LEARNING [VOICE]

Joshua Ian Tauberer

A DISSERTATION

in

Linguistics

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2010

Jiahong Yuan
Assistant Professor, Department of Linguistics
Supervisor of Dissertation

Gene Buckley
Associate Professor, Department of Linguistics
Graduate Group Chair

Mark Liberman, Trustee Professor of Phonetics, Department of Linguistics
Committee Member

Daniel Swingley, Associate Professor, Department of Psychology
Committee Member

— 'Tricky Linguistics', A Bit of Fry and Laurie (1989) —

Hugh: So let's talk instead about flexibility of language, uh, linguistic elasticity if you like.

Stephen: Yes, I think that I said earlier that I said that our language English—

Hugh: As spoken by us.

Stephen: As we speak it, yes, certainly. —defines it. We are defined by our language, if you will.

Hugh: (to the audience) Hello, we're talking about language.

Stephen: Perhaps I can illustrate my point. Let me at least try. Here's a question.

Hugh: What is it?

Stephen: Ah. Well my question is this: Is our language, English, capable, is English capable of sustaining demagoguery?

Hugh: Demagoguery?

Stephen: Demagoguery.

Hugh: And by demagoguery you mean . . . ?

Stephen: By demagoguery I mean demagoguery.

Hugh: I thought so.

. . .

Stephen: There's language, and there's speech. There's chess, and there's a game of chess. Mark the difference for me, mark it please . . . Imagine a piano keyboard. Eighty-eight keys, only 88, and yet, and yet, hundreds of new melodies, new tunes, new harmonies are being composed upon hundreds of different keyboards in Dorset alone. Our language, Tiger, our language: hundreds of thousands of available words, thrillions of legitimate new ideas, so that I can say the following sentence and be utterly sure that no one has ever said it before in the history of human communication:

Acknowledgements

A dissertation is not just a degree or the culmination of years of effort. If you look close enough you will probably see rings for the many seasons that have come and gone, representing the deeply emotional states it went through — I went through — with among other things a confusing and protracted transformation in self-image from student to professional. Let me start at the beginning.

I sincerely thank my advisor, Professor Jiahong Yuan, for his direction and encouragement. Stumbling out of syntax and into phonetics in my third year, things could have taken a more terrible turn, but Jiahong brings everyone quickly up to speed to the latest in acoustic and corpus phonetics. I also thank the remainder of my committee, Professors Mark Liberman, whose insights never cease to amaze, and Daniel Swingley, for valuable feedback. I'd also like to thank Professor Gene Buckley for serving on my proposal committee and Professors Tony Kroch, Charles Yang, and Maribel Romero for their unique ways they influenced my thinking and keeping me in the game. And I'd like to thank Amy Forsyth, our department's administrative coordinator.

The last six+ years would not have been the same without my cohort, Laia Mayol, Lucas Champollion, Keelan Evanini, Jonathan Gress-Wright, and Jean-François Mondon. I have fond memories of studying in the GSC and burgers at New Deck, at least until New Deck got old. Oh and I especially note — and once again apologize for — the incident referred to in Mayol (2009, acknowledgments). I greatly enjoyed collaborating with Keelan on Tauberer and Evanini (2009), a portion of which made its way into this thesis. The p.lab and its many furniture arrangements would not have been the same without my friends and classmates Giang Nguyen (who suggested my thesis have something to do with babies), Laurel MacKenzie ("gellatto"), Yanyan Sui (tennis and soup), Aviad Eilam (trips to New York and gossip I soon forgot), Stefanie Brody (C3PO's trills), Tanja Scheffler (many years of XTAG meetings and PWPL), and Michael Friesner (my first year buddy).

I wish my memories of Penn were not inextricably linked with some very painful years which I would just as soon forget. Without my friends, especially Andrew Clausen (who, while I'm here, I also thank for help with maximum likelihood estimation), David Robinson, Aliza Wasserman, and again Laia, Laurel, and Yanyan, I could not have made it through life, let alone the dissertation. I couldn't overstate how important you all have been to me. (I also thank the TV show 'House'.)

Much of the data used here was derived from existing corpora. The largest and most important was the Providence corpus, by Demuth, Culbertson, and Alter — thank you for making it available — and I thank the parents and children involved in the corpus as well. I also thank the two LibriVox readers whose works I used in Chapter 7 and the participants in the study in Section 5.2.

Finally I thank my family, most of all my parents Gale and Peter. No parents could be more encouraging, supportive, and interested, and best of all always ready for a game of scrabble. And really finally, I thank the remainder of my family for their page-counting support, especially my Bubby and late Zayde, my aunts and uncles, and my cousins (hugs).

ABSTRACT

LEARNING [VOICE]

Joshua Ian Tauberer

Supervisor: Jiahong Yuan

The [voice] distinction between homorganic stops and fricatives is made by a number of acoustic correlates including voicing, segment duration, and preceding vowel duration. The present work looks at [voice] from a number of multidimensional perspectives.

This dissertation's focus is a corpus study of the phonetic realization of [voice] in two English-learning infants aged 1;1–3;5. While preceding vowel duration has been studied before in infants, the other correlates of post-vocalic voicing investigated here — preceding F_1, consonant duration, and closure voicing intensity — had not been measured before in infant speech. The study makes empirical contributions regarding the development of the production of [voice] in infants, not just from a surface-level perspective but also with implications for the phonetics-phonology interface in the adult and developing linguistic systems. Additionally, several methodological contributions will be made in the use of large sized corpora and data modeling techniques.

The study revealed that even in infants, F_1 at the midpoint of a vowel preceding a voiced consonant was lower by roughly 50 Hz compared to a vowel before a voiceless consonant, which is in line with the effect found in adults. But while the effect has been considered most likely to be a physiological and nonlinguistic phenomenon in adults, it actually appeared to be correlated in the wrong direction with other aspects of [voice] here, casting doubt on a physiological explanation. Some of the consonant pairs had statistically significant differences in duration and closure voicing. Additionally, a preceding vowel duration difference was found and as well a preliminary indication of a developmental trend that suggests the preceding vowel duration difference is being learned.

The phonetics of adult speech is also considered. Results are presented from a dialectal corpus study of North American English and a lab speech experiment which clarifies the relationship between preceding vowel duration and flapping and the relationship between [voice] and F_1 in preceding vowels. Fluent adult speech is also described and machine learning algorithms are applied to learning the [voice] distinction using multidimensional acoustic input plus some lexical knowledge.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Phoneticians and phonologists use features such as [voice], [high], and [round] to categorize segments. The [±voice] or [tense/lax] feature separates pairs of homorganic phonemes into two groups, the [+voice] "voiced" phonemes /b,d,v/ etc. and the [-voice] "voiceless" phonemes /p,t,f/ etc. At the level of phonology, these sorts of distinctions play a role in phonological rules such as voice agreement in the '-s' morpheme in English ('cats' [s] versus 'dogs' [z]), final devoicing in languages including German, or feature spreading such as in a case in New Julfa discussed later. At the level of phonetics, there are generalizations as well: all of the [+voice] stops share similar properties when it comes to the glottal state, aspiration, (shorter) stop closure duration, and (longer) preceding vowel duration (PVD), among other dimensions. But whereas phonological phenomena are investigated within a formal framework such as ordered rules or constraints, considerably less attention has been paid to the formal system that connects phonological features to their phonetic counterparts or correlates.

With that as a starting point, this dissertation concerns the production of the phonological [voice] contrast and its development in infants. The primary contribution is a corpus study of the phonetic realization of [voice] in two English-learning infants aged 1;1–3;5. While preceding vowel duration has been studied before in infants, other correlates of post-vocalic voicing investigated here — preceding F_1, consonant duration, and closure voicing intensity — have not been measured before in infant speech. The study will make empirical contributions regarding the development of the production of [voice] in infants, not just from a surface-level perspective but also with implications for the phonetics-phonology interface in the adult and developing linguistic systems. Additionally, several methodological contributions will be made in the use of large sized corpora and in model fitting for segment durations. Several new corpus studies of adult speech are also presented relating to the distributions of acoustic cues in fluent speech, machine learning of the [voice] distinction in fluent speech, a replication of past lab-speech research with new details, and dialectal variation in acoustic cues.

The multidimensional nature of the acoustics of features has been recognized time and again in the speech perception literature, especially as it relates to cue trading (even to some extent in infant speech perception, Simon and Fourcin 1978; Mayo and Turk 2005). In cue trading, listeners give weights to the different cues to the identity of an ambiguous stimulus. A strong VOT cue might outweigh a weak vowel duration cue in the categorization of a segment as voiced or voiceless. The production and language development literature on the other hand has focused more on single phonetic dimensions of utterance-initial stops and considerably little on the interaction of the correlates of [voice] in post-vocalic position. The present work looks at [voice] from a number of multidimensional perspectives and in particular attempts to classify phonetic dimensions as either being a part of a *multidimensional linguistic competence* or as a physiological consequence of some other aspect of [voice] which actually is specified in the grammar.

Organization of this thesis

This dissertation is organized into six main chapters.

In Chapter 2 the acoustic dimensions of [voice] will be explored through a thorough review and critique of the relevant speech production literature. It has long been known that [voice] can be measured on many acoustic dimensions, even when we look to the left or right of the obstruent segment itself. Lisker (1986) identified sixteen acoustic differences between voiced and unvoiced word-medial /b/ and /p/. As he pointed out, not all are necessarily under separate linguistic control of the speaker. They may not all be free to vary, owing to physiological constraints. In other words, while some phonetic (articulatory or acoustic) target may be specified as part of linguistic competence, other phonetic observations we may make about [voice] may be an almost accidental byproduct of the physical speech apparatus. This distinction is discussed for each acoustic dimension. A focus is given to acoustic dimensions relevant in non-initial context, especially preceding vowel duration.

Chapter 3 reviews the development of [voice] in infants from a surface perspective, meaning what phones appear in children's inventories, and from the perspective of the phonetics-phonology interface, meaning whether the child has learned the relationships between acoustic dimensions, phones, and phonemes.

With this in mind, I discuss a model of the phonetics-phonology interface in Chapter 4. The model is something that perhaps many phoneticians and phonologists implicitly assume, but the details seem to not have been discussed before. As research at the interface moves forward, we need a precise operational theory under which our work, such as this dissertation, can be carried out and evaluated. The focus on the model in Chapter 4 is where the phonetic aspects correlated with a phonological feature are stored, and the refinements I propose subsume all of the linguistically specified aspects of [voice] within essentially a [voice] cover-feature.

The phonetics of [voice] in adult speech is considered in Chapter 5. Section 5.1 presents the results of a dialectal corpus study of the preceding vowel duration difference aspect of [voice] in North American English, based on joint work with Keelan Evanini that was first reported in Tauberer and Evanini (2009). A lab experiment collecting lab speech is presented in Section 5.2 which clarifies the relationship between the preceding vowel duration effect and flapping and the relationship between [voice] and F_1 in preceding vowels.

The new corpus study of infant speech is presented in Chapter 6. In this chapter I also present a novel application of parameter estimation for segment durations in a linguistically plausible model of segment production, with full R code in Appendix B.

Chapter 7 returns to adult speech, but this time fluent adult speech. The chapter had two goals. First, it provides a descriptive account of the acoustics of [voice] in fluent speech on a number of acoustic dimensions. Compared to past work, the emphasis here is less on whether acoustic differences exist (e.g. shown by statistical significance) but whether the acoustic differences are large enough to be useful for discrimination, based on how separated the voiced and voiceless distributions are. Additionally in this chapter, machine learning algorithms are applied to the task of learning the [voice] distinction using multidimensional input.

Appendix C lists a number of experimental directions that are raised by this dissertation (which I do not at this time have any intention of carrying out, but may serve useful for others).

How to read this thesis

Computing technology underlies much advancement throughout science. Technology makes new investigations possible. As it becomes faster to process data, such as with forced alignment tools as discussed several times in this dissertation, analyses of large-scale corpora reveal patterns that older technology was too insensitive to find. But technology has a second effect on the scientific method by making it easier for scientists to share work in a way that makes validation and replication easier (so-called reproducible research). The movement toward open access electronic journals, based on the receding costs of production and distribution, is a part of this side to technology. Another aspect is in how technology is used in the writing process.

This thesis was written using LaTeX and, for the results section, Sweave by Friedrich Leisch[1]. Sweave is a tool that integrates the R statistics program with LaTeX so that R code can be embedded within the LaTeX document. This is a benefit for the reader, who can validate that numeric results are justified by the analyses that generated them, and it is a benefit to the writer who can ensure that the numeric results reported are those in fact produced by the R analysis. In Section 6.3.2 the reader will see R code and R output in `typewriter`-style font. R code is marked by lines starting with > and + symbols.

As for terminology, the reader should note that I use [voice] exclusively to denote a phonological contrast, except in Chapter 4 when discussing alternative feature theories. I use “voiced” and “voiceless” or “unvoiced” (interchangeably) to refer to the feature values, but sometimes to the physical glottal state when it is clear. At other times I use “glottal state”, “glottal signal”, or “phonetic voicing” when discussing the phonetic (articulatory, acoustic, or otherwise physical) aspect of voicing. As is traditional, angled lines /.../ indicate phonemes (i.e. grammatical contrasts) while square brackets [...] indicate phones (i.e. acoustic patterns).

[1] http://www.stat.umn.edu/∼charlie/Sweave/

Chapter 2

The Acoustics of [voice]

The distinctive feature [voice] has largely been associated with voice onset time in the acoustic phonetics literature. This is due in part to the focus on utterance-initial segments. In this context, voice onset time is the most significant acoustic dimension for separating [+voice] from [-voice]. But it has long been known that [voice] can be measured on many acoustic dimensions, even when we look to the left or right of the obstruent segment itself. Lisker (1986) identified sixteen acoustic differences between voiced and unvoiced word-medial /b/ and /p/. As he pointed out, not all are necessarily under separate linguistic control of the speaker. They may not all be free to vary, owing to physiological constraints, but certainly some are.

In this chapter I review the literature on the acoustic dimensions of [voice] in English, with a focus on production (versus perception) and on non-initial context. In particular, the effect on preceding vowel duration is given extensive focus. The review starts within the temporal bounds of the obstruent's constriction and then moves to the properties of preceding and following vowels. Along the way it is evaluated whether each acoustic correlate of [voice] is merely a physiological consequence of some other articulatory dimension of [voice], or whether it is specified in the grammar separately from other aspects of [voice].

2.1 Glottal signal, aspiration, and pharyngeal cavity expansion

The most obvious and in some cases primary cue to voicing is the timing and intensity of the glottal signal (i.e. voice at the phonetic level), so it is the obvious place to begin. But the differences between voiced and voiceless goes well beyond Voice Onset Time.

Pre-vocalic stops

For pre-vocalic stops, Voice Onset Time (VOT; Lisker and Abramson 1964b), the time from release to the initiation of the glottal signal, is commonly considered the most salient distinguishing feature of voiced versus voiceless segments. The mapping from abstract [voice] to VOT differs across languages. In English utterance-initially, as reported in Keating (1984), the contrast is manifest primarily as a difference between short-lag VOT, meaning VOT mostly between 0 and 40 ms, and long-lag VOT, meaning between 50 and 90 ms. There is also occasional pre-voicing of voiced stops, that is, voicing during closure just before the release, which is represented by a negative VOT. Still, the production of the voicing categories is not perfectly separated. Flege and Brown (1982) reported

	Pre-voicing	Short-Lag VOT	Long-Lag VOT
English	optional	0–40 ms	50–90 ms
Polish	-150 – -30 ms	20–70 ms	not used
Thai	<-40 ms	0–30 ms	>20 ms

Table 2.1: Keating (1984): Voice onset time values for voice contrasts in English, Polish, and Thai.

15–25% of utterance-initial /b/ showing pre-voicing, only 5–10 percentage points more than the rate of pre-voicing in *voiceless* /p/.

Languages like English are called aspirating languages or are sometimes said to employ the [tense/lax] distinction rather than [voice]. In Polish, on the other hand, the contrast manifests as a difference between pre-voicing (-150 to -30 ms VOT) and short-lag VOT (20 to 70 ms). Languages like Polish are said to be pre-voicing or employ the [voice] feature rather than [tense/lax]. Thai, finally, makes a three-way distinction between pre-voicing (less than -40 ms VOT), short-lag, and long-lag VOT (Keating 1984). See Table 2.1. Dutch, Puerto Rican Spanish, Hungarian, and Tamil pattern with Polish; Cantonese patterns with English and German; and Eastern Armenian and Korean pattern with Thai (Lisker and Abramson 1964a).

Because of the tight relationship between VOT and aspiration in languages like English the terms are often used interchangeably when referring to the voice contrast (Lisker and Abramson 1964a). There is also a tight articulatory relationship between the two as aspiration noise can only be heard if the time of maximal glottal opening is synchronous with the time of oral release (for a summary see Fuchs 2005). Jessen (2001) noted the terminological distinction due to Repp (1979) