Acoustic segment durations in prosodic research: a practical guide

CHAPTER · JANUARY 2006

CITATIONS
56

READS
225

3 AUTHORS, INCLUDING:

Alice Turk
The University of Edinburgh
65 PUBLICATIONS **1,706** CITATIONS
SEE PROFILE

Satsuki Nakai
Queen Margaret University
26 PUBLICATIONS **365** CITATIONS
SEE PROFILE

Methods in Empirical Prosody Research

Edited by
Stefan Sudhoff, Denisa Lenertová, Roland Meyer,
Sandra Pappert, Petra Augurzky, Ina Mleinek,
Nicole Richter, Johannes Schließer

Offprint

Walter de Gruyter · Berlin · New York

Alice Turk, Satsuki Nakai (Edinburgh) &
Mariko Sugahara (Kyoto)*

Acoustic Segment Durations in Prosodic Research: A Practical Guide

1 Introduction

Carefully designed durational experiments are promising tools for testing and formulating theories of prosodic structure, its relationship with grammar, and its phonetic implementation. If properly designed, they allow for tight control of prosodic variables of interest, and can yield reliable durational measurements. Results from these tightly controlled experiments can then be used to form hypotheses about the way segment durations vary in more natural speech situations, or can be used to test hypotheses based on observations of natural speech corpora.

In this paper, we discuss methodological issues relating to such studies. In the first part of the paper, we outline principles of reliable and accurate acoustic speech segmentation that allow us to make inferences about the durations of consonantal constrictions and surrounding, mostly vocalic, intervals. In doing so, we discuss the relative segmentability of a range of segment types, in the hope that this will help researchers to design materials with the maximum likelihood of accurate segmentation. In the second part of the paper, we discuss additional methodological issues relating to the design of durational experiments. These include ways of designing materials to control for sources of known durational variability, and methods for eliciting prosodic contrasts.

* We thank Matthew Aylett, Simon King, Peter Ladefoged, Jim Scobbie, Laurence White, Ivan Yuen, and especially Stefanie Shattuck-Hufnagel and Jim Sawusch for discussion of ideas presented here, and Bert Remijsen for detailed comments on a pre-final version of this chapter. We are also grateful to two anonymous reviewers for their useful comments, to Sari Kunnari for help in collecting the Finnish data, and to Kari Suomi and Richard Ogden for helpful information regarding Finnish phonology and phonetics. This work was supported by Leverhulme, and British Academy grants to the first and second authors, and an AHRC grant to the first and third authors.

2 Principles of acoustic speech segmentation

Segmenting the speech signal into phone-sized units is somewhat of an artificial task, since the gestures used to produce successive speech sounds overlap to a great degree, as illustrated in Browman and Goldstein (1990) and elsewhere. For example, the closing gesture tongue movement for /g/ in the phrase *Say guide walls* (Figure 1) begins before the end of the preceding vowel /e/, as evidenced by the rising F2 formant transition for this vowel.[1] This situation of articulatory overlap makes it difficult to determine the point in the acoustic signal where the vowel ends and the consonant begins. Nevertheless, there are often salient acoustic landmarks that correspond straightforwardly to recognisable articulatory events (Stevens, 2002). In particular, although we know that movement towards consonantal constriction begins earlier, abrupt spectral changes coincide with the onsets and releases of oral consonantal constrictions for the production of stops, fricatives, and affricates, as illustrated in Figure 1 (/s, g, d, z/) and Figure 2 (/s, p/).

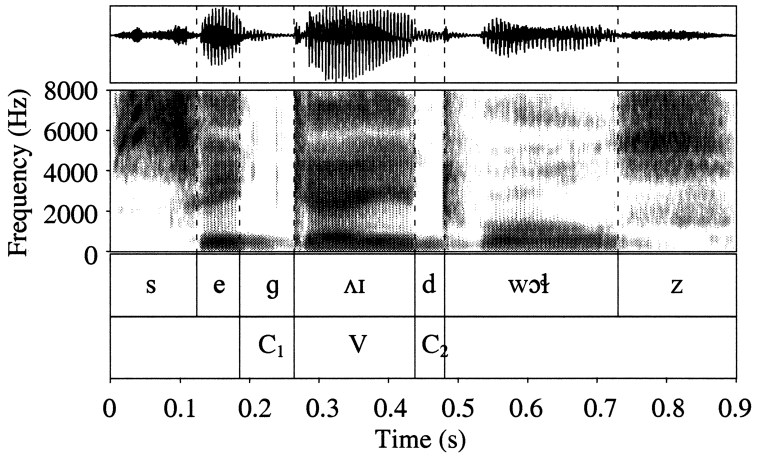

Figure 1: *Say guide walls,* spoken by a female Scottish English speaker

[1] /e/ is monophthongal in Scottish English.

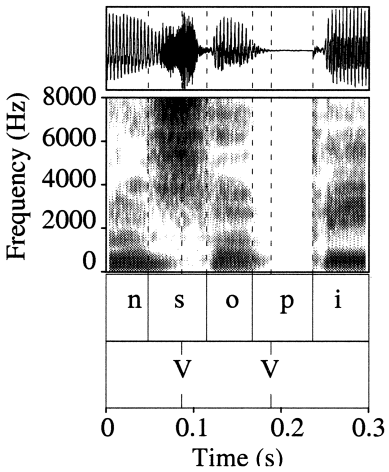

Figure 2: A fragment (underlined) from *MINUSTA "san" sopii kohtaan tuhat-kaksisataa* 'I THINK "san" fits [#] 1200', spoken by a female Northern Finnish speaker. *San* is a nonsense word. *V* in the second label tier indicates the offset of voicing for [s], and [p].

We propose that acoustic segment durations should be determined by the intervals that these oral consonantal constriction events define. Oral constriction criteria are preferable to criteria based on the onset or offset of voicing, since oral constriction criteria can be used comparably for many different classes of speech sounds, including voiced and voiceless oral stops, fricatives, affricates, and nasal stops. Although oral constriction and voicing criteria might be thought to be interchangeable in some cases, e.g. at the onsets of voiceless obstruent constrictions in vowel-voiceless obstruent sequences, e.g. as between word-medial [o] and [p] in *sopii*, voicing often persists after the onset of the phonologically voiceless constriction (e.g. [s] and [p] in Figure 2). In situations of this type, the oral constriction onset criterion is clearly preferable.

As can be clearly seen in Figure 2, oral constriction criteria can yield very different segment durations than criteria based on voicing. Similar discrepancies are observed in situations where aspirated voiceless stop offsets are measured. These potential differences also make a strong case for being explicit about segmentation criteria in reports, and above all for application of consistent segmentation criteria.

The duration of an interval between a C_1 constriction release landmark and a following C_2 constriction onset landmark in a C_1VC_2 sequence (e.g. the [ʌɪ] interval in Figure 1) is often described as the duration of a "vowel". We follow this convention here. However, this interval is not exclusively vocalic, since the so-called "vowel" duration includes formant transitions and burst noise that cue the identity of the surrounding consonants, in addition to any aspiration

from preceding voiceless aspirated stops. This point should be kept in mind when interpreting labels for such intervals.

We argue that the judicious choice of experimental materials can yield reliable, accurate durational measurements, if the materials contain alternations of salient oral consonantal constrictions and sonorant segments such as vowels. In particular, constriction onsets and releases are relatively easy to identify in: (1) stop consonants, e.g. [p, t, k, b, d, g], sibilants, e.g. [s, ʃ, z, ʒ], and affricates, e.g. [tʃ, dʒ] in VCV contexts and (2) non-homorganic clusters (clusters containing consonants of different places of articulation) differing in manner. We will discuss segmentation criteria for these sequences below.

