4
Lenition of /h/ and glottal stop

JANET PIERREHUMBERT and DAVID TALKIN

4.1 Introduction

In this paper we examine the effect of prosodic structure on how segments are pronounced. The segments selected for study are /h/ and glottal stop /ʔ/. These segments permit us to concentrate on allophony in source characteristics. Although variation in oral gestures may be more studied, source variation is an extremely pervasive aspect of obstruent allophony. As is well known, /t/ is aspirated syllable-initially, glottalized when syllable-final and unreleased, and voiced throughout when flapped in an intervocalic falling stress position; the other unvoiced stops also have aspirated and glottalized variants. The weak voiced fricatives range phonetically from essentially sonorant approximants to voiceless stops. The strong voiced fricatives exhibit extensive variation in voicing, becoming completely devoiced at the end of an intonation phrase. Studying /h/ and /ʔ/ provides an opportunity to investigate the structure of such source variation without the phonetic complications related to presence of an oral closure or constriction. We hope that techniques will be developed for studying source variation in the presence of such complications, so that in time a fully general picture emerges.

Extensive studies of intonation have shown that phonetic realization rules for the tones making up the intonation pattern (that is, rules which model what we do as we pronounce the tones) refer to many different levels of prosodic structure. Even for the same speaker, the same tone can correspond to many different F_0 values, depending on its prosodic environment, and a given F_0 value can correspond to different tones in different prosodic environments (see Bruce 1977; Pierrehumbert 1980; Liberman and Pierrehumbert 1984; Pierrehumbert and Beckman 1988). This study was motivated by informal observations that at least some aspects of segmental allophony

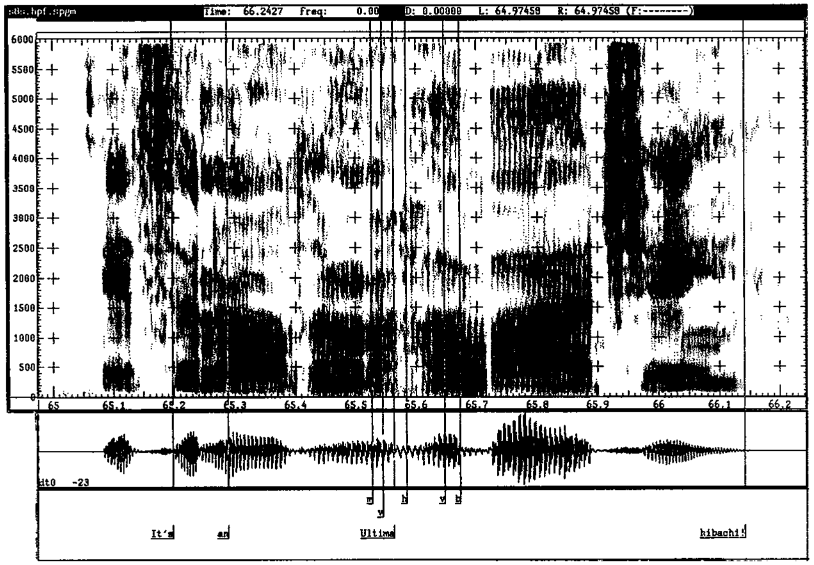

Figure 4.1 Wide-band spectrogram and waveform of the word *hibachi* produced with contrastive emphasis. Note the evident aspiration and the movement in F_1 due to the spread glottis during the /h/. The hand-marked segment locators and word boundaries are indicated in the lower panel: m is the /m/ release; v marks the vowel centers; h the /h/ center; b the closure onset of the /b/ consonant. The subject is DT

behave in much the same way. That is, we suspected that tone has no special privilege to interact with prosody; phonetic realization rules in general can be sensitive to prosodic structure. This point is illustrated in the spectrograms and waveforms of figures 4.1 and 4.2. In figure 4.1 the word *hibachi* carries contrastive stress and is well articulated. In figure 4.2, it is in postnuclear position and the /h/ is extremely lenited; that is, it is produced more like a vowel than the /h/ in figure 4.1. A similar effect of sentence stress on /h/ articulation in Swedish is reported in Gobl (1988).

Like the experiments which led to our present understanding of tonal realization, the work reported here considers the phonetic outcome for particular phonological elements as their position relative to local and nonlocal prosodic features is varied. Specifically, the materials varied position relative to the word prosody (the location of the word boundary and the word stress) and relative to the phrasal prosody (the location of the phrase boundary and the phrasal stress as reflected in the accentuation). Although there is also a strong potential for intonation to affect segmental source characteristics (since the larynx is the primary articulator for tones), this issue is not substantially addressed in the present study because the difficul-

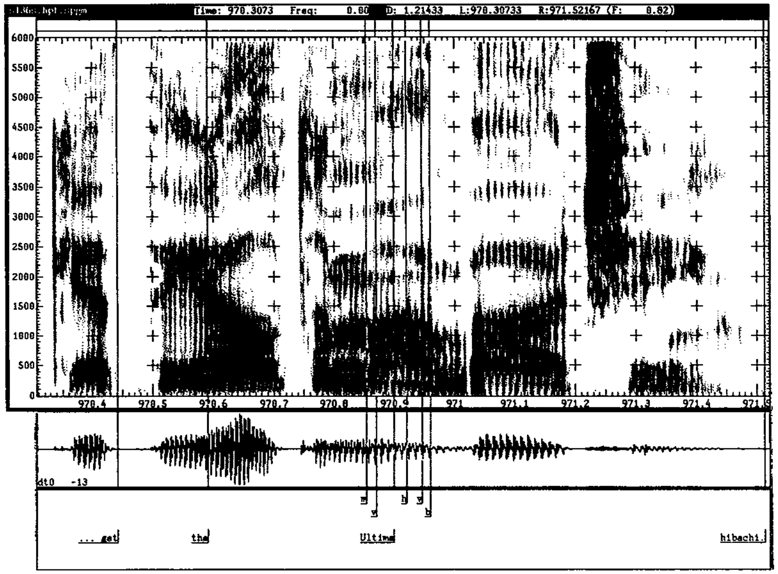

Figure 4.2 Wide-band spectrogram and waveform of the word *hibachi* in postnuclear position. Aspiration and F_1 movement during /h/ are less than in figure 4.1. The subject is DT

ties of phonetic characterization for /h/ and /ʔ/ led us to an experimental design with Low tones on all target regions. Pierrehumbert (1989) and a study in progress by Silverman, Pierrehumbert, and Talkin do address the effects of intonation on source characteristics directly by examining vocalic regions, where the phonetic characterization is less problematic.

The results of the experiment support a parallel treatment of segmental source characteristics and tone by demonstrating that the production of laryngeal consonants depends strongly on both word- and phrase-level prosody. Given that the laryngeal consonants are phonetically similar to tones by virtue of being produced with the same articulator, one might ask whether this parallel has a narrow phonetic basis. Studies such as Beckman, Edwards, and Fletcher (this volume) which reports prosodic effects on jaw movement, indicate that prosody is not especially related to laryngeal articulations, but can affect the extent and timing of other articulatory gestures as well. We would suggest that prosody (unlike intonational and segmental specifications) does not single out any particular articulator, but instead concerns the overall organization of articulation.

A certain tradition in phonology and phonetics groups prosody and intonation on the one hand as against segments on the other. Insofar as segments behave like tones, the grouping is called into question. We would like instead to argue for a point of view which contrasts structure (the prosodic pattern) with content (the substantive demands made on the articulators by the various tones and segments). The structure is represented by the metrical tree, and the content by the various autosegmental tiers and their decomposition into distinctive features. This point of view follows from recent work in metrical and autosegmental phonology, and is explicitly put forward in Pierrehumbert and Beckman (1988). However, many of its ramifications remain to be worked out. Further studies are needed to clarify issues such as the degree of abstractness of surface phonological representations, the roles of qualitative and quantitative rules in describing allophony, and the phonetic content of distinctive features in continuous speech. We hope that the present study makes a contribution towards this research program.

4.2 Background

4.2.1 /h/ and glottal stop

Both /h/ and glottal stop /ʔ/ are produced by a laryngeal gesture. They make no demands on the vocal-tract configuration, which is therefore determined by the adjacent segments. They are both less sonorous than vowels, because both involve a gesture which reduces the strength of voicing. For /h/, the folds are abducted. /ʔ/ is commonly thought to be produced by adduction (pressing the folds together), as is described in Ladefoged (1982), but examination of inverse filtering results and electroglottographic (EGG) data raised doubts about the generality of this characterization. We suggest that a braced configuration of the folds produces irregular voicing even when the folds are not pressed together (see further discussion below).

4.2.2 Source characterization

The following broad characteristics of the source are crucial to our characterization. (1) For vowels, the main excitation during each pitch period occurs at the point of contact of the vocal folds, because the discontinuity in the glottal flow injects energy into the vocal tract which can effectively excite the formants (see Fant 1959). This excitation point is immediately followed by the "closed phase" of the glottal cycle during which the formants have their most stable values and narrowest bandwidths. The "open phase" begins when the vocal folds begin to open. During this phase, acoustic interaction at the glottis results in greater damping of the formants as well as shifts in their location. (2) "Softening" of vocal-fold closure and an increase in the open quotient is associated with the "breathy" phonation in /h/. The abduction gesture (or gesture of spreading the vocal folds) associated with this type of phonation brings about an increase in the frequencies and bandwidths of the formants, especially F_1; an increase in the overall rate of spectral roll-off; an additional abrupt drop in the magnitude of the second and higher harmonics of the excitation spectrum; and an increase in the random noise component of the source, especially during the last half of the open phase. For some speakers, a breathy or soft voice quality is found during the postnuclear region of declaratives, as a reflex of phrasal intonation. (3) A "pressed" or "braced" glottal configuration is used to produce /ʔ/. This is realized acoustically as period-to-period irregularities in the timing and spectral content of the glottal excitation pulses. A full glottal stop (with complete obstruction of airflow at the glottis) is quite unusual. Some speakers use glottalized voicing, rather than breathy voicing, during the postnuclear region of declaratives.

4.2.3 Prosody and intonation

We assume that the word and phrase-level prosody is represented by a hierarchical structure along the lines proposed by Selkirk (1984), Nespor and Vogel (1986), and Pierrehumbert and Beckman (1988) (see Ladd, this volume). The structure represents how elements are grouped phonologically, and what relationships in strength obtain among elements within a given grouping. Details of these theories will not be important here, provided that the representation makes available to the phonetic realization rules all needed information about strength and grouping.

