THE EFFECT OF VOWEL REDUCTION ON LANDMARK DETECTION

Sharlene A. Liu
Research Lab of Electronics, Dept. Electrical Engineering & Computer Science
MIT, Cambridge, MA 02139, USA
email: liu@lexic.mit.edu

ABSTRACT

In the speech waveform, *landmarks* guide the search for the underlying distinctive features. The landmark detection rate by an automatic algorithm was 94%. An analysis of the prosodic environments in which the landmark detector failed showed that a right-reduced vowel environment caused more misses than other prosodic environments. Consonantal duration was shorter and energy change was smaller in this environment.

1. INTRODUCTION

The proposed model of lexical access uses *landmarks* to guide the search for *distinctive features* [1]. Fig. 1 shows a flow diagram of the lexical access system. In the speech waveform, landmarks are salient points around which important acoustic cues identify the underlying distinctive features. They appear to be perceptual foci, and specify times when certain articulatory targets are to be achieved [2]. After landmarks are detected, distinctive features are extracted in the vicinity of each landmark. The feature specifications associated with each landmark are then organized into a sequence of segments, and the lexicon is accessed by features.

A landmark detection algorithm was developed to automatically locate acoustically-abrupt landmarks [3]. Fig. 2 shows a spectrogram with the acoustically-abrupt landmarks indicated. Acoustically-abrupt landmarks are typically consonantal closures and releases, and other spectral discontinuities caused by velopharyngeal port and vocal fold activity. The algorithm detected most of the desired landmarks, but missed some. In order to understand the circumstances under which it misses landmarks and to improve on the landmark detector, a study of the effect of vowel reduction on landmark detection was conducted. This paper presents the landmark detection algorithm, results of landmark detection, reduced vowel effects, and an acoustic analysis of various reduced vowel environments.

2. LANDMARK DETECTION

This section describes a landmark detection experiment. The database used, the details of the algorithm, and the results will be presented.

2.1 Database

Four speakers (2 female, 2 male) read sentences naturally and clearly. The utterances were recorded with an omnidirectional microphone and digitized at 16 kHz. The signal-to-noise ratio was 30 dB. The acoustically-abrupt landmarks in the utterances were hand-labeled according to the phonetic type of the segments in the vicinity of the landmark (e.g. vowel-stop, vowel-nasal), and whether the landmark designated a closure or release. The reduced vowels (typically /ə/s, syllabic /l/s, syllabic nasals, and sometimes /ɚ/) were also labeled. All other vowels, stressed or otherwise, were considered unreduced.

2.2 Detection Algorithm

The landmark detection algorithm relies on spectral discontinuity and acoustic-phonetic knowledge. It is divided into two stages: general processing and landmark type-specific processing. The output of the algorithm is a series of landmarks specified by time and type.

In general processing, a short-time Fourier transform magnitude (STFTM) is computed and smoothed over 20 ms to remove variations due to glottal pulses and random noise. The spectrum is divided into six bands: 0-0.4, 0.8-1.5, 1.2-2, 2-3.5, 3.5-5, and 5-8 kHz. Band 1 (0-0.4 kHz) keeps track

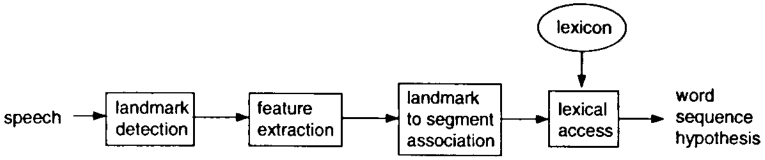

Figure 1: *Flow diagram of the proposed lexical access system.*

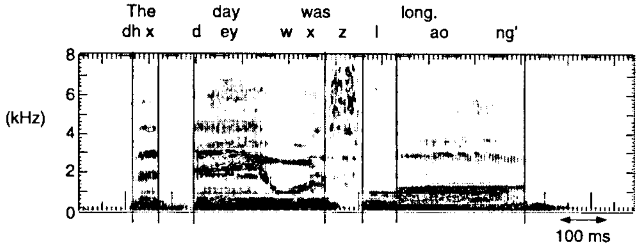

Figure 2: *Spectrogram with acoustically-abrupt landmarks indicated by vertical lines.*

of the turning on and off of voicing. The mid-frequency bands keep track of spectral changes due to sonorant consonantal segments, as well as bursts and the cessation of noise after bursts. In each band, the energy and its derivative are calculated. The peaks in the derivative represent times of abrupt spectral change in a band.

