant interval is facilitated. The increased slackening also carriers over into the following vowel, leading to a lowered fundamental frequency near the beginning of the vowel. The opposite effect occurs for a consonant that is [+stiff vocal folds] (i.e., a voiceless consonant). The vocal-fold stiffening inhibits glottal vibration during the obstruent interval. This increased stiffening extends into the following vowel, leading to a raised fundamental frequency near the beginning of the vowel. Again we see a coincidence between the articulatory manipulation for the feature and the acoustic consequences of this manipulation. In this example there is an acoustic correlate of the feature in the constricted interval for the consonant and a different acoustic correlate in the vowel following the consonantal release. (As will be noted later, in certain phonetic or prosodic environments in English, additional gestures may be introduced to enhance the saliency of the contrast between consonants that differ in the feature [stiff vocal folds].)

These and other examples provide support for the view that for each articulator-bound distinctive feature there is a defining articulatory manipulation and a distinctive acoustic pattern that results from this manipulation. We will observe later, however, that, in addition to the defining acoustic attribute for a feature, there may be added articulatory gestures and further acoustic cues that can contribute to identification of the feature, depending on the context in which it appears.

The acoustic correlates of the articulator-bound features depend to some extent on the type of segment—whether it is a vowel, a glide, or a consonant. In general, the acoustic manifestations of these features are most salient in the vicinity of the landmarks. For example, to estimate the values of the features for a vowel or glide, attributes of the acoustic signal in the region that is within a few tens of milliseconds of the vowel or glide landmarks must be examined. In general, the acoustic properties in the vicinity of these landmarks for vowels and glides change in a smooth and continuous manner, and the time course of these changes provides the information needed to uncover the features for these segments. Acoustic information that can help to determine articulator-bound features for consonants resides in regions that are a few tens of milliseconds on either side of the consonant landmarks. Since the consonant landmarks are defined by acoustic discontinuities, the acoustic properties near one side of a landmark are usually quite different from the properties near the other side of the landmark, as in the example above relating to stiff vocal folds (Stevens, 2000a). The diverse information from these two sets of properties, including the nature of the abrupt changes in certain acoustic properties, must be integrated to determine the intended articulator-bound features.

The features listed in Tables I–IV can be organized in a geometrical, treelike hierarchy to represent more closely their relation to the articulatory structures that are used to implement them. This treelike structure highlights the fact that the various features in Tables I–IV are not simply a list, but are structured into groups. As has been noted elsewhere (Clements, 1985; McCarthy, 1988; Keyser and Stevens, 1994), the phonological patterning of the sounds in language can be described succinctly in terms of these groupings. In the model described in the present article, however, the lexical representations will not utilize this formal hierarchical structure.

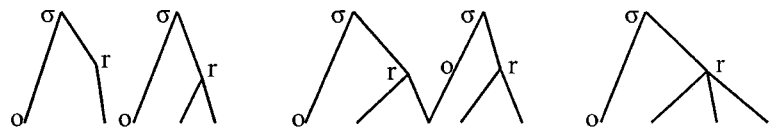
III. LEXICAL REPRESENTATION AND PLANNING STAGE

The stored items in the lexicon consist of sequences of segments, each of which is represented by a list of features. The features are of several kinds: membership in one of three broad classes of vowel, glide, and consonant; in the case of vowels and glides, a listing of articulator-bound features; and in the case of consonants, a specification of the articulator-free features (continuant, sonorant, strident), one of three possible primary articulators (tongue body, tongue blade, and lips), and a set of articulator-bound features.

It is also necessary to specify in the lexicon the structure of the syllables. In English, there are particular constraints on the contexts in which consonants can occur. These constraints for a given consonant can be described conveniently in terms of the position of the consonant in the syllable and on the features of the vowel and consonants within the syllable. A conventional way of representing syllable structure in English characterizes the syllable as an onset and a rime, with the rime consisting of a nucleus and a coda (cf. Clements and Keyser, 1983). A consonant or consonant sequence can be appended after the coda as an affix. If there is a glide or a sonorant consonant in the onset or the coda, this segment is always adjacent to the nucleus, or vowel (with rare exceptions, as in the sequences /mju/ or /nju/ in words like mute and new). With some exceptions, a consonantal segment (i.e., a segment that is implemented with abrupt landmarks) is always adjacent to either a vowel or a glide. (The exceptions are when the consonant is /s/ or is adjacent to /s/, as for the /s/ in spot or the /k/ in desk. Other exceptions are in affixes following the rime, as in asked or fifth, where the affixes are always consonants produced with the tongue blade.)

Knowledge of the syllable structure is important during speech production because details of the implementation of a

TABLE V. Lexical representations for the words debate, wagon and help. The syllable structure of each word is schematized at the top (σ=syllable, o=onset, r=rime).



	d	ə	b	e	t		w	æ	g	ə	n		h	ɛ	l	p
Vowel		+		+				+		+				+		
Glide							+						+			
Consonant	+		+		+				+		+				+	+
Stressed		-		+				+		-				+		
Reducible		+		-				-		+				-		
Continuant	-		-		-				-		-				-	-
Sonorant	-		-		-				-		+				+	-
Strident																
Lips			+													+
Tongue blade	+				+						+				+	
Tongue body									+							
Round			-				+									-
Anterior	+				+						+				+	
Lateral															+	
High		+		-			+	-	+	-				-		
Low		-		-			-	+	-	-				-		
Back		-		-			+	-	-	+				-		
Adv. tongue root				+			+	-						-		
Spread glottis													+			
Nasal											+					
Stiff vocal folds	-		-		+				-							+

consonant, including the timing of the articulator movements, are dependent to some extent on its position within the syllable. The timing of implementation of the syllable nucleus or vowel is also influenced by the syllable structure. For a listener, the ability to estimate syllable structure in running speech can be helpful in determining word boundaries, since the onset of a word is always the onset of a syllable in the lexicon.

Words in English containing more than one syllable always mark one of the syllables as carrying primary stress. Furthermore, some vocalic nuclei in a word may be marked in the lexicon as being potentially reducible. Thus, three kinds of vocalic nuclei are identified in the lexicon: stressed, reducible, and neither stressed nor reducible. The acoustic correlates of primary stress and of reduction are not well defined, but some combination of vowel amplitude, duration, fundamental frequency contour, glottal source waveform, and vowel formant pattern contributes to judgments of these attributes. In the lexical representation, the status of a syllable as being stressed, reducible, or neither stressed nor reducible is indicated by two features that are attached to each vowel. These features are [stress] and [reducible], and three combinations of plus or minus are used. The combination [+stress, +reducible] is not allowed.

Examples of the representation of three words in memory in the proposed lexical access model are given in Table V. A schematic representation of the syllable structure for each word is given at the top of the table. The syllable is marked by σ and the onset and rime by o and r. One of the words has one syllable and the other two have two syllables. In the case of the word wagon, the consonant /g/ is shared between the two syllables of the word. This consonant is said to be ambisyllabic (Clements and Keyser, 1983).

The first three rows of the table below the syllable structure specify the general category of vowel, glide, or consonant. Stress and potential for reduction for the vowels are marked in the next two rows, followed by the three articulator-free features that apply to consonants. The primary articulator for each consonant is given by a "+" placed in one of the next three rows. The remaining rows list the articulator-bound features. The last two features are separated out from the others since they can be considered as

secondary articulator-bound features for consonants in English.

The conventions used for marking the features or descriptors for lexical items in Table V depart somewhat from lexical representations proposed, for example, by Chomsky and Halle (1968) or by Clements (1985). For example, the designation of vowel, glide, or consonant, while marked in the lexicon by "+" in the appropriate row, is not strictly a feature representation. Any one segment is classified by one "+," and such a segment is automatically not a member of either of the other classes. Similar comments could be made about the use of "+" to specify the primary articulator for a consonant. A consonant (at least in English and in most other languages) can have just one primary articulator.

It is evident that the representation of any one segment of a lexical item is rather sparse. All but one of the segments in Table V are represented in terms of 6 or 7 features (except /h/), whereas the entire list of features in the table numbers 21. The specification of some features is conditional on the values of the articulator-free features and, for consonants, on the designation of the primary articulator. If one requires that each feature be truly distinctive, in the sense that changing the value of the feature creates a potentially different word, then the feature representations in Table V can be made even more sparse. For example, a segment that is [+consonant, +tongue body] is automatically [+high], and for such a segment the features [−low, +back] are not distinctive. Thus the average number of distinctive features for each segment is smaller than the six or seven implied in Table V, especially for consonants.

In the representation of the words debate and wagon in Table V, the reducible syllables are assigned a full set of vowel features. The first syllable in debate has the features for the high front vowel /i/, and the second syllable in wagon has the features for the nonhigh back vowel /ʌ/. These are postulated to be the features underlying the vowels, and the acoustic manifestations of these features would be well defined in some situations if the words were spoken clearly. In this case, the vowels would be implemented as [−reduced]. However, the presence of the feature [+reducible] is an indication that these vowel features are not distinctive in the sense that no minimal pairs are generated when the values of these features are changed. Thus these features do not contribute to identification of the words. In the more common production of these vowels, the durations would be decreased.