2.1 Relative segmentability

There is a clear relationship between segmentation reliability and the strength of conclusions that can be drawn from experiments that use segmentation as part of their methodology. In order to ensure confidence in results of durational experiments, we recommend that materials be designed with the highest possible number of target segments whose durations can be reliably and accurately estimated.

In the following sections, we present detailed segmentation criteria for sequences of segments shown in Table 1. These criteria derive from the theory of the relationship between articulation and acoustics (see Stevens, 2002), and from our experience in segmenting American English, Standard Scottish English, and, to a lesser extent, Southern Standard British English, Standard Dutch, Northern Finnish and Standard Japanese. Although grounded in general acoustic theory as it relates to speech production, there may be language specific factors, such as allophonic variation, assimilation or coarticulation patterns that may make some of these specific criteria less applicable for particular languages or language varieties.

Table 1 includes 1) phones that we have found to be reliably segmentable in most contexts, 2) phones which we have found to be reliably segmentable in restricted contexts, 3) phones which we have found to be less reliably segmentable in most contexts, and 4) others which are to be avoided whenever possible. The phone classes mentioned in Table 1 conform to definitions given in Ladefoged (2001); we have defined additional terms not explicitly mentioned there. Not all phone types are included; we only discuss cases that we have had sufficient experience with to describe with confidence.

It should be noted that nasal stops are the most appropriate class of segments for experiments where both duration and F0 are of interest. Obstruents are known to raise or lower F0 in adjacent pitch periods depending on their voicing specification, and are therefore less appropriate for F0 analyses.

	Boundary between consonant and vowel in CV or VC sequences, where consonants are:	Boundary between two members of a consonant cluster, where phones in clusters are:
Reliably segmented in most contexts	Oral stops, e.g. [p, b, t, d, k, g] Sibilants, e.g. [s, ʃ, z, ʒ] Affricates, e.g. [tʃ, dʒ]	Oral stops, nasal stops, sibilants, and affricates in the following sequences, when these differ in place and manner of articulation: Sonorant consonant*-oral stop Sonorant consonant-sibilant Sonorant consonant-affricate Oral stop-sonorant consonant Sibilant-sonorant consonant Affricate-sonorant consonant Sibilant-oral stop Oral stop-sibilant Nasal stop-sibilant Sibilant-nasal stop *Sonorant consonant = approximants or nasal stops, e.g. [l, j, w, m]
Reliably segmented in some contexts	Nasal stops, e.g. [n, m] Weak voiceless fricatives, e.g. [f, θ]	
Less reliably segmented		Weak voiceless fricatives Nasal or voiceless stops in homorganic nasal-stop or stop-nasal clusters, e.g. [mp, pm]
To be avoided	Central and lateral approximants, e.g. [w,l]; [h] Weak voiced fricatives, e.g. [v, ð]	Voiceless and voiced consonants in homorganic clusters, e.g. [st], [mb] Consonants in clusters sharing manner of articulation, e.g. [pk], [bt], [mn], [sʃ] Stop-affricate clusters

Table 1: Relative segmentability of consonants in VCV and cluster contexts

2.2 Segmentation criteria

The detailed segmentation criteria that we present in the following sections are all based on the more general strategy of finding constriction onsets and re-

leases, as described above. Most of the criteria we discuss are based on spectral characteristics most easily seen in spectrograms. Waveforms can also be useful for segmentation since they show dips and rises in amplitude, which often correspond to the onsets of constrictions and their release. However, amplitude dips can sometimes be gradual on waveform displays, particularly when constrictions are voiced, and some types of frication noise can be difficult to distinguish from aspiration noise on waveform displays. For these reasons, we prefer to rely primarily on spectrograms for first-pass segmentation decisions within an accuracy of 5-10 ms, and on waveforms for more fine-grained segmentation decisions, once general boundary regions have been defined.

Note that segmentation accuracy will necessarily depend on factors other than segmentation criteria, namely 1) the sampling rate used in digitising the signal, and 2) the spectrogram analysis window size, assuming that spectrograms are used for segmentation, and 3) the degree to which each successive analysis window overlaps (frame shift). For example, a sampling rate of 16,000 Hz will yield accuracy within .0625 ms if segmenting on the waveform, but a spectrogram analysis window size of 5 ms (for 200 Hz wideband analysis) will limit the accuracy of spectrogram-based criteria to within this 5 ms window. The reduced accuracy of spectrogram-based criteria as compared to those based on the waveform supports the use of the waveform for final fine-grained segmentation once segment boundaries have already been determined within 5-10 ms.

When using visual displays for segmentation, it is easier to see gross spectral changes when these are zoomed out, or contain longer stretches of speech. We recommend more zoomed out spectrogram displays to determine general boundary regions, and more zoomed in waveform displays for determining exact boundary locations.

In the following sections, we discuss segmentation criteria in rough order of relative segmentability, as organised in Table 1.

2.2.1 Consonants in VCV contexts

Oral stops

In our experience, canonical variants of oral stops are generally easy to segment (see [g, d] in Figure 1, [p] in Figure 2). The onset of stop closures in VCV contexts are associated with 1) a decrease in overall amplitude, and 2) cessation of all but the lowest formant and harmonic energy. Although some stop closures are also accompanied by the cessation of voicing, many voiced stops and even some phonemically voiceless stops have voicing that continues through part or all of the stop closure (see [t] and the second [p] in Figure 3). In addition, for some vowel-voiceless stop sequences, voicing can stop earlier

than the oral closure, resulting in pre-aspiration (see also pre-aspiration before a fricative in Figure 6). These phenomena (cf. Lucero, 1999) highlight the importance of non-F0 based criteria when identifying stop closures. In many cases the offset of F2 energy coinciding with an overall dip in amplitude is the best criterion because F2 and higher frequency energy is often critically damped when the oral tract is closed. Using F2 to identify stop closure is preferred over using F1, because F1 energy is less often critically damped and is often confusable with F0.

It should be noted that for some speakers under very sensitive recording conditions, even F2 energy can continue into closure. In these cases, the offset of F3 and higher frequency energy coinciding with a drop in overall amplitude would provide a better criterion than F2.

In syllable-final position, vowels before English voiceless stops are often glottalised, as shown in Figure 3 ([a] and [e]); in these cases, the end of formant (e.g. F2) energy can still be used as a criterion for finding oral closure onset if a full glottal stop has not occurred before the oral constriction has been made. The absence of a full glottal stop before oral closure can be diagnosed by 1) voicing that continues after the end of formant energy, and/or formant values at closure that are appropriate for the place of articulation of the stop. In Figure 3, although the [e] before the second [p] in *paper* is glottalised, there is no evidence that a glottal stop precedes the closure: Voicing continues after the end of formant energy, and the F2 value at the onset of [p] closure looks appropriate for a bilabial stop. Evidence for the lack of a glottal stop before the onset of [k] closure in *tax* comes from low amplitude voicing during the closure. Note that closure onset criteria cannot be applied where glottal stop has fully replaced an oral closure, as for glottal stop variants of /t/ in English.

In Figure 3, weak high frequency frication noise occurs at the beginning of the first /p/ constriction, and throughout the second /p/ constriction. This frication noise is evidence of incomplete closure.