Substantive elements, both tones and segments, are taken to be autosegmentally linked to nodes in the prosodic tree. The tones and segments are taken to occur on separate tiers, and in this sense have a parallel relationship to the prosodic structure (see Broe, this volume). In this study, the main focus is on the relationship of the segments to the prosodic structure. The relationship of the tones to the prosodic structure enters into the study indirectly, as a result of the fact that prosodic strength controls the location of pitch accents in English. In each phrase, the last (or nuclear) pitch accent falls on the strongest stress in the phrase, and the prenuclear accents fall on the strongest of the preceding stresses. For this reason, accentuation can be used as an index of phrasal stress, and we will use the word "accented" to mean "having sufficient phrasal stress to receive an accent." "Deaccented" will mean "having insufficient phrasal stress to receive an accent"; in the present study, all deaccented words examined are postnuclear.

Rules for pronouncing the elements on any autosegmental tier can reference the prosodic context by examining the position and properties of the node the segment is linked to. In particular, studies of fundamental

frequency lead us to look for sensitivity to boundaries (Is the segment at a boundary or not? If so, what type of boundary?) and to the strength of the nodes above the segment.

Pronunciation rules are also sensitive to the substantive context. For example, in both Japanese and English, downstep or catathesis applies only when the tonal sequence contains particular tones. Similarly, /h/ has a less vocalic pronunciation in a consonantal environment than in a vocalic one. Such effects, widely reported in the literature on coarticulation and assimilation, are not investigated here. Instead, we control the segmental context in order to concentrate on the less well understood prosodic effects.

Although separate autosegmental tiers are phonologically independent, there is a strong potential for phonetic interaction between tiers in the case examined here, since both tones and laryngeal consonants make demands on the laryngeal configuration. This potential interaction was not investigated, since our main concern was the influence of prosodic structure on segmental allophony. Instead, intonation was carefully controlled to facilitate the interpretation of the acoustic signal.

4.3 Experimental methods

4.3.1 Guiding considerations

The speech materials and algorithms for phonetic characterization were designed together in order to achieve maximally interpretable results. Source studies such as Gobl (1988) usually rely on inverse filtering, a procedure in which the effects of vocal-tract resonances are estimated and removed from the signal. The residue is an estimate of the derivative of the flow through the glottis. This procedure is problematic for both /ʔ/ and /h/. For /ʔ/, it is difficult to establish the pitch periods to which inverse filtering should be applied. (Inverse filtering carried out on arbitrary intervals of the signal can have serious windowing artifacts). Inverse filtering of /h/ is problematic because of its large open quotient. This can introduce subglottal zeroes, rendering the all-pole model of the standard procedure inappropriate, and it can increase the frequency and bandwidth of the first formant to a point where its location is not evident. The unknown contribution of noise also makes it difficult to evaluate the spectral fit to the periodic component of the source. These considerations led us to design materials and algorithms which would allow us to identify differences in source characteristics without first estimating the transfer function.

Specific considerations guiding the design of the materials were: (1) F_1 is greater than three times F_0. This minimizes the effects of F_1 bandwidth and location on the low-frequency region, allowing it to reflect source characteristics in a more straightforward fashion. (2) Articulator movement in the upper vocal tract is minimal during target segments. (3) The consonants under study are produced by glottal gestures in an open-vowel environment to facilitate interpretation of changes to the vocal source.

4.3.2 Materials

In the materials for the experiment, the position of /h/ and /ʔ/ relative to word-level and phrase-level prosodic structure is varied. We lay out the full experimental design here, although we will only have space to discuss some subsets of the data which showed particularly striking patterns.

In the materials /h/ is exemplified word-initially and -medially, before both vowels with main word stress and vowels with less stress:

mahogany tomahawk hogfarmers hawkweed
hibachi Omaha

The original intention was to distinguish between a secondary stress in *tomahawk* and an unstressed syllable at the end of *Omaha*, but it did not prove possible to make this distinction in the actual productions, so these cases will be treated together as "reduced" /h/s.

Intervocalic /ʔ/ occurs only word-initially. /ʔ/ as an allophone of /t/ is found before syllabic nasals, as in "button," but not before vowels.) So, for /ʔ/ we have the following sets of words, providing near minimal comparisons to the word-initial /h/s:

August awkwardness abundance Augustus
augmentation

This set of words was generated using computerized searches of several online dictionaries. The segmental context of the target consonant was designed to have a high first formant value and minimize formant motion, in order to simplify the acoustic phonetic analysis. The presence of a nasal in the vicinity of the target consonant is undesirable, because it can introduce zeroes which complicate the evaluation of source characteristics. This suboptimal choice was dictated by the scarcity of English words with medial /h/, even in the educated vocabulary. We felt that it was critical to use real words rather than nonsense words, in order to have accurate control over the word-level prosody and in order to avoid exaggerated pronunciations.

Words in which the target consonant was word-initial were provided with a /ma/ context on the left by appropriate choice of the preceding word. Phrases such as the following were used:

Oklahoma August lima abundance figures
plasma augmentation

The position of the target words in the phrasal prosody were also manipulated. The phrasal positions examined were (1) accented without special focus, (2) accented and under contrast, (3) accented immediately following an intonational phrase boundary, and (4) postnuclear. In order to maximize the separation of F_0 and F_1, the intonation patterns selected to exemplify these positions all had L tones (leading to low F_0) at the target location. This allows the source influences on the low-frequency region to be seen more clearly. The intonation patterns were also designed to display a level (rather than time-varying) intonational influence on F_0, again with a view to minimizing artifacts. The accented condition used a low-rising question pattern (L^* H H% in the transcription of Pierrehumbert [1980]):

(1) Is he an Oklahoma hogfarmer?

Accent with contrast was created by embedding the "contradiction" pattern (L^* L H%) in the following dialogue:

(2) A: Is it mahogany?
B: No, it's rosewood.
C: It's mahogany!

In the "phrase boundary" condition, a preposed vocative was followed by a list, as in the following example:

(3) Now Emma, August is hot here, July is hot here, and even June is hot here.

The vocative had a H^* L pattern (that is, it had a falling intonation and was followed by an intermediate phrase boundary rather than a full intonation break). Non-final list items had a L^* H pattern, while the final list item (which was not examined) had a H^* L L% pattern. The juncture of the H^* L vocative pattern with the L^* H pattern of the first list item resulted in a low, level F_0 configuration at the target consonant. Subjects were instructed to produce the sentences without a pause after the vocative, and succeeded in all but a few utterances (which were eliminated from the analysis). No productions lacked the desired intonational boundary.

In the "postnuclear" condition, the target word followed a word under contrast:

(4) They're Alabama hogfarmers, not Oklahoma hogfarmers.

In this construction, the second occurrence of the target word was the one analyzed.

4.3.3 *Recording procedures*

Since pilot studies showed that subjects had difficulty producing the correct patterns in a fully randomized order, a blocked randomization was used. Each block consisted of twelve sentences with the same phrasal intonation pattern; the order within each block was randomized. The four blocks were then randomized within each set. Six sets were recorded. The first set was discarded, since it included a number of dysfluent productions for both speakers.

The speech was recorded in an anechoic chamber using a 4165 B&K microphone with a 2231 B&K amplifier, and a Sony PCM-2000 digital audio tape recorder. The speakers were seated. A distance of 30 cm from the mouth to the microphone was maintained. This geometry provides intensity sensitivity due to head movement of approximately 0.6 dB per cm change in microphone-to-mouth distance. Since we have no direct interface between the digital tape recorder and the computer, the speech was played back and redigitized at 12 kHz with a sharp-cutoff anti-alias filter set at 5.8 kHz, using an Ariel DSP-16 rev. G board, yielding 16 bits of precision. The combined system has essentially constant amplitude and phase response from 20 Hz to over 5 kHz. The signal-to-noise ratio for the digitized data was greater than 55 dB. Electroglottographic signals were recorded and digitized simultaneously on a second channel to provide a check for the acoustically determined glottal epochs which drive the analysis algorithms.

4.4 Analysis algorithms and their motivation

The most difficult part of the study was developing the phonetic characterization, and the one used is not fully successful. Given both the volume of speech to be processed and the need for replicability, it is desirable to avoid measurement procedures which involve extensive fitting by eye or other subjective judgment. Instead, we would argue the need for semi-automatic procedures, in which the speech is processed using well-defined and tested algorithms whose results are then scanned for conspicuous errors.

4.4.1 *Pitch-synchronous analysis*

The acoustic features used in this study are determined by "pitch-synchronous" analyses in which the start of the analysis window is phase-locked on the time of glottal closure and the duration of the window is determined by the length of the glottal cycle. Pitch-synchronous analysis is desirable because it offers the best combination of physical insight and time resolution. One glottal cycle is the minimum period of interest, since it is difficult to draw

conclusions about the laryngeal configuration from anything less. When the analysis window is matched to the cycle in both length and phase, the results are well behaved. In contrast, when analysis windows the length of a cycle are applied in arbitrary phase to the cycle, extensive signal-processing artifacts result. Therefore non-pitch-synchronous moving-window analyses are typically forced to a much longer window length in order to show well-behaved results. The longer window lengths in turn obscure the speech events, which can be extremely rapid.

Pitch-synchronous analysis is feasible for segments which are voiced throughout because the point of glottal closure can be determined quite reliably from the acoustic waveform (Talkin 1989). We expected it to be applicable in our study since the regions of speech to be analyzed were designed to be entirely voiced. For subject DT, our expectations were substantially met. For subject MR, strong aspiration and glottalization in some utterances interfered with the analysis.

Talkin's algorithm for determining epochs, or points of glottal closure, works as follows: speech, recorded and digitized using a system with known amplitude and phase characteristics, is amplitude- and phase-corrected and then inverse-filtered using a matched-order linear predictor to yield a rough approximation to the derivative of the glottal volume velocity (U'). The points in the U' signal corresponding to the epochs have the following relatively stable characteristics: (1) Constant polarity (negative), (2) Highest amplitude within each cycle, (3) Rapid return to zero after the extremum, (4) Periodically distributed in time, (5) Limited range of inter-peak intervals, and (6) Similar size and shape to adjacent peaks. A set of peak candidates is generated from all local maxima in the U' signal. Dynamic programming is then used to find the subset of these candidates which globally best matches the known characteristics of U' at the epoch. The algorithm has been evaluated using epochs determined independently from simultaneously recorded EGG signals and was found to be quite accurate and robust. The only errors that have an impact on the present study occur in strongly glottalized or aspirated segments.

4.4.2 *Measures used*

Pitch-synchronous measurements used in the current study are (1) root mean square of the speech samples in the first 3 msec. following glottal closure expressed in dB re unity (RMS), (2) ratio of per-period energy in the fundamental to that in the second harmonic (HR), and (3) local standard deviation in period length (PDEV). RMS and HR are applied to /h/, and PDEV is applied to /ʔ/.