In addition to stem morphemes, the lexicon also contains affixes which can be appended to stems to produce new words. This process can lead to modification of certain features in the stem or to different forms for the affix depending on features of segments at or near the stem boundary. Rules for generating these new formatives involve modification of features or groups of features in environments that are also specified in terms of features. An example is the generation of the plural forms of nouns. The process for generating these plural forms can be described as a sequence of two ordered rules. The first rule is to add an unstressed schwa vowel at the ends of words that terminate in strident consonants produced with the tongue blade. The second step in forming the plural adds /z/ in all cases except when the noun terminates in a voiceless segment, in which case the voiceless /s/ is added. These rules lead to the standard plural forms such as bus/busses, bag/bags, and bat/bats, the plural being represented by the affix /z/ in the first two examples and /s/ in the third example. Similar rules that manipulate features or groups of features apply to a large variety of affixes.

It is assumed that a representation of words in the memory of a speaker or listener takes the form shown in Table V. One type of evidence for this assumption comes in part from the fact that a language user appears to have access to phonological rules of the type just described for appending the plural phoneme to nouns in English. There is a large number of such rules (e.g., Chomsky and Halle, 1968), involving essentially all the features that are contrastive in English, i.e., for which the change in the value of a feature creates a possible new word. The rules can be expressed efficiently in terms of manipulations of these features or groups of features.

When an utterance is to be produced, there is an initial planning process. For a multi-word utterance, a sequence of words, possibly constructed from morpheme sequences, is selected from the lexicon. Retrieval processes and phonological rules operate on the morpheme sequences to produce a sequence of segments and their associated features, which will subsequently be translated into articulatory commands. The syllabification of this segment/feature sequence may be adjusted somewhat from the syllabification in the lexicon. This planning stage also involves prosodic markers in the form of a metrical template which, at a minimum, indicates phrasally prominent syllables, full vowel syllables without phrasal prominence, and reduced syllables. In addition, prosodic specifications for the F0 contour, duration pattern, and amplitude contour of the utterance must be generated, to reflect both word-level and phrase-level prosodic constituents and prominences. The details of these prosodic aspects of the planning stage are not central to the aim of this article, which is concerned instead with the listener's task of estimating the speaker's planned sequence of segments and features. These issues about the nature of the planning stage have also been addressed in Levelt (1989), Shattuck-Hufnagel (1992), Levelt *et al.* (1999).

IV. MULTIPLE ACOUSTIC CUES FOR FEATURES AND THEIR DEPENDENCE ON CONTEXT

A. Articulatory interactions and multiple acoustic cues

From the point of view of the speaker, the array of segments and features in the planning stage provides a sketch of the instructions to the articulators specifying how a word is to be produced. It describes which articulators are to be manipulated and how they are to be shaped and positioned. These movements of the articulators in turn give rise to the sound pattern.

Although each of the distinctive features has certain defining articulatory and acoustic correlates, there are additional acoustic properties that are biproducts of the principal

articulatory and acoustic requirements for the feature. These properties can arise through articulatory actions that are not specified directly by the feature. Some of these actions may be automatic consequences of the primary articulatory gesture for the feature, and others may be introduced to enhance the perceptual contrast defined by the feature when the feature occurs in certain phonetic or prosodic contexts (cf. Keyser and Stevens, 2001). Such enhancing gestures often give rise to new acoustic properties, as well as strengthen the primary acoustic correlate of the feature. Of these enhancing gestures, some are obligatory in the sense that they are required for proper acoustic implementation of the primary acoustic property, whereas others may be optional. These additional articulatory gestures and their acoustic properties are not specified in terms of distinctive features since they do not by themselves define contrasts in the language. They can, however, provide acoustic and perceptual cues that potentially help the listener in the estimation of the distinctive features. This view that enhancing gestures are introduced by a speaker to strengthen a perceptual contrast is consistent with similar views expressed by Diehl (1991), by Kingston and Diehl (1994), and by Diehl *et al.* (2001).[4]

The enhancing gestures are presumably introduced when the defining acoustic correlate for a particular contrast is not sufficiently salient. That is, the use of enhancing gestures is driven by perceptual requirements (Diehl, 1991; Keyser and Stevens, 2001). Since the inventory of features (and contrasts) is language dependent, the gestures that may be used for enhancing the perceptual saliency of a feature may be language dependent. It is possible, then, to observe differences in the acoustic manifestation of the same feature in different languages. Such variability is well documented in the literature (cf. Ladefoged, 1980).

The articulatory actions that are automatic consequences of the implementation of particular features include (1) the stiffening of the vocal folds during the production of a high vowel, leading to a higher fundamental frequency for high vowels than for low vowels (House and Fairbanks, 1953; Whalen *et al.*, 1998); (2) the increased duration of low vowels relative to high vowels (House and Fairbanks, 1953); and (3) the different duration of the frication noise burst at the release of different articulators in producing a stop consonant (Cho and Ladefoged, 1999; Hanson and Stevens, 2000), the duration being shortest for labials, longest for velars, and intermediate for tongue-blade consonants. These (and possible other) consequences of particular feature-related gestures are determined by anatomical and physiological factors over which the speaker has little control.

We consider next some examples of active secondary articulatory gestures that are required if the primary acoustic correlate of a feature is to be realized. In each of these examples, the primary feature is an articulator-free feature. (1) Any consonant that is classified as [−sonorant] must be produced with a closed velopharyngeal port, since by definition pressure is built up in the oral cavity for such a consonant. (2) The production of a consonant that is [+continuant] requires significant airflow through the oral constriction, and usually this airflow can only be achieved when the glottal opening is greater than it would normally be for a vowel. (3) A [+strident] consonant (in English) is produced by directing the airstream against the lower incisors. This action requires that the tongue blade be shaped in a way that properly directs the jet of air, and therefore requires that the jaw be raised so that the lower incisors are properly positioned to provide an obstacle for the jet. The contrasting [−strident] consonant requires a tongue blade position and shape that avoids an airstream that impinges on the lower incisors downstream from the constriction. For each of these examples, the secondary articulatory gestures have acoustic consequences that can provide cues to aid the listener in uncovering the feature.

Recruitment of articulators that are not directly specified by the features for a segment may also be motivated by the need to enhance the acoustic and perceptual consequences of one of the features of the segment. Some of these enhancement actions are reviewed in Keyser and Stevens (2001). We restrict our discussion here to two examples where the enhancing gesture creates not only a strengthened primary acoustic cue for the feature, but also introduces additional acoustic properties or perceptual cues that can contribute to a listener's estimation of the feature.

There are two primary acoustic correlates of the feature [+stiff vocal folds]. During the consonantal interval (while there is a buildup of intraoral pressure), a segment with the feature [+stiff vocal folds] shows essentially no glottal vibration during the constricted interval for the consonant. The contrasting segment with [−stiff vocal folds] does exhibit glottal vibration at some time during the constricted interval for the consonant. In the initial part of the vowel following the consonant the fundamental frequency is higher for a [+stiff vocal folds] segment than for the [−stiff vocal folds] cognate, reflecting the carryover of vocal-fold stiffness into the following vowel. Several types of gestures are used to enhance the feature [stiff vocal folds] depending on the syllable position of the consonant. In syllable-initial position for a stop consonant before a stressed vowel, aspiration is introduced by maintaining a spread configuration for the glottis for a few tens of milliseconds following the consonant release, leading to a delay in the onset of glottal vibration following the consonant release. This action presumably increases the separation between frication noise at the consonant release and the onset of glottal vibration, and hence enhances the voiceless character of the consonant. The vowel preceding a syllable-final voiceless consonant is often shortened relative to its duration preceding a voiced consonant, particularly if the syllable is phrase-final, thereby reducing the amount of glottal vibration in the syllable and enhancing the perception of the feature [+stiff vocal folds]. Also, when a voiceless stop consonant is in syllable-final position in certain phonetic contexts (particularly for an alveolar consonant), the vocal folds are often adducted to form a glottal closure, leading to an abrupt termination of glottal vibration. This glottalization enhances the voiceless character of the consonant in this syllable-final position (Keyser and Stevens, 2001).

Another articulatory action that can be interpreted as an enhancing gesture is the positioning of the tongue body for a tongue-blade stop consonant. In the case of a [+anterior]

stop consonant in English, for example, the tongue body is adjusted to a somewhat fronted position, presumably to assist in positioning the tongue blade constriction and hence to enhance the contrast with labial consonants, as seen in the spectrum shape of the burst. This tongue-body gesture is reflected in the formant movements in the following vowel. The fronted tongue-body position also leads to formant transitions that are different from those of labial consonants, and therefore contributes to the perceptual saliency of the alveolar consonant.

It is evident, then, that several acoustic cues in combination can lead to identification of a feature and hence of the word for which the feature constitutes a component. In running speech, each of these cues can be present with various degrees of strength, depending on several factors such as speaking style, the syllable position of the segment, and the phonetic and prosodic environment in which the feature occurs. In some environments and for some styles of speaking, all of the cues for a particular feature might be strongly present, and identification of the feature is robust and reliable. In other situations, including speech in noise or in the presence of other distorting influences (such as the degree of casualness), some of the cues may be weak or absent. For example, in rapid speech, the articulatory movements for two adjacent segments may overlap in a way that obscures the acoustic consequences of some of the movements (Browman and Goldstein, 1990). In such cases, identification of the feature from the available cues may be unreliable and may depend strongly on knowledge of the context in which the feature occurs. (See Sec. IV C.) Situations can arise, for example, in which the defining acoustic cues for a feature in some contexts are not available, but cues associated with enhancing gestures remain.

We turn now to a discussion of cases where (1) feature identification is robust, and (2) cues may be severely weakened.

B. Word-initial segments and prominent syllables

There is some evidence that features for consonants in word-initial position exhibit a stronger set of acoustic cues than consonants in other positions (Cutler and Carter, 1987; Manuel, 1991; Gow *et al.*, 1996). This statement is especially true when the word is not a function word, in which the vowel can be reduced. For the most part, these initial consonants are adjacent to vowels, or at least they usually precede vowels, glides, or liquids. Thus there is an opportunity for cues to be present both in the interval preceding the release when the consonant constriction is in place, and in the sonorant or vocalic interval immediately following the release. These cues tend to be modified minimally by a segment at the end of a preceding word. It is not uncommon for some of the cues for features in a word-initial segment to spread their influence to regions of the sound that might normally be associated with the final segment in a preceding word (cf. Zsiga, 1994). In some sense, then, this spreading of cues enhances the identification of the features for the word-initial consonant. Thus, for example, in a sequence like ''his five sisters,'' voicing in the segment /z/ in his or /v/ in five may be only weakly represented near the time of closure for the fricative, because of the influence of the word-initial voiceless /f/ or /s/, respectively. Or, to put it another way, the voicelessness of word-initial /f/ and /s/ is strengthened.

There are, however, exceptions to this word-initial robustness principle. For example, some of the acoustic characteristics of word-initial /ð/ can be influenced by a preceding word-final consonant, so that /ð/ may appear to have the characteristics of a noncontinuant or sonorant consonant (Manuel, 1995). These effects can be observed in sequences like ''win those cups,'' or ''at those classes,'' where there may be little direct acoustic evidence for the features [+continuant] and [−sonorant] which are normally associated with /ð/. It is noted, however, that the features [+anterior, +distributed] (i.e., the place features for /ð/) appear to be represented robustly in the signal independent of the context.[5] Another exception is word-initial /h/ when it occurs before a reduced vowel, as in ''will he go.'' In casual speech, there may be little or no evidence for /h/ in this phonetic environment.

When there is a word-initial consonant cluster, the same robustness principle applies to the consonant that is immediately adjacent to the vowel. Since there are constraints on the features of the consonant or consonants preceding this vowel-adjacent consonant, the features for these other components of the initial cluster can also be identified reliably (cf. Fujimura, 1997).

In running speech, some syllables are produced with greater prominence than others. This greater prominence is manifested in the sound by increased amplitude, increased vowel duration, and increased duration of the initial consonant. In such a prominent or accented syllable, the cues for the features of the initial consonant or consonants are present in the sound with greater strength and possibly with greater number than are the cues for consonants in other environments. For example, voiceless stop consonants are aspirated in this position, thereby enhancing identification of the feature [+stiff vocal folds], as noted above. And the increased vowel and consonant durations permit a clearer representation of consonantal place of articulation in the vicinity of the consonant release, with minimal influence of vowels and consonants other than the immediately following vowel. The increased vowel duration for an accented syllable also reduces the influence of adjacent consonants and vowels on the formant frequencies near the middle of the vowel. Consequently, the cues for place of articulation for the vowel as well as for the initial consonant are more robust, and the vowel features can be estimated with minimal reliance on contextual information.

It frequently happens, of course, that accented syllables are also word-initial syllables (Cutler and Carter, 1987). In this case, there is more than one reason why the features for initial consonants are represented by robust cues in the sound.

C. Sources of weakening or modification of cues for features

We have just shown that there is a set of environments for consonants and vowels in which cues for the distinctive features tend to be robust. Cues for certain features of seg-

ments in other contexts are subject to weakening or to elimination in running speech. This modification of cues for features of a particular segment generally arises because of the influence of adjacent segments. The articulatory movements needed to implement an adjacent segment may prevent some of the cues for the features from appearing in the signal, or may weaken the cues because of overlap of these movements with those required for the features (Browman and Goldstein, 1990). We discuss here a few examples of such cases in English.

In running speech, it is common to classify the vowels into three categories: accented vowels, full vowels but not accented, and reduced vowels. For example, in the word potato, the first vowel is normally reduced, the second vowel is accented, and the third vowel is a full vowel but nonaccented. Reduced vowels are inherently produced with weakened cues for place features. It is expected that when a vowel is reduced, it is sufficient to specify simply that the vowel is present, with no identification of place features. The more difficult issue is to determine the presence of a reduced vowel. It is normally expected that each vowel or syllabic nucleus in running speech is characterized by a peak in low-frequency amplitude in the waveform, and this peak defines a landmark for the vowel. This peak may be small, however, in the case of a reduced vowel.

Some phonetic environments in which evidence for a reduced vowel may not appear as a separate low-frequency peak in amplitude of the oral acoustic output are listed as follows:

(1) If a reduced vowel immediately follows another vowel, without an intervening consonant, the presence of the vowel may not be manifested as a low-frequency peak separate from the peak for the preceding vowel. An example is the sequence saw a dog where the sequence /ɑə/ usually gives rise to only a single peak in low-frequency amplitude, with the peak occurring at the time the mouth opening is largest. The lack of a separate low-frequency amplitude peak can also be observed for a sequence of two vowels when neither vowel is reduced. In an utterance like ''he saw eight dogs,'' a separate vowel landmark may not be observed in the sequence /ɑe/. This merging of two vowel landmarks into one cannot occur when the vowel is [−tense], since such vowels must be followed by consonants.

(2) When a reduced vowel is surrounded by voiceless obstruent consonants, the glottal spreading and vocal-fold stiffening that accompanies the consonants could spread into the vowel, and the vowel would then become voiceless. Examples of this kind of consonant reduction are sometimes observed in words such as potato and classical.

(3) When a nasal consonant follows a reduced vowel, there are some phonetic environments in which the vowel–consonant sequence reduces to a syllabic nasal. The words button or lesson are examples where such a reduction can occur. In the sequence /ən/ in these examples, the nasalization of the vowel preceding the nasal consonant extends back over the entire length of the vowel and into the end of the closure for the preceding consonant. This preceding consonant has the same place of articulation as /n/, so that the consonant is terminated by the opening of the velopharyngeal port and is followed immediately by the syllabic nasal. A similar merging can occur when a reduced vowel is followed by a liquid, in which case the peak in low-frequency amplitude occurs within the syllabic /r/ or /l/. The syllabic /r/ could in some cases be considered as a vowel segment (rather than being derived from a sequence /ər/), but in other cases it is a reduction, as in an utterance like come for dinner, where /ər/ in for reduces to syllabic /r/. Similar comments could be made about /l/, for example in the word legal. In all of these cases, it is usually possible to detect a low-frequency amplitude peak or landmark indicating the presence of the syllabic liquid or nasal, and hence to detect that a syllable is present, but there is no direct evidence on the surface for a sequence of a vowel and a consonant.

In reduced vowels, the formant frequencies can be influenced strongly by the consonant context and by the phonetic characteristics of vowels in adjacent syllables. Some influence of context can also be seen in vowels that are not reduced. This influence can be sufficiently strong that estimation of the features underlying a nonreduced vowel must take into account the consonantal context as well as the formant frequencies within the vowel.

As has been observed, the acoustic cues for a consonant are most robust when the consonant is in word-initial position preceding a stressed vowel. For consonants in a number of other phonetic environments, the inventory of cues for the various consonantal features is often more sparse. That is, in some other environments there is less opportunity to generate acoustic cues for the consonant in the vicinity of the times of closure and release for the consonant. The articulatory maneuvers that are specified by the features for the consonant are implemented, but the gestures for nearby segments prevent the acoustic cues for the consonantal features from appearing in the signal. This effect of overlapping gestures is more prevalent when there is an increased speaking rate.

One common omission of a cue for a consonant is an acoustic record of both closing and opening movements when a consonant occurs in a sequence of two or more consonants. For example, in a sequence like up to, the closure for /t/ often occurs before the opening movement for /p/, so that neither the labial release nor the alveolar closure creates an acoustic landmark. There is an articulatory opening and closing gesture for each consonant, but only two landmarks are evident in the sound. It may also happen that the overlap in gestures for the two consonants in a sequence is sufficiently great that the acoustic manifestations for some features of the first consonant are influenced by the second consonant. Thus there can be a weakening of cues for certain features of the first consonant. Examples are the weakening of place cues for a tongue-blade consonant that is followed by a consonant with a different place of articulation (e.g., the sequence note closely) or the weakening of acoustic cues for

voicing of the first consonant in a sequence like his cake (Gow, 2002).

In some versions of words like button or cotton, a cue for the presence of the /t/ following the first vowel may be the glottalization that terminates the vowel. The actual alveolar closure may occur after this glottalization, and consequently there is no direct acoustic evidence of this closure event. The underlying features for the stop consonant (i.e., [−continuant, −sonorant, +tongue blade, +anterior, +stiff vocal folds]) must be inferred from the sequence of three events: the formant movements in the preceding vowel, the glottalization, and the abrupt onset of syllabic /n/ resulting from opening of the velopharyngeal port. All of the gestures for the segments and features are implemented, and sufficient acoustic cues are available to reconstruct the sequence of gestures and hence the segments and features.

Another modification of the acoustic cues for a consonant is the flapping of a coronal stop consonant in certain phonetic environments (Zue and Laferriere, 1979). Examples are in the sequences writer, rider, atom, and bad apple. In these examples, tongue blade contact is made on the hard palate, as specified for the features of an alveolar stop consonant, but the closure and release are so close together that the corresponding acoustic landmarks essentially merge. This reduction in closure time for the consonant may interfere with the implementation of cues for the voicing feature.

These various modifications of the landmarks and acoustic cues for the features of vowels and consonants in English have implications for the lexical access process. It is of some significance that many if not all of the modifications of landmarks and of cues for features are a consequence of influences from other articulatory gestures associated with nearby segments. Usually the articulatory gestures specified by the inventory of features for a segment, including gestures introduced to provide enhancement, are in fact implemented, but the gestural context or the rapidity of the gestures may influence the acoustic end result, leading to reduction in the strength of the cues or of the landmarks or to elimination of acoustic cues or landmarks. Uncovering of the segments and features that underlie the words in an utterance, then, involves using the acoustic data to make inferences about the gestures that the speaker uses to implement these features, since the gestures tend to bear a closer relation to the features than do the acoustic patterns.

V. DERIVING A SEGMENT- AND FEATURE-BASED REPRESENTATION FROM THE SPEECH SIGNAL

A. Introduction: Steps in the derivation

When presented with an utterance, the task of the human listener, or of a speech recognizer that simulates the human listener, is to derive a discrete or symbolic representation similar to that in Table V through appropriate analysis of the acoustic signal, and, ultimately, to derive the sequence of words. In the proposed model, analysis of the acoustic signal leads to extraction of cues that can be interpreted in terms of articulatory movements. From these cues, the segments and features are estimated. We outline in this section the initial process of deriving from the acoustic signal a representation

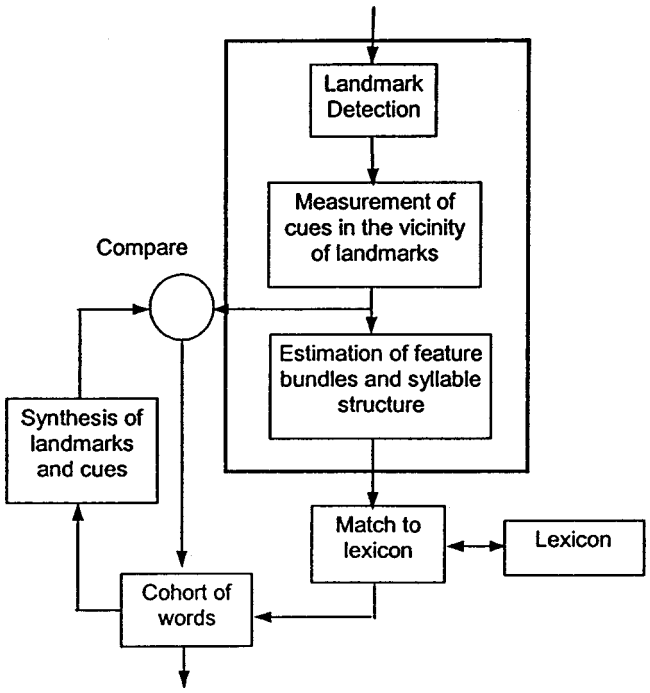


FIG. 2. Block diagram of model of human lexical access. The three components in the box marked by heavy lines are the principal concern of the proposed research (from Stevens, 2000b).

based on segments, features, and syllabic structure, and we illustrate how this derivation proceeds with some examples. Three steps are involved in this process. These steps are schematized by components within the box represented by bold lines in the block diagram in Fig. 2.

In the first step, described in Sec. V B, the locations and types of the basic acoustic landmarks in the signal are established. These acoustic landmarks are identified by the locations of low-frequency energy peaks, energy minima with no acoustic discontinuities, and particular types of abrupt acoustic events. From these acoustic landmarks certain articulatory events can be hypothesized: the speaker produced a maximum opening in the oral cavity, or a minimum opening without an abrupt acoustic discontinuity, or a narrowing in the oral cavity sufficient to create several types of acoustic discontinuity. Estimates of the articulatory-free features are made based on these hypotheses. The second step (Sec. V C) consists of the extraction of acoustic cues from the signal in the vicinity of the landmarks. These cues are derived by first measuring the time course of certain acoustic parameters such as the frequencies of spectral peaks or spectrum amplitudes in particular frequency ranges, and then specifying particular attributes of these parameter tracks. The selection of the acoustic cues is guided by a requirement that they be directly related to the movements or states of various articulators in this time region in the vicinity of the landmarks. The output of this component is a sequence of landmarks, labeled with times and with preliminary estimates of articulator-free features at each landmark, and with the values of a set of acoustic cues attached. Within this component, estimates are also made of the syllable affiliation of the consonants, to the extent that this information is revealed in the acoustic signal. As a part of this second step, the parameters that are derived

from the signal near the landmarks are also examined to determine whether this pattern is consistent with the landmark sequence derived in the initial step of the model. Based on this analysis, new landmarks might be hypothesized, or it might be recognized that some landmarks do not signal the presence of a segment.

In the third step (Sec. V D), the landmarks and acoustic cues derived in the signal processing stage in Fig. 2 are consolidated, and a sequence of feature bundles is derived, as shown by the third block in the figure. The acoustic cues are combined in a way that leads to estimates of the articulator-bound features corresponding to each of the landmarks identified in the first step. A further task in this step is to convert the landmark representation derived in the first step into a representation in terms of underlying segments. For most landmarks there is one-to-one conversion of landmarks to segments for vowels and glides. For consonants, there are two abrupt landmarks (corresponding to an articulatory closure and release). These two landmarks must be converted to a single consonant segment. At this level, the derived segments, the features, and their combinations must be consistent with the patterns that are present in syllables in words in the lexicon, although acoustic evidence for some segments and features may be missing due to casual speaking or external interference, for example, by noise.

The part of the model outside the highlighted box in Fig. 2 looks ahead to the problem of using the estimated segments and features for an utterance, together with the lexicon, to propose a sequence of words. There could be more than one hypothesis concerning the words or word sequences, particularly since there may be a low confidence in the estimates for some features. A final step in the lexical access process would be to test whether each hypothesized sequence is consistent with the original acoustic pattern. This testing involves an internal synthesis of the main acoustic landmarks and cues that would be exhibited by each hypothesized sequence. In particular, the possible sequences of landmarks, together with acoustic parameters around these landmarks, are synthesized. These internally generated landmarks and parameters are matched against the observed landmarks and parameters, and the sequence that gives the best match is selected. The synthesis of the acoustic landmarks and parameters from a hypothesized word or word sequence can be done with full knowledge of the context in which the sequence occurs, and consequently more information is available to help in estimating the landmarks and parameters. In the bottom-up analysis that leads to the hypothesized cohort, this context may not be available, and consequently the feature estimates may be less robust. The process of deriving potential word sequences and the details of the final analysis-by-synthesis step (Stevens and Halle, 1967) are not developed in this article. The derivation of word sequences and their meaning from the acoustic signal also utilizes linguistic knowledge at a level higher than the phonological representation of lexical items. The role of these higher-level sources of linguistic knowledge is not addressed here. The principal concern of this article is the operations in the blocks within the rectangular box in Fig. 2.

B. Detection of acoustic landmarks and estimation of articulator-free features

The initial step in the analysis is to determine the locations of the vowel, glide, and consonant landmarks. Roughly speaking, the vowel landmarks are the places in the signal where there is a maximum in amplitude in a frequency band in the range of the first formant frequency. These places are usually where the first formant frequency is maximally high, and correspond roughly to the times when the oral cavity is maximally open during a syllable. Additional cues for landmark locations come from the fact that there are constraints on the locations of these vowel landmarks. For example, the time between the vowel landmarks is rarely shorter than 120 ms, and, except when there is a pause, is rarely longer than about 350 ms in running speech. Thus there are certain rhythmic aspects of running speech that place some limits on the spacing of the vowel landmarks. The relative amplitude of the maximum for a vowel landmark can be useful for estimating the degree of prominence of the vowel. Examples of the amplitude changes within a vowel have been shown in Fig. 1(a). Procedures for estimating vowel landmarks have been developed and evaluated by Howitt (2000). Earlier work on the detection of syllable nuclei has been reported by Mermelstein (1975).

It has already been observed that any method based on detection of maxima of parameters like amplitude and first-formant frequency will not always generate a landmark for each syllabic nucleus in the underlying representation of running speech. For certain vowel sequences, particularly when one vowel is reduced, only one landmark may be detected. Further analysis of acoustic parameters within the sequence is needed to uncover the presence of more than one vowel. This analysis examines the formant movements primarily of $F1$ and $F2$ on either side of the putative landmark to determine whether the direction and magnitude of these movements are sufficient to suggest the presence of an adjacent vowel. (See Sec. V C.)

The consonantal landmarks are places in the sound where particular types of spectral discontinuities occur. These discontinuities are the result of consonantal closure or release movements by one of the three articulators: lips, tongue blade, and tongue body. The constrictions that give rise to these landmarks are always formed in the oral cavity. A landmark caused by forming a consonantal constriction in the oral cavity has certain types of spectral change that result from rapid changes in the cross-sectional area of the constriction near the landmark.

Other types of acoustic discontinuities can occur as a consequence of closings or openings of the soft palate or of laryngeal or pharyngeal movements. These discontinuities have a different acoustic character, and do not qualify as consonantal landmarks in the sense defined here, since they are not produced by a closure or release of one of the oral articulators and therefore do not mark the occurrence of a consonant segment. Further analysis of the signal in the vicinity of the preliminary estimates of consonantal landmark locations is required to separate the bona fide landmarks from these other discontinuities. This analysis examines the formant movements on the vocalic side of each discontinuity.

If these movements are small and/or slow, it is assumed that the discontinuity is not produced by the formation of a closure or release by an oral articulator. The existence of these discontinuities can, however, provide cues for some of the features and for the existence of certain segments. At this stage in the analysis, then, it is assumed that the presence of these ''nonconsonantal'' discontinuities is marked.[6]

Examples of the locations of the consonantal landmarks are shown at the top of the spectrogram in Fig. 1(a). For certain consonant sequences, an acoustic landmark may not be manifested in the sound for every consonant release and closure, particularly when two stop consonants with different places of articulation occur in sequence.

Each consonantal landmark can readily be further categorized in terms of the articulator-free features that underlie the consonant—the features [sonorant], [continuant], and [strident]. The decision concerning the feature [sonorant] is based largely on the presence or absence of strong vocal-fold vibration on both sides of the landmark, since there is no increase in intraoral pressure, and the pressure across the glottis is the full subglottal pressure. One landmark for a sonorant consonant will always be adjacent to a vowel or a glide. In the case of a landmark produced by a nonsonorant consonant, the consonant is identified as [−continuant] or as [+continuant] depending on whether frication noise is continuous on one side of the landmark. For consonants identified as [+continuant], the main cue for the feature [strident] is the high-frequency amplitude of the frication noise in relation to an adjacent vowel. Some initial progress in developing algorithms for detecting the consonant landmarks and for estimating the associated articulator-free features has been reported by Liu (1996). Extensions of this work for nasal consonants have been reported by Chen (2000).

A glide can occur only immediately adjacent to a vowel in English. Thus if a glide landmark is to be detected it will always occur adjacent to a vowel landmark, with no intervening consonant landmark. In prevocalic position, a glide is usually characterized by a narrowing of the airway in the oral cavity, and hence by a low first-formant frequency, a reduced low-frequency amplitude in relation to the following vowel, and smooth transitions of formant frequencies and of the amplitudes of the formant peaks between the glide and the vowel. Examples of glide landmarks are labeled on the spectrogram in Fig. 1(b). A preliminary algorithm for locating these glide landmarks has been developed by Sun (1996).

Glides can also be implemented immediately following a vocalic nucleus. Such an ''offglide'' is often seen in diphthongs such as /ɑj/ and /ɑw/, where there is a relatively slow and continuous movement of the first and second formant frequencies after the vocalic nucleus /ɑ/. The landmarks for these offglides are placed toward the end of this formant movement where the narrowing of the airway in the oral cavity is most extreme. At the initial landmark-detecting stage, offglides (and some prevocalic glides) may often not be detected through simple measures of amplitude and first-formant frequency. The presence of these glides must be determined at a later stage when articulator-bound features for the vowel are estimated, based on other parameters that are extracted from the signal.

Estimation of the locations of acoustic landmarks is, for the most part, a process that examines local acoustic properties, and is not generally influenced by more remote acoustic events. However, associating these landmarks with underlying segments and their articulator-free features, and postulating segments not directly signaled by the landmarks, may require additional analysis that examines a broader context. (See Sec. V C.)

C. Toward estimation of acoustic cues relevant to articulator-bound features

The detection of a peak landmark or a valley landmark is usually acoustic evidence that a vowel or a glide occurred in the linguistic representation underlying the speaker's utterance. Acoustic cues that provide evidence for the positions and movements of the articulators that can lead to estimation of the features for a vowel or a glide are expected to be found in the vicinity of these respective landmarks.

The detection of a particular type of abrupt event in the signal is evidence that a closure or narrowing is made by a primary consonant articulator or that a release has been made by such an articulator. Each consonant has an articulatory closure and release (except for consonant sequences produced with the same major articulator). Thus there are two articulatory events when a consonant is produced—a closure and a release—although, as already noted, one of these two events may not always be evident in the signal. Acoustic cues for the articulatory states and movements on which the articulator-bound features for the consonant are based reside in the vicinity of the acoustic landmarks.

The landmarks, then, define the centers or regions in the signal where acoustic parameters are examined, and, based on this examination, cues are extracted. Interpretation of these cues leads to hypotheses as to what the various articulators are doing in the vicinity of the landmarks. We propose here an inventory of acoustic parameters that should be extracted from the signal. The acoustic cues are derived by sampling these parameters (or changes in the parameters) at particular times in the vicinity of landmarks. The parameters fall into ten categories that provide evidence for relevant configurations and movements of the supralaryngeal and laryngeal structures and the state of the respiratory system. Parameters in the first five categories relate to regions of the signal where there is no major constriction in the vocal tract and the sound source is at or near the glottis. We refer to such regions loosely as vocalic regions. Parameters in the next three categories provide information about the supraglottal and laryngeal states during time intervals when there is a consonantal constriction. The ninth category describes parameters that can potentially lead to estimates of changes in subglottal pressure in an utterance, particularly at the initiation and termination of a phrase, and the tenth category is concerned with temporal characteristics based on times at which landmarks are located.

Here is the proposed list of parameters:

(1) Acoustic parameters related to the position of the tongue body and to lip rounding. These parameters are measured when there is an acoustic source at the glottis,

there is no nasalization, and the vocal tract above the glottis is not sufficiently constricted to create a significant increase in supraglottal pressure. In the vicinity of vowel, glide, and consonant landmarks, these parameters are the formant frequencies and their bandwidths (or, equivalently, the relative amplitudes of the formant prominences). Interpretation of these formant parameters in terms of tongue-body positions and lip rounding may depend on nasalization [item (2) below] and on the glottal configuration [item (5)].

(2) Parameters that are related to the presence of a velopharyngeal opening in a vocalic region as in (1). These parameters include the amplitude of the F1 prominence in relation to the spectrum amplitude at low frequencies (200–300 Hz) and in the 1000-Hz range (Hattori *et al.*, 1958; Chen, 1997).

(3) Parameters that describe the spectrum change at times when there is rapid motion of articulators, particularly the lips and the tongue blade. In this same region within which there is an acoustic source at the glottis, there are times that the formant frequencies move very rapidly, particularly near a release or closure for a constant. Cues for the consonant place of articulation reside in these rapid spectral changes, and may be different from those in item (1) above.

(4) The frequency of glottal vibration. This is the principal parameter indicating vocal-fold stiffness when the vocal folds are vibrating. This parameter is present when the vocal folds are vibrating during a vowel, a glide, or a sonorant consonant.

(5) Parameters from which the state of the glottis in a vocalic region can be estimated. These parameters include measures of the low-frequency spectrum shape (such as H1 − H2, where H1 and H2 are the amplitudes of the first two harmonics), the spectrum tilt (such as H1 − A3, where A3 is the amplitude of the third formant prominence), the bandwidth of $F1$ (as inferred from H1 − A1), and a measure of the amount of noise in the spectrum at high frequencies (Klatt and Klatt, 1990; Hanson, 1997). These parameters also indicate whether or not there is glottal vibration and whether aspiration noise is present.

(6) Cues for the place of articulation for an obstruent consonant, as determined from the spectrum shape of the frication noise, and by its amplitude in relation to an adjacent vowel. When there is a raised intraoral pressure, as for an obstruent consonant, frication noise may be generated in the vicinity of the oral constriction, either as a brief burst (for a stop consonant) or as continuous noise (for a fricative).

(7) Parameters relating to the state of the glottis and of the vocal folds within the region when there is an increased supraglottal pressure. Continued vocal-fold vibration throughout the obstruent region is evidence for slackened vocal folds, while lack of vocal-fold vibration is evidence for stiffened vocal folds near consonant release and for spread or constricted glottis near consonant closure following a vowel.

(8) Parameters that help to determine the state of the velopharyngeal opening and place of articulation for a nasal consonant, or, for a liquid, the special state of the tongue blade. These parameters are measured during the nasal murmur or the constricted region for a liquid, where there is a low-frequency first formant.

(9) Parameters providing information concerning the subglottal pressure. These parameters are especially important near the beginning and end of an utterance or adjacent to a pause, where the acoustic cues for certain vowel and consonant articulations are likely to be modulated because of the changing subglottal pressure.

(10) Parameters based on measurements of the times between landmarks or between landmarks and other events such as onset of glottal vibration. These parameters provide information about timing, and these in turn provide cues for certain features of vowels and consonants.

In the vicinity of a landmark defined by a peak in low-frequency amplitude (a vowel landmark), the acoustic analysis must lead to cues derived from parameters of the type in item (1), with additional help from items (2), (4), and (5) and possibly (8) in the case of syllabic nasals and liquids. The same inventory applies to so-called glide landmarks. In the vicinity of an abrupt landmark that signals a consonantal closure or release, there are cues that are derived from parameters measured in the vocalic region adjacent to the landmark as well as in the region on the other side of the landmark where there is a consonantal constriction (Stevens, 2000a). The cues, then, are based on the types described in items (1)–(5) in the vocalic regions, and (6)–(8) in the constricted regions, depending on the articulator-free features. As noted, cues derived from item (9) are invoked at phrase boundaries. Parameters related to timing [item (10)] form the basis for several cues for vowel and consonant place features and to voicing for consonants.

There are a number of acoustic cues that might be extracted from the parameters listed above in order to contribute to identification of the various features in running speech. The details of these acoustic cues and their effectiveness in identifying the features are beyond the scope of this article. We give one illustration of this cue-selection process, however, by listing some of the cues for one class of features—the features that define the place of articulation for stop consonants. These cues include measures of the first two or three formant frequencies and their movements in the vowel adjacent to the consonant landmark (e.g., Kewley-Port, 1982; Sussman *et al.*, 1991; Manuel and Stevens, 1995); the spectrum amplitude of the release burst at high frequencies relative to midfrequencies (e.g., Fant, 1960, 1973; Stevens, 1998); the spectrum amplitude of the burst in particular frequency ranges in relation to the spectrum amplitude of the vowel immediately adjacent to the landmark (e.g., Ohde and Stevens, 1983); and the duration of the frication noise burst (e.g., Lisker and Abramson, 1964; Fant, 1973; Hanson and Stevens, 2000). All of these cues in one way or another define attributes that are closely related either to (1) the location of the constriction along the oral cavity (e.g., formant transitions for velars and labials, burst spectra for all places

of articulation), (2) the identity of the articulator that forms the constriction (e.g., burst duration and rates of movement of the formants), or (3) an enhancing gesture that characterizes the forward movement of the tongue body adjacent to the landmark for an alveolar consonant (e.g., *F*2 and *F*3 movements adjacent to the landmark). A similar catalog of cues for other features could be listed.

Derivation of the time course of parameters in the vicinity of landmarks has two functions other than providing a basis for specifying cues that can be combined to estimate distinctive feature values. One of these functions is to allow preliminary estimates of the syllable affiliation of consonants, and the other is to verify or to modify the initial assignment of articulatory-free features to landmarks.

Although syllable affiliation cannot always be estimated with confidence from acoustic analysis, some cues are available in the signal. One example is the delay in onset of glottal vibration following the release of a stop consonant. This delay, together with the presence of aspiration during the delay interval, can be seen in the acoustic parameters that are extracted following the release of the stop consonant. When these attributes are present in the signal, the stop consonant is in syllable-initial position. On the other hand, if the acoustic parameters provide evidence for glottalization at the closure landmark for a stop consonant, the consonant is probably syllable-final. In a vowel preceding the closure landmark for a nasal consonant, nasalization can be observed, and this nasalization begins a few tens of milliseconds preceding the consonant closure if the consonant is syllable final. When the nasal consonant is affiliated with the following vowel, the extent of nasalization in the preceding vowel is considerably less extensive (Krakow, 1993). Differences in the time course of formant parameters and formant amplitude changes can also be observed in syllable-initial and syllable-final liquid consonants (Sproat and Fujimura, 1993).

Tracking of acoustic parameters related to articulation can also provide evidence either for additional segments not detected by simple landmark estimation or for landmarks that are inadvertently inserted but which do not mark the presence of a vowel, consonant, or glide segment. For example, when a sequence of two vowels shows only one low-frequency peak, and hence only one acoustic landmark, the trajectories of the first two formants over a few tens of milliseconds preceding and following the landmark should be sufficient to determine whether or not a sequence of two vowels gave rise to these trajectories. In this case, a second segment is hypothesized, and a ''pseudo-landmark'' is inserted in the appropriate region.

An example of an inadvertent insertion of a landmark is the production of a glottal stop at the onset of a vowel-initial word (e.g., at the onset of the second word that sometimes occurs in the sequence two apples). A glottal onset here might be detected as a landmark representing a consonant release. Further analysis at the beginning of the following vowel would show, however, the presence of glottalization, coupled with the lack of consonantal formant movements immediately following the onset. This acoustic pattern is not the product of the release of a narrowing formed by an oral articulator, and hence is not a bona fide consonant landmark. However, the presence of the glottal stop should be noted, since it provides a possible cue for word onset—a cue that can be useful at a higher level when lexical candidates are determined.

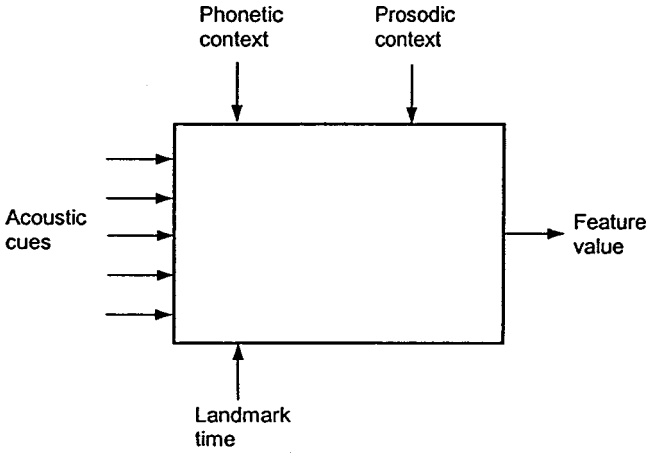


FIG. 3. Schematic representation of a module for estimating an articulator-bound feature (from Stevens, 2000b).

D. Estimating the underlying segments and the corresponding articulator-bound features

Once the sequence of landmarks has been identified, and a set of acoustic cues has been evaluated in the vicinity of each landmark, the next step is to transform this graded landmark/cue pattern into a symbolic or quantal description consisting of a sequence of feature bundles. In the vicinity of each landmark, certain acoustic parameters are extracted depending on the type of landmark. The values of these parameters, or the changes that occur in these parameters, provide acoustic cues that are relevant to estimation of the articulator-bound features for the segments. These cues are combined or weighted to obtain estimates of the features. The particular combination of cues and their weighting usually depends on the context in which a landmark occurs. That is, the value of a particular feature (+ or −) associated with a landmark-derived segment depends not only on what the articulators are doing around the time of the landmark but also on the effect of instructions to the articulators from segments that precede or follow the segment that underlies the landmark.

We propose that the estimation of each feature is carried out in a specialized module, as represented schematically in Fig. 3. The inputs to the module are the acoustic cues that are relevant to the feature, and the output is an estimate of the feature as + or −. In addition, the output gives an estimate of the confidence of this value of the feature. (A simple method for specifying the confidence of a feature estimate would use a two-point scale: if a value of + or − for a feature is estimated with high confidence, this value is marked in the output. If the estimate of the feature value has low confidence, no value is entered for this feature.) The module also has inputs that specify the phonetic and prosodic context of the feature for which the model is designed. It is assumed that modules that perform these types of functions

are present in the listener's brain. It is supposed that such modules exist in some elementary form in the child that is learning language, and that experience helps to fill in details such as additional cues and the effects of context, so that the modules become more elaborated based on continued exposure to language.

There is a module of the type shown in Fig. 3 for each articulator-bound feature or group of features. A list of these modules for English includes (1) place of articulation for consonants, (2) the voicing feature for obstruent consonants, (3) the feature [nasal], (4) identification of liquid consonants, (5) the feature [back] for vowels and glides, (6) the features [high] and [low] for vowels, and (7) the feature [advanced tongue root] (or tenseness) for vowels. It is noted that a module is activated for each consonant landmark, although when a consonant is in intervocalic position there are two landmarks for the same consonant. If the feature estimates derived from the modules for the closure and release landmarks are the same (and if the time between landmarks is appropriate for a single segment rather than a geminate), then these two landmarks are merged, and a single segment is proposed. It is possible that the confidence in the feature estimates for this combined segment is greater than the confidence based on each landmark separately. The gain achieved by combining landmarks has been examined for the voicing feature by Choi (1999).

From the listings of features in Tables III–V, it can be seen that, once the articulator-free features associated a landmark have been established, the number of articulator-bound features that is needed to completely specify a segment is relatively small. For example, in the case of consonants that are [−sonorant, −continuant] (i.e., stop consonants), the task of the modules is simply to specify place of articulation ([lips], [tongue blade], or [tongue body]) and the voicing feature ([stiff vocal folds]). Consonants that are [+sonorant] also require just two or three modules to estimate the articulator-bound features. In the case of vowels three modules are necessary, and for glides the number is even fewer. We are dealing, therefore, with a relatively sparse specification of articulator-bound features, and the number of modules that are called into action for a given landmark with its attendant articulator-free features is generally in the range 2 to 4.

The details of each of these modules, including the selection and design of acoustic cues that form possible inputs for each module, are beyond the scope of this article. Some progress toward implementing the modules for some consonant features is reported elsewhere (Choi, 1999; Stevens *et al.*, 1999; Chen, 2000).

VI. SUMMARY AND DISCUSSION

The central concept of the proposed model of lexical access is the representation of words in memory as sequences of segments, each consisting of a bundle of binary distinctive features. Each word in a language generally has just one such representation in the lexicon. The distinctive features define the contrasts in the language: a change in one feature in one segment can potentially generate a different word. Independent of what language is involved, there is a universal inventory of features that are determined by the properties of the vocal tract as a generator of sounds with perceptually distinctive acoustic properties. Each feature has a defining acoustic and articulatory correlate (although other correlates may also be employed). The features that are used contrastively in a given language are a subset of this universal inventory.

There are two kinds of distinctive features: articulator-free and articulator-bound. Articulator-free features specify classes of articulatory actions but are not tied to particular articulators. They give rise to several types of acoustic landmarks that indicate the presence of segments, and establish regions in the signal where acoustic evidence for the articulator-bound features can be found. Articulator-bound features specify which articulators are involved in producing the landmarks, and how these articulators are positioned and shaped. When the articulator-free features have been established, the number of articulator-bound features that are needed to fill out the bundles of segments is relatively sparse (three to four features, on average).