Stop releases can be easily identified in the presence of a release burst, whose onset can be taken as the release (end of [t], [k] and both [p]'s in Figure 3). Velar stops are often accompanied by multiple bursts (e.g. both [k]'s in Figure 4, at times .2 and .8). Any of the multiple bursts (e.g. the first, last, or most salient burst) could potentially be used to mark the offset of the stop, as long as the choice is used consistently where measurements are to be compared, but the first burst arguably conforms best to our criterion of marking the constriction release, since subsequent bursts are produced through the (uncontrolled) Bernoulli effect.[2]

---

[2] Thanks to Mark Jones for pointing this out on the Phonet discussion list.

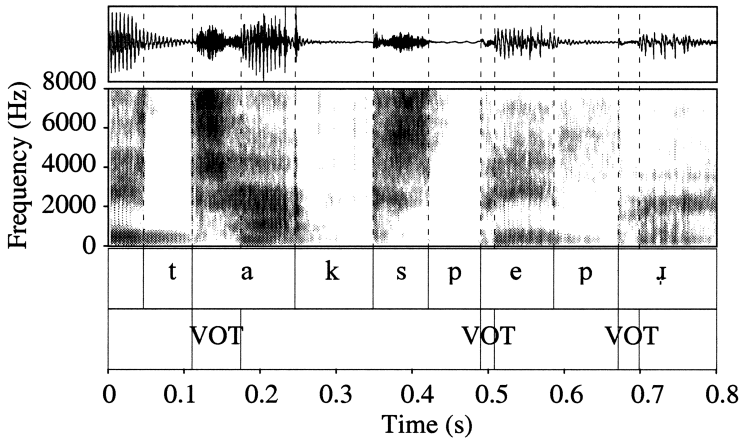

Figure 3: *Tax paper,* spoken by a female Scottish English speaker. The boundaries for the offsets of /a/ and /e/ are placed on the last glottal pulse peak in the intervals delimited by continuous F2.

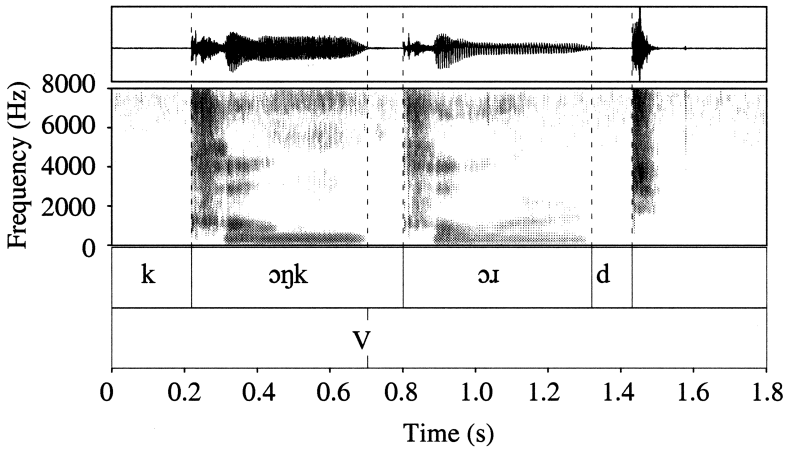

Figure 4: *Concord,* spoken by a female Scottish English speaker. *V* in the second label tier indicates the offset of voicing for [k].

Bursts are sometimes not evident on the spectrogram for voiced stops that are [+anterior], i.e. produced at or in front of the alveolar ridge (e.g. [b, d], as well as the American English flap allophone of /t, d/); in these cases stop release can be considered to occur near the point of F2 onset.

The releases of aspirated voiceless stops in VCV contexts are followed by voiceless aspiration which ends at the onset of voicing (VOT) of the following vowel (see VOT for [t] and [p] in Figure 3). One might wonder whether to

include this interval as a part of the duration of the so-called "stop", where the following "vowel" would begin at VOT. In these cases, it is useful to note that segmentation criteria for VOT are based on acoustic correlates of the onset of vocal fold vibration, and differ qualitatively from the acoustic correlates of oral constriction onsets and releases. It is our view that if segmentation of voiceless stops is to be comparable with that of fricatives and voiced stops, what we measure as voiceless stop durations should also correspond to their oral constriction durations, and should therefore end at oral release. On this view, vowels following voiceless stops would begin at consonantal release, rather than VOT. In our studies, we are careful to apply consonantal constriction criteria consistently across segment types, but do often measure intervals from consonantal release to VOT for other purposes. Note in this regard that waveforms are often more useful than spectrograms for determining voicing onset or offset.

One drawback in using stops in experimental materials is that they can exhibit considerable variability in their allophonic and phonetic realisations. English /t/ is a notorious example in this regard. Many varieties of English frequently use glottal stop in syllable-final position, e.g. *wha*[ʔ] (*what*), as well as in words like *city,* where American English uses a tap. In addition, glottal constrictions before final stops, e.g. before some renditions of [k] in *tack*, can make the onset of oral closure difficult to identify. And finally, /g/ closures in American English and Japanese can be realised as voiced fricatives or approximants (see different phonetic realisations of Japanese /g/ in Figure 5). In American English, these non-canonical /g/ closures generally occur in intervocalic contexts where the second vowel is unstressed, e.g. *ogre.*

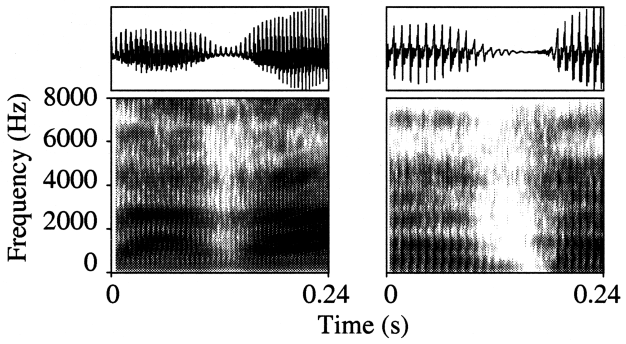

Figure 5: Japanese /aga/ where /g/ is realised as an approximant (left panel; spoken by a female Standard Japanese speaker), and as a fricative (right panel; spoken by a male Standard Japanese speaker). /aga/ is a fragment from *Sei-wa "gansani"-o totemo yorokobu* 'Sei is very pleased with "gansani"'. *Wa* is a topic marker; *gansani* is a nonsense word.

Sibilants

In general, we have found sibilant fricatives to be particularly useful for cross-linguistic studies of segment durations, since they appear to show little allophonic or other phonetic variation within a single language.

In our experience, the onset and offset of frication energy appropriate for a sibilant of its place of articulation is the most useful criterion for sibilant constriction segmentation in VCV contexts (see [z] in Figure 1 and [s] in Figure 2). The presence of fricative noise is unambiguous evidence for an oral constriction, but the onset or offset of frication noise can sometimes appear gradual, or can be confused with breathiness or aspiration noise, making accurate segmentation challenging.[3] The offset/onset of vowel formant energy (e.g. F2, see above discussion of oral stops) can sometimes also be used to identify fricative constriction boundaries, but is often less useful than the fricative noise criterion, since formant energy can often be seen in the presence of frication (see [se] in Figure 1). A small silent gap can sometimes precede or follow frication noise, particularly in high vowel or sonorant consonant contexts; these silent gaps are often due to the change in noise source from just behind the constriction to the vocal folds, or *vice versa* (Stevens, 1998).