Given the relatively constant intertoken phonetic context, RMS provides an intertoken measure closely related to the strength of the glottal excitation in corresponding segments. The integration time for RMS was held constant to minimize interactions between F_0 and formant bandwidths. RMS was not a useful measure for /ʔ/, since even a strongly articulated /ʔ/ may have glottal excitation as strong as the neighboring vowels; the excitation is merely more irregular. RMS is relatively insensitive to epoch errors in /h/s, since epoch-location uncertainty tended to occur when energy was more evenly distributed through the period, which in turn renders the measurement point for RMS less critical.

HR is computed as the ratio (expressed in dB) of the magnitudes of the first and second harmonics of an exact DFT (Discrete Fourier Transform) computed over one glottal period. The period duration is from the current epoch to the next, but the start time of the period is taken to be at the zero crossing immediately preceding the current epoch. This minimizes periodicity violations introduced by cycle-to-cycle excitation variations, since the adjusted period end will usually also fall in a low-amplitude (near zero) region of the cycle. HR is a relevant measure because the increase in open quotient of the glottal cycle and the lengthening of the time required to accomplish glottal closure associated with vocal-fold abduction tends to increase the power in the fundamental relative to the higher harmonics. This increase in the average and minimum glottal opening also changes the vocal-tract tuning and sub- to superglottal coupling. The net acoustic effect is to introduce a zero in the spectrum in the vicinity of F_1 and to increase both the frequency and bandwidths of the formants, especially F_1. Since our speech material was designed to keep F_1 above the second harmonic, these effects all conspire to increase HR with abduction. The reader is referred to Fant and Lin (1988) for supporting mathematical derivations. Figure 4.3 illustrates the behavior of the HR over the target intervals which include the two /h/s shown in figures 4.1 and 4.2.

Fant and Lin's derivations do not attempt to model the contribution of aspiration to the spectral shape, and the relation of abduction to HR indeed becomes nonmonotonic at the point at which aspiration noise becomes the dominant power source in the second-harmonic region. One of the subjects, DT, has sufficiently little aspiration during the /h/s that this nonmonotonicity did not enter substantially into the analysis, but for subject MR it was a major factor, and as a result RMS shows much clearer patterns. HR is also sensitive to serious epoch errors, rendering it inapplicable to glottalized regions.

PDEV is the standard deviation of the seven glottal period lengths surrounding the current epoch. This measure represents an effort to quantify the irregular periodicity which turned out to be the chief hallmark of /ʔ/. It was adopted after detailed examination of the productions failed to support

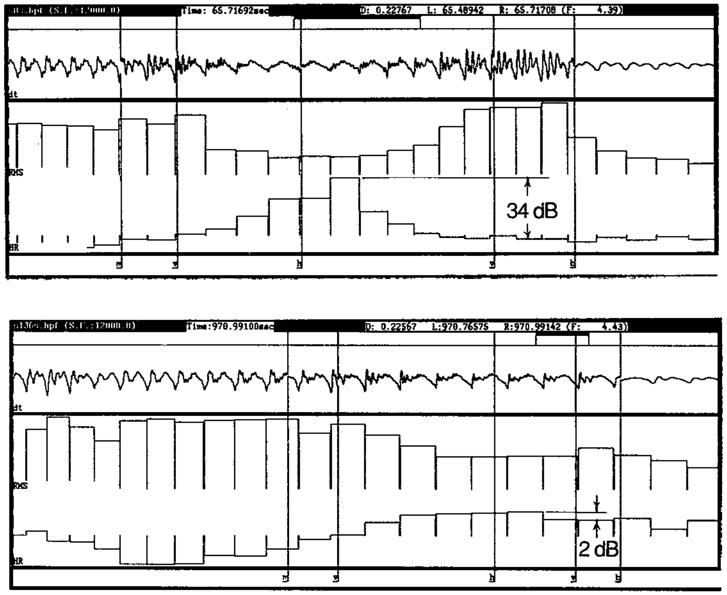

Figure 4.3 HR and RMS measured for each glottal period throughout the target intervals of the utterances introduced in figures 4.1 and 4.2. Note that the difference of ~34 dB between the HR in the /h/ and following vowel for the well-articulated case (top) is much greater than the ~2dB observed in the lenited case (bottom). The RMS values discussed in the text are based on the (linear) RMS displayed in this figure

the common understanding that /ʔ/ is produced by partial or complete adduction of the vocal folds. This view predicts that the spectrum during glottalization should display a lower HR and a less steep overall spectral roll-off than are found in typical vowels. However, examination of the EGG signal in conjunction with inverse-filtering results showed that many tokens had a large, rather than a small, open quotient and even showed aspiration noise during the most closed phase, indicating that the closure was incomplete. The predicted spectral hallmarks were not found reliably, even in utterances in which glottalization was conspicuously indicated by irregular periodicity. We surmise that DT in particular produces /ʔ/ by bracing or tensing partially abducted vocal folds in a way which tends to create irregular vibration without a fully closed phase.

4.4.3 Validation using synthetic signals

In order to validate the measures, potential artifactual variation due to F_0–F_1 interactions was assessed. A six-formant cascade synthesizer excited by a

Liljencrants–Fant voice source run at 12 kHz sampling frequency was used to generate synthetic voiced-speech-like sounds. These signals were then processed using the procedures outlined above. F_1 and F_0 were orthogonally varied over the ranges observed in the natural speech tokens. F_1 bandwidth was held constant at 85 Hz while its frequency took on values of 500 Hz, 700 Hz and 800 Hz. For each of these settings the source fundamental was swept from 75 Hz to 150 Hz during one second with the open quotient and leakage time held constant. The bandwidths and frequencies of the higher formants were held constant at nominal 17 cm, neutral vocal-tract values. Note that the extremes in F_1 and F_0 were not simultaneously encountered in the natural data, so that this test should yield conservative bounds on the artifactual effects.

As expected, PDEV did not vary significantly throughout the test signal. The range of variation in HR for these test signals was less than 3 dB. The maximum peak-to-valley excursion in RMS due to F_0 harmonics crossing the formants was 2 dB with a change in F_0 from 112 Hz to 126 Hz and F_1 at 500 Hz. This is small compared to RMS variations observed in the natural speech tokens under study.

4.4.4 Analysis of the data

Time points were established for the /m/ release, the first vowel center, the center of the /h/ or glottal stop, the center of the following vowel, and the point of oral constriction for the consonant. This was done by inspection of the waveform and broad-band spectrogram, and by listening to parts of the signal.

The RMS, HR and PDEV values for the vowel were taken to be the values at the glottal epoch nearest to the measured vowel center.

RMS was used to estimate the /h/ duration, since it was almost invariably lower at the center of the /h/ than during the following vowel. The /h/ interval was defined as the region around the minimum RMS observed for the /h/ during which RMS did not exceed a locally determined threshold. Taking RMS(C) as the minimum RMS observed and RMS(V2) as the maximum RMS in the following vowel, the threshold was defined as RMS(C) + 0.25*[RMS(V2) − RMS(C)]. The measure was somewhat conservative compared to a manual segmentation, and was designed to avoid spurious inclusions of the preceding schwa when it was extremely lenited. The consonantal value for RMS was taken to be the RMS minimum, and its HR value was taken to be the maximum HR during the computed /h/ interval.

The PDEV value for the /ʔ/ was taken at the estimated center, since the

intensity behaviour of the /ʔ/s did not support the segmentation scheme developed for the /h/s. Durations for /ʔ/ were not estimated.

4.5 Results

After mentioning some characteristics of the two subjects' speech, we first present results for /h/ and then make some comparisons to /ʔ/.

4.5.1 Speaker characteristics

There were some obvious differences in the speech patterns of the two subjects. When these differences are taken into account, it is possible to discern strong underlying parallels in the effects of prosody on /h/ and /ʔ/ production.

MR had vocal fry in postnuclear position. This was noticeable both in the deaccented condition and at the end of the preposed vocative *Now Emma.* He had strong aspiration in /h/, leading to failure of the epoch finding in many cases and also to nonmonotonic behavior of the HR measure. As a result, the clearest patterns are found in RMS (which is more insensitive than HR to epoch errors) and in duration. In general, MR had clear articulation of consonants even in weak positions.

DT had breathiness rather than fry in postnuclear position. Aspiration in /h/ was relatively weak, so that the epoch finder and the HR measure were well behaved. Consonants in weak positions were strongly reduced.

4.5.2 Effects of word prosody and phrasal stress on /h/

Both the position in the word and the phrasal stress were found to affect how /h/ was pronounced. In order to clarify the interpretation of the data, we would first like to present some schematic plots. Figure 4.4 shows a blank plot of RMS in the /h/ against RMS in the vowel. Since the /h/ articulation decreases the RMS, more /h/-like /h/s are predicted to fall towards the left of the plot while more vowel-like /h/s fall towards the right of the plot. Similarly, more /h/-like vowels are predicted to fall towards the bottom of the plot, while more vowel-like vowels should fall towards the top of the plot.

The line y = x, shown as a dashed diagonal, represents the case where the /h/ and the vowel had the same measured RMS. That is, as far as RMS is concerned, there was no local contrast between the /h/ and the vowel. Note that this case, the case of complete neutralization, is represented by a wide range of values, so that the designation "complete lenition" does not actually fix the articulatory gesture. In general, we do not expect to find /h/s which are

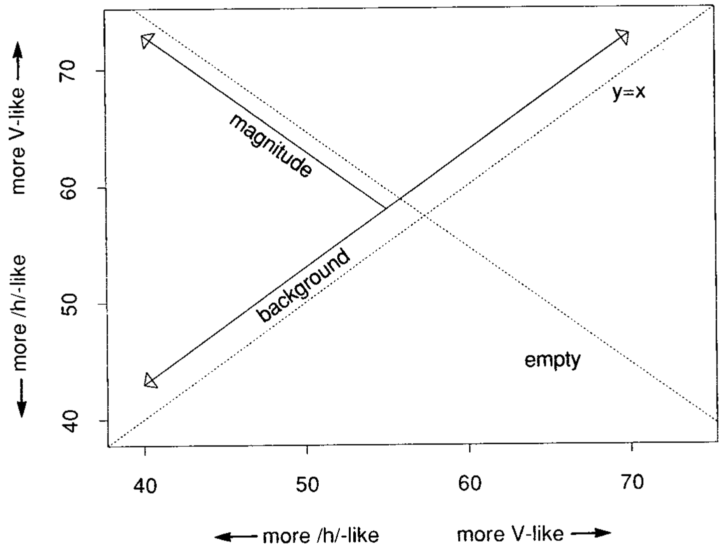

Figure 4.4 A schema for interpreting the relation of RMS in the /h/ to RMS in the following vowel. Greater values of RMS correspond to more vowel-like articulations, and lesser values correspond to more /h/-like articulations. The line y = x represents the case in which the /h/ and the following vowel do not contrast in terms of RMS. Distance perpendicular to this line represents the degree of contrast between the /h/ and the vowel. Distance parallel to this line cannot be explained by gestural magnitude, but is instead attributed to a background effect on both the /h/ and the vowel. The area below and to the right of y = x is predicted to be empty