When a particular articulator-bound feature is specified in a segment within a word that a speaker is producing, variability can occur in the acoustic pattern for that feature. This variability arises for at least two reasons. One type of variability occurs because, in addition to the primary articulatory action specified by the feature, the speaker recruits additional articulatory actions to enhance the perceptual contrast defined by the feature when the segment is in a particular phonetic or prosodic environment. These enhancing gestures may not only strengthen the perceptual salience of the defining acoustic pattern for the feature, but may also introduce additional acoustic properties that can provide a listener with further acoustic cues for the presence of the feature. There may also be other articulatory and acoustic attributes that are biomechanical consequences of implementation of the feature, and that provide additional enhancement to the defining acoustic correlates of the feature. Thus the enhancing gestures can contribute a number of acoustic cues to the identification of a feature depending on the syllable position and the prosodic environment in which the segment occurs. A second type of variability occurs because of overlap in the articulatory gestures that are involved in producing adjacent segments, causing a weakening or obscuring of some of the acoustic cues that would otherwise help to uncover the underlying features.

Estimation of the features from acoustic processing of the speech signal is often straightforward. The defining acoustic properties for the features, together with some additional cues, are evident in the signal, and the effects of overlapping of gestures for nearby segments are minimal. Frequently, however, this is not the case. The listener must be aware of cues for both the defining and the enhancing gestures, and must be able to account for the fact that some of these acoustic cues may be weakened or eliminated due to gestural overlap. From a processing point of view, the inventory of acoustic cues for a feature must be selected to reflect the articulatory actions that created the acoustic pattern, since the enhancements and the overlapping are best defined in terms of articulatory gestures. Once the cues for an

articulator-bound feature have been extracted, these cues must be weighted or combined in some way to yield an estimate of the feature. These combinations of cues must be learned by a speaker of the language.

The lexical-access model described here has several attributes that differ somewhat from those of other models, particularly models that are more data driven. Three of these attributes are listed here.

(1) There is generally only one representation of each word in the lexicon, and that representation is in terms of bundles of distinctive features. For a given segment, the output of the model is required to postulate a pattern of distinctive features consistent with the patterns that are used in the lexicon. While variability can lead to many ways of producing words, these multiple representations do not appear in the lexicon. Other models tend to deal with variability by proposing several possible pronunciations for words, these pronunciations being specified in terms of acoustic units called phones or phonemelike units (Zue *et al.*, 1990; Rabiner and Juang, 1993; O'Shaughnessy, 2000).
(2) The analysis of the signal proceeds by identifying a sequence of landmarks, which provide evidence for the sequence of underlying segments. The concept of an acoustic segmentation in which sequences of pieces of the waveform are labeled with phones is not consistent with the proposed model. It is suggested that labeling of an utterance be done at least in terms of landmarks, since it is proposed in the model that landmark identification is a critical first step.
(3) The selection of acoustic cues in the model is based on the view that variability in articulation is guided by a few principles, and is rather constrained. These acoustic cues are designed to provide information about relevant articulatory states and movements. A spectral representation of the signal based on acoustic patterns that are not specifically oriented to the estimation of articulation is expected to show substantially greater variability than one that is oriented to this goal. Use of such an acoustic representation requires considerable training from a database of utterances in order to describe variability that exists in the acoustic patterns.

In the model proposed here, words in running speech are accessed by assuming a mental representation of words in terms of segments and features, and identifying the words through an analysis that uncovers the segments in the word and the features that define the segments. Such a view of lexical access has not been universally accepted (e.g., Klatt, 1979). A major cause for the skepticism for the proposed approach is the apparent variability in the acoustic manifestation of a feature or of a word, and hence the apparent lack of correspondence between acoustic parameters and the distinctive features. As we begin to acquire an understanding of the sources of variability in the acoustic manifestation of a segment, it is hoped that the link between acoustic patterns and features will become more transparent, and that principles governing variability can be better defined.

Much further research is needed to fill out the details of the proposed model. For example, while some illustrations of the use of enhancing gestures have been given in Keyser and Stevens (2001), a more thorough study of enhancement both in English and across a variety of languages is needed. Some principles governing the selection of acoustic parameters which must be involved in feature identification have been described here, but it will be necessary to specify in more detail the acoustic cues that are to be derived from these parameters and how these cues should be combined in a variety of listening situations, particularly in noise.

ACKNOWLEDGMENTS

Development of the ideas in this lexical-access model was strongly influenced by discussions with a number of colleagues and students. Acknowledged with special thanks are the contributions of Stefanie Shattuck-Hufnagel, Sharon Manuel, Jay Keyser, Morris Halle, Marilyn Chen, Elizabeth Choi, and Wil Howitt. This research was supported in part by Grant No. DC02978 from the National Institutes of Health.

[1]Throughout this article we will frequently use the terms ''segment'' and ''feature.'' These terms refer strictly to the abstract units in terms of which words are represented in the lexicon of a speaker/listener. The term ''segment'' does not refer directly to a portion of the speech waveform and does not have temporal characteristics. A ''feature'' is a linguistic entity, and does not refer directly to an attribute of the acoustic signal. Landmarks provide evidence for underlying segments, and acoustic properties in the vicinity of landmarks provide cues for the features of these segments.

[2]The feature [advanced tongue root] is used here to distinguish between vowels such as /i ɪ/ or /e ɛ/. In some formulations, a feature [tense] is used to capture this distinction: /i/ is [+tense] and /ɪ/ is [−tense].

[3]The feature [stiff vocal folds] is related to the feature fortis/lenis as described by Kohler (1984), but is defined here somewhat differently.

[4]The role of perceptual distinctiveness in shaping the acoustic attributes of segment inventories in language has also been persuasively argued by Lindblom (1990) and by Liljencrants and Lindblom (1972). The discussion of enhancement theory in the present article and in Keyser and Stevens (2001) has attempted to draw on this concept of perceptual saliency while retaining (in somewhat modified form) the traditional view of a universal inventory of distinctive features based on primary articulatory and acoustic correlates.

[5]Strictly speaking, it is not necessary to use the feature [distributed] for /ð/ in English, since /ð/ is distinguished from /s/ by the feature [strident], as noted earlier. The shaping of the tongue blade for /ð/ and /θ/, with a concomitant adjustment of the tongue body to a more backed position than for /s/, can be considered as a gesture that prevents the airstream from impinging on the lower incisors, and hence enhances the distinction between the [+strident] and [−strident] fricatives.

[6]The use of acoustic discontinuities of both types (i.e., those formed by constricting the vocal tract with an oral articulator and those formed in other ways) has been developed by Glass (1988) as an effective way of segmenting the acoustic stream into a sequence of units. This type of acoustic segmentation has been incorporated into the front end of an existing automatic speech recognition system (Zue *et al.*, 1990).

Browman, C. P., and Goldstein, L. (**1990**). ''Tiers in articulatory phonology,'' in *Papers in Articulatory Phonology I: Between the Grammar and Physics of Speech*, edited by J. Kingston and M. E. Beckman (Cambridge U.P., Cambridge), pp. 341–376.

Chen, M. Y. (**1997**). ''Acoustic correlates of English and French nasalized vowels,'' J. Acoust. Soc. Am. **102**, 2360–2370.

Chen, M. Y. (**2000**). ''Nasal detection module for a knowledge-based speech recognition system,'' in Proceedings 6th International Conference on Spoken Language Processing (ICSLP 2000), Vol. IV, pp. 636–639, Beijing, China.