Relatively long periods of aspiration noise (equivalent to a partially voiceless vowel) or breathiness can sometimes occur before or after the onset of voiceless frication, e.g. the 'asp' intervals before [ʃ] in *tosh* in Figure 6, and after [s] in Figure 7 (see also Gordeeva and Scobbie, 2004 for a discussion of this phenomenon in Scottish English). This aspiration noise is spectrally different from the adjacent fricative noise, and often contains voiceless formant energy. In these cases it is particularly important not to rely on the cessation of voicing (F0) as a cue to the onset or offset of the fricative. However, it should be noted that at times it may be difficult to find a clear spectral discontinuity between the aspiration noise and fricative noise, see Section 2.4 for a discussion of segmentation uncertainties. In our experience, American English does not appear to have heavily pre-aspirated voiceless consonants, whereas many British varieties do, at least for some speakers.

Phonologically voiced fricatives in intervocalic position can also be problematic, if voicing continues throughout the frication and frication amplitude is relatively low, due to e.g. reduced airflow across the glottis during voicing. Even though the spectral changes due to the onset or offset of frication may be visible on the spectrogram, precise segmentation points are often difficult to determine on a waveform when low amplitude frication noise occurs simultaneously with voicing.

---

[3] Some palatalised fricatives in high, front vowel contexts, e.g. Japanese /zi/ (realised as [ʒi]) can be difficult to segment, as frication can occur simultaneously with the high vowel articulation.

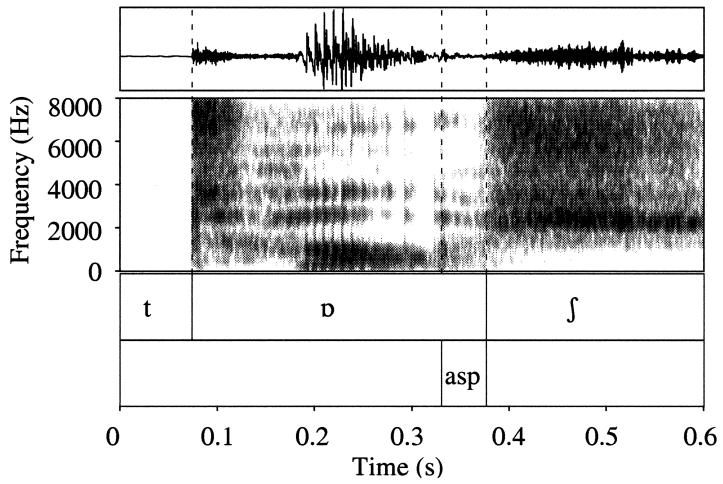

Figure 6: *Tosh,* spoken by a male Southern Standard British English speaker

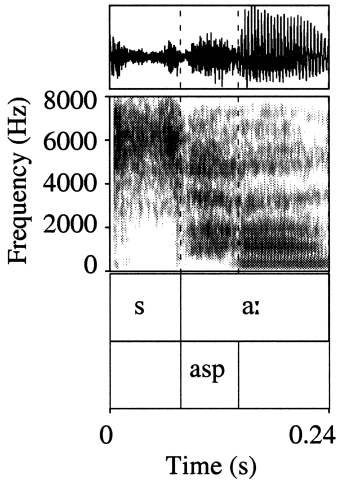

Figure 7: A fragment from *Buun-sensei ICHI-BAN-ga "saasaa"-tte ittakedo* 'Mr. Boone said NUMBER ONE is "saasaa"', spoken by a female Standard Japanese speaker. *Saasaa* is a nonsense word.

Affricates

Since affricates can be considered as sequences of a stop + fricative, criteria for identifying the onsets of affricates will be identical to those for identifying

the onsets of stops. Similarly, criteria for identifying their offsets will be the same as those for identifying the offsets of fricatives.

Nasal stops

Although oral stops and sibilants provide some of the most salient acoustic cues to constriction onsets and releases, other types of segments can also be reliably segmented in some contexts. The oral closures associated with nasal stops are often accompanied by abrupt spectral changes at closure onset and release as illustrated in Figure 8, [m] in Scottish English *max tapes* and Figure 9, [n] in Finnish *san*. In addition, these abrupt spectral changes often coincide with a brief v-like dip-followed-by-a-rise in the waveform as shown in these figures. In contrast to syllable-initial nasals, nasals in syllable-final or word-final position can be difficult to segment in many languages, since the onset of oral closure is often obscured by heavy nasalization on the preceding vowel, and in some cases oral closure can be absent altogether (see Figure 10 [n]). However, there may be some cross-linguistic differences in this regard: Unlike English and Japanese, Finnish is reported to have little anticipatory nasalization of coda nasal stops in word-final positions and consequently often has clear onsets of oral constrictions for word-final nasal stops (Lehiste, 1964, see also Figure 9). We find this observation to be generally true, although some Finnish speakers do nasalise vowels before coda nasal stops in phrase-medial positions.

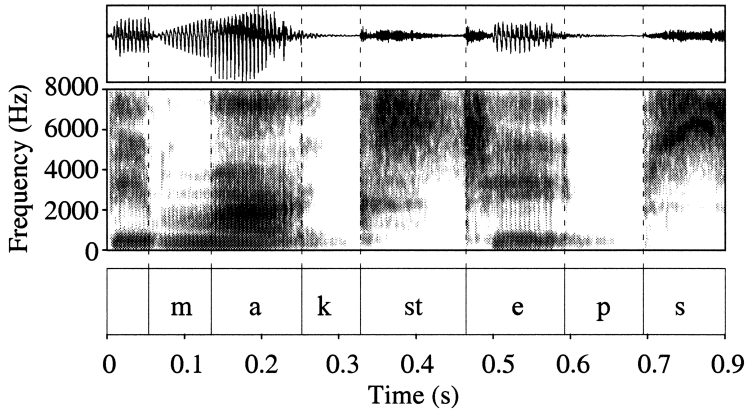

Figure 8: *Max tapes,* spoken by a female Scottish English speaker. The boundaries for the offsets of /a/ and /e/ are placed on the last glottal pulse peak in the intervals delimited by continuous F2.

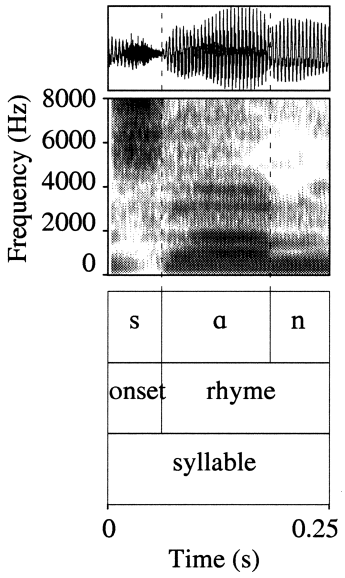

Figure 9: A fragment from *MINUSTA "san" sopii kohtaan tuhatkaksisataa* 'I THINK "san" fits [#] 1200', spoken by a female Northern Finnish speaker. *San* is a nonsense word.

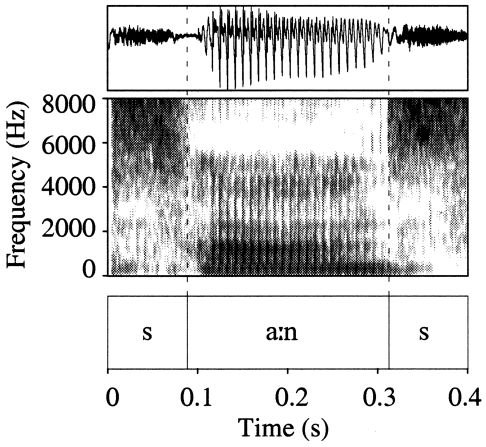

Figure 10: A fragment from *Toujou-sensei-ni kii-tara ICHI-BAN-ga "saansa"* 'According to Mr. Tojo, NUMBER ONE is "saansa"', spoken by a male Standard Japanese speaker. *Saansa* is a nonsense word.