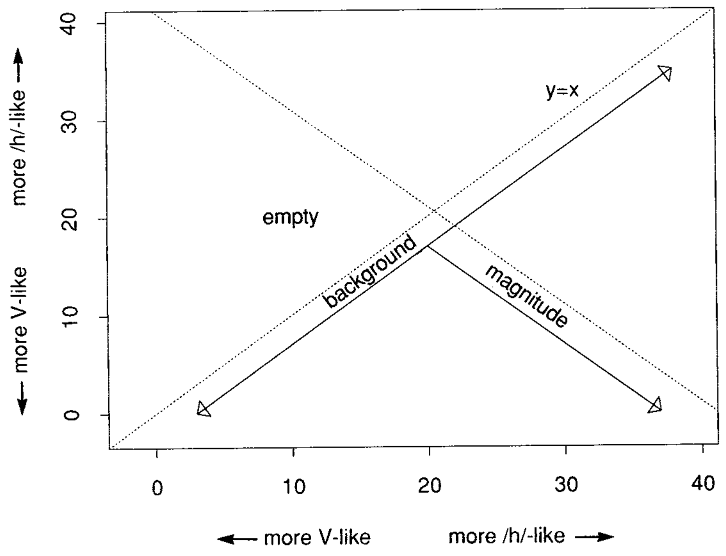

Figure 4.5 A schema for interpreting the relation of HR in the /h/ to HR in the following vowel. It has the same structure as in figure 4.4, except that greater rather than lesser values of the parameter represent more /h/-like articulations

more vocalic than the following vowel, so that the lower-right half is expected to be empty. In the upper-left half, the distance from the diagonal describes the degree of contrast between the /h/ and the vowel. Situations in which both the /h/ and the vowel are more fully produced would exhibit greater contrast, and would therefore fall further from the diagonal. Note again that a given magnitude of contrast can correspond to many different values for the /h/ and vowel RMS.

Figure 4.5 shows a corresponding schema for HR relations. The structure is the same except that higher, rather than lower, x and y values correspond to more /h/-like articulations.

In view of this discussion, RMS and HR data will be interpreted with respect to the diagonal directions of each plot. Distance perpendicular to the y = x line (shown as a dotted line in each plot) will be related to the strength or magnitude of the CV gesture. Location parallel to this line, on the other hand, is *not* related to the strength of the gesture, but rather to a background effect on which the entire gesture rides. One of the issues in interpreting the data is the linguistic source of the background effects.

Figures 4.6 and 4.7 compare the RMS relations in word-initial stressed /h/, when accented in questions and when deaccented. The As are farther from the y = x line than the Ds, indicating that the magnitude of the gesture is greater when /h/ begins an accented syllable. For subject DT, the two clouds of points can be completely separated by a line parallel to y = x. Subject MR shows a greater range of variation in the D case, with the most carefully articulated Ds overlapping the gestural magnitude of the As.

Figures 4.8 and 4.9 make the same comparison for word-medial /h/ preceding a weakly stressed or reduced vowel. These plots have a conspicuously different structure from figures 4.6 and 4.7. First, the As are above and to the right of the Ds, instead of above and to the left. Second, the As and Ds are not as well separated by distance from the y = x line; whereas this separation was clear for word-initial /h/s, there is at most a tendency in this direction for the medial reduced /h/s.

The HR data shown for DT in figures 4.10 and 4.11 further exemplifies this contrast. Word-initial /h/ shows a large effect of accentuation on gestural magnitude. For medial reduced /h/ there is only a small effect on magnitude; however, the As and Ds are still separated due to the lower HR values during the vowel for the As. HR data is not presented for MR because strong aspiration rendered the measure a nonmonotonic function of abduction.

Since the effect of accentuation differs depending on position in the word, we can see that both phrasal prosody and word prosody contribute to determining how segments are pronounced. In decomposing the effects, let us first consider the contrasts in gestural magnitude, that is perpendicular to the x = y line. In the case of *hawkweed* and *hogfarmer*, the difference between As

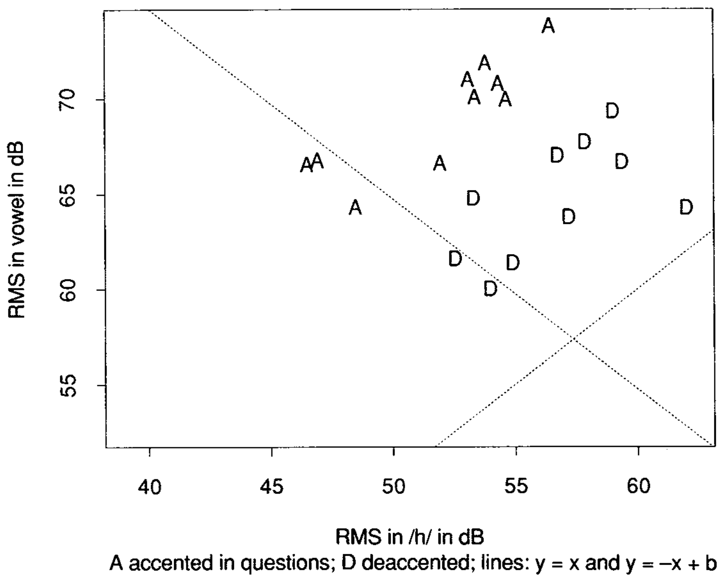

Figure 4.6 RMS in /h/ of *hawkweed* and *hogfarmer* plotted against RMS in the following vowel, when the words are accented in questions (the As) and deaccented (the Ds). The subject is DT

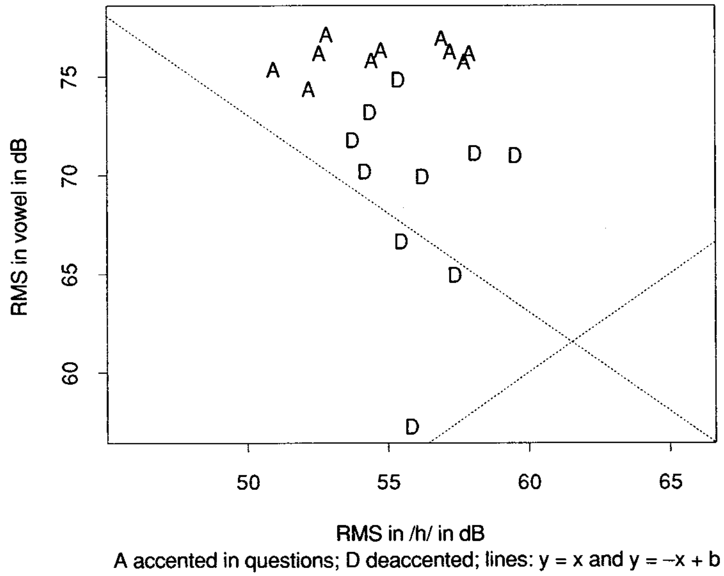

Figure 4.7 RMS in /h/ of *hawkweed* and *hogfarmer* plotted against RMS in the following vowel, when the words are accented in questions (the As) and deaccented (the Ds). The subject is MR

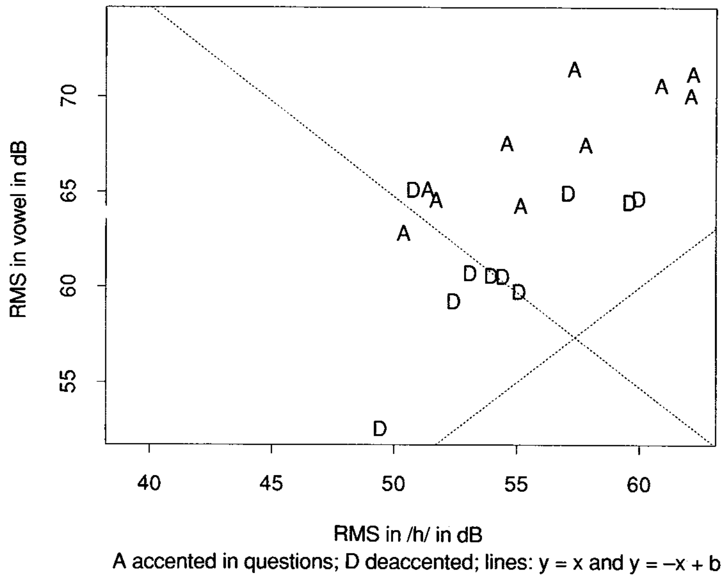

Figure 4.8 RMS in /h/ of *Omaha* and *tomahawk* plotted against RMS in the following vowel, when the words are accented in questions and deaccented. The subject is DT

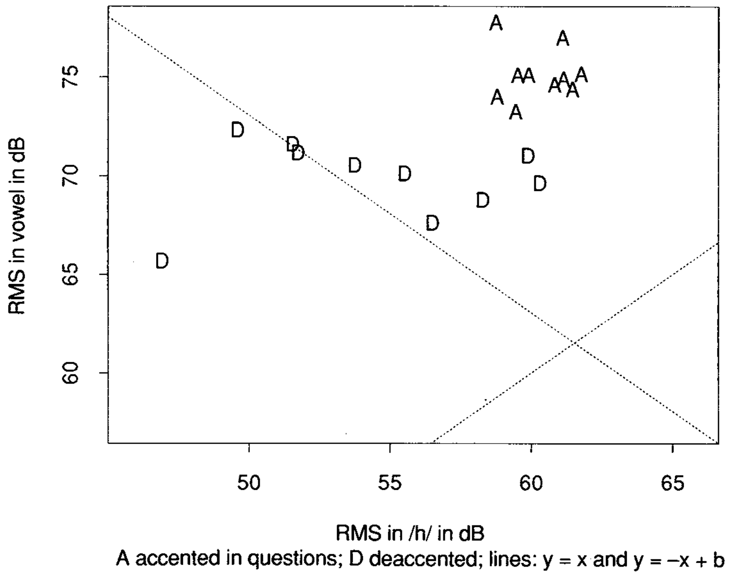

Figure 4.9 RMS in /h/ of *Omaha* and *tomahawk* plotted against RMS in the following vowel, when the words are accented in questions and deaccented. The subject is MR