Chistovich, L. A., and Lublinskaya, V. V. (**1979**). ''The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli,'' Hear. Res. **1**, 185–195.
Cho, T., and Ladefoged, P. (**1999**). ''Variation and universals in VOT: evidence from 18 languages,'' J. Phonetics **27**, 207–229.
Choi, J-Y. (**1999**). ''Detection of consonant voicing: A module for a hierarchical speech recognition system,'' Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.
Chomsky, N., and Halle, M. (**1968**). *The Sound Pattern of English* (Harper and Row, New York).
Clements, G. N. (**1985**). ''The geometry of phonological features,'' Phonology Yearbook **2**, 225–252.
Clements, G. N., and Keyser, S. J. (**1983**). *CV Phonology* (MIT, Cambridge, MA).
Cranen, B., and Boves, L. (**1987**). ''On subglottal formant analysis,'' J. Acoust. Soc. Am. **81**, 734–746.
Cutler, A., and Carter, D. M. (**1987**). ''The predominance of strong initial syllables in the English language,'' Comput. Speech Lang. **2**, 133–142.
Delgutte, B., and Kiang, N. Y. S. (**1984**). ''Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics,'' J. Acoust. Soc. Am. **75**, 897–907.
Diehl, R. L. (**1991**). ''The role of phonetics within the study of language,'' Phonetica **48**, 120–134.
Diehl, R. L., Molis, M. R., and Castleman, W. A. (**2001**). ''Adaptive design of sound systems,'' in *The Role of Speech Perception in Phonology*, edited by E. Hume and K. Johnson (Academic, San Diego), pp. 123–139.
Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague).
Fant, G. (**1973**). *Speech Sounds and Features* (MIT, Cambridge, MA).
Fujimura, O. (**1997**). ''Syllable features and the temporal structure of speech,'' in *Proc. of LP'96: Typology: Prototypes, Item Orderings and Universals*, edited by B. Palek (Charles U.P., Prague), pp. 53–93.
Glass, J. R. (**1988**). ''Finding acoustic regularities in speech: Applications to phonetic recognition,'' Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA.
Gow, Jr., D. W. (**2001**). ''Assimilation and anticipation in continuous spoken word recognition,'' J. Memory Lang. **45**, 133–159.
Gow, Jr., D. W. (**2002**). ''Does English coronal place assimilation create lexical ambiguity?'' J. Exp. Psychol. Hum. Percept. Perform. **28**, 163–179.
Gow, D., Melvold, J., and Manuel, S. Y. (**1996**). ''How word onsets drive lexical access and segmentation: Evidence from acoustics, phonology, and processing,'' in *Proc. 1996 International Conference on Spoken Language Processing* (University of Delaware and Alfred I. duPont Institute, Philadelphia, PA), pp. 66–69.
Halle, M. (**1992**). ''Features,'' in *Oxford International Encyclopedia of Linguistics*, edited by W. Bright (Oxford U.P., New York).
Halle, M. and Stevens, K. N. (**1971**). ''A note on laryngeal features,'' MIT Research Laboratory of Electronics Quarterly Progress Report 101, pp. 198–213.
Halle, M. and Stevens, K. N. (**1991**). ''Knowledge of language and the sounds of speech,'' in *Music, Language, Speech and Brain*, edited by J. Sundberg, L. Nord, and R. Carlson (MacMillan, London), pp. 1–19.
Hanson, H. M. (**1997**). ''Glottal characteristics of female speakers: Acoustic correlates,'' J. Acoust. Soc. Am. **101**, 466–481.
Hanson, H. M., and Stevens, K. N. (**2000**). ''Modeling stop-consonant releases for synthesis,'' J. Acoust. Soc. Am. **107**, 2907(A).
Hattori, S., Yamamoto, K., and Fujimura, O. (**1958**). ''Nasalization of vowels in relation to nasals,'' J. Acoust. Soc. Am. **30**, 267–274.
House, A. S., and Fairbanks, G. (**1953**). ''The influence of consonantal environment upon the secondary acoustical characteristics of vowels,'' J. Acoust. Soc. Am. **25**, 105–113.
Howitt, A. W. (**2000**). ''Automatic syllable detection for vowel landmarks,'' Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA.
Kewley-Port, D. (**1982**). ''Measurement of formant transitions in naturally produced stop consonant-vowel syllables,'' J. Acoust. Soc. Am. **72**, 379–389.
Keyser, S. J., and Stevens, K. N. (**1994**). ''Feature geometry and the vocal tract,'' Phonology **11**, 207–236.
Keyser, S. J., and Stevens, K. N. (**2001**). ''Enhancement revisited,'' in *Ken Hale: A Life in Language*, edited by M. Kenstowicz (MIT, Cambridge, MA), pp. 271–291.
Kingston, J., and Diehl, R. L. (**1994**). ''Phonetic knowledge,'' Language **70**, 419–454.
Klatt, D. H. (**1979**). ''Speech perception: A model of acoustic-phonetic analysis and lexical access,'' J. Phonetics **7**, 279–312.
Klatt, D. H., and Klatt, L. C. (**1990**). ''Analysis, synthesis, and perception of voice quality variations among female and male talkers,'' J. Acoust. Soc. Am. **87**, 820–857.
Kohler, K. J. (**1984**). ''Phonetic explanation in phonology: The feature fortis/lenis,'' Phonetica **41**, 150–174.
Krakow, R. A. (**1993**). ''Nonsegmental influences on velum movement patterns: Syllables, sentences, stress, and speaking rate,'' in *Phonetics and Phonology: Nasals, Nasalization, and the Velum*, edited by M. K. Huffman and R. A. Krakow (Academic, San Diego), pp. 87–116.
Ladefoged, P. (**1980**). ''What are linguistic sounds made of?'' Language **65**, 485–502.
Levelt, W. J. M. (**1989**). *Speaking* (MIT, Cambridge, MA).
Levelt, W. J. M., Roeloff, A., and Meyer, A. (**1999**). ''A theory of lexical access in speech production,'' Brain Behav. Sci. **22**, 1–75.
Liljencrants, J., and Lindblom, B. (**1972**). ''Numerical simulation of vowel quality systems: The role of perceptual contrast,'' Language **48**, 839–862.
Lindblom, B. (**1990**). ''Explaining phonetic variation: A sketch of the H & H theory,'' in *Speech Production and Speech Modeling*, edited by W. J. Hardcastle and A. Marchal (Kluwer, Dordrecht), pp. 403–439.
Lisker, L., and Abramson, A. S. (**1964**). ''A cross-language study of voicing in initial stops: Acoustical measurements,'' Word **20**, 527–565.
Liu, S. A. (**1996**). ''Landmark detection for distinctive feature-based speech recognition,'' J. Acoust. Soc. Am. **100**, 3417–3430.
Manuel, S. Y. (**1991**). ''Some phonetic bases for the relative malleability of syllable-final versus syllable-initial consonants,'' in Proceedings 12th International Congress of Phonetic Sciences, Aix-en-Provence, Vol. V, pp. 118–121.
Manuel, S. Y. (**1995**). ''Speakers nasalize /ð/ after /n/, but listeners still hear /ð/,'' J. Phonetics **23**, 453–476.
Manuel, S. Y., and Stevens, K. N. (**1995**). ''Formant transitions: Teasing apart consonant and vowel contributions,'' in Proceedings International Conference on Phonetic Sciences, Stockholm, Vol. 4, pp. 436–439.
McCarthy, J. J. (**1988**). ''Feature geometry and dependency: a review,'' Phonetica **45**, 84–108.
Mermelstein, P. (**1975**). ''Automatic segmentation of speech into syllabic units,'' J. Acoust. Soc. Am. **58**, 880–883.
Ohde, R. N., and Stevens, K. N. (**1983**). ''Effect of burst amplitude on the perception of stop consonant place of articulation,'' J. Acoust. Soc. Am. **74**, 706–714.
O'Shaughnessy, D. (**2000**). *Speech Communications: Human and Machine* (IEEE, New York).
Rabiner, L., and Juang, B-H. (**1993**). *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ).
Shattuck-Hufnagel, S. (**1992**). ''The role of word structure in segmental serial ordering,'' Cognition **42**, 213–259.
Sproat, R., and Fujimura, O. (**1993**). ''Allophonic variation in English /l/ and its implications for phonetic implementation,'' J. Phonetics **21**, 291–311.
Stevens, K. N. (**1972**). ''The quantal nature of speech: Evidence from articulatory-acoustic data,'' in *Human Communication: A Unified View*, edited by P. B. Denes and E. E. David, Jr. (McGraw-Hill, New York), pp. 51–66.
Stevens, K. N. (**1989**). ''On the quantal nature of speech,'' J. Phonetics **17**, 3–46.
Stevens, K. N. (**1998**). *Acoustic Phonetics* (MIT, Cambridge, MA).
Stevens, K. N. (**2000a**). ''Diverse acoustic cues at consonantal landmarks,'' Phonetica **57**, 139–151.
Stevens, K. N. (**2000b**). ''From acoustic cues to segments, features and words,'' in Proceedings 6th International Conference on Spoken Language Processing (ICSLP2000), Beijing, China, Vol. 1, pp. A1–A8.
Stevens, K. N. (**2001**). ''The properties of the vocal-tract walls help to shape several phonetic distinctions in language,'' in *Travaux du Cercle Linguistique de Copenhague*, Vol. XXXI, pp. 285–297.
Stevens, K. N., and Halle, M. (**1967**). ''Remarks on analysis by synthesis and distinctive features,'' in *Models for the Perception of Speech and Visual Form*, edited by W. Wathen-Dunn (MIT, Cambridge, MA), pp. 88–102.
Stevens, K. N., Manuel, S. Y., and Matthies, M. (**1999**). ''Revisiting place of

articulation measures for stop consonants: Implications for models of consonant production,'' in Proceedings 14th International Congress of Phonetic Sciences (ICPhS'99, San Francisco), Vol. 2, pp. 1117–1120.
Sun, W. (**1996**). ''Analysis and interpretation of glide characteristics in pursuit of an algorithm for recognition,'' MS thesis, Massachusetts Institute of Technology, Cambridge, MA.
Sussman, H. M., McCaffrey, H. A., and Matthews, S. A. (**1991**). ''An investigation of locus equations as a source of relational invariance for stop place categorization,'' J. Acoust. Soc. Am. **90**, 1309–1325.
Syrdal, A. K., and Gopal, H. S. (**1986**). ''A perceptual model of vowel recognition based on the auditory representation of American English vowels,'' J. Acoust. Soc. Am. **79**, 1086–1100.
Whalen, D. H., Gick, B., Kumada, M., and Honda, K. (**1998**). ''Cricothyroid activity in high and low vowels: Exploring the automaticity of intrinsic *F*0,'' J. Phonetics **27**, 125–142.
Zsiga, E. C. (**1994**). ''Acoustic evidence for gestural overlap in consonant sequences,'' J. Phonetics **22**, 121–140.
Zue, V. W., and Laferriere, M. (**1979**). ''An acoustic study of medial /t, d/ in American English,'' J. Acoust. Soc. Am. **66**, 1039–1050.
Zue, V. W., Glass, J. R., Goodine, D., Phillips, M., and Seneff, S. (**1990**). ''The SUMMIT speech recognition system: Phonological modeling and lexical access,'' in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 49–52.