Landmark detection for distinctive feature-based speech recognition

Sharlene A. Liu[a)]
Room 36-511, Research Lab of Electronics and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 7 August 1995; revised 15 October 1995; accepted 30 May 1996)

This work is a component of a proposed knowledge-based speech recognition system which uses *landmarks* to guide the search for *distinctive features*. In the speech signal, landmarks identify times when the acoustic manifestations of the linguistically motivated distinctive features are most salient. This paper describes an algorithm for automatically detecting acoustically abrupt landmarks. Some examples of acoustically abrupt landmarks are stop closures and releases, nasal closures and releases, and the point of cessation of free vocal fold vibration due to a velopharyngeal port closure at a nasal-to-obstruent juncture. As a consequence of landmark detection, the algorithm provides estimates of the broad phonetic class (articulator-free features) of the underlying segment. The algorithm is hierarchically structured, and is rooted in linguistic and speech production theory. It uses several factors to detect landmarks: energy abruptness in five frequency bands and at two levels of temporal resolution, segmental duration, broad phonetic class constraints, and articulatory constraints. Tested on a database of continuous, clean speech of women and men, the landmark detector has detection rates over 90%. A large majority of the detections were within 20 ms of the landmark transcription, and almost all were within 30 ms. The results are analyzed by landmark type and phonetic class.

PACS numbers: 43.72.Ne [JS]

INTRODUCTION

A. A knowledge-based approach

In a knowledge-based approach to speech recognition, knowledge about speech and the operating environment is explicitly built into the recognizer by the system designer. It can model parts of the human perception process. Since humans are the best speech recognizers known to us, a knowledge-based approach is a worthwhile endeavor. If successful, it should have the flexibility to work in many operating environments. A knowledge-based system is dedicated to processing speech (versus signals in general) and therefore is efficient in that sense. Because knowledge sources are explicitly incorporated, improvements to the system can be made in a directed, meaningful manner. Prosodic and segmental level knowledge can usually be added in a straightforward way. The knowledge necessary to design such a system, however, should be as comprehensive as possible and the desired acoustic parameters need to be automatically extractable.

Rather than explicitly specifying speech knowledge in a recognition system, a statistical approach builds models by training on speech data, thereby implicitly acquiring knowledge on its own. Automatic learning during the training phase makes statistical methods powerful, since much of our present ignorance about speech can be overcome in this way. Statistical methods have been successful for large-vocabulary, speaker-independent speech recognition [e.g., the work by Lee (1989)]. In the training phase, large amounts of data are used to cover all the possible contextual variations of phonetic units. Frequently occurring speech units can be modeled well, but infrequently occurring units will not be as well specified. Because of their heavy reliance on data, statistical methods do not generalize easily to tasks for which they are not explicitly trained. For example, if the operating environment does not match the training environment (e.g., a different microphone is used, background noise is added, or speech is bandlimited to the telephone line), recognition accuracy can decrease severely. To accommodate a new operating environment, retraining or adaptation is usually required. In adverse conditions, like noisy or telephone quality environments, building models the statistical way does not always give satisfactory performance [e.g., the work by Das *et al.* (1993)].[1]

Speech recognition systems may use a combination of both knowledge-based and statistical approaches. In much of speech recognition research since the late 1980s, hidden Markov models (HMMs), a popular statistical tool, form the foundation of most systems. Increasingly, artificial neural networks (ANNs), another statistical tool, are becoming popular as well. Hybrid HMM/ANN systems are being built. In these fundamentally statistical systems, knowledge sources are added whenever suitable. Examples of such knowledge sources are phone duration information, an auditory model front end, or, more simply, the mel-frequency scale to approximate the frequency warping performed by the basilar membrane. The fundamental philosophy underlying the speech recognition system adopted in this paper, however, is knowledge based instead of statistical. The system employs some statistics as a guide to making decisions but the foundation of the system is not statistical.

Figure 1 shows a block diagram of the proposed

[a)]E-mail address: liu@lexic.mit.edu

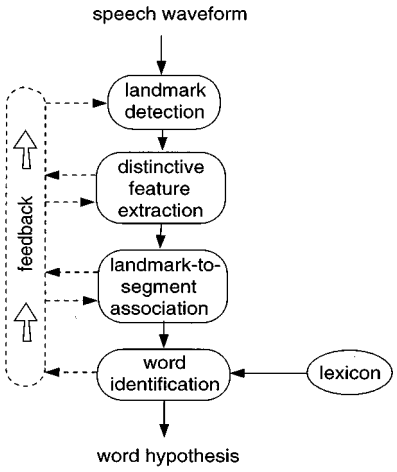

FIG. 1. The proposed speech recognition system.

knowledge-based speech recognition system (Stevens *et al.*, 1992). This system sets the framework for the research presented in this paper. The first step is to locate *landmarks*, which are points in an utterance around which information about the underlying *distinctive features* may be extracted. Next, the distinctive features at a landmark are identified based on acoustic measurements made in the vicinity of the landmark. The landmarks and associated features are related to the underlying segments[2] and a sequence of segments is hypothesized. This sequence is matched to a lexicon whose words are directly defined in terms of features, and word hypotheses are made. Feedback allows information from an advanced stage to be used to correct mistakes made at an earlier stage.

B. Distinctive features

The speech recognition system of Fig. 1 has two noteworthy properties. The first is the chosen unit of speech, the distinctive feature. Distinctive features concisely describe the sounds of a language at a subsegmental level. They have a relatively direct relation to acoustics and articulation. They take on binary values and form a minimal set which can distinguish each segment from all others in a language (Jakobson *et al.*, 1952). Another advantage of distinctive features is that they can concisely describe many of the contextual variations of a segment. These contextual variations could be due to individual speaking styles and phonological assimilation across word boundaries. If the unit of modeling were the phoneme rather than the distinctive feature, a separate model would be required for every modification of the phoneme, resulting in an explosion of the number of speech units required, as Lee (1989) experienced with triphone modeling. The number of units is even larger if syllable or word models are used.

C. Landmark detection

The second noteworthy property of the proposed speech recognition system is landmarks. They are a guide to the presence of underlying segments, which organize distinctive features into bundles. Landmarks define regions in an utterance when the acoustic correlates of distinctive features are most salient. They mark perceptual foci and articulatory targets. Stevens (1985) has suggested that, for some phonetic contrasts, a listener focuses on landmarks to get the acoustic cues necessary for deciphering the underlying distinctive features. Furui (1986) and Ohde (1994) have made this same observation for Japanese syllables and children's speech, respectively. In order to exploit these information-rich areas of the speech waveform, the proposed speech recognition system first finds landmarks in the speech waveform so that subsequent processing can focus on relevant signal portions, instead of treating each part of the signal equally importantly. Based on the kind of landmark found, certain distinctive features will be relevant and others will not. This very directed approach minimizes the amount of processing necessary. Landmarks can also be found somewhat independently of timing factors, like speaking rate and segmental duration. On the other hand, landmarks can give timing information to aid in later processing. The vertical lines in Fig. 2 denote some examples of landmarks. The types of landmark shown will be explained in Sec. I.

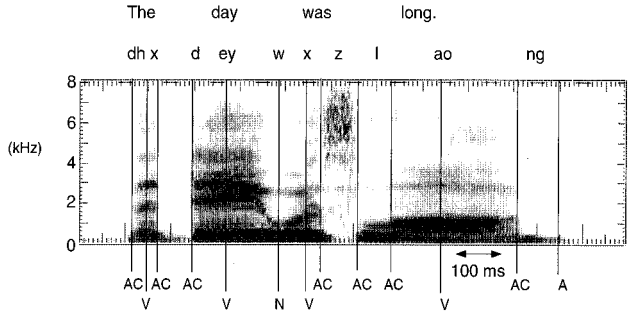

FIG. 2. An illustration of landmarks; **AC**=abrupt-consonantal, **A**=abrupt, **N**=nonabrupt, **V**=vocalic.

Landmark detection is just one way to organize the speech waveform. Frame-based processing and segmentation are two other possibilities. All three methods begin with slicing the speech waveform into equal-length, likely overlapping frames; their difference lies in how they organize subsequent processing. Figure 3 illustrates the differences among these three techniques.

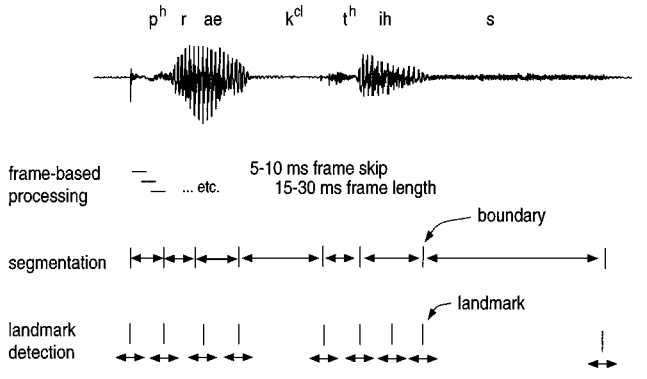

FIG. 3. The differences among frame-based processing, segmentation, and landmark detection. The waveform is of the word ''practice.''