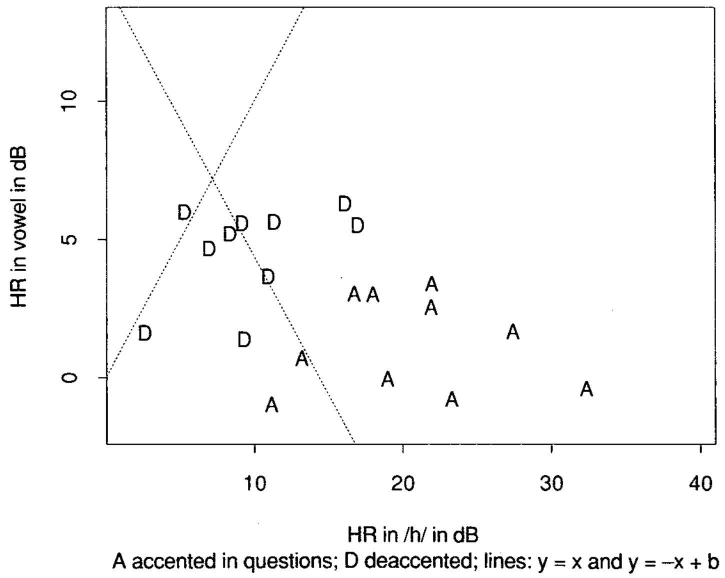

Figure 4.10 HR in /h/ of *hawkweed* and *hogfarmer* plotted against HR in the following vowel, when the words are accented in questions and deaccented. The subject is DT

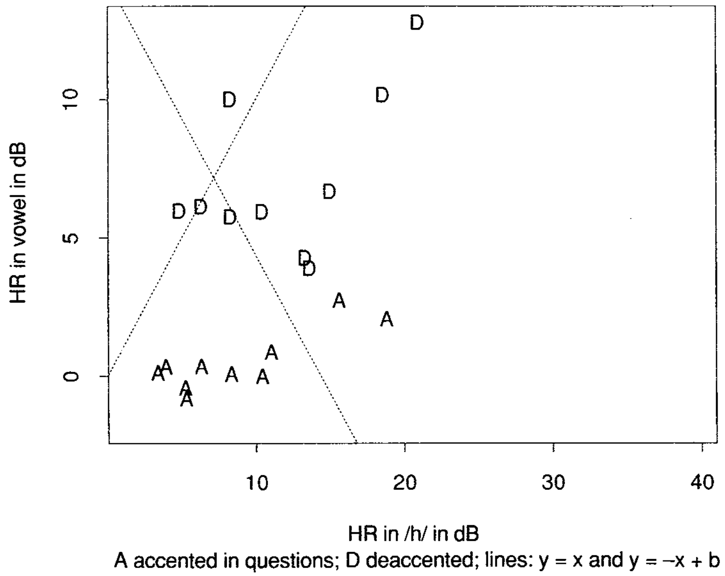

Figure 4.11 HR in /h/ of *Omaha* and *tomahawk* plotted against HR in the following vowel, when the words are accented in questions and deaccented. The subject is DT

and Ds is predominately in this direction. The *Omaha* and *tomahawk* As and Ds exhibit a small difference in this direction, though this is not the most salient feature of the plot. From this we deduce that accentuation increases gestural magnitude, making vowels more vocalic and consonants more consonantal. The extent of the effect depends on location with respect to the word prosody; the main stressed word-initial syllable inherits the strength of accentuation, so to speak, more than the medial reduced syllable does. At the same time we note that in *tomahawk* and *Omaha*, the As are shifted relative to the Ds parallel to the y = x line. That is, both the consonant and the vowel are displaced in the vocalic direction, as if the more vocalic articulation of the main stressed vowel continued into subsequent syllables. The data for *tomahawk* and *Omaha* might thus be explicated in terms of the superposition of a local effect on the magnitude of the CV gesture and a larger-scale effect which makes an entire region beginning with the accented vowel more vocalic.

The present data do not permit a detailed analysis of what region is affected by the background shift in a vocalic direction. Note that the effect of a nuclear accent has abated by the time the deaccented postnuclear target words are reached, since these show a more consonantal background effect than the accented words do. In principle, data on the word *mahogany* would provide critical information about where the effect begins, indicating, for example, whether the shift in the vocalic direction starts at the beginning of the first vowel in an accented word or at the beginning of the stressed vowel in a foot carrying an accent. Unfortunately, the *mahogany* data showed considerable scatter and we are not prepared at this point to make strong claims about their characterization.

4.5.3 The Effect of the phrase boundary on /h/

It is well known that syllables are lengthened before intonational boundaries. Phrase-final voiced consonants are also typically devoiced. An interesting feature of our data is that it also demonstrated lengthening and suppression of voicing after an intonational boundary, even in the absence of a pause. Figures 4.12 and 4.13 compare duration and RMS in word-initial /h/ after a phrase boundary (that is, following *Now Emma* with word-initial /h/) in accented but phrase-medial position, and in deaccented (also phrase-medial) position. In both plots, the " %" points are below and to the right of the A and D points, indicating a combination of greater length and less strong voicing.

DT shows a strong difference between A and D points, with Ds being shorter and more voiced than As. MR shows at most a slight difference between As and Ds, reflecting his generally small degree of lenition of

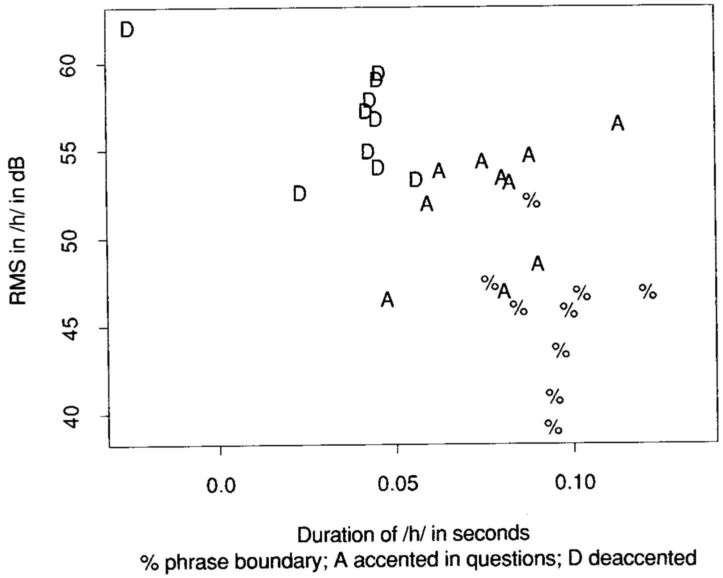

Figure 4.12 Duration vs. RMS in /h/ for *hawkweed* and *hogfarmer* when accented at a phrase boundary, accented but phrase-medial in questions, and deaccented. The subject is DT

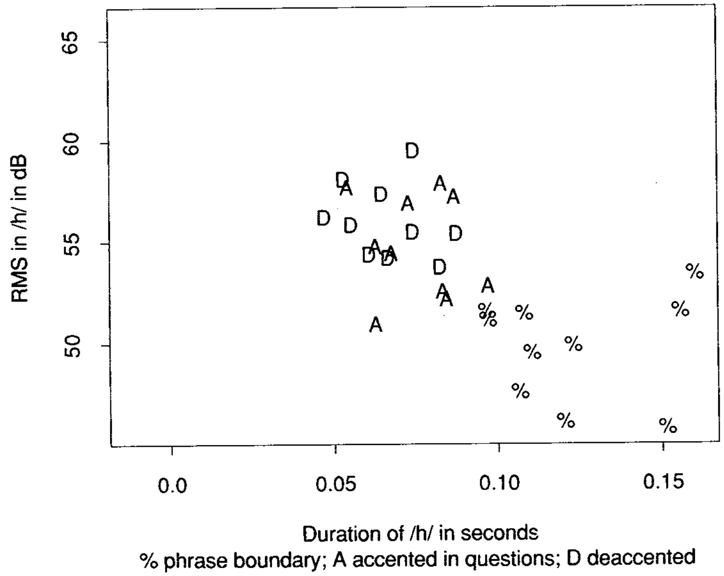

Figure 4.13 Duration vs. RMS in /h/ for *hawkweed* and *hogfarmer* when accented at a phrase boundary, accented but phrase-medial in questions, and deaccented. The subject is MR

consonants in weak positions. For MR, the effect of the phrase boundary is thus a more major one than the effect of accentual status.

A subset of the data set, the sentences involving *tomahawk*, make it possible to extend the result to a nonlaryngeal consonant. The aspiration duration for the /t/ was measured in the four prosodic positions. The results are displayed in figures 4.14 and 4.15. The lines represent the total range of observations for each condition, and each individual datum is indicated with a tick. For DT, occurring at a phrase boundary approximately doubled the aspiration length, and there was no overlap between the phrase-boundary condition and the other conditions. For MR, the effect was somewhat smaller, but the overlap can still be attributed to only one point, the smallest value for the phrase-boundary condition. For both subjects, a smaller effect of accentuation status can also be noted.

The effect of the phrase boundary on gestural magnitude can be investigated by plotting the RMS in the /h/ against RMS in the vowel, the word-initial accented /h/ in phrase-initial and phrase-medial position. This comparison, shown in figures 4.16 and 4.17, indicates that the gestural magnitude was greater in phrase-initial position. The main factor was lower RMS (that is, a more extreme consonantal outcome) for the /h/ in phrase-initial position; the vowels differed slightly, if at all. Returning to the decomposition in terms of gestural-magnitude effects and background effects, we would suggest that the phrase boundary triggers both a background shift in a consonantal direction (already observed in preboundary position in the "deaccented" cases) and an increase in gestural magnitude. The effect on gestural magnitude must be either immediately local to the boundary, or related to accentual strength, if deaccented words in the middle of the postnuclear region are to be exempted as observed.

It is interesting to compare our results on phrase-initial articulation with Beckman, Edwards, and Fletcher's results (this volume) on phrase-final articulation. Their work has shown that stress-related lengthening is associated with an increase in the extent of jaw movement while phrase-final lengthening is not, and they interpret this result as indicating that stress triggers an underlying change in gestural magnitude while phrase-final lengthening involves a change in local tempo but not gestural magnitude. Given that our data do show an effect of phrase-initial position on gestural magnitude, their interpretation leads to the conclusion that phrase-initial and phrase-final effects are different in nature.