Weak fricatives

Weak voiceless fricative constrictions (e.g. for [f] and [θ]) can often be identified by the onset and offset of frication noise, offset and onset of surrounding vowels' F2, and corresponding dips and rises in overall amplitude. However, at times their frication noise is too weak for reliable identification, e.g. they can sometimes be difficult to distinguish from pause in phrase final position.

Weak voiced fricatives can be even less reliable to segment due to the difficulty of detecting the onset of low amplitude frication in the presence of voicing on a waveform, and to their phonetic variability. For example, English [ð] realisations can sometimes show frication energy similar to [θ], but in other cases this segment can appear very similar to a coronal stop on a spectrogram, while being heard as a clear rendition of [ð] (Zue, 1985).

R-sounds

R-sounds are known to exhibit great variability in their realisation both across and within languages (Ladefoged and Maddieson, 1996). Some of the non-approximant tap or fricative variants can be segmented using criteria similar to those described for stops and sibilant fricatives above. However, variable realisations of /r/ in many languages, including Finnish, Japanese, Dutch, and Scottish English, can make /r/ less useful than other segments for durational experiments. For example, in Scottish English, /r/ can be realised as a tap, an approximant, or can be absent, depending on sociolinguistic and contextual factors (Romaine, 1978). Segmentation of tap variants in this variety is often straightforward, but approximant segmentation is prohibitively difficult, as discussed below.

Central approximants and [h]

The onsets and offsets of constrictions for central approximants (glides, [r, ɻ]) and [h] are notoriously difficult to identify, as shown in Figure 1, [wɔ]). Some researchers suggest using the midpoint of transitions from a preceding vowel to the glide, and the midpoint of transitions from the glide to a following vowel, as criteria for segmentation. We find that these criteria are difficult to implement in many contexts, e.g. where vowels lack reliable steady states. In addition, it is not entirely clear to which articulatory events these transition midpoints correspond, since they do not correspond clearly to points of constriction onset and release that serve to define stop and sibilant constrictions.

Laterals

Although laterals such as [l] can sometimes be associated with clear spectral discontinuity at constriction onset and release, these discontinuities can often be absent. In our experience, there is enough speaker and contextual variability for us to be wary of relying on their segmentation. *Walls* in Figure 1 contains an example of a syllable-final velarised lateral whose oral constriction is extremely difficult to identify.

2.2.2 Clusters

The boundary between consonants in non-homorganic clusters can be identified using criteria similar to those described above. For instance, sibilant-stop and stop-sibilant sequences which involve differences in place of articulation, e.g. [ks], [sp], are relatively easy to segment into fricative constriction vs. oral stop closure intervals, see [ksp] in Figure 3 *tax paper.*

Homorganic clusters, on the other hand, present many segmentation difficulties, in spite of the fact that many contain rather marked differences in manner of articulation. For example, for homorganic nasal-stop clusters, e.g. Figure 4 [ŋk] in *Concord*, our principle of identifying oral constrictions and releases for each segment cannot be used, since these two phonological segments are produced with a single oral constriction.

For these homorganic cases, e.g. [ŋk], other acoustic landmarks can sometimes be identified, e.g. the boundary between voicing and voicelessness in the [ŋk] cluster in Figure 4, and/or the acoustic correlates of velic closure. If these landmarks are to be used to infer segmental durations, it should be noted that their articulatory origin is different from that of other acoustic landmarks relating to oral constrictions. Clusters that are completely voiced are more difficult to segment, e.g. [ŋg] in Figure 11.

We have in the past recorded materials containing [st] clusters, hoping to be able to identify the boundary between frication noise associated with [s] and complete closure associated with the stop [t]. But in many of these cases, speakers produce incomplete closures for the oral stops (see Figure 8, [st] in *max tapes*). Although there is often evidence of a change in degree of stricture, this change can appear gradual, making it difficult to reliably segment these phones. This case contrasts with the [sp] cluster in Figure 3 (*tax paper*), where the onset of weak frication was clearly related to the onset of constriction for /p/.

Note that word or phrase boundaries can sometimes intervene between two members of a cluster. In these cases there is a probability of a pause at the boundary which should be taken into consideration when segmenting these phones. For example, if a voiceless stop occurs phrase-initially, its voiceless closure is acoustically indistinguishable from pause.

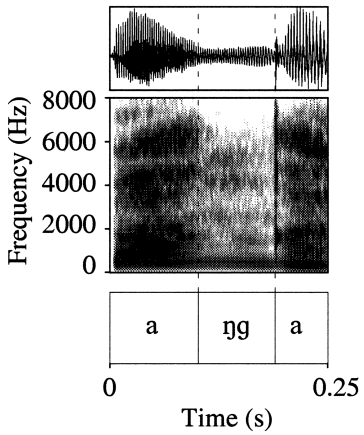

Figure 11: A fragment from *Buun-sensei ICHI-BAN-ga "saasaa"-tte ittakedo* 'Mr. Boone said NUMBER ONE is "saasaa"', spoken by a female Standard Japanese speaker. *Ban* 'number' is a suffix; *ga* is a nominative particle.

2.3 Segmentation uncertainties[4]

It is inevitable that there will be cases where boundaries cannot be found with certainty, even for speech materials whose phones have been carefully chosen to give the highest probability of reliable segmentation. These cases can be grouped into different classes: 1) cases where it is clear that the boundaries occur somewhere within a short window of uncertainty (roughly the duration of a single pitch period, i.e. 5-10 ms), 2) cases where boundaries are known to exist within a wider window of uncertainty, 3) cases where boundaries are completely obscured. One way of dealing with cases like 1) is to annotate them (e.g. with *?*), and to segment them according to a chosen policy of either "when in doubt, place the boundary earlier", or "when in doubt, place the boundary later," to be applied throughout the dataset. Researchers can then choose whether to include these measurements in their analyses, depending on the expected size of effects. Although measurements from cases like 2) and 3) should never be included in data analysis, there may be other ways of salvaging the data for these segments. These may include 1) applying other types of segmentation criteria that might be more reliable in particular cases, e.g. voicing (or laryngeal) criteria, while keeping the implications of these choices in mind, and 2) grouping segments together to yield reliable durations for segment sequences, e.g. as in Figure 4.

[4] The ideas in this section were developed in collaboration with Stefanie Shattuck-Hufnagel.

2.4 Using criteria consistently

Using criteria consistently across materials to be compared is one of the most important principles of acoustic speech segmentation. Target materials and carrier sentences should be chosen carefully with this principle in mind. In particular, segments which are known to have different allophonic variants should be avoided in compared conditions. There are some cases where it is impossible to keep phonetic context constant across conditions, i.e. in comparisons of phrase-final vs. phrase-medial materials, where it is likely that a pause will occur after a phrase-final word. In these cases, the choice of segmentation criteria may have drastic implications for conclusions about the presence and magnitude of prosodic effects. Our study on Japanese and Finnish utterance-final lengthening provides interesting examples in this regard. In Japanese, utterance-final vowels often end in creaky phonation. In an extreme case shown in Figure 12, the utterance ends with widely spaced glottal pulses that give the auditory impression of [a], although they lack continuous formant structure. In cases like this, a segmentation criterion based on continuous F2 yields a much shorter vowel than one based on laryngeal activity. In this particular example, where the vocalic interval based on the laryngeal criterion is labelled as *'a max'* and the vocalic interval based on continuous F2 is labelled as *'a'*, the choice of segmentation criterion makes a difference of 227 ms in the estimated duration of the final vowel.

