Robustness of Acoustic Landmarks in Spontaneously-Spoken American English

Stefanie Shattuck-Hufnagel, Nanette M.Veilleux

Speech Group, RLE, Massachusetts Institute of Technology
stef@speech.mit.edu, veilleux@simmons.edu

ABSTRACT

Acoustic landmarks (abrupt changes associated with consonant closures and releases, vowels and glides) play an important role in some models of lexical access (e.g. Stevens 1998, 2002), so it is important to determine how often they actually survive the rigors of articulatory overlap and weakening in spontaneous speech production. A corpus of spontaneous American English speech was collected from 8 adult female speakers and hand labeled for the occurrence of landmarks. Preliminary results for one conversation (240 secs., 610 words, analysis completed for 1003 of 2750 predicted landmarks) show that 86% of landmarks were realized overall, with a sharply lower rate for coronal stops /t/ and /d/. These results suggest that the majority of landmarks are available for detection both by human listeners and automatic recognition algorithms. Ongoing analyses are comparing the rate of automatic detection of these acoustic events with the hand labels, and tabulating the relatively limited set of contexts in which predicted landmarks are lost or changed.

Keywords: landmarks, distinctive features, lexical access, feature cues, articulatory overlap.

1. INTRODUCTION

Acoustic speech signals contain regions of abrupt change caused by actions of the articulatory tract, such as consonant closures and releases. These regions, which can be called landmarks, play an important role in many models of speech processing. For example, Stevens (1998, 2002) proposes a distinctive-feature-based model of human speech processing in which landmarks play several critical roles. First, they provide information about a particularly important class of distinctive features: the articulator-free features (Halle 1990, Stevens and Keyser to appear), which correspond roughly to the manner features. By specifying the nature and serial order of the closures and releases for consonants, and the extremum landmarks for glides and vowels, the landmarks allow the listener to formulate an initial representation of the CV segmental structure of the utterance or phrase based on very early processing of the incoming signal. Second, this preliminary estimate of the articulator-free features, e.g. [consonant], [sonorant], [continuant], [strident], [vowel] and [glide], places strong constraints on the set of remaining features that must be recognized for each segment. For example, if the landmark is a vowel landmark, no analysis for acoustic cues to the feature [strident] need be carried out in that region. Thus landmark detection constrains the type of further information that the processor must look for in the signal, by specifying the articulator-free features. Third, the landmark string specifies locations where the signal is particularly rich in information about those additional features i.e. about the articulator-bound features of voicing and place. This knowledge facilitates efficient further processing for feature cues, making it unnecessary to compute values for every parameter at every location. Finally, the initial CV representation that the listener forms on the basis of the landmarks, although incomplete, can serve as the organizing framework for the listener's processing of other aspects of the utterance, such as words, phrasal groupings and prominence patterns. This may provide information (e.g. about lexical stress) that can help to constrain lexical access processing, and in addition may allow some kinds of higher-level prosodic processing to begin before a complete representation of the features, segments and words of the utterance has been formulated.

Because landmarks play such a critical role in this feature-cue-based processing model, it is important to know how often they are present in the signal in order to evaluate the model. This is a particularly significant issue for informal continuous speech, because this kind of speech shows reduction and articulatory overlap leading to phonetic variation that has often been described as omission or assimilation of entire segments as well as other changes, some of them quite extreme. If as a result a large proportion of the predicted acoustic landmarks for the words of an utterance are not implemented, this feature-based model will face challenges. On the other hand, if most of the landmarks are implemented in recognizable form, it may be possible to develop ways to deal with the small proportion that are missing, particularly if landmark loss occurs in a small number of predictable contexts, and cues to other features are preserved in adjacent regions of the signal defined by landmarks for nearby segments. This paper reports initial results from a study of how often speakers implement the predicted landmarks in a corpus of task-directed American English speech, and of the contexts where landmarks are either changed from their predicted form, or lost.

2. METHODS

The larger study (of which this paper reports a part) involves hand- and automatic labeling of the predicted landmarks in a corpus of more than one hour of task-directed spontaneous speech, elicited using the Maptask method (Anderson et al. 1991).

2.1. Recording the corpus

The corpus consists of 16 4-7 minute dialogues in which an instruction giver describes a path on the map that is visible only to her. The instruction follower has a similar map, on which she must sketch the path according to the verbal instructions provided by the giver. The experiment is designed to encourage speech with a natural quality by the fact that both speakers are well acquainted with each other before the experiment begins, and by small differences between the two maps which require some interaction to resolve.

2.1. Labelling

2.2.1 Word labels

The 16 conversations were orthographically transcribed by Dr. Lisa Lavoie. To date, only the instruction givers' speech is included in the corpus, largely because (perhaps due to the nature of the task) the giver did most of the talking. The transcribed words were roughly aligned with the speech signal by creating an interval Textgrid in Praat (www.fon.hum.uva.nl/praat/). Alignments were only approximate, because the articulations for two adjacent words often overlap (as for *in the*, Manuel 1995), and because word boundaries sometimes occur when there is no acoustic signal (e.g. between 2 stops, as in *get to*). Events such as silences, breaths, laughter, speech by the other speaker and overlapping speech were also marked in this tier. Consistent with Shattuck-Hufnagel and Veilleux (2000), 45% of the 610 words were content words (nouns, verbs, adjectives and regular adverbs) while 55% were function words or interjections (the interjections, such as *oh*, *yeah* and *um,* made up 10% of the words).

2.2.2 Landmark labels by hand

For each phonemic segment of each transcribed word, the landmarks predicted for a citation form of the word were tabulated: for each consonant, a closure and release landmark; for each glide, a single landmark corresponding to the articulatory extremum of narrowing, and for each vowel, a single landmark corresponding to the articulatory extremum of widening. Each of these predicted landmarks was annotated as i) implemented in the speech signal as predicted, ii) changed to a different type of landmark, or iii) not implemented. An example of a changed landmark would be a /k/ realized as a fricative rather than as a stop (Lavoie 2002); this articulation provides an acoustic cue for a different value of an articulator-free feature, i.e. [+continuant] rather than [-continuant]. An example of a non-implemented landmark would be a /d/ in an /nd/ sequence, such as *and a*, where the intervocalic oral closure is nasalized throughout, shows no diminution of amplitude as expected in the voice bar of a stop, and there is no evidence for a stop release noise (which would indicate pressure buildup behind a constriction followed by release). To determine whether landmarks were implemented as expected, labelers used a combination of visual inspection of the wave form and spectrogram, and listening. This labeling method relies on the assumption that Stevens' (1998, 2002) acoustic definitions for landmarks align with visual discontinuities in these displays.

Landmark labeling by hand was carried out over a period of several years by a number of participants in the MIT Undergraduate Research Opportunities

Program. Most were undergraduates in the Department of Electrical Engineering and Computer Science with some training in signal processing methods. Initial training sessions were supplemented by weekly meetings with the first author, to discuss difficult cases. Each conversation was labeled by one labeler and checked by a second, to ensure that all predicted landmarks were accounted for and conventions followed. In ongoing analyses, each predicted landmark label is checked by the first author; data presented here have been checked in this way.

Predicted landmarks were labeled in one of four ways. If in the opinion of the labeler the signal (as viewed in the wave form and spectrographic displays available in Praat) showed good evidence for the landmark in its expected form, the landmark was labeled as 'implemented'. If the signal showed some possible evidence for the landmark, it was labeled as 'probably implemented'. Both of these types of observations were recorded on the Landmarks labeling tier, which as a result contained only labels for predicted landmarks for which there was appropriate evidence in the signal. A second tier, the Comments Tier, was reserved for the other two cases. If there was no discernable evidence for a predicted landmark visible in the displays and the segment was not audible to the transcriber, the landmark was labeled as 'missing'. Finally, if the listener could hear the segment but could not find a reasonable-looking candidate acoustic discontinuity for the landmark, it was labeled as 'possibly not implemented'.

These four basic types of labels were supplemented by a number of additional diacritics. For example, if a pair of closure and release landmarks for a consonant were both missing, but were replaced by a different type of landmark (such as a fricative closure and release for a /k/ in place of the predicted stop closure and release, or a glide extremum for a /g/ instead of a stop closure and release), the nature of the replacement landmarks was noted in the Comments tier. Similarly, if a stop closure landmark was invisible in the signal because it was preceded by a silence, but the burst at release clearly indicated that closure and subsequent pressure buildup had occurred, the closure landmark was labeled as missing but known to have occurred. Other such diacritics are described in the conventions for hand-labelling of landmarks (in preparation).

2.2.3 *Automatic landmark labels*

Because it is important to compare automatic detection of acoustic landmarks with these hand labels, we have also applied Liu's (1996) algorithm for consonant landmark detection to the corpus dialogues. These autodetected landmarks are currently being compared with the hand labels.

2.2.4 *Other labels*

As part of the ongoing study of the contexts in which landmark change or loss can occur, we are labeling the 16 conversations for prosodic phrasing and prominence (using the ToBI system of Tones and Break Indices, www.ling.ohio-state.edu/~tobi/); for syntactic word type (using a quasi-exhaustive list of function words drawn from the Brown Corpus, Shattuck-Hufnagel and Veilleux 2000); and for word frequency. Because the results reported below do not make use of these labels, they will not be discussed further here.

Preliminary results for the instruction giver's speech in the first conversation are reported here. This conversation occupied 4 minutes (including pauses and responses from the instruction follower), and contained 610 words spoken by the instruction giver, with 2750 predicted landmarks. Results for the first 1003 landmarks are described below. These landmarks occurred in 210 words, for a mean number of landmarks per word of 4.8.

3. RESULTS

Of the first 1003 predicted landmarks in this conversation, 858 or 86% showed strong or probable evidence in the spoken signal. Of the remainder, 6% were changed to a different type of landmark (e.g. a stop closure became a fricative closure or a glide), and 8% were apparently not implemented. Thus, a very substantial proportion of the landmarks are available to specify locations for further acoustic processing, estimate the CV structure of the utterance, and provide cues to the articulator-free features of those candidate segments, if they can be detected. However, 14% of the predicted landmarks are not available for these purposes. Thus, it is important to know whether the 14% of missing landmarks fall into a small number of classes, which could perhaps be predicted and dealt with in a recognition model or algorithm. In the next two sections we summarize the contexts in which landmarks were either changed or apparently not implemented.

3.1. Landmarks that were changed

Changes leading to different types of landmarks than those predicted from the articulator-free lexical features of a word usually involved some kind of reduction or lenition. The most common processes included glottalization of final /t/ (affecting 18 landmarks), lenition of /g/ to a glide (10 landmarks), and flapping of a /t/ (8 landmarks). Another common process, which may reflect the difficulty of maintaining the precise articulation necessary for a fricative, was the production of the inter-dental voiced fricative /dh/ as a stop (9 tokens). This process is rather common in American English (Zhao 2006). Other changes included frication of /k/, flapping of /d/, and production of /d/ or /dh/ as a glide (i.e. with a gradual diminution and increase in amplitude, without a discretely-identifiable closure period). Each of these changes is well-attested in American English; the ongoing analysis of this corpus is aimed at specifying the segmental, prosodic and frequency contexts in which they occur.

3.2. Landmarks that were omitted

Like processes that changed one type of landmark into another, processes that led to the omission of landmarks by this speaker appear limited in number. Most common was the loss of /d/ landmarks in word-final /nd/ sequences (eliminating 26 landmarks). Others included the loss of medial release and closure landmarks in a stop-stop sequence such as *get to*, and loss of reduced-vowel syllables or rimes, such as the *–ow-* in *towards* or the *be-* in *because*. These latter pronunciations might be described as alternative lexical entries, but we have chosen to describe them as processes that omit landmarks, in order to provide a more stringent test of the hypothesis that most of the landmarks predicted from the most complete lexical representation are implemented.

4. CONCLUSIONS

This analysis of more than 1000 predicted consonant, glide and vowel landmarks in spontaneous speech produced by a single speaker showed that most landmarks were implemented as predicted; those that were changed or lost fell into a small number of predictable classes. Ongoing analyses of automatic detection of these abrupt acoustic changes (and of the contexts in which their change and loss are likely) in the remaining corpus will test the generality of the findings. These early data are consistent with Stevens' (1998, 2002) proposal that landmarks are robust even in continuous communicative speech, so that human listeners can use them as cues to the articulator-free features of the speaker's intended segments and words. Additional data from other speakers, as well as from speakers of other languages, will be required to provide a fully-adequate test of the hypothesis that acoustic landmarks provide reliable cues to articulator-free features in natural speech.

5. REFERENCES

[1] Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S. and Weinert, R. 1991. The HCRC Map Task Corpus. *Language and Speech* 34, 351-366

[2] Halle, M. 1990. Features. In W. Bright, Oxford International Enchclopedia of Linguistics. New York: Oxford University Press

[3] Lavoie, L. 2002. Subphonemic and suballophonic consonant variation: The role of the phoneme inventory. ZAS Papers in Linguistics 28

[4] Liu, S. 1996. Landmark detection for distinctive feature-based speech recognition. *J. Acoust. Soc. Amer.*100, 3417-3430

[5] Manuel, S.Y. 1995. Speakers nasalize /ð/ after /n/ but listeners still hear /ð/. *J. Phonetics* 43, 453-476

[6] Stevens, K. 1998. *Acoustic Phonetics*. Cambridge, MA: MIT Press

[7] Stevens, K.N. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *J. Acoust.Soc. Amer.* 111, 1872-1891

[8] Stevens, K.N. and Keyser, S.J. (to appear). Quantal theory, enhancement and overlap. In Clements, N. and Ridouane, R., Proceedings of the Meeting on Quantal Theory, The Sorbonne, Paris, July 2006

[9] Zhao, S. Y. (2006). Contextual Effects on the Continuancy of /ð/. *J. Acoust. Soc. America* 119(5), 3300.

Acknowledgments: We thank Dr. Robert Ladd and his colleagues at Edinburgh University for design of the Maptask stimuli and maps; Dr. Olga Goubanova for eliciting the speech and organizing the database, Dr. Liaa Lavoie for creating orthographic transcriptions, Chi-Yun Park for providing the automatically detected landmarks, and Alejna Brugos of Boston University's Applied Linguistics Program as well as Yelena Yasinnik, Rob Speer and many other participants in the MIT Undergraduate Research Program for their tireless efforts in hand-labelling the words, prosody and acoustic landmarks in this database.