However, we feel that the possibility of a unified treatment of phrase-peripheral effects remains open, pending the resolution of several questions about the interpretation of the two experiments. First, it is possible that the gestural-magnitude effect observed in our data is an artifact of the design of the materials, since the *Now Emma* sentences may have seemed more unusual

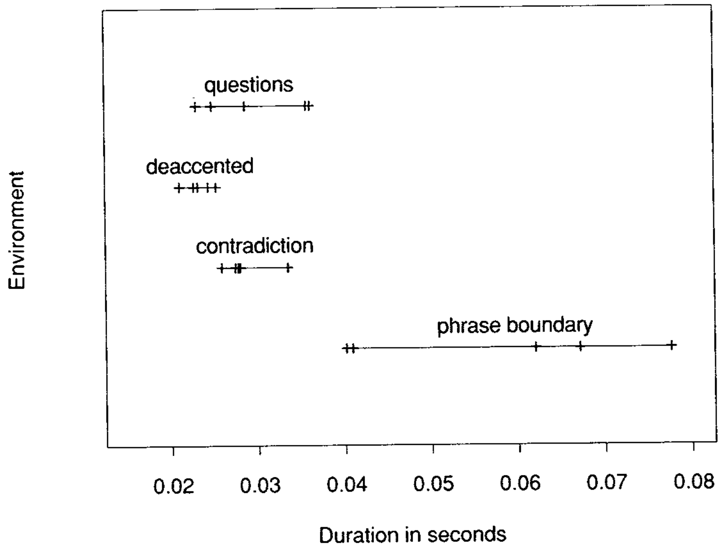

Figure 4.14 Voice-onset time in /t/ of *tomahawk* for all four prosodic positions; subject DT. Ticks indicate individual data points

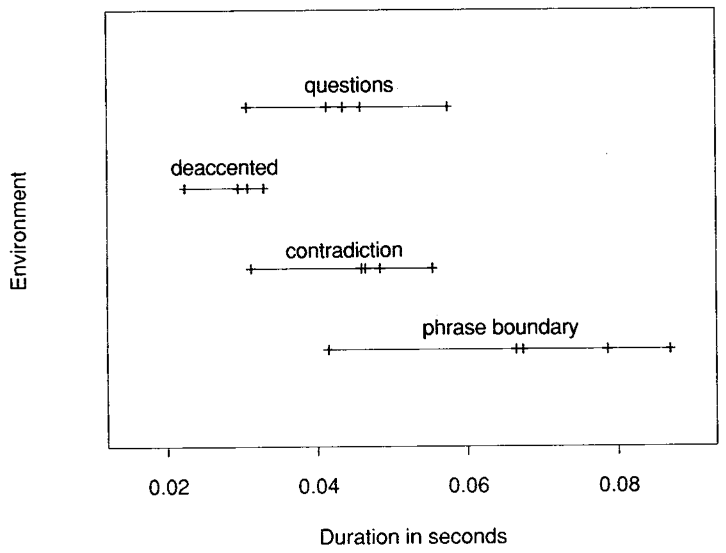

Figure 4.15 Voice-onset time in /t/ of *tomahawk* for all four prosodic positions; subject MR. Ticks indicate individual data points

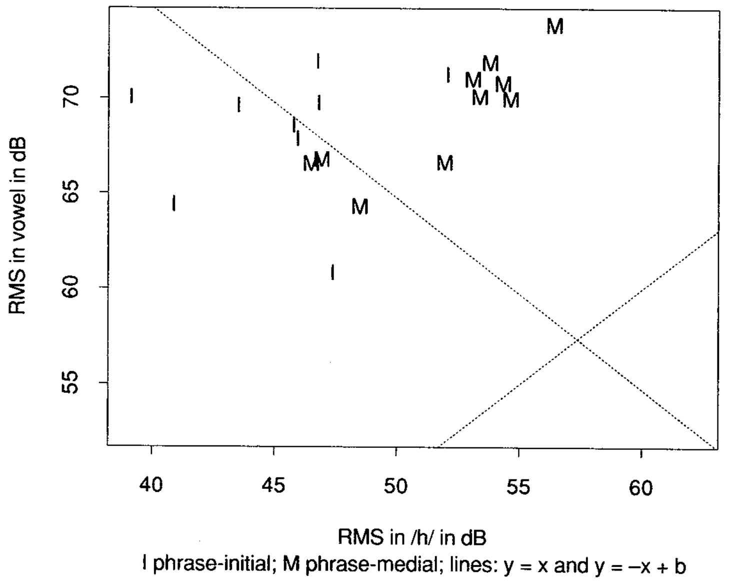

Figure 4.16 RMS of /h/ and final vowel for subject DT in accented phrase-initial (I) and phrase-medial (M) question contexts

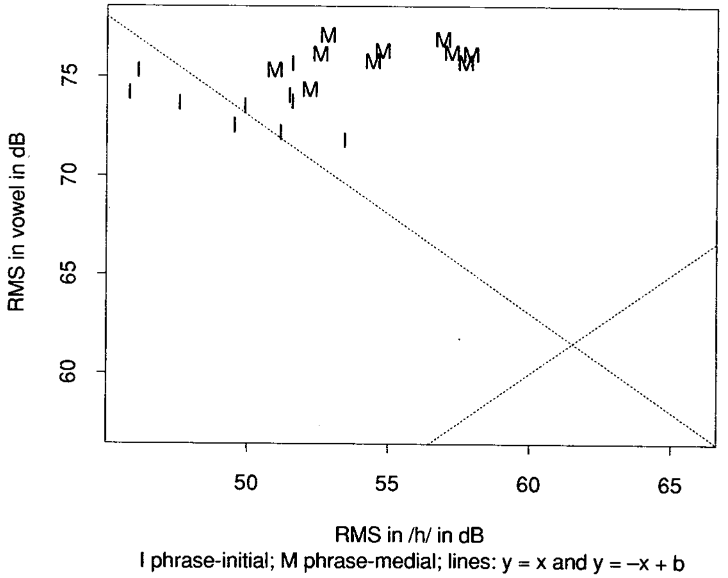

Figure 4.17 RMS of /h/ and final vowel for subject MR in accented phrase-initial (I) and phrase-medial (M) question contexts

or semantically striking than those where the target words were phrase-internal. If this is the case, semantically matched sentences would show a shift towards the consonantal direction in the vowel following the consonant, as well as the consonant itself. Second, it is possible that an effect on intended magnitude is being obscured in Beckman, Edwards, and Fletcher's (this volume) data by the nonlinear physical process whose outcome serves as an index. Possibly, jaw movement was impeded after the lips contacted for the labial consonant in their materials, so that force exerted after this point did not result in statistically significant jaw displacement. If this is the case, measurements of lip pressure or EMG (electromyography) might yield results more in line with ours. Third, it is possible that nonlinearities in the vocal-fold mechanics translate what is basically a tempo effect in phrase-initial position into a difference in the extent of the acoustic contrast. That is, it is possible that the vocal folds are no more spread for the phrase-initial /h/s than for otherwise comparable /h/s elsewhere but that maintaining the spread position for longer is in itself sufficient to result in greater suppression of the oscillation. This possibility could be evaluated using high-speed optical recording of the laryngeal movements.

4.5.4 Observations about glottalization

Although all /h/s in the study had some noticeable manifestation in the waveform, this was not the case for /ʔ/. In some prosodic positions, glottalization for /ʔ/ appeared quite reliably, whereas in others it did not. One might view /ʔ/ insertion as an optional rule, whose frequency of application is determined in part by prosodic position. Alternatively, one might take the view that the /ʔ/ is always present, but that due to the nonlinear mechanics involved in vocal-fold vibration, the characteristic irregularity only becomes apparent when the strength and/or duration of the gesture is sufficiently great. That is, the underlying control is gradient, just as for /h/, but a nonlinear physical system maps the gradient control signal into two classes of outcomes. From either viewpoint, an effect of prosodic structure on segmental production can be demonstrated; the level of representation for the effect cannot be clarified without further research on vocal-fold control and mechanics.

Table 4.1 summarizes the percentage of cases in which noticeable glottalization for /ʔ/ appeared. The columns represent phrasal prosody and the rows indicate whether the target syllable is stressed or reduced in its word. The most striking feature of the table is that the reduced, non-phrase-boundary entries are much lower than the rest, for both subjects. That is, although stressed syllables had a high percentage of noticeable /ʔ/s in all positions, reduced syllables had a low percentage except at a phrase boundary. This

Table 4.1 *Percentage of tokens with noticeable /ʔ/*

Subject	Stress	%–boundary	Accented	Deaccented
MR	stressed	100	85	100
	reduced	93	33	44
DT	stressed	90	95	80
	reduced	97	17	27

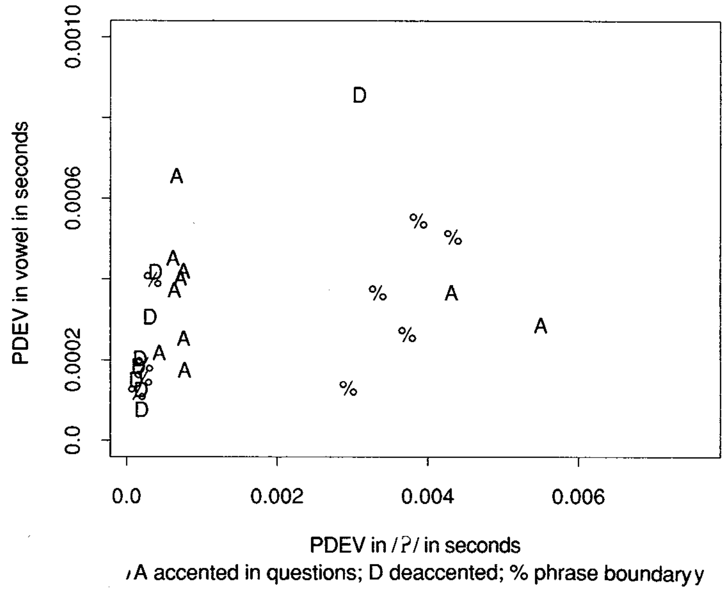

Figure 4.18 PDEV for /ʔ/ beginning *August* and *awkwardness* plotted against PDEV in the following vowel, for subject DT

result shows that word-level and phrase-level prosody interact to determine the likelihood of observed glottalization. It does not provide any information about the degree of glottalization in cases where it was equally likely. In figure 4.18, PDEV is used to make this comparison for subject DT, for syllables with word stress. Only utterances in which glottalization was observed are included. In the deaccented tokens, PDEV during /ʔ/ was overall much lower than in the accented phrase-medial tokens or the phrase-initial tokens.

4.6 Discussion and conclusions

The experiment showed that the pronunciation of both /h/ and /ʔ/ depends on word- and phrase-level prosody. We decompose these effects into effects on gestural magnitude and background effects. An overall shift in a vocalic direction was associated with accent, beginning at the rhyme of the accented syllable and affecting even later syllables in the same word. The phrase boundary was found to shift articulation on both sides in a more consonantal direction; related phrase-initial lengthening of the consonant, analogous to the phrase-final lengthening observed by many other researchers, was also observed. Superimposed on the background effects we observe effects on gestural magnitude related to the strength of a segment's prosodic position in the word and in the phrase. Accent affected the gestural magnitude both for main stressed and reduced syllables within the accented word, but it affected the stressed syllables more. There is also some evidence for a phrase-boundary effect on gestural magnitude, although further investigation is called for.