In frame-based processing, presently the most popular way of dividing up the speech waveform, the frames *themselves* are the centers of subsequent processing. A frame is typically 15–30 ms long and occurs every 5–10 ms. A fixed set of speech attributes is measured at each frame.

More structured than frame-based processing, segmentation finds *boundaries* in the speech waveform. These boundaries delimit unequal-length, semi-steady-state, abutting regions, with each region corresponding to a phone or subphone unit.[3] Subsequent processing focuses on these regions, typically acquiring averages across a region [e.g., see Gish and Ng (1993)] and sometimes measuring attributes near the boundaries [e.g., see Zue *et al.* (1990b)]. In various tasks, Flammia *et al.* (1992) and Marcus (1993) have found that a segmentation approach performs better than or comparably to a frame-based approach while reducing the computational load in training and testing by a significant amount. In continuous speech recognition systems of the 1970s through the mid-1980s, segmentation was a popular method of organizing the speech waveform. Segmentation was compatible with acoustic–phonetic processing, which was then widely used [e.g., see Weinstein *et al.* (1975)]. A problem arises with segmentation, however, when parts of the waveform do not have sharp boundaries, like those corresponding to diphthongs and semivowels. In order to accommodate different degrees of abruptness, either oversegmentation is required to find all the pieces of interest (Andre-Obrecht, 1988) or a multilevel representation is needed (Glass, 1988).

Landmark detection is distinctly different from frame-based processing and segmentation, as illustrated in Fig. 3. As mentioned before, landmarks are *foci*, so speech processing is done around a landmark rather than in between two landmarks. While many of the boundaries in segmentation are also the landmarks for a landmark-based system, not all boundaries are landmarks and not all landmarks are boundaries. The problem of delimiting semivowels and diphthongs is avoided altogether by landmark detection, as will be explained in Sec. I. Landmark detection is typically more hierarchical and involves more than one acoustic measure. Also, landmarks are associated with bundles of distinctive features whereas segmentation is associated with phones.

D. Objective and outline

As the first step in the lexical access process, landmark detection is of primary importance. The most numerous types of landmarks are acoustically abrupt. An estimate based on a phonetically balanced subset of sentences in the TIMIT corpus (Zue *et al.*, 1990a) shows that acoustically abrupt landmarks comprise approximately 68% of the total number of landmarks in speech. These landmarks are often associated with consonantal segments, e.g., a stop closure or release. This paper describes a knowledge-based algorithm for automatically detecting acoustically abrupt landmarks. Acoustically abrupt landmarks comprise ''abrupt-consonantal'' and ''abrupt'' landmarks, which are described in Sec. I. The process of landmark detection provides information about the articulator-free features [consonantal], [sonorant], and [continuant]. The landmark detector is designed and tested on clean, broadband, phonetically controlled speech. Other experiments involving the TIMIT corpus, telephone speech, and speech in noise are reported by Liu (1995b).

The organization of this paper is as follows. In Sec. I we describe landmarks in greater detail. In Sec. II we present the database used in a landmark detection experiment. In Sec. III we detail the landmark detection algorithm. In Sec. IV we present the results of running the algorithm on the database. Finally, in Sec. V we summarize the paper, showing how the articulator-free features can be deduced from the landmark type, and give suggestions on how the landmark detector can be improved.

I. LANDMARKS

Landmarks are categorized into four groups: abrupt-consonantal, abrupt, nonabrupt, and vocalic. Figure 2 shows examples of these landmark groups for the utterance ''The day was long.'' This section will describe these landmark groups. The focus of this paper is on detecting the abrupt-consonantal and abrupt landmarks.

Phonologically, segments can be classified as [+consonantal] or [−consonantal]. A [+consonantal] segment involves a *primary articulator* forming a tight constriction in the midline of the vocal tract (Sagey, 1986). Only the articulators in the oral cavity (i.e., lips, tongue blade, and tongue body) can be primary articulators. Segments not involving a tight constriction or implemented by articulators outside of the oral cavity (e.g., soft palate and glottis) are [−consonantal]. Speech is formed by a series of articulator narrowings and releases. The most salient of these narrowings and releases are acoustically abrupt. An acoustically abrupt constriction involving a primary articulator is typically tight and is a consequence of implementing a [+consonantal] segment. An *abrupt-consonantal* (**AC**) landmark marks the closure and another marks the release of one of these constrictions. The clearest manifestation of an **AC** landmark is when the constriction occurs adjacent to a [−consonantal] segment. A pair of these landmarks, one on either side of the constriction, will be referred to as the *outer* **AC** landmarks. An example of a pair of outer **AC** landmarks is the [b] closure and release in ''able.'' Other landmarks can occur within or outside of the pair of outer **AC** landmarks. These are described below.

A common sequence of landmarks is one in which the outer **AC** landmarks are governed by the same underlying segment and, thus, are implemented by the same articulator (e.g., the [b] closure and release in able). In consonantal clusters, however, the two outer **AC** landmarks are often not governed by the same articulator (e.g., the [p] closure and [d] release in ''tap dance''). The release by the first articulator or the formation of the constriction by the second articulator may or may not be manifested in the acoustic signal. If the articulatory event is observable in the acoustic signal, then it is marked as an *intraconsonantal* **AC** landmark. In the ''tap dance'' example, if the [p] release is evident in the sound, then the [p] burst and the [d] closure are intraconsonantal **AC** landmarks.

The configuration of the glottis and the soft palate are articulatorily independent of the occurrence of an **AC** land-

mark. Just as for abrupt primary articulator movement, the movements of the glottis or the soft palate can independently cause an abrupt change in the acoustic signal. For example, as the glottis moves from a spread to a modal configuration when air is passing through, vocal-fold vibration begins. This onset of vocal-fold vibration is observed as a rapid change in the characteristics of the sound. Likewise, if the velopharyngeal port closes when the oral cavity is already closed, there is an abrupt increase in intraoral pressure, resulting in a reduction in the amplitude of glottal pulses. These abrupt changes in the sound caused by glottal or velopharyngeal activity but without accompanying primary articulator movement are labeled as *abrupt* (**A**) landmarks. The **AC** and **A** landmarks differ in that **A** landmarks do not involve primary articulator movement. The **A** landmarks can occur outside of a pair of outer **AC** landmarks (called *intervocalic* **A** landmarks; e.g., the voice onset after the [p] burst in ''paint'') or within the pair (called *intraconsonantal* **A** landmarks; e.g., the [n]–[t] transition in ''canteen''). Together, **AC** and **A** landmarks comprise approximately 68% of the total number of landmarks, as noted in the Introduction.

For semivowels, the constriction that is formed is not narrow enough to create an abrupt change in the spectrum of the sound. The narrowing of the constriction, however, does reach some articulatory extreme out of which it gradually releases. The narrowing of the constriction usually causes a decrease in the first-formant frequency, F_1, and in the amplitude of the sound. If the consonant occurs between two vowels, a minimum in F_1 and in waveform amplitude denotes the narrowest point in the constriction (e.g., the middle of [w] in ''away''). This point is a *nonabrupt* (**N**) landmark and occurs outside of a pair of outer **AC** landmarks. The narrowest point in the production of a semivowel can be coincident with an acoustically abrupt part of speech; in such a case, the landmark is both an **N** and an acoustically abrupt landmark (e.g., the [dw] release in ''dwell''). The **N** landmarks comprise approximately 3% of the total number of landmarks, as estimated from the TIMIT database.

Finally, vowels have their own landmarks. When the vocal tract is at an open extreme for a vowel, a local maximum occurs in both F_1 and waveform amplitude (e.g., the middle of [ae] in ''bat''). This point is a *vocalic* (**V**) landmark and occurs outside of a pair of outer **AC** landmarks. The **V** landmarks make up approximately 29% of the total number of landmarks, as estimated from the TIMIT database.

For lexical access, a landmark-to-segment relation must be specified. Sometimes, a one-to-one relation holds, like the relation between a **V** landmark and a vowel. Sometimes, a two-to-one relation holds; for example, an intervocalic fricative can have a landmark at closure and a landmark at re-

TABLE I. Detection rates of the development and test sets, by phonetic category. An * next to a detection rate means that the number of tokens is less than 30 so the detection rate is unreliable. The number of tokens is given in parentheses.