In the landmark type-specific processing stage, the peaks in the derivative direct processing to find three types of landmarks. These three types are: **g**(lottis), which marks the beginning and end of glottal vibration, **s**(onorant), which marks sonorant consonantal closures and releases, and **b**(urst), which designates stop or affricate bursts and points where aspiration or frication ends due to a following stop closure. The **g** landmarks are found from Band 1 peaks. Pairing of landmarks at voicing onset and offset and a minimum syllable requirement are imposed. The **s** landmarks are found from Bands 2-5 peaks during voiced regions delimited by **g** landmarks. A steady-state requirement during the closure of a sonorant consonantal segment and a sufficient high-frequency abruptness are imposed. The **b** landmarks are found from Bands 2-5 peaks during the unvoiced regions delimited by **g** landmarks. A silence period during the closure of a [-continuant] segment, is required.

Table 1: *Results of landmark detection.*

# Tokens	1597
Deletion	4%
Substitution	2%
Insertion	10%
Total error	16%

2.3 Results of Detection

Table 1 shows the results of running the landmark detection algorithm on the database. A landmark is considered correctly detected if it is within 30 ms of the hand-labeled landmark and is of the correct type (**g**, **s**, **b**). A *deletion* is a missed landmark. A *substitution* is within 30 ms of the hand-labeled landmark but misidentified by type. An *insertion* is a false landmark. The rates in Table 1 were calculated by dividing by the number of tokens.

From the deletion and substitution rates, one sees that 94% of the landmarks were correctly detected. In terms of phonetic category, almost 100% of the unvoiced obstruents were detected. Voiced obstruents were somewhat more problematic, because voice bars reduce energy abruptness in

Band 1. The algorithm's detection of nasals and [l]s was also lower. One reason is that they are often implemented in a glide-like fashion, so that spectral change is not very abrupt. This is especially true for [l]s. Another reason is that the s detector is somewhat context-dependent. Sonorant consonantal segments next to high, back vowels did not always produce a sufficiently large change in energy. The b landmarks were detected well for the most part; however, weak bursts and noisy stop closure intervals caused some b deletions.

3. PROSODIC EFFECTS

In this section, the effect of vowel reduction on landmark detection in VCV sequences is considered. Table 2 organizes the landmark detection rate by prosodic context. Landmarks occur singularly or in clusters between two vowels. A landmark in *left-reduced* environment means that the preceding vowel is reduced while the succeeding vowel is unreduced. A landmark in *right-reduced* environment is the opposite. *Both reduced* means that both vowels are reduced. *Neither-reduced* means that both vowels are unreduced. The largest error rate occurred for landmarks in right-reduced position, while the smallest error rate occurred for landmarks in left-reduced position. Because the right-reduced environment is the flapping environment for alveolar stops in American English, there is reason to believe this environment causes consonants, in general, to be reduced.

3.1 Constriction duration

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 harder the landmark is to detect. The landmark detector relies on detecting energy change. If a constriction is too short, the energy change may be de-emphasized by the smoothing during the 6-band energy calculation. For voiced obstruents, voice bars deemphasize the energy change in Band 1 even more. The first row in Table 3 shows the average constriction duration of singleton consonants in the four prosodic environments. The constriction duration in right-reduced environment is shortest, explaining in part why landmarks in this environment have the lowest detection rate. The constriction duration in left-reduced environment is longest, resulting in a higher detection rate. The durations in both-reduced and neither-reduced environments are in between, in agreement with the detection rates. Of relevance, Turk [4] showed that, within a word, stop consonant constriction durations are shorter in left-stressed environments than in right-stressed environments.