The interaction of effects on gestural magnitude and background effects is highly reminiscent of the interactions between local and large-scale effects which have proved critical for modeling the manifestations of tone in F_0 contours. The effects on gestural magnitude for /h/ and /ʔ/ are broadly analogous to the computations involved in mapping tones into F_0 target levels or excursions, while the background effects are reminiscent of effects such as declination and final lowering which affect the F_0 values achieved for tones in an entire region. Thus, the experimental results support a parallel treatment of segments and tones in terms of their phonological representation and the phonetic realization rules which interpret them. They argue against the view which segregates tone and intonation into a "suprasegmental" component, a view which still underlies current speech technology (Lea 1980; Allen, Hunnicutt, and Klatt 1987; Waibel 1988). This view provides for prosodic effects on F_0, intensity, and duration, but does not support the representations or rules needed to describe prosodic effects on segmental allophony of the kind observed here.

Our observations about /h/ and /ʔ/ production broadly support the ideas about phonetic representation expressed in Browman and Goldstein (1990) and Saltzman and Munhall (1989), as against the approach of the International Phonetic Association or *The Sound Pattern of English* (Chomsky and Halle 1968). Gradient or n-ary features on individual segments would not well represent the pattern of lenition observed here; for example, equally lenited /h/s can be pronounced differently in different positions, and equally voiced /h/s can represent different degrees of lenition in different positions. An intrinsically quantitative representation, oriented towards critical aspects of articulation, appears to offer more insight than the traditional fine phonetic transcription. At the same time, the present results draw attention to the need for work on articulatory representation to include a proper treatment of hierarchical structure and its manifestations. A quantitative articulatory description will still fail to capture the multidimensional character of lenition if it handles only local phonological and phonetic properties.

Comments on chapter 4

OSAMU FUJIMURA

First of all, I must express my appreciation of the careful preparation by Pierrehumbert and Talkin of the experimental material. Subtle phonetic interactions among various factors such as F_0, F_1, and vocal-tract constriction are carefully measured and assessed using state-of-the-art signal-processing technology. This makes it possible to study subtle but critical effects of prosodic factors on segmental characteristics with respect to vocal-source control. In this experiment, every technical detail counts, from the way the signals are recorded to the simultaneous control of several phonological conditions.

Effects of suprasegmental factors on segmental properties, particularly of syntagmatic or configurational factors, have been studied by relatively few investigators beyond qualitative or impressionistic description of allophonic variation. It is difficult to prepare systematically controlled paradigms of contrasting materials, partly because nonsense materials do not serve the purpose in this type of work, and linguistic interactions amongst factors to be controlled prohibit an orthogonal material design. Nevertheless, this work exemplifies what can be done, and why it is worth the effort. It is perhaps a typical situation of laboratory phonology.

The general point this study attempts to demonstrate is that so-called "segmental" aspects of speech interact strongly with "prosodic" or "suprasegmental" factors. Paradoxically, based on the traditional concept of segment, one might call this situation "segmental effects of suprasegmental conditions." As Pierrehumbert and Talkin note, such basic concepts are being challenged. Tones as abstract entities in phonological representations manifest themselves in different fundamental frequencies. Likewise, phonemes or distinctive-feature values in lexical representations are realized with different phonetic features, such as voiced and voiceless or with and without articulatory closure, depending on the configuration (e.g. syllable- or word-initial vs. final) and accentual situations in which the phoneme occurs. The same phonetic segments, to the extent that they can be identified as such, may correspond to different phonological units. Pierrehumbert and Talkin, clarifying the line recently proposed by Pierrehumbert and Beckman (1988), use the terms 'structure' and 'content' to describe the general framework of phonological/phonetic representations of speech. The structure, in my interpretation (Fujimura 1987, 1990), is a syntagmatic frame (the skeleton) which Jakobson, Fant, and Halle (1952) roughly characterized by configurational features. The content (the melody in each autosegmental tier) was described in more detail in distinctive (inherent and prosodic) features in the same classical treatise. Among different aspects of articulatory features, Pierrehumbert and Talkin's paper deals with voice-source features, in particular with glottal consonants functioning as the initial margin of syllables in English.

What is called a glottal stop is not very well understood and varies greatly. The authors interpret acoustic EGG signal characteristics to be due to braced configurations of the vocal folds. What they mean by "braced" is not clear to me. They "surmise" that the subject DT in particular produces the glottal stop by bracing or tensing partially abducted vocal folds in a way that tends to create irregular vibration without a fully closed phase. Given the current progress of our understanding of the vocal-fold vibration mechanism and its physiological control, and the existence of advanced techniques for direct and very detailed optical observation of the vocal folds, such qualitative and largely intuitive interpretation will, I hope, be replaced by solid knowledge in the near future. Recent developments in the technique of high-speed optical recording of laryngeal movement, as reported by Kiritani and his co-workers at the University of Tokyo (RILP), seem to promise a rapid growth of our knowledge in this area.

A preliminary study using direct optical observation with a fiberscope (Fujimura and Sawashima 1971) revealed that variants of American English /t/ were accompanied by characteristic gestures of the false vocal folds. Physiologically, laryngeal control involves many degrees of freedom, and EGG observations, much less acoustic signals, reveal little information about specific gestural characteristics. What is considered in the sparse distinctive-feature literature about voice-source features tends to be grossly impressionistic or even simply conjectural with respect to the production-control mechanisms. The present paper raises good questions and shows the right way to carry out an instrumental study of this complex issue. Particularly in this context, Pierrehumbert and Talkin's detailed discussion of their speech material is very timely and most welcome, along with the inherent value of the very careful measurement and analysis of the acoustic-signal characteristics. This combination of phonological (particularly intonation-theoretical) competence and experimental-phonetic (particularly speech-signal engineering) expertise is a necessary condition for this type of study, even just for preparing effective utterances for examination. Incidentally, it was in these carefully selected sample sentences that the authors recently made the striking discovery that a low-tone combination of voice-source characteristics gives rise to a distinctly different spectral envelope (personal communication).

One of the points of this paper that particularly attracts my attention is the apparently basic difference between the two speakers examined. In recent years, I have been impressed by observations that strong interspeaker variation exists even in what we may consider to be rather fundamental control strategies of speech production (see Vaissière 1988 on velum movement strategies, for example). One may hypothesize that different production strategies result in the same acoustic or auditory consequence. However, I do not believe this principle explains the phenomena very well, even though in some cases it is an important principle to consider. In the case of the "glottal stop," it represents a consonantal function in the syllable onset in opposition to /h/, from a distributional point of view. Phonetically (including acoustically), however, it seems that the only way to characterize this consonantal element of the onset (initial demisyllable) is that it lacks any truly consonantal features. This is an interesting issue theoretically in view of some of the ideas related to nonlinear phonology, particularly with respect to underspecification. The phonetic implementation of unspecified features is not necessarily empty, being determined by coarticulation principles only, but can have some *ad hoc* processes that may vary from speaker to speaker to a large extent. In order to complete our description of linguistic specification for sound features, this issue needs much more attention and serious study.

In many ways this experimental work is the first of its kind, and it may open up, together with some other pioneering work of similar nature, a new epoch in speech research. I could not agree more with Pierrehumbert and Talkin's conclusion about the need for work on articulatory representation to include a proper treatment of hierarchical structure and its manifestations. Much attention should be directed to their assertion that a quantitative articulatory description will still fail to capture the multidimensional character of lenition if it handles only the local phonological and phonetic properties. But the issue raised here is probably not limited to the notion of lenition.

Comments on chapters 3 and 4

LOUIS GOLDSTEIN

Introduction

The papers in this section, by Pierrehumbert and Talkin and by Beckman, Edwards, and Fletcher, can both be seen as addressing the same fundamental question: namely, how are the spatiotemporal characteristics of speech gestures modulated (i.e., stretched and squeezed) in different prosodic environments?* One paper examines a glottal gesture (laryngeal abduction/adduction for /h/– Pierrehumbert and Talkin), the other an oral gesture (labial closure/opening for /p/– Beckman, Edwards, and Fletcher). The results are similar for the different classes of gestures, even though differences in methods (acoustic analysis vs. articulator tracking) and materials makes a point-by-point comparison impossible. In general, the studies find that phrasal accent increases the magnitude of a gesture, in both space and time, while phrasal boundaries increase the duration of a gesture without a concomitant spatial change. This striking similarity across gestures that employ anatomically distinct (and physiologically very different) structures argues that general principles are at work here. This similarity (and its implications) are the focus of my remarks. I will first present additional evidence showing the generality of prosodic effects across gesture type. Second, I will examine the oral gestures in more detail, asking how the prosodic effects are distributed across the multiple articulators whose motions contribute to an oral constriction. Finally, I will ask whether we yet have an adequate understanding of the general principles involved.

Generality of prosodic effects across gesture type

The papers under discussion show systematic effects of *phrasal* prosodic variables that cut across gesture type (oral/laryngeal). This extends the parallelism between oral and laryngeal gestures that was demonstrated for word stress by Munhall, Ostry, and Parush (1985). In their study, talkers produced the utterance /kakak/, with stress on either the first or second syllable. Tongue-lowering and laryngeal-adduction gestures for the intervocalic /k/ were measured using pulsed ultrasound. The same effects were observed for the two gesture types: words with second-syllable stress showed larger gestures with longer durations. In addition, their analyses showed that the two gesture types had the same velocity profile, a mathematical characterization of the shape of curve showing how velocity varies over time in the course of the gestures. On the basis of this identity of velocity profiles, the authors conclude that "the tongue and vocal folds share common principles of control" (1985: 468).

Glottal gestures involving laryngeal abduction and adduction may occur with a coordinated oral-consonant gesture, as in the case of the /k/s analyzed by Munhall, Ostry, and Parush, or without such an oral gesture, as in the /h/s analyzed by Pierrehumbert and Talkin. It would be interesting to investigate whether the prosodic influences on laryngeal gestures show the same patterns in these two cases. There is at least one reason to think that they might behave differently, due to the differing aerodynamic and acoustic consequences. Kingston (1990) has argued that the temporal coordination of laryngeal and oral gestures could be more tightly constrained when the oral gesture is an obstruent than when it is a sonorant, because there are critical aerodynamic consequences of the glottal gesture in obstruents (allowing generation of release bursts and frication). By the same logic, we might expect the size (in time and space) of a laryngeal gesture to be relatively more constrained when it is coordinated with an oral-obstruent gesture than when it is not (as in /h/). The size (and timing) of a laryngeal gesture coordinated with an oral closure will determine the stop's voice-onset time (VOT), and therefore whether it is perceived as aspirated or not, while there are no comparable consequences in the case of /h/. On the other hand, these differences may prove to be irrelevant to the prosodic effects.