	Phonetic category	Landmark type	Detection rate (Development)	Detection rate (Test)
Outer **AC**	+v fric clos	**g**(lottis)	96% (97)	96% (47)
	+v fric rel	⋯	95% (109)	98% (58)
	−v fric clos	⋯	100% (58)	*100% (28)
	−v fric rel	⋯	100% (107)	*100% (26)
	flap clos	⋯	*88% (26)	*88% (8)
	flap rel	⋯	*85% (27)	*75% (8)
	+v stop clos	⋯	95% (94)	98% (48)
	unasp'd stop rel	⋯	97% (104)	95% (44)
	−v stop clos	⋯	99% (126)	100% (62)
	asp'd stop rel	**b**(urst)	95% (111)	89% (72)
	nasal clos	**s**(onorant)	90% (170)	72% (53)
	nasal rel	⋯	90% (105)	77% (35)
	[l] clos	⋯	*71% (21)	*33% (9)
	[l] rel	⋯	81% (36)	*76% (17)
Intraconsonantal **AC**	stop clos	**b**(urst)	*100% (20)	*30% (10)
	stop rel	⋯	*95% (19)	*100% (10)
	fric clos	⋯	*100% (8)	*90% (29)
	affric rel	⋯	84% (56)	95% (40)
	nasal→fric	**g**(lottis)	*100% (26)	*100% (2)
	fric→nasal	⋯	*100% (13)	*100% (4)
Intraconsonantal **A**	velophar clos	**g**(lottis)	95% (57)	*92% (25)
	velophar rel	⋯	*25% (4)	*100% (8)
Intervocalic **A**	[ʔ] clos	**g**(lottis)	90% (41)	*100% (23)
	[ʔ] rel	⋯	100% (51)	*100% (23)
	−v [h] onset	⋯	⋯ (0)	⋯ (0)
	−v [h] offset	⋯	97% (111)	97% (72)
Total			94% (1597)	91% (761)

lease. In some cases, a three-to-one relation holds, as in [p], which has one landmark at closure, one at the labial release, and one at voice onset. Or, there may be a one-to-two relation, as in ''bright,'' where the landmark at the [b] release serves both [b] and [r].

For the purpose of evaluating the performance of a landmark detection algorithm, the **AC** and **A** landmarks are classified into phonetic categories. The second column of Table I lists these categories. The outer **AC** landmarks are the closures and releases associated with fricatives, flaps, stops, nasals, and [l]'s next to [−consonantal] segments. For example, the closure and release of the [b] in [ebe] are outer **AC** landmarks. Flaps are classified as obstruents even though they are often too short to allow much pressure buildup. Eventually, they will need a specialized detector dedicated to finding them. Intraconsonantal **AC** and **A** landmarks are those in consonant clusters. For illustration, if the release of [d] in ''tadpole'' is evident in the sound, then it is an intraconsonantal **AC** stop release and the [p] closure is an intraconsonantal **AC** stop closure. In the word ''ski,'' the end of /s/ at the /k/ closure is an intraconsonantal **AC** fricative closure. An affricate release as in [č] of ''church'' is an intraconsonantal **AC** affricate release. The landmark between the [s] and [m] in ''small'' is a fricative→nasal landmark. The landmark between the [m] and [z] in ''plums'' is a nasal→fricative landmark. A velopharyngeal closure or release occurs for nasal/stop combinations. For example, the end of glottal vibration after the [n] in ''bending'' is a velopharyngeal closure. The beginning of glottal vibration for the [m] in ''batman'' is a velopharyngeal release. The intervocalic **A** landmarks are caused by the glottis. These are the onsets and offsets of glottal stops and aspirated consonants. The remaining columns of Table I will be discussed later.

II. DATABASE

A. Speech recording

Utterances were tape-recorded in a quiet room using an Electrovoice omnidirectional microphone dangling 25 cm in front of and 5 cm above the speaker's mouth. This placement was roughly equidistant to the nose, mouth, and throat, so the microphone could pick up signals from all three radiating sources. The recordings were passed through an antialiasing filter with a cutoff frequency of 7.5 kHz before being digitized at 16 kHz. The 7.5-kHz cutoff frequency allowed relevant high-frequency frication noise in female speech to be captured. The signal-to-noise ratio (SNR) for speech recorded in this condition was about 30 dB.

B. Lexicon and data sets

A lexicon of 250 words was used to construct 40 syntactically correct sentences. The words were mostly monosyllabic (69%), some bisyllabic (30%), and very few trisyllabic (1%). Fifteen percent of the words had consonant clusters (e.g., [sp] in ''sport,'' [nd] in ''and,'' or [fy] in ''few''). Two data sets—development and test—were constructed. The development set was used during algorithm development: the design and parameter values of the algorithm were modified by hand based on knowledge accumulated from preliminary results with the development set. It consisted of four speakers (two women, two men) speaking the first 20 sentences. The test set was independent from the development set to see how well the algorithm generalized to new speakers and new sentences. It consisted of two new speakers (one woman, one man) speaking the last 20 sentences.

C. Labeling convention

Because a complete speech recognition system which employs landmark detection was not available, the landmark detector could not be evaluated in terms of a word recognition score. Instead, an intermediate level of evaluation was necessary. This intermediate level was a landmark detection score, which is highly dependent on the landmark labeling convention. Thus, care was taken to objectively label utterances with landmarks.

Landmark labeling was done using phonological and acoustic rules. Three decisions had to be made: (1) the existence of an **AC** or **A** landmark, (2) the time placement of the landmark, and (3) the categorization of the landmark. Waveform and spectrogram displays and listening were used to guide landmark labeling.

An **AC** or **A** landmark exists if the underlying segment is acoustically evident and acoustically abrupt. The intended sentences were used to guide the labeling. Most of the time, the intended phonological targets were manifested in the speech signal. Sometimes, however, phonological targets were modified in the speech signal. For example, a [ð] following a nasal, as in ''in the,'' often shows little evidence of diminished voicing, so instead of labeling it as a velopharyngeal closure followed by a fricative release, it was labeled simply as a sonorant consonantal release. Even if a phonological target is manifested in the speech signal, it may not be acoustically abrupt. A typical example is [l], which can be acoustically abrupt or not. An [l] closure or release was considered acoustically abrupt if 90% of its broadband energy transition in dB occurred within 40 ms.

Regarding the time placement of a landmark, an **AC** or **A** landmark was put at the time of the articulator movement which caused the landmark, as inferred from the acoustic data. Time placement for most **AC** and **A** landmarks is straightforward because of the abrupt spectral change that takes place. The landmark for a voiced obstruent closure was placed at the disappearance of high-frequency formant energy in the spectrogram.

The landmarks were grouped into the phonetic categories listed in Table I. A flap was defined to have a 35 ms or less closure interval; otherwise it was categorized as a [t] or [d]. This criterion is in accord with Zue and Laferriere's (1979) acoustic study, in which they found that the average closure period of a flap is 26–27 ms, with a standard deviation of 10–12 ms. Outer **AC** landmarks associated with stop closures and fricatives were further divided into voiced and unvoiced. To avoid interpreting the various acoustic manifestations of voicing at these landmarks, a voiced/unvoiced decision was made based on the underlying phonology alone. Stop releases, on the other hand, were categorized as aspirated or unaspirated based on the voice onset time (VOT).

Almost all of the aspirated stops were unvoiced, and almost all of the unaspirated stops were voiced. Unaspirated stop releases were labeled with one landmark while aspirated stop releases were labeled with two landmarks (one at the burst and one at the voice onset). A stop release was considered unaspirated if the VOT was 20 ms or less. It was considered aspirated if the VOT was more than 20 ms. This criterion agrees with Lisker and Abramson's (1964) finding that English unvoiced aspirated stops in word-initial position have an average VOT of 28 ms or more, depending on the place of articulation, and the voiced unaspirated counterparts have an average VOT of 17 ms or less. Affricates also had a 20-ms criterion: separate burst and fricative release landmarks were assigned only if the VOT exceeded 20 ms. Nasal onsets and offsets at the beginnings and ends of utterances were labeled as velopharyngeal releases and closures, respectively.

In order to assess the effect of vowel reduction on landmark detection, vowels next to **AC** and **A** landmarks were labeled as either reduced or unreduced. A reduced vowel is short in duration [typically less than 50 ms (Klatt, 1976)] and low in intensity [typically 12 dB or more down from a neighboring stressed vowel (Beckman, 1986)]. Schwas by definition are reduced; syllabic nasals, syllabic [l]'s, and some [ɚ]'s are typically reduced. All other vowels, stressed or otherwise, were classified as unreduced.

These labeling conventions agree for the most part with those of TIMIT for acoustically abrupt points in the speech waveform (Zue and Seneff, 1990). TIMIT labels were similarly motivated by phonemics and acoustics. Abrupt acoustic changes were always marked. If no acoustic evidence existed for a certain phoneme, then no label was put there. Spectrogram displays and listening were also used to decide on labels. One difference between the database presented here and TIMIT was the labeling of stop releases. Whereas only one label was placed at an unaspirated stop release in this database, two labels were placed in TIMIT, one for the burst and one for the voice onset, regardless of the VOT.

III. LANDMARK DETECTION ALGORITHM

Figure 4 shows a flow diagram of the algorithm for landmark detection. Speech input goes through a general signal processing stage, whose outputs feed a landmark type-specific processing stage. The output of type-specific processing is a series of landmarks specified by time and type. In general processing, a spectrogram is computed and divided into six frequency bands. Then, a coarse- and a fine-processing pass are executed. In each pass, an energy waveform is constructed in each of the six bands, the derivative of the energy is computed, and peaks in the derivative are detected. Localized peaks in time are found by matching peaks from the coarse- and fine-processing passes. These peaks represent times of abrupt spectral change in the six bands. In type-specific processing, the localized peaks direct processing to find three types of landmarks. These three types are the following:

(1) **g**(lottis), which marks a time when there is a transition of freely vibrating vocal folds (with no increase in intraoral pressure) to a condition where the vocal folds are not freely vibrating, or vice versa;
(2) **s**(onorant), which marks sonorant consonantal closures and releases;
(3) **b**(urst), which designates stop or affricate bursts and points where aspiration or frication ends due to a stop closure.

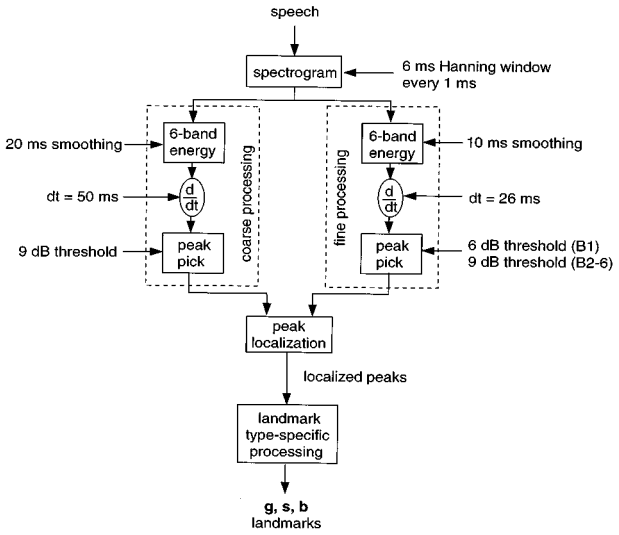

FIG. 4. The landmark detection algorithm. The input is speech and the output is three types of landmarks: **g**(lottis), **s**(onorant), **b**(urst).

Table I associates landmarks in each phonetic category with a landmark type. The rest of this section presents the steps just outlined in greater detail.

A. General processing

A broadband spectrogram is computed with a 6-ms Hanning window every 1 ms. Each 6-ms frame is zero-padded out to 512 points before a discrete Fourier transform (DFT) is taken. The top panel of Fig. 5 shows a spectrogram for the utterance: ''The money is coming today.'' The spacing between points for the DFT is 31.2 Hz, so that spectral peak amplitudes in later energy computations can be estimated reasonably well. The high frame rate allows quick acoustic changes to be monitored. Some acoustic changes happen very quickly, particularly the ones associated with obstruent segments as articulators move from one quantal state to another. The short Hanning window produces a broadband spectrum, which gives broad spectral information while suppressing harmonic detail.

The resulting spectrogram is then divided into the following six frequency bands:

Band 1:	0.0–0.4 kHz
2:	0.8–1.5
3:	1.2–2.0
4:	2.0–3.5
5:	3.5–5.0
6:	5.0–8.0

Band 1 monitors the presence or absence of glottal vibration.

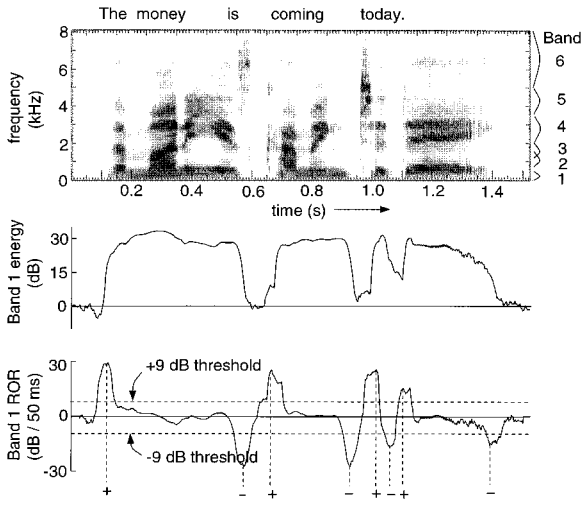

FIG. 5. General processing as the first part of the landmark detection algorithm. The top panel shows the spectrogram for the utterance ''The money is coming today.'' The middle panel is the band 1 energy and the bottom panel is the band 1 ROR, both from coarse processing. The two dotted horizontal lines are thresholds for peak picking. The peaks detected are shown with ± signs indicating the polarity of the peaks. Band 1 peaks are also likely candidates for **g**(lottis) landmarks.

It does not extend beyond 400 Hz in order to reduce the chance of picking up low-frequency burst energy. Closures and releases for sonorant consonants are detected using bands 2–5. These bands approximate the frequency ranges for the spectral prominences of sonorant consonants. For intervocalic sonorant consonantal segments, a large spectral change usually occurs in the 0.8- to 2-kHz range, and this change is often due to the introduction of a zero in the vocal-tract transfer function in that range. In order to capture this change, bands 2 and 3 span this range and are chosen to overlap. Each of these two bands is not guaranteed to contain one spectral prominence, but at least one band is expected to capture at least one spectral prominence. At a sonorant consonantal closure, band 1 energy remains strong because glottal vibration continues; however, spectral prominences above F_1 show a marked abrupt decrease in energy because of increased acoustic losses, the rapid change of zeros in pole–zero pairs, and the fall in F_1. The onsets and offsets of aspiration and frication noise associated with stops, fricatives, and affricates can also be found from bands 2–5. Noise energy will lie in at least one of these four bands. Band 6 spans the remaining frequency range up to 8 kHz, and is one of the bands used for silence detection for stops.

Following the computation of the spectrogram, energy changes in the six bands are found using a two-pass strategy, as indicated by the two parallel branches coming out of the spectrogram block in Fig. 4. Both passes employ the same processing steps except that the first pass uses coarse parameter values to find the general vicinity of a spectral change, and the second pass uses fine parameter values to localize it in time. The processing strategy will be described with the first pass parameter values, and then the second pass parameter values will be given.

In the coarse-processing pass, an energy waveform in each of the six bands is calculated. The middle panel of Fig. 5 gives an example of the band 1 energy waveform. An energy waveform should be able to resolve abrupt acoustic changes due to sudden changes in formant frequency amplitudes, but ignore glottal pulse variations and noise fluctuations. To smooth out the unwanted characteristics, a 20-ms average of the squared magnitude of the spectrogram, centered about the time of interest, is computed every 1 ms. Within each band, the maximum in the smoothed spectrogram at each time is chosen to represent the energy in that band. Energy is then recorded in dB. Provided a band encompasses exactly one spectral prominence, picking the maximum energy in the band as a function of time is the same as following the spectral prominence amplitude in time.

Once the six-band energy is computed, a six-band *rate-of-rise* (ROR) is found by taking an overlapping dB first difference of the energy in each band. The ROR waveform of a band indicates how quickly the energy is changing in that band. Working with dB differences automatically considers relative values so that gain normalization is not necessary across utterances. The first difference is computed every 1 ms using a 50-ms time step, centered about the time of interest. The 50-ms time step is chosen to span energy transitions of abrupt closures and releases, including voiced obstruent closures, taking into account the 6-ms Hanning window and 20-ms smoothing. The third panel of Fig. 5 shows the ROR calculated from the band 1 energy waveform above it.

The positive and negative peaks in the ROR waveform are the points of abrupt acoustic change in a band. Mermelstein's (1975) peak-picking algorithm was tailored to find the ROR ±peaks whose absolute value is greater than 9 dB. The 9-dB threshold is motivated by the difference in glottal source amplitude between an obstruent and a neighboring vocalic segment and by empirical evidence. The two dotted horizontal lines in the third panel of Fig. 5 show the ±9-dB thresholds. The peaks that are detected by the peak-picker are shown with ± signs indicating the polarity of the peaks.

In the parallel fine-processing pass shown in Fig. 4, some parameter values are modified in order to localize energy changes in time. A 10-ms smoothing interval is used on the spectrogram instead of 20 ms; a time step of 26 ms is used for the ROR calculation instead of 50 ms; and the peak threshold for band 1 is reduced from 9 to 6 dB. This 3-dB reduction is made to accommodate smaller peaks due to the reduction of the time step. The peak thresholds in bands 2–6 are kept at 9 dB to prevent too many spurious peaks in those bands from occurring.

As shown in Fig. 4, the ROR peaks resulting from the coarse and fine passes come together at a ''peak localization'' block. Here, ROR peaks from the coarse pass are used to guide the search for corresponding ROR peaks in the same band from the fine pass. Within ±30 ms of a coarse pass peak, the biggest fine pass peak (in absolute terms) with the same sign as the coarse pass peak is chosen as the localized peak. Localized peaks are the inputs into the landmark type-specific processing stage. The type-specific detectors for **g**, **s**, and **b** landmarks are explained next.

B. G(lottis) detector

A **g**(lottis) landmark pinpoints a time the vocal folds start or stop free vibration. The factors causing free vocal-fold vibration to cease are a buildup of intraoral pressure due to a supraglottal constriction, vocal-fold spreading or glottal closure, or a reduction of subglottal pressure. The localized band 1 ROR peaks from general processing are initially all candidates for **g** landmarks. These candidates must pass a series of criteria. A +peak indicates the turning on of glottal vibration; a −peak indicates the turning off of glottal vibration. When glottal vibration turns on, it must turn off some time later. Thus each +peak should be followed by a −peak. Peaks are inserted wherever necessary to satisfy this condition. The point of insertion is guided by the shape of the band 1 energy contour. After the peaks are paired, a minimum vowel requirement is imposed on each ± pair. A ± pair should span at least the vowel part of a syllable, the minimum vowel being a schwa. In acoustic terms, this requirement is that band 1 energy between a ± pair of peaks must be no less than 20 dB below the maximum band 1 energy in the utterance for at least 20 ms. The 20-dB threshold agrees with Stevens (1994) study, in which he showed that the F_1 amplitude of a reduced vowel can regularly be as low as 17 dB below that of the pitch-accented vowel in the utterance, provided the reduced vowel is not devoiced. The 20-ms duration requirement is a lower bound on the length of a schwa, which can regularly be about 30 ms in duration (van Beinum, 1994). If the region between the ± pair of peaks does not satisfy the vowel requirement, then that pair is most likely due to creaky voicing or a low-frequency burst, and is deleted.

C. S(onorant) detector

An **s**(onorant) landmark is caused by the closure or release of a nasal or [l]. Figure 6 illustrates **s** landmark detection. A −**s** landmark designates a closure and a +**s** landmark designates a release. As the vocal tract constricts for a sonorant consonant, energy in the F_2–F_4 range decreases. At a release, this energy increases. If the constriction is tight enough and occurs sufficiently quickly, the energy will change rapidly and simultaneously in all bands. During the constricted interval for a nasal or an [l], a primary articulator has made a complete closure, the vocal folds continue to vibrate, and the vocal-tract shape is relatively constant. Thus the spectrum should remain relatively steady, especially at low frequencies.

To find **s** landmarks, only voiced regions, bounded by a +**g** landmark on the left and a −**g** landmark on the right, are considered. Within a voiced region, any peaks having the same sign in bands 2–5 are grouped together. The biggest absolute peak in each group is designated the ''pivot'' and is a likely candidate for an **s** landmark. The pivot then has to pass a steady-state test and an abruptness test. The steady-state attribute is measured by examining the spectral magnitude in the 0- to 600-Hz range of the spectrogram; higher frequencies are not used because pole and zero movement may cause some variation in the higher frequencies. At closure and release, high-frequency abruptness is measured by checking for sufficient change in a high-pass energy signal calculated from the 1.3- to 8-kHz range of a preemphasized version of the spectrogram. Spanning this broad frequency range allows the detection of high-frequency energy changes while avoiding false alarms that would occur had high-frequency abruptness been measured on narrower bands susceptible to semivowel formants moving in and out of the bands. As a further measure of high-frequency abruptness, the bands 2–5 peaks that were grouped together by sign earlier must occur somewhat coincidentally with the pivot in that group. The steady-state and abruptness tests that a pivot must pass are designed to exclude pivots caused by semivowels, which are generally not steady state and not acoustically abrupt.

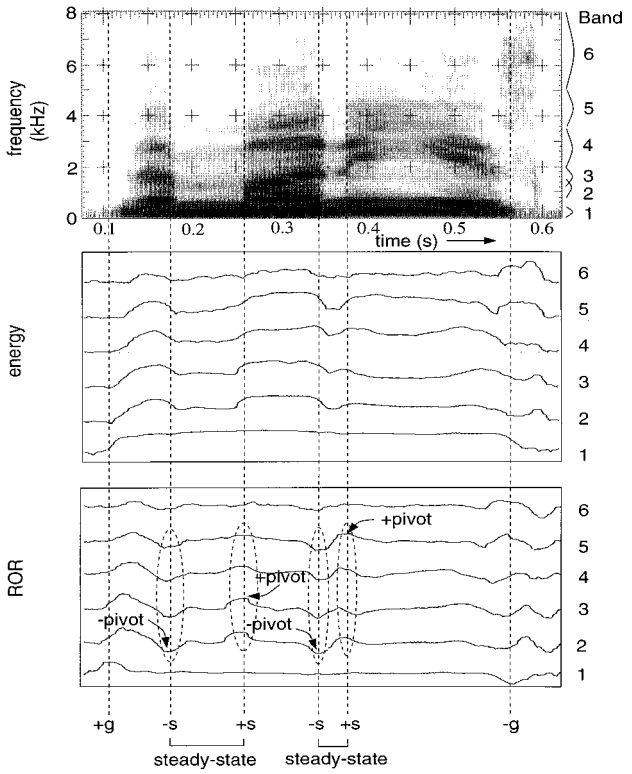

FIG. 6. **S**(onorant) landmark detection. ''The money is ...'' portion of Fig. 4 is shown. The top panel is the spectrogram. The middle panel is the six-band energy waveform and the bottom panel is the six-band ROR waveforms, both from fine processing. The energy waveforms' vertical range is approximately 35 dB for bands 1, 5, and 6, and 50 dB for bands 2–4. The ROR waveforms' vertical range is each approximately ±30 dB/time step.

Heavily voiced obstruents (e.g., [d, v]) are often difficult to detect with the **g** detector described in the previous section because Band 1 energy may not change sufficiently. They usually do, however, exhibit clear higher frequency energy changes. If they are missed by the **g** detector, they can be detected by a variation of the **s** detector. In this variation, pivots are found and high-frequency abruptness is required as before; however, the low-frequency energy in between the closure and release of the obstruent must not be steady state.

D. B(urst) detector

Figure 7 illustrates **b**(urst) landmark detection. A +**b** landmark signifies an affricate or aspirated stop burst. The

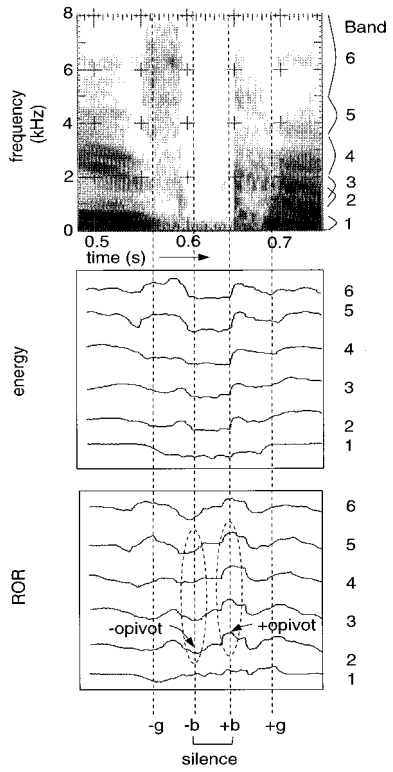

FIG. 7. **B**(urst) landmark detection. The ''... is coming ...'' portion of Fig. 4 is shown. See Fig. 5 caption for a description.

acoustic correlates for a $+\mathbf{b}$ landmark are a silence interval followed by a sharp increase in energy in high frequencies. Since **b** landmarks can occur only during regions without glottal vibration, only the regions delimited by a $-\mathbf{g}$ on the left and a $+\mathbf{g}$ on the right are searched. First, opivots (for ''obstruent pivots'') are found in an analogous manner to pivots, as shown in Fig. 7. An opivot is a candidate for a **b** landmark. Next, silence is measured around an opivot. For a +opivot, a silence interval must exist to the left. This silence period is measured with bands 3–6 energy, using the background energy levels in each band as reference. Bands 1 and 2 are not used in order to allow voice bars of voiced obstruent closures to persist without upsetting the detection of an obstruent closure.

A $-\mathbf{b}$ landmark signifies the offset of frication or aspiration noise due to a stop closure. The acoustic correlates for a $-\mathbf{b}$ landmark are a sharp decrease in high-frequency energy followed by a silence interval. This silence is measured using all the bands, including bands 1 and 2, since a voice bar preceding the following stop consonant is unlikely in English.

IV. RESULTS AND DISCUSSION

A. Scoring

Results of landmark detection are presented in terms of deletion, substitution, insertion, neutral, and detection rates. These rates were determined by comparing the output of the landmark detector with the labeled landmarks. A landmark was considered correctly detected if it was of the same sign and type (**g**,**s**,**b**) as the hand-labeled landmark, and was within ±30 ms of this hand-labeled landmark. The role of a landmark is to identify approximate times around which further detailed processing needs to be performed to extract the desired acoustic cues. Depending on the acoustic cue to be extracted, the signal processing may be concentrated more to the left or more to the right of the hand-labeled landmark. For example, to find the presence of a voice bar at an obstruent closure, processing would be concentrated to the *right* of the hand-labeled landmark, which would be at the oral closure. To determine the place of articulation at this same landmark, processing would be concentrated to the *left* of the hand-labeled landmark. Voiced obstruent closures were sometimes detected more than 30 ms beyond the oral closure because the voice bar could push the maximum band 1 energy change to the right by tens of milliseconds. However, if a $-\mathbf{g}$ landmark was obviously due to the cessation of the voice bar, the voiced obstruent closure was considered detected.

A deletion is a missed landmark. It occurs when no landmark of the correct sign, regardless of type, is detected in the vicinity of the hand-labeled landmark. The deletion rate is calculated by dividing the number of deletions by the total number of landmarks in the category of interest.

A substitution is a landmark of the correct sign but wrong type. The substitution rate is found by dividing the number of substitutions by the total number of landmarks in the category of interest.

The detection rate represents the number of hand-labeled landmarks correctly identified by sign and type. It can be deduced from the substitution and deletion rates:

$$\text{Detection rate} = 100\% - \text{Substitution rate} - \text{Deletion rate}. \tag{1}$$

An insertion is a false landmark. It is not in the hand-labeled set and should not have been detected. The insertion rate is calculated by dividing the number of insertions by the total number of landmarks in the category of interest. A landmark of incorrect sign found near a hand-labeled landmark, regardless of type, is an insertion.

A neutral landmark is also not in the hand-labeled set but, because it can be useful in acoustic–phonetic decoding, is not counted as an insertion. Neutral landmarks do not contribute to error. Examples of neutral landmarks are $\pm\mathbf{g}$ landmarks at creaks,[4] a $-\mathbf{s}$ landmark at the closure of a voiced obstruent, and a $+\mathbf{b}$ landmark at an unaspirated stop burst. The neutral rate is computed by dividing the number of neutrals by the total number of landmarks in the category of interest.

Conventionally, the error rate refers to the sum of the deletion, substitution, and insertion rates. This rate will be called E_1. Note that the error rate can exceed 100% because there is no limit to the number of insertions the landmark detector can produce. Here E_1 is a conservative representation of the results, as it requires that the sign *and* the type of the landmark be correct in order not to add to the error. A substitution, though, is clearly not as serious an error as a deletion or an insertion. In some cases, a substitution would not even be considered an error. To reflect this interpretation of the results, another error rate, E_2, is defined as the sum of

TABLE II. Results of the development set, by landmark type.

Landmark type	No. Tokens	Del	Subs	Ins	Neut	E_1	E_2
g(lottis)	1052	1%	3%	7%	0%	11%	8%
s(onorant)	332	10%	2%	21%	9%	33%	31%
b(urst)	213	5%	1%	6%	33%	12%	11%
Total	1597	4%	2%	10%	7%	16%	14%

the deletion and insertion rates. Here E_2 excludes the substitution rate from the error. In other words, a detected landmark of the right sign but wrong type is not considered an error in E_2.

B. Overall results

Tables II and III show the results by landmark type for the development and test data sets, respectively. The total error rates, E_1 and E_2, in Table II are approximately equal to those in Table III. In particular, the total E_2 is 14% in both tables. The two tables differ only in the details of the error types. The total deletion rate is lower in Table II than in Table III, while the total insertion rate is higher. This difference illustrates the trade-off between deletions and insertions.

The temporal precision with which the landmark detector finds landmarks is of interest. Of the total number of landmarks, 44% were detected within 5 ms of the hand-labeled transcription; 73% were within 10 ms; 83% were within 20 ms; and 88% were within 30 ms. A small percentage were beyond 30 ms, due to voiced obstruent closures. The rest (9%) were either substituted or deleted. Since the precise placement of a landmark is dependent on the acoustic cue to be extracted, tuning the landmark detector to be more precise than it is at this stage would serve no real purpose. Nevertheless, even though the algorithm was not designed specifically to be precise, its time accuracy is still high.

Although the databases for development and test are small—a total of six speakers and 40 sentences—the trends shown in Tables II and III are believed to be representative of the behavior of the landmark detector. To test this hypothesis, a more comprehensive database was used. This new database consists of a subset of the TIMIT corpus, a phonetically and dialectally comprehensive database of American English (Fisher *et al.*, 1986; Zue *et al.*, 1990a). The experiment involving the TIMIT database is briefly presented here for the purpose of illustrating the generalizability of the landmark detector beyond the original database; further details can be found in the work by Liu (1995b). The development set is composed of 16 speakers speaking a total of 80 utterances. The test set is composed of 16 new speakers speaking a total of 48 utterances. Every sentence in the two data sets is unique, and the speakers, balanced by gender, span all the dialectal regions of the United States.

TABLE III. Results of the test set, by landmark type.

Landmark type	No. Tokens	Del	Subs	Ins	Neut	E_1	E_2
g(lottis)	486	2%	1%	2%	0%	5%	4%
s(onorant)	114	29%	1%	27%	10%	57%	56%
b(urst)	161	10%	2%	2%	8%	14%	12%
Total	761	8%	1%	6%	3%	15%	14%

TABLE IV. Results of the TIMIT development and test sets.

Data set	No. Tokens	Del	Subs	Ins	Neut	E_1	E_2
Development	2144	6%	4%	8%	6%	18%	14%
Test	1267	5%	5%	9%	8%	19%	14%

Because the microphone type and the SNR of TIMIT differ from those of the original database, some modifications were made to customize the landmark detection algorithm to the TIMIT conditions. The overall idea of finding abrupt spectral change and applying phonetic knowledge at the points of spectral change, however, remains the same. Eventually, in a fully automated system, the algorithm would discover the change in operating environment and modify its parameter values automatically.

Results of the TIMIT experiment are shown in Table IV. The E_2 rate is again 14% for both development and test sets, as was the case in the original experiment. This striking similarity in E_2 rates lends credence to the original experiment involving the smaller database. The substitution rates in the TIMIT experiment are higher than in the original experiment mainly because of **g** substitutions for **s** landmarks.

An interesting point to note regarding the TIMIT experiment is that there is no degradation in performance between the development and test sets. A relatively small development set to achieve equal performance in the test set is a trademark of knowledge-based engineering. In statistical systems, a large amount of training still leads to inferior performance in the test set.

The remainder of this section will center on the original experiment.

C. G(lottis) landmarks

As demonstrated by Tables II and III, the **g** detector was robust for most of the phonetic contexts in which **g** landmarks occurred. Table I shows the detection rates of the development and test sets by phonetic category. The **g** deletions were due mostly to voiced obstruents. Voicing in the low frequencies reduced abruptness in band 1 energy. Flaps were hard to detect because they were heavily voiced and had a short closure interval. If the closure interval was less than 20 ms, then the smoothing in the general processing stage of the landmark detection algorithm obscured the spectral discontinuity. In contrast to voiced obstruents, voiceless obstruents had a near 100% detection rate.

Figure 8 shows a spectrogram and some typical behaviors of the landmark detector. The bulk of the **g** insertions were due to semivowels. When a semivowel was implemented with a tight constriction, there was an abrupt weakening of high-frequency energy, and sometimes low-frequency energy as well. The ROR peaks that occurred were then picked up as **g** landmarks. Figure 8 shows that the two [w]'s near 0.5 and 0.9 s caused **g** insertions. There are two **g** neutrals at the creak at 0.8 s.

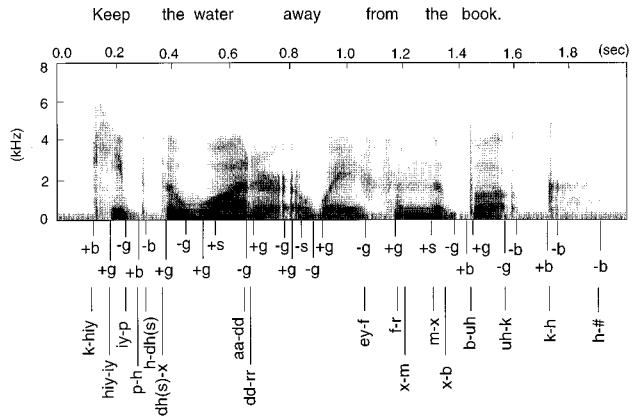

FIG. 8. A spectrogram with the output of the landmark detector and the hand-labeled landmarks below.

D. S(onorant) landmarks

Compared to **g** landmarks, **s** landmarks were more difficult to detect. **S** landmarks relied on the energy change in bands 2–5. In these bands, resonance peak amplitudes and resonance frequency ranges were dependent on phonetic context and can vary quite a bit. Most of the **s** deletions were next to semivowels, high vowels, back vowels, or reduced vowels. In these contexts, spectral change was small at the consonantal closures and releases. Table I shows that, of the **s** landmarks, nasals were detected better than [l]'s. On average, [l]'s were implemented less abruptly than nasals. The [m] closure at 1.2 s in Fig. 8 was missed by the **s** detector.

Most of the **s** insertions were due to semivowels. A semivowel's high-frequency abruptness and low-frequency steady-state voicing could be mistaken for sonorant consonantal segments. In Fig. 8, an **s** insertion occurred for each of the [w]'s, at 0.55 and 0.83 s. Another cause of **s** insertions was the nonsimultaneous change between high- and low-frequency energy at obstruent constrictions. At an obstruent closure, for example, the disappearance of high-frequency energy before low-frequency energy may look like a vowel–nasal closure to the **s** detector.

E. B(urst) landmarks

The **b** deletion rate was higher than the **g** deletion rate. Being low-amplitude signals, bursts were often obscured by noise. For example, background noise or speech noise could mar the silence interval preceding a burst. Also, the energy changes associated with **b** landmarks were sometimes too weak to cause an energy discontinuity. For example, the frication noise followed by a stop closure, as in the syllable boundary of ''bathtub,'' sometimes did not cause a −opivot.

The **b** insertions were mainly due to extralingual noise during stop closures. A −**b** insertion just before 1.8 s in Fig. 8 marks the drop in energy after the [k] burst. A +**b** neutral marks a voiced stop burst for [b] at 1.4 s, and another marks the creak at 1.6 s.

TABLE V. Detection rates of the development set, by position with respect to reduced vowels. V=unreduced vowel, v=reduced vowel. The number of tokens is given in parentheses.

	Left-reduced vCV	Right-reduced VCv
altogether	98% (444)	87% (367)
+v fric	100% (41)	81% (59)
+v stop	100% (52)	80% (49)

F. Influence of vowel reduction

Some insight into the effect of prosody on landmark detection can be gained by considering the position of a landmark in relation to reduced vowels. Reduced vowel effects on the development set are summarized in Table V and described more fully by Liu (1995a). Closure and release landmarks in left-reduced position (reduced vowel to the left, unreduced vowel to the right) had a higher detection rate than landmarks in right-reduced position (unreduced vowel to the left, reduced vowel to the right). This contrast was especially pronounced in voiced fricatives and voiced stops. Flaps were not counted because there were not enough tokens.

The finding that landmarks in left-reduced environment have a higher detection rate than landmarks in right-reduced environment suggests that voiced obstruents are likely to be reduced in right-reduced environment. Consistent with this observation, the American English flapping rule states that alveolar stops are likely to be flapped in right-reduced environment.

An acoustic analysis of the various prosodic environments shows why landmarks in right-reduced environment are harder to detect. One acoustic factor that affects landmark detection is constriction duration. The shorter the duration, the more likely the acoustic changes at a constriction will be smoothed out. The average constriction duration of singleton consonants in left-reduced position is 89±13 ms, while in right-reduced position it is only 63±28 ms. Another acoustic factor affecting landmark detection is the amount of energy change at closure and release. The bigger the change, the taller the ROR peak will be, so the easier the detection. The change in the 20-ms smoothed, band 1 energy at closure and release was measured for all voiced obstruents. The band 1 energy change at landmarks in left-reduced position was 21±5 dB, while in right-reduced position it is only 16±6 dB.

G. Comparison to related work

Other researchers have tried to perform tasks similar to landmark detection for such purposes as phonetic recognition, automatic phone label alignment, and concatenative speech synthesis. These tasks are often referred to as ''segmentation,'' which was described in the Introduction. Direct comparison of landmark detection with segmentation is not possible because the philosophy and goals of the two tasks are different. Nevertheless, the results of some segmentation

work will be presented to set the results of landmark detection in context. The comparison will be made with the test results in Table III.

Some researchers have employed a single-level segmenter to generate a single hypothesis of what they consider to be the phonetic boundaries in a speech waveform. Their goal is to maximize the detection of phonetic boundaries while minimizing insertions. Because the sounds of speech have varying levels of abruptness, with consonantal segments being the most abrupt, semivowels being less abrupt, and vowels being essentially steady state, a single-level segmenter using one acoustic measure cannot increase detections without increasing insertions as well. Glass and Zue (1986) have demonstrated this trade-off. They used a segmentation algorithm which generated a boundary whenever the feature vector generated from the broadband signal changed sufficiently between two frames. Regardless of signal representation (hair cell response, critical band representation, or linear predictive coding), the segmentation error rate hovered around 30%. The hair cell response results, which Glass and Zue favored, gave a deletion rate of 23%, an insertion rate of 6%, and thus a total error of 29%. Their task required finding semivowel and diphthong boundaries, as well as voiced stop releases and subsequent onset of full glottal vibration. In order to compare their results to the landmark detection results, the following adjustments on the landmark detection results have to be made: (1) include semivowel and diphthong ''boundary'' markers, voicing onsets of voiced stops, and creaks in the hand-labeled set (these labels comprise about 25% of the total set); (2) count semivowel/diphthong insertions, creak insertions, and short stop burst neutrals as detections; (3) count voice bar neutrals as insertions; and (4) not count substitutions in the error rate. With these adjustments, the landmark error rate is 26% (deletion rate=22%, insertion rate=4%), which is slightly better than Glass and Zue's 29%. The deletion rate of the landmark detector could even be lower if the algorithm had been designed to find semivowels. Although this comparison is only a rough one, it shows that the landmark detector performs as well as, if not better than, a single-level segmenter.

V. SUMMARY AND CONCLUSIONS

Landmark detection is the first step in a proposed knowledge-based speech recognition system based on identifying distinctive features. A landmark detection algorithm was designed to find the abrupt-consonantal and abrupt landmarks. It incorporates measures of spectral abruptness and acoustic–phonetic information in a knowledge-based, hierarchical fashion to detect landmarks. The algorithm classified these landmarks into three types: **g**(lottis), **s**(onorant), and **b**(urst).

An experiment was performed using clean speech. Error rates were relatively low. The **g** landmarks, which correspond to stops and fricatives among others, were detected with a 98% detection rate. The **b** landmarks, which correspond to bursts and the cessation of noise, were detected with a 90% detection rate. The **b** detector was sensitive to nonspeech noise, such as lip smacks. Most of the deletions and insertions were due to **s** landmarks, which correspond to nasals and [l]'s. The large variation in the acoustic manifestations of vowel contexts and the sometimes glidelike nature of sonorant consonants made **s** landmarks difficult to detect.

A. Error analysis

Because landmark detection is a critical first step in the lexical access system, some way must be devised to take care of the errors in landmark detection. Of the errors, insertions can be removed when further processing for distinctive features in the vicinity of an insertion determines that the insertion is not a valid landmark. Many of these insertions are due to semivowel and diphthong transitions; therefore, they are a clue to the placement of nonabrupt and vocalic landmarks.

Substitutions are not real errors since they point to a desired landmark but assign the wrong type to it. Again, further processing can determine the true type of a substitution.

Deletions are the most serious kinds of error. The deletion rate was 8%. Mislabeling of the database may account for up to 2% in this 8%. This 2% figure was estimated from the number of hand-labeled landmarks whose existence was questionable. An example of a hand-labeled landmark in this 2% is an [n] closure in ''want.'' The closure may be coincident with the velopharyngeal port closure, but if these two events are mislabeled with separate landmarks, then a deletion is likely to result. The presence of nasality in the preceding vowel would indicate an underlying nasal segment whether or not a $-$**s** landmark is detected. As another case, segments which are phonologically [+consonantal] but realized nonabruptly may be mislabeled with abrupt-consonantal landmarks. This type of mislabeling is common for nasals and [l]'s. To some extent, the neutral category of landmarks reduces mislabeling errors by not counting certain kinds of insertions and deletions as errors when decisions about landmark labeling is ambiguous.

Another kind of deletion comes from not finding heavily voiced obstruents, flaps, nasals, and [l]'s. One percent in the deletion rate was from missing heavily voiced obstruents and flaps, and 2% was from missing nasals and [l]'s. The landmarks of these segments are hard to find because energy abruptness in the expected frequency bands may be compromised by a voice bar, the vowel context, or the speed of closure and release. These deletions can potentially be captured by a nonabrupt landmark detector, which puts a landmark near the energy minimum of a constriction for semivowels. Further analysis in the area, to find nasality based on pole–zero pairs or widened first-formant bandwidth, for example, may then be able to identify the underlying segment.

Missing **b** landmarks accounted for some deletions. One percent out of the 8% deletion rate was from not finding sufficient silence during closure. This error can be reduced by using a broadband measure for silence with one common reference level rather than using a different reference for each band.[5] Another 1% in the deletion rate was due to not finding sufficient energy abruptness at the bursts or closures. A better SNR, attained possibly through noise-cancelling preprocessing, could prevent these deletions.

B. Use of prosody

Adding prosodic information could improve performance. If the duration between two detected landmarks is too long, suggesting that a landmark was missed in between, then the ROR peak threshold could be lowered to detect more subtle spectral changes in that region. Knowing where reduced vowels are could help since reduction affects the acoustics around a neighboring landmark, as shown in Sec. IV F. Prosodic phrase boundary information could be used to customize the landmark detector to a change in fundamental frequency, a decreased waveform intensity, and a lengthened syllable.

C. Articulator-free features

In the process of landmark detection, inferences about some of the articulator-free features at a landmark can be made. Constraints on the articulator-free features, and thus broad phonetic class, arise. An **s** landmark carries with it the features [+consonantal] and [+sonorant]. A **b** landmark carries with it the features [+consonantal], [−sonorant], and [−continuant]. A **g** landmark is ambiguous, since it could be [+consonantal] or [−consonantal]. If a stop closure or release is causing the **g** landmark, for example, then the landmark would carry the features [+consonantal], [−sonorant], and [−continuant]; however, if an [h] or glottal stop is causing the landmark, then it would carry the feature [−consonantal]. The presence or absence of formant movements around the **g** landmark would resolve this ambiguity.