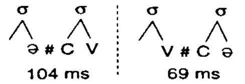

Figure 3: *The effect of vowel reduction and syllable affiliation on consonant duration.*

In the above analysis, the syllable affiliation of the consonant was not taken into account. It has been hypothesized that reduction is syllable-affiliated, so that the effect of vowel reduction on a consonant is greater if that consonant belongs to the same syllable as the reduced vowel than if it belongs to a different syllable. To test out this hypothesis, the constriction durations of Table 3 were grouped according to the word affiliation of the consonant. Consonants in word-medial position were not used because of the difficulty of deciding their syllable affiliation. The duration of consonants in left-reduced and right-reduced environments was noted when the consonant was affiliated with the right vowel. Fig. 3 illustrates the two cases considered. Consonants in left-reduced environment had an average duration of 104 ± 30 ms, while in right-reduced environment the average duration was 69 ± 25 ms. The average difference is 35 ms, which is bigger than the 26 ms difference when syllable affiliation was not considered. This finding supports the hypothesis that consonant reduction is affected not only by neighboring vowel reduction, but by the syllable affiliation of the consonant to the neighboring vowels as well.

3.2 Energy change

In addition to constriction duration, the amount of energy change at closure and release also affects landmark detection. The bigger the change, the easier the detection, and vice versa. The change in the 20 ms–smoothed, Band 1 energy at closure and release was measured for all voiced obstruents. An energy change at closure was measured by subtracting the lowest energy level (in dB) during the constriction from the energy at a point directly preceding the closure transition in the vowel. At release, the measurement is made with the succeeding vowel. The second row in Table 3 shows that the energy change is 5 dB less, on average, for voiced obstruents in right-reduced environment than in left-reduced environment. The energy changes in the other prosodic environments were in between. This gradation in energy change is consistent with the landmark detector's performance in the four prosodic environments.

Table 2: *Landmark detection rate, grouped by position with respect to reduced vowels. The number of tokens is given in parentheses.*

	Left-reduced *vCV*	*Right-reduced* *VCv*	*Both-reduced* *vCv*	*Neither-reduced* *VCV*
altogether	98% (444)	87% (367)	96% (134)	89% (390)
+v fric	100% (41)	81% (59)	100% (13)	87% (77)
+v stop	100% (52)	80% (49)	100% (22)	95% (59)

Table 3: *Constriction duration and voiced obstruent low-frequency energy change at constriction, grouped by position with respect to reduced vowels. The number of tokens is given in parentheses.*

	Left-reduced *vCV*	*Right-reduced* *VCv*	*Both-reduced* *vCv*	*Neither-reduced* *VCV*
constriction duration	89 ± 13 ms (151)	63 ± 28 ms (111)	76 ± 21 ms (38)	67 ± 28 ms (106)
+v obstruent energy change	21 ± 5 dB (120)	16 ± 6 dB (144)	18 ± 5 dB (54)	17 ± 6 dB (144)

4. CONCLUSION

In this paper, the effect of neighboring reduced vowels on landmark detection was studied. Landmark detection is the first step of a proposed lexical access system. It was found that landmarks in right-reduced environment tended to be missed more often than in other prosodic environments, notably the left-reduced environment. An acoustic analysis showed that, in right-reduced environments, the consonantal constriction duration tended to be shorter and the amplitude change smaller than in other prosodic environments. The effect is amplified when syllable affiliation of the consonant to the neighboring vowels is considered.

ACKNOWLEDGEMENTS

I thank Ken Stevens and Stefanie Shattuck-Hufnagel for their guidance. This research was partially supported by NSF.

REFERENCES

[1] K. N. Stevens, S. Y. Manuel, S. Shattuck-Hufnagel, S. Liu, "Implementation of a model for lexical access based on features", *Proc. ICSLP*, 1992, pp. 499-502.

[2] K. N. Stevens, "Evidence for the role of acoustic boundaries in the perception of speech sounds", *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, ed. V. Fromkin, Academic Press, New York, 1985, pp. 243-255.

[3] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition", *Ph.D. Thesis*, Dept. Electrical Engineering and Computer Science, MIT, May 1995.

[4] A. Turk, "The American English flapping rule and the effect of stress on stop consonant durations", *Working Papers of the Cornell Phonetics Laboratory*, March 1992, pp. 103-134.