In order to test whether there are differences in the behavior of the laryngeal gesture in these two cases, I compared the word-level prosodic effects in Pierrehumbert and Talkin's /h/ data (some that were discussed by the authors and others that I estimated from their graphs) with the data of a recent experiment oy Cooper (forthcoming). Cooper had subjects produce trisyllabic words with varying stress patterns (e.g. *percolate, passionate, Pandora, permissive, Pekingese*), and then reproduce the prosodic pattern on a repetitive /pipipip/ sequence. The glottal gestures in these nonsense words were measured by means of transillumination. I was able to make three comparisons between Cooper and Pierrehumbert and Talkin, all of which showed that the effects generalized over the presence or absence of a coordinated oral gesture. (1) There is very little difference in gestural magnitude between word-initial and word-medial positions for a stressed /h/, *hawkweed* vs. *mahogany*. (2) There is, however, a word-position effect for unstressed syllables (*hibachi* shows a larger gesture than *Omaha* or *tomahawk*). (3) The laryngeal gesture in word-initial position is longer when that syllable is stressed than when it

*This work was supported by NSF grant BNS 8820099 and NIH grants HD-01994 and HD-13617 to Haskins Laboratories.

is unstressed (*hawkweed* vs. *hibachi*). All of these effects can be seen in Cooper's data.

In addition, Cooper's data show a very large reduction of the laryngeal gesture in a reduced syllable immediately following the stressed vowel (in the second /p/ of utterances modeled after *percolate* and *passionate*). In many cases, no laryngeal spreading was observable at all. While this environment was not investigated by Pierrehumbert and Talkin, they note that such /h/s have been considered by phonologists as being deleted altogether (e.g. *vehicle* vs. *vehicular*). The coincidence of these effects is again striking. Moreover, this is an environment where oral gestures may also be severely reduced: tongue-tip closure gestures reduce to flaps (Kahn 1976). Thus, there is strong parallelism between prosodic effects on laryngeal gestures for /h/ and on those that are coordinated with oral stops. This similarity is particularly impressive in face of the very different acoustic consequences of laryngeal gesture in the two cases: generation of breathy voice (/h/) and generation of voiceless intervals. It would seem, therefore, that it is the gestural dynamics themselves that are being directly modulated by stress and position, rather than the output variables such as VOT. The changes can be stated most generally at the level of gestural kinematics and/or dynamics.

Articulatory locus of prosodic effects for oral gestures

The oral gestures analyzed by Beckman, Edwards, and Fletcher are bilabial closures and openings into the following vowel. Bilabial closures are achieved by coordinated action of three separate articulatory degrees of freedom: jaw displacement, displacement of the lower lip with respect to the jaw, and displacement of the upper lip with respect to the upper teeth. The goal of bilabial closure can be defined in terms of the vertical distance between the upper and lower lips, which needs to be reduced to zero (or to a negative value, indicating lip compression). This distance has been shown to be relatively invariant for a bilabial stop produced in different vowel contexts (Sussman, MacNeilage, and Hanson 1973: Macchi 1988), while the contributions of the individual articulators vary systematically – high-vowel contexts show both higher jaw positions and less displacement of the lower lip with respect to the jaw than is found in low-vowel contexts. Token-to-token variation shows a similarly systematic pattern (Gracco and Abbs 1986). This vertical interlip distance, or **lip aperture**, is an example of what we call "vocal-tract variables" within the computational gestural model being developed at Haskins Laboratories (Browman and Goldstein 1986, 1990; Saltzman 1986; Saltzman *et al.* 1988a). Gestures are the primitive

phonological units; each is modeled as a dynamical system, or control regime, whose spatial goals are defined in terms of tract variables such as these. When a given gesture is active, the individual articulatory components that can contribute to a given tract variable constitute a "coordinative structure" (Kelso, Saltzman, and Tuller, 1986) and cooperate to achieve the tract-variable goal. Individual articulators may compensate for one another, when one is mechanically restrained or involved in another concurrent speech gesture. In this fashion, articulatory differences in different contexts are modeled (Saltzman 1986). With respect to the prosodic effects discussed by Beckman, Edwards, and Fletcher, it is important to know whether the gesture's tract-variable goals are being modified, or rather if some individual articulator's motions are being amplified or reduced, and if so, whether these changes are compensated for by other articulators.

Since Beckman, Edwards, and Fletcher present data only for the jaw, the question cannot be answered directly. However, Macchi (1988) has attempted to answer this question using data similar to the type employed by them. Macchi finds that prosodic variation (stress, syllable structure) influences primarily the activity of the jaw, but that, unlike variation introduced by vowel environment, the prosodic effects on the jaw are not compensated for by the displacement of the lower lip with respect to the jaw. The displacement remains invariant across prosodic contexts. Thus, the position of the lower lip in space does vary as a function of prosodic environment, but this variation is caused almost exclusively by jaw differences. That is, the locus of the prosodic effects is the jaw.

Unifying oral and laryngeal effects

If Macchi's analysis is correct (and generalizes beyond the speakers and utterances analyzed), it fits well with Beckman, Edwards, and Fletcher's characterization of prosodic effects in terms of a **sonority** variable. It suggests that it would be possible to develop a model in which "segmental" structure is modeled by gestures defined in the space of tract variables, such as Lip Aperture, while prosodic effects would be modeled by long-term "gestures" defined in the space of Sonority, which would be related directly to jaw position. The association between the jaw height and sonority (or vocal-tract "openness") is an attractive one (although Keating [1983] presents some problems with using it as a basis for explaining why segments order according to the sonority hierarchy).

A major problem with this view emerges, however, if we return to the laryngeal data. Here, as we have seen, the effects parallel those for oral gestures, in terms of changes in gesture magnitude and duration. Yet it would be hard to include the laryngeal effects under the rubric of sonority, at least as traditionally defined. Phrasal accent results in a more open glottis, but a more open glottis would result in a less sonorous output. Thus the sonority analysis fails to account, in a unified way, for the parallel prosodic modulations of the laryngeal and oral gestures. An alternative analysis would be to examine the effects in terms of the overall amount of energy expended by the gestures, accented syllables being more energetic. However, this would not explain why the effects on oral gestures seem to be restricted to the jaw (which *is* explained by the sonority account). Finding a unified account of laryngeal and oral effects remains an exciting challenge.

Comments on chapters 3 and 4

IRENE VOGEL

Prosodic structure

Chapters 3 and 4, in common with those in the prosody section of this volume, all view the structure of phonology as consisting of hierarchically arranged phonological, or prosodic, constituents.* The phonetic phenomena under investigation, furthermore, are shown to depend crucially on such constituents in the correct characterization of their domains of application. Of particular importance is the position of specific elements within the various constituents. As Pierrehumbert and Talkin suggest, "phonetic realization rules in general can be sensitive to prosodic structure," whether they deal with tonal, segmental, or, presumably, durational phenomena. In fact, a large part of the phonology–phonetics interface seems to involve precisely the matching up of the hierarchical structures – the phonology – and the physical realizations of the specific tonal, segmental, and durational phenomena – the phonetics.

This issue – the role of phonological constituents in phonetics implementation – leads directly to the next point: precisely, what are the phonological constituents that are relevant for phonetics? A common view of phonology groups speech sounds into the following set of constituents (from the word up):

(1) phonological utterance;
intonational phrase;
phonological phrase;
clitic group;
phonological word;

Phonological constituents referred to in this volume, however, include the following:

(2) (a) Pierrehumbert and Talkin:
(intonational) phrase
(phonological) word
(b) Beckman, Edwards, and Fletcher:
(intonation) phrase
(phonological word)
(c) Kubozono:
major phrase
minor phrase
(d) van den Berg, Gussenhoven, and Rietveld:
association domain
association domain′

Given such an array of proposed phonological constituents, it is important to stop and ask a number of basic questions. First of all, do we expect any, or possibly all, of the various levels of phonological structure to be universal? If not, we run the risk of circularity: a phenomenon P in some language is found to apply within certain types of strings which we thus define as a phonological constituent C; C is then claimed to be motivated as a phonological constituent of the language because it is the domain of application of P. In so doing, however, we lose any predictive power phonological constituents may have, not to mention the fact that we potentially admit an infinite number of language types in terms of their phonological constituent structure. It would thus be preferable to claim that there is some finite, independently motivated, universal set of phonological constituents. But what are these?

The constituents in (1) were originally proposed and motivated as such primarily on the basis of phonological rules (e.g. Selkirk 1978; Nespor and Vogel 1986). In various papers in this volume we find phonetic data arguing in favor of phonological constituents, but with some different names. Pierrehumbert and Talkin as well as Beckman, Edwards, and Fletcher assume essentially the structures in (1), though they do not state what

*This was presented at the conference as a commentary on several papers, but because of the organization of the volume appears here with chapters 3 and 4.

definitions they are using for their constituents. In Kubozono's paper, however, we find major phrase and minor phrase, and, since neither is explicitly defined, we do not know how they relate to the other proposed constituents. Similarly, van den Berg, Gussenhoven, and Rietveld explicitly claim that their association domain and association domain′ do not coincide with any phonological constituents proposed elsewhere in the literature. Does this mean we are, in fact, adopting the position that essentially anything goes, where we just create phonological constituent structure as we need it? Given the impressive cross-linguistic insights that have been gained in phonology by identifying a small finite set of prosodic constituents along the lines of (1), it would be surprising, and dismaying, if phonetic investigation yielded significantly different results.

Phonology and phonetics

It might be said that anything that is rule-governed and thus predictable is part of competence and should therefore be considered phonology. If this is so, one could also argue that "phonetic implementation rules" are phonology since they, too, follow rule-governed patterns (expressed, for example, in parametric models). Is phonetics, then, just the "mechanical" part of the picture? This is probably too extreme a position to defend, but it is not clear how and where exactly we are to draw the line between what is phonological and what is phonetic.

One of the stock (if simplified) answers to this question is something like the following: phonology deals with unique, idealized representations of speech, while phonetics deals with their actual manifestations. Since in theory infinite variation is possible for any idealized phonological representation, a question arises as to how we know whether particular variations are acceptable for a given phenomenon. Which variations do we consider in our research and which may/must we exclude? Some of the papers in this volume report that it was necessary to set aside certain speakers and/or data because the speakers were unable to produce the necessary phenomena. This is all the more surprising since the data were collected in controlled laboratory settings where we would expect there to be less variation that usual. Furthermore, in more than one case, it seems that the data found to be most reliable and crucial to the study were produced by someone involved in the research. This is not meant necessarily as a methodological criticism, since much can be gained by examining constrained sets of data. It does, however, raise serious questions about interpretation of the results. Moreover, if in phonetic analyses, too, we have to pull back from the reality of variation, we blur the distinction between phonology and phonetics, since abstraction and idealization may no longer be considered

defining characteristics of phonology. Of course, if the goal of phonetics is to model specific phenomena, as is often the case, we do need to end up with abstractions again. We must still ask, though, what is actually being modeled when this model itself is based on such limited sets of data.