Outer **AC** landmarks come in closure–release pairs except at the beginning and end of an utterance. Once the value of the feature [consonantal] is ascertained, the outer **AC** landmarks can be identified and paired with each other. Knowing the time duration between two **AC** landmarks will aid in distinguishing between singleton consonants and consonant clusters.

D. Nonabrupt and vocalic landmarks

The landmark detector described here finds abrupt and abrupt-consonantal landmarks. Algorithms to detect the nonabrupt landmarks, for semivowels, and the vocalic landmarks, for vowels, still need to be developed. The landmark detector presented here can help find the remaining landmarks. The **g** landmarks already delimit the voiced regions within which the nonabrupt and vocalic landmarks have to occur. Pairs of pivots found by the **s** detector could be surrounding a nonabrupt landmark, especially if these pivots fail the steady-state and abruptness criteria.

E. Lexical access

Once the articulator-free features are completely identified around each landmark, further processing can be customized to the broad phonetic environment to identify the articulator-bound features at each landmark. Following feature extraction, the landmarks with their feature bundles need to be collapsed together in some cases and expanded in other cases so that a one-to-one relation exists between a segment and a bundle of features. An example in which two feature bundles at two landmarks would have to be collapsed into one feature bundle to represent one segment is the closure and release landmarks of an intervocalic [b]. An example in which one feature bundle at one landmark would have to be expanded into two bundles is the [br] release in ''bright,'' in order to account for the two segments [b] and [r]. Procedures would have to be introduced to account for the different ways one could pronounce a word. One approach would be to construct a pronunciation network for each word having multiple pronunciations. This pronunciation network would be described at the distinctive feature level, since one of the advantages of distinctive features is their ability to describe modifications concisely. Another way to describe modification is by identifying all modifiable features in the lexicon and the contexts in which modification occurs, and then placing reduced weight on discrepancies in these features during the lexical matching process, based on the context. Once the pronunciation network is in place or the modifiable features are marked, lexical access using a sequence of feature bundles is a straightforward process.

ACKNOWLEDGMENTS

I am grateful to Professor Kenneth N. Stevens for his guidance throughout this research and the writing of this paper. I also thank Dr. James R. Glass and two anonymous reviewers for their helpful suggestions. This work was partially supported by a grant from NSF and by the Clarence J. LeBel fund.

[1]Knowledge-based systems also suffer from changes in operating environment. However, the difference is that a knowledge-based approach directly measures the critical changes that have occurred (e.g., F_0 changes, new microphone characteristics) and then modifies its processing parameters to suit the new environment. This approach is conceivably faster than a purely statistical approach, in which the system would have to be presented with substantial data from the new environment.

[2]A *segment* is used in this paper to refer to a bundle of distinctive features which describe a speech sound and which have acoustic correlates. It does not refer to a physical slice of the acoustic signal.

[3]Segmentation does not involve the further step of associating any phonetic significance to these pieces. The assignment of phonetic labels is done by a subsequent classification stage.

[4]Creaks have linguistic significance. For example, they can signal word boundaries (Umeda, 1978).

[5]In the TIMIT experiment, this technique was employed and did indeed decrease the error.

Andre-Obrecht, R. (**1988**). ''A new statistical approach for the automatic segmentation of continuous speech signals,'' IEEE Trans. Acoust. Speech Signal Process. **36**, 29–40.

Beckman, M. E. (**1986**). *Stress and Non-stress Accent* (Foris, Dordrecht, The Netherlands).

Das, S., Bakis, R., Nadas, A., Nahamoo, D., and Picheny, M. (**1993**). ''Influence of background noise and microphone on the performance of the IBM Tangora speech recognition system,'' IEEE Proc. Int. Conf. Acoust. Speech Signal Process. **2**, 71–74.

Fisher, W. M., Doddington, G. R., and Goudie-Marshall, K. M. (**1986**). ''The DARPA speech recognition research database: specifications and status,'' in *DARPA Speech Recognition Workshop Proceedings* (Science Applications International Corporation), pp. 93–99.

Flammia, G., Dalsgaard, P., Andersen, O., and Lindberg, B. (**1992**). ''Segment based variable frame rate speech analysis and recognition using spectral variation function,'' in *Proceedings of the International Conference on Spoken Language Processing*, edited by J. Ohala, T. Nearey, B. Derwing, M. Hodge, and G. Wiebe (Personal Publishing Ltd., Edmonton, Canada), Vol. 2, pp. 983–986.

Furui, S. (**1986**). ''On the role of spectral transition for speech perception,'' J. Acoust. Soc. Am. **80**, 1016–1025.
Gish, H., and Ng, K. (**1993**). ''A segmental speech model with applications to word spotting,'' IEEE Proc. Int. Conf. Acoust. Speech Signal Process. **2**, 447–450.
Glass, J. R. (**1988**). ''Finding acoustic regularities in speech: applications to phonetic recognition,'' Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.
Glass, J. R., and Zue, V. W. (**1986**). ''Signal representation for acoustic segmentation,'' in *Proceedings of the First Australian Conference on Speech Science and Technology* (Australian National University Printing Service, Canberra, Australia), pp. 124–129.
Jakobson, R., Fant, G., and Halle, M. (**1952**). ''Preliminaries to speech analysis,'' Technical Report 13, MIT Acoustics Laboratory.
Klatt, D. H. (**1976**). ''Linguistic uses of segmental duration in English: Acoustic and perceptual evidence,'' J. Acoust. Soc. Am. **59**, 1208–1221.
Lee, K.-F. (**1989**). *Automatic Speech Recognition: The Development of the SPHINX System* (Kluwer, Norwell, MA).
Lisker, L., and Abramson, A. S. (**1964**). ''A cross-language study of voicing in initial stops: Acoustical measurements,'' Word **20**, 385–422.
Liu, S. A. (**1995a**). ''The effect of vowel reduction on landmark detection,'' Proc. Int. Congr. Phonet. Sci. **4**, 136–139.
Liu, S. A. (**1995b**). ''Landmark detection for distinctive feature-based speech recognition,'' Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.
Marcus, J. (**1993**). ''Phonetic recognition in a segment-based HMM,'' IEEE Proc. Int. Conf. Acoust. Speech Signal Process. **2**, 479–482.
Mermelstein, P. (**1975**). ''Automatic segmentation of speech into syllabic units,'' J. Acoust. Soc. Am. **58**, 880–883.
Ohde, R. N. (**1994**). ''The developmental role of acoustic boundaries in speech perception,'' J. Acoust. Soc. Am. **96**, 3307.
Sagey, E. (**1986**). ''The representation of features and relations in nonlinear phonology,'' Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.
Stevens, K. N. (**1985**). ''Evidence for the role of acoustic boundaries in the perception of speech sounds,'' in *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, edited by V. Fromkin (Academic, New York), pp. 243–255.
Stevens, K. N. (**1994**). ''Prosodic influences on glottal waveform: preliminary data,'' in *Proceedings of the International Symposium on Prosody* (Japan Society for the Promotion of Science, Yokohama, Japan), pp. 53–64.
Stevens, K. N., Manuel, S. Y., Shattuck-Hufnagel, S., and Liu, S. (**1992**). ''Implementation of a model for lexical access based on features,'' in *Proceedings of the International Conference on Spoken Language Processing*, edited by J. Ohala, T. Nearey, B. Derwing, M. Hodge, and G. Wiebe (Personal Publishing Ltd., Edmonton, Canada), Vol. 1, pp. 499–502.
Umeda, N. (**1978**). ''Occurrence of glottal stops in fluent speech,'' J. Acoust. Soc. Am. **64**, 88–94.
van Beinum, F. J. K. (**1994**). ''What's in a schwa?,'' Phonetica **51**, 68–79.
Weinstein, C. J., McCandless, S. S., Mondshein, L. F., and Zue, V. W. (**1975**). ''A system for acoustic-phonetic analysis of continuous speech,'' IEEE Trans. Acoust. Speech Signal Process. **23**, 54–67.
Zue, V. W., and Laferriere, M. (**1979**). ''Acoustic study on medial /t,d/ in American English,'' J. Acoust. Soc. Am. **66**, 1039–1050.
Zue, V. W. and Seneff, S. (**1990**). ''Transcription and alignment of the TIMIT database,'' in *Recent Research Toward Advanced Man-Machine Interface through Spoken Language*, edited by H. Fujisaki (Steering Group of the Priority Area Research, Tokyo), pp. 464–473.
Zue, V., Seneff, S., and Glass, J. (**1990a**). ''Speech database development at MIT: TIMIT and beyond,'' Speech Commun. **9**, 351–356.
Zue, V. W., Glass, J. R., Goodine, D., Leung, H., Phillips, M., Polifroni, J., and Seneff, S. (**1990b**). ''Recent progress on the SUMMIT system,'' presented at the Third DARPA Speech and Natural Language Workshop.