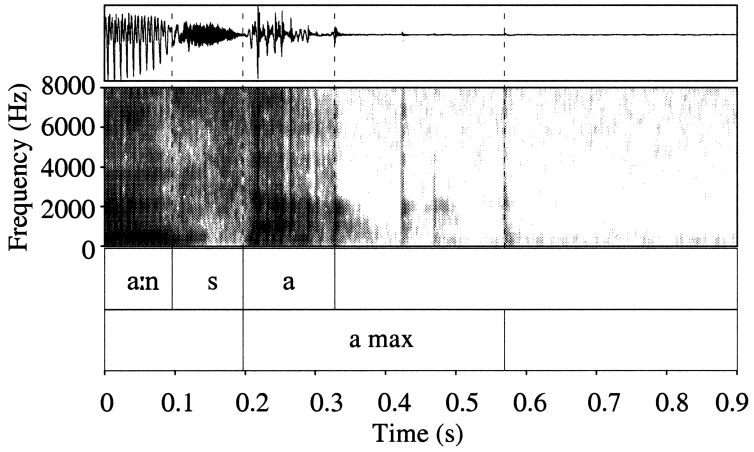

Figure 12: A fragment from *Toujou-sensei-ni kii-tara ICHI-BAN-ga "saansa"* 'According to Mr. Tojo, NUMBER ONE is "saansa"', spoken by a male Standard Japanese speaker. *Saansa* is a nonsense word. The boundary for the offset of /a/ is placed on the last glottal pulse peak in the interval delimited by continuous F2.

In Finnish, one of the most striking characteristics of many utterance-final words is their breathy ending (cf. Figure 13, see also Ogden, 2004). The breathy phonation at the end of an utterance often results in voiceless formant structure, which would be included in the vowel if the vowel were judged to end with the apparent end of formant structure. If the end of the vowel were judged to coincide with the end of voicing, the voiceless formant structure would be excluded. In this example, the choice of segmentation criterion makes a difference of 47 ms in the estimated duration of the final vowel.

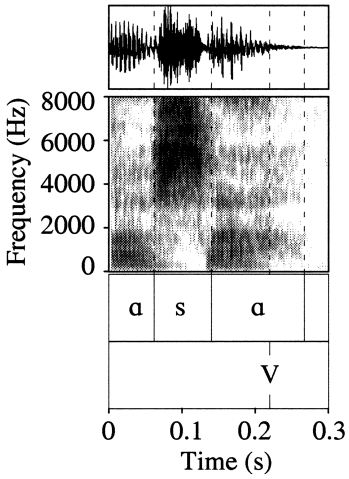

Figure 13: A fragment from *Kohtaan SEITSEMÄNSATAA voisi vastata "sasa"* 'For [#] SEVEN-HUNDRED [you] could answer "sasa"', spoken by a female Northern Finnish speaker. *Sasa* is a nonsense word. *V* in the second label tier indicates the offset of voicing for [ɑ].

3 Experimental design

One of our main goals in discussing relative segmentability in the first part of the paper was to facilitate the design of experimental materials with the maximum likelihood of reliable segmentation. In the following sections, we discuss additional methodological issues for prosodic studies of duration, including ways of controlling for durational effects unrelated to prosodic structure, elicitation methods, and analysis tools.

3.1 Controlling for the influence of multiple factors on duration: Corpus studies vs. controlled experiments

It is well known that segment durations are influenced by a variety of factors including talker, intrinsic segmental properties, segmental context, prosodic context, and global rate of speech (see Klatt, 1976 for a review). These multiple sources of durational variability make it especially challenging to make inferences about the influence of individual prosodic factors on segment durations. There are currently two main approaches to this problem: 1) to study prosodic effects on duration in very large corpora, so that the effects of non-prosodic factors can be modelled statistically to allow for segment duration normalisation (e.g. Campbell and Isard, 1991; Wightman, Shattuck-Hufnagel, Ostendorf, and Price, 1992), and 2) to study prosodic effects in tightly controlled experiments.

The main advantage of the corpus approach is that prosodic effects can be studied in natural speech situations. However, this advantage is outweighed to some extent by the following: First, segment duration estimates are dependent on the accuracy of the automatic segmentation algorithms whose use is inevitable given the volume of data (see discussion of automatic segmentation in Section 4). Also, some of the more subtle prosodic effects have the potential to be obscured, primarily because of difficulties in establishing accurate normalisation procedures. Finally, some prosodic effects cannot be easily isolated in natural corpora, due to the fact that variables affecting duration are often correlated (the "Data Sparsity" problem, van Santen, 1994). For example, Campbell (1992) examines Kaiki, Takeda, and Sagisaka's (1990) corpus and demonstrates that their surprising finding of sentence-final vowel shortening in Japanese could be explained by the correlation of sentence-finality with the occurrence of the past-tense marker *-ta* in their corpus. Campbell argues that the 'sentence-final shortening' is likely to be a result of this correlation, and not a prosodic effect. In other words, [a] in the past tense marker may be short due to various other reasons, e.g. informational redundancy. Kaiki et al.'s corpus therefore lacked enough variation in sentence-final vowels for appropriate comparisons with sentence-medial vowels.

Although data from tightly controlled laboratory studies cannot compare in naturalness with data from spontaneous speech corpora, experimental design can be used very effectively to control for the effects of confounding factors such as intrinsic segment durations, segmental context effects, rate of speech and inter-talker variability. In addition, problems of accurate segmentation can be dealt with through careful experimental design, as discussed at length above.

In controlled experiments, materials are designed to be as alike as possible across conditions while varying the predictor, independent variables of interest, e.g. phrasal stress or constituent boundary placement. In these experi-

ments, talkers are typically asked to read required materials embedded in carrier sentences. In order to ensure that findings can be generalised, it is advisable to include several (6-12) test words or phrases in each experimental condition. Especially if subtle prosodic effects are expected, word frequency and other factors such as morphological structure and orthographic representation (number of letters for each sound) should be controlled, or appropriately varied (Walsh and Parker, 1983; Warner, Jongman, Sereno, and Kemps, 2004).

To a large extent, numbers of talkers and repetitions tend to be dictated by practical issues, but results will inevitably be more reliable in studies where more talkers are studied. When planning a study, researchers would do well to estimate the time it takes to conduct acoustic segmentation realistically. In our experience, only approximately 6-15 disyllabic words can be segmented per hour, depending on the number of segments in each word and on the speed of the segmenter. In addition, it is important to plan for reliability checks when more than one person segments materials from a single experiment.

In the following sections, we discuss ways of controlling known sources of durational variability in more detail.

3.1.1 Control of intrinsic segment durations and segmental context

Different segment types are known to have different intrinsic durations, e.g. low vowels tend to be longer than high vowels (Peterson and Lehiste, 1960; Klatt, 1976). For this reason, and because coarticulation from surrounding segments or gestures has the potential to affect target segment durations, prosodic effects should ideally be tested on identical target words. The choice of materials adjacent to the target word is also important, particularly if word-initial or word-final segments are of interest. For example, if word-final vowel durations are to be compared with word-medial vowel durations, the word following the target word should be chosen so that phonetic context is identical in both cases, e.g. [i] in *beef arm* vs. *bee farm*. In these situations, it is particularly important to avoid eliciting pauses between the target and a following word.

3.1.2 Control of rate of speech

In controlled studies, initial practice sessions and randomisation of test materials are used to control for rate of speech increases that are a frequent consequence of talker experience as s/he progresses through the experimental session. If test materials are presented in blocks, it is advisable to randomise presentation of materials within blocks. If these are blocked by experimental condition, the order of block presentation should be counterbalanced across talkers. As a check that experimental control of rate of speech has succeeded, it is

important to include a control sequence across the experimental conditions for comparison. For example, in a carrier sentence such as *Say the word ___ again*, an appropriate, segmentable control might be *aythewor*, that is the stretch of speech from the release of the [s] to the onset of closure for the [d].

3.1.3 Control of inter-talker variability

Variation in segment durations between talkers can be quite large, due in large part to inter-talker differences in rate of speech. Where within-subject differences are of primary interest (as is often the case), these effects of inter-talker variability can be dealt with effectively using within-subjects experimental design, and appropriate statistical analyses, (e.g. single subject analyses, paired t-tests, or repeated measures analyses of variance, Loftus and Loftus, 1988). Readers are referred to Raaijmakers, Schrijnemakers, and Gremmen (1999) and Baayen (2004) for current debate about how best to statistically analyse designs involving both multiple subjects and multiple items.

3.2 Elicitation of prosodic contrasts

Elicitation of prosodic contrasts can be challenging, especially since talkers normally have many options for the prosodic realisation of a single utterance (see discussion in Shattuck-Hufnagel and Turk, 1996). Reliable methods for eliciting desired contrasts and knowing when the desired contrasts have been achieved are crucial if durational patterns are to be directly related to these contrasts.

Desired phrasal stress patterns, intonation contours, and prosodic phrasing can be encouraged in several ways through the use of 1) syntactic manipulations, 2) orthographic manipulations, 3) precursor sentences, and 4) explicit instructions, as detailed below.

3.2.1 Varying the syntactic structure of read materials

Although prosodic constituent boundaries correspond only indirectly to the boundaries of syntactic constituents, it is very often the case that different syntactic structures are produced with different prosodic structures (e.g. Lehiste, 1973), and in many cases prosodic constituent boundaries are aligned with either the left or right edge of particular types of syntactic constituents (Selkirk, 1986; Truckenbrodt, 1999). Varying the syntactic structure of read materials has been used successfully to encourage systematic variation in prosodic constituent structure in many types of experimental study (e.g. Lehiste, 1973; Scott, 1982; Cambier-Langeveld, 1997; Fougeron and Keating, 1997). In most studies, every effort is made for compared sentences to contain the same, or at

least similar segmental material, and similar numbers of syllables within each utterance. For example, Cambier-Langeveld (1997) used sentences shown in (1) to elicit a range of prosodic boundary strengths after the word *rododendron*, which we have typed in bold (see Cambier-Langeveld, 1997 for a definition of each type of prosodic boundary):

(1) Prosodic Word-boundary: *Piet wil die rare* ***rododendron****planten, gek als hij is.*
'Piet wants those strange rhododendron plants, crazy as he is.'
Phonological-Phrase boundary: *Piet wil die rare* ***rododendron*** *planten, gek als hij is.*
'Piet wants to plant that strange rododendron, crazy as he is.'
Intonational-Phrase-boundary: *Piet wil die rare* ***rododendron****, plantengek als hij is.*
'Piet wants that strange rododendron, plant-crazy as he is.'
Utterance –boundary: *Plantengek als hij is wil Piet die rare* ***rododendron****.*
'Plant-crazy as he is, Piet wants that strange rhododendron.'

One factor to be wary of in syntactic category manipulations is that in some languages, there are morpho-phonetic/phonological alternations associated with words of particular grammatical categories. For example, in Finnish word-initial consonants are long following words of certain grammatical categories that end in vowels, e.g. infinitives and the second person imperative. Thus, in *Haluan tulla sinne* ('I want to come there.'), the /s/ in '*sinne*' is longer than in other, 'non-lengthening' contexts (see Sulkala and Karjalainen, 1992).

3.2.2 Orthographic manipulations

Punctuation can be used successfully to signal intended syntactic structures, e.g. Cambier-Langeveld's presentation of *rododendron* and *planten* with and without an intervening space and comma in the prosodic conditions shown above.

In addition, capital letters are sometimes used to indicate the presence of contrastive phrasal stress, e.g. *I said BUY cakes, not MY cakes* (Turk and Sawusch, 1997). However, capitals should be used with caution because they can encourage talkers to put unnatural degrees of emphasis on target words.

3.2.3 Precursor sentences

Our current preferred way of controlling the placement of phrasal stress is to use precursor sentences that suggest the likely location of new or unpredictable information, on the assumption that unpredictable words are likely to bear phrasal stress (e.g. Aylett and Turk, 2004). For example, Sugahara and Turk

(in prep.) elicited phrasal stress on *baking* and the lack of phrasal stress on *pan* through the use of the sentence set shown in (2):

(2) A pan used in the kitchen.
Say "baking pan" for me.

3.2.4 Explicit instructions

We recommend avoiding the use of explicit instructions to elicit desired prosodic structures, where these instructions could have the undesired effect of getting talkers to exaggerate or to produce patterns or contrasts that they would not normally produce. However, in some cases, explicit instructions seem the most effective in eliciting the prosodic structure in question. For instance, we often ask talkers to avoid pausing within a target utterance to discourage the insertion of phrase boundaries before or after target words, as is often tempting in *Say-X-for-me*-type carrier sentences.

3.3 Influencing the expected magnitude of prosodic effects

The hierarchical level of prosodic and/or morpho-syntactic constituents is expected to have an influence on the magnitude of durational effects that signal them. For example, in many languages, final lengthening is known to be greater in magnitude for Full Intonational Phrases than for smaller phrases (e.g. Wightman et al. 1992). In languages of this type, durational evidence for constituents near the bottom of the hierarchy, e.g. feet or words, is consequently expected to be subtle at best. In designing experiments to test for these constituents, it may be advisable to try to use contexts and elicitation methods that are likely to yield the largest possible effects. A growing body of experimental evidence suggests that phrasal stress, rate of speech, and, in some languages, the number of syllables in a word can be manipulated to yield larger or smaller durational differences between prosodic conditions of interest. For example, Turk and Shattuck-Hufnagel (2000) found that durational differences between segments and syllables in English, e.g. *tune #acquire* vs. *tuna #choir* sequences were more marked when one of the test words bore phrasal stress (e.g. phrasal stress on [tun(ə)] or [kwaɪɹ]). In addition, Beckman and Edwards (1990) and Sugahara and Turk (in prep.) found that durational differences related to word and morphological structure are magnified at slow rates of speech. And thirdly, Cambier-Langeveld (2000) and White (2002) found that effects of phrasal stress in Dutch and English were proportionally larger on monosyllabic than on polysyllabic words.

4 Summary and conclusion

In this paper, we discussed types of acoustic landmarks that define intervals whose durations can be straightforwardly correlated with the durations of recognisable articulatory events, namely oral constrictions and the vocalic intervals between them. We presented a list of segment types whose constrictions can be accurately and reliably measured in VCV and cluster contexts, and proposed that experimental materials should be designed as far as possible to include these measurable segment sequences.

In addition, we discussed issues of experimental design and analysis such as encouraging the elicitation of desired prosodic conditions, and controlling for non-prosodic factors on duration. While the issues we discuss are far from exhaustive, it is our hope that other researchers will benefit from these practical considerations derived from our collective experience using acoustic segment durations in prosodic research.

In particular, it is hoped that these issues might be taken into consideration in the design of automatic segmentation tools that would allow researchers to interactively specify criteria along the lines presented here. The main advantages of automatic over manual techniques are that automatic techniques are more objective and consistent, and can be used in a fraction of the time required for manual segmentation. These tools often involve alignment of a transcription with an acoustic signal using Hidden Markov Model techniques, and can involve subsequent boundary correction using, e.g. spectral discontinuity detection to improve accuracy up to less than 20 ms for all segment types (cf. Kim and Conkie, 2002). Currently, these types of auto-segmentation tools are being perfected for use in creating inventories for optimal unit selection text-to-speech synthesis systems, and not specifically for phonetic investigations into segmental durations. However, they do appear to have the potential to detect landmarks of the type discussed in this paper, and would be useful especially if their accuracy were close to 5-10 ms for segments used in durational research. Their accuracy in segmenting a variety of experimental materials will need to be thoroughly evaluated before they can usefully be improved and customised to replace manual segmentation in cross-linguistic phonetic studies of duration.

References

Aylett, M. and A. Turk (2004): The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. Language and Speech 47(1), 31-56.

Baayen, R. H. (2004): Statistics in Psycholinguistics: A critique of some current gold standards. Mental Lexicon Working Papers 1, Edmonton, 1-45.

Beckman, M. E. and J. Edwards (1990): Lengthening and shortening and the nature of prosodic constituency. In: J. Kingston and M. E. Beckman (eds.): Laboratory Phonology I. Cambridge: Cambridge University Press, 152-178.

Boersma, P. and D. Weenink (2005): Praat: doing phonetics by computer (version 4.3). http://www.praat.org.

Browman, C. and L. Goldstein (1990): Tiers in articulatory phonology, with some implications for casual speech. In: J. Kingston and M. E. Beckman (eds.): Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech. Cambridge: Cambridge University Press, 341-376.

Cambier-Langeveld, T. (1997): The domain of final lengthening in the production of Dutch. In: H. de Hoop and J. Coerts (eds.): Linguistics in the Netherlands. Amsterdam: John Benjamins, 13-24.

Cambier-Langeveld, T. (2000): Temporal Marking of Accents and Boundaries. PhD dissertation. Holland Institute of Generative Linguistics; Netherlands Graduate School of Linguistics. The Hague: Holland Academic Graphics.

Campbell, N. (1992): Segmental elasticity and timing in Japanese speech. In: Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (eds.): Speech Perception, Production and Linguistic Structure. Amsterdam: IOS Press, 403-418.

Campbell, W. N., and S. D. Isard (1991): Segment durations in a syllable frame. Journal of Phonetics 19, 37-47.

Fougeron, C. and P. Keating (1997): Articulatory strengthening at edges of prosodic domains. Journal of the Acoustical Society of America 101, 3728-3740.

Gordeeva, G. and J. M. Scobbie (2004): Non-normative preaspiration of voiceless fricatives in Scottish English: A comparison with Swedish preaspiration. Colloquium of the British Association of Academic Phoneticians, University of Cambridge, http://www.qmuc.ac.uk/sls/pg/ogordeeva/BAAP2004_GORDEEVA_SCOBBIE.ppt

Kaiki, N., K. Takeda and Y. Sagisaka (1990): The control of segmental duration in speech synthesis using linguistic properties. In: Proceedings of the 1st ESCA Workshop on Speech Synthesis, Autrans, France, 18-22.

Kim, Y.-J. and Conkie, A. (2002): Automatic segmentation combining an HMM-based approach and spectral boundary correction. In: Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP), Denver, 145-148.

Klatt, D. H. (1976): Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. Journal of the Acoustical Society of America 59(5), 1208-1220.

Ladefoged, P. (2001): A course in phonetics. Fourth edition. Fort Worth: Harcourt College Publishers.

Ladefoged, P. and I. Maddieson (1996): The sounds of the world's languages. Malden, M.A: Blackwell.

Lehiste, I. (1964): Juncture. In: E. Zwirner and W. Bethge (eds.): Proceedings of the Fifth International Congress of Phonetic Sciences, Münster, 172-100.

Lehiste, I. (1973): Rhythmic units and syntactic units in production and perception. Journal of the Acoustical Society of America 54, 1228-1234.

Loftus, G. R. and E. F. Loftus (1988): Essence of Statistics (Second ed.). New York: Alfred A. Knopf.

Lucero, J. C. (1999): A theoretical study of the hysteresis phenomenon at vocal fold oscillation onset-offset. Journal of the Acoustical Society of America, 105(1), 423-431.

Ogden, R. (2004): Non-modal voice quality and turn-taking in Finnish. In: E. Couper-Kuhlen and C. E. Ford (eds.): Sound Patterns in Interaction. Amsterdam: Benjamins, 29-62.

Peterson, G. E. and I. Lehiste (1960): Duration of syllable nuclei in English. Journal of the Acoustical Society of America 32, 693-703.

Raaijmakers, J., Schrijnemakers, J. and A. Gremmen (1999): How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. Journal of Memory and Language 41, 416-426.

Remijsen, B. (2005): http://www.ling.ed.ac.uk/~bert/praatscripts.html

Romaine, S. (1978): Post-vocalic /r/ in Scottish English: Sound change in progress? In: P. Trudgill (ed.): Sociolinguistic Patterns in British English. London: Edward Arnold, 144-158.

Scott, D. R. (1982): Duration as a cue to the perception of a phrase boundary. Journal of the Acoustical Society of America 71(4), 996-1007.

Selkirk, E. O. (1986): On derived domains in sentence phonology. Phonology Yearbook 3, 371-405.

Shattuck-Hufnagel, S. and A. E. Turk (1996): A prosody tutorial for investigators of auditory sentence processing. Journal of Psycholinguistic Research 25(2), 193-247.

Stevens, K. N. (1998): Acoustic Phonetics. Cambridge, Massachusetts: MIT Press.

Stevens, K. N. (2002): Toward a model for lexical access based on acoustic landmarks and distinctive features. Journal of the Acoustical Society of America 111(4), 1872-1891.

Sugahara, M. and A. Turk (in prep.): The influence of morphological structure on acoustic segment durations. Manuscript.

Sulkala, H. and M. Karjalainen (1992): Finnish. London: Routledge.

Truckenbrodt, H. (1999): On the relation between syntactic phrases and phonological phrases. Linguistic Inquiry 30(2), 219-255.

Turk, A. E. and J. R. Sawusch (1997): The domain of accentual lengthening in American English. Journal of Phonetics 25, 25-41.

Turk, A. E. and S. Shattuck-Hufnagel (2000): Word-boundary-related duration patterns in English. Journal of Phonetics 28, 397-440.

van Santen, J. P. H. (1994): Assignment of segmental duration in text-to-speech synthesis. Computer Speech and Language 8, 95-128.

Walsh, T. and F. Parker (1983): The duration of morphemic and non-morphemic /s/ in English. Journal of Phonetics 11, 201-206.

Warner, N., A. Jongman, J. Sereno, and R. Kemps (2004): Incomplete neutralization and other sub-phonemic durational differences in production and perception: Evidence from Dutch. Journal of Phonetics 32, 251-276.

White, L. (2002): English speech timing: a domain and locus approach. Unpublished Ph.D. dissertation, U. of Edinburgh, Edinburgh, UK.

Wightman, C. W., S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price (1992): Segmental durations in the vicinity of prosodic phrase boundaries. Journal of the Acoustical Society of America 91(3), 1707-1717.

Zue, V. (1985): Notes on spectrogram reading. In Speech Spectrogram Reading: An Acoustic Study of English Words and Sentences. Special Summer Course. Lecture Notes and Spectrograms. Massachussetts Institute of Technology, 1988. N1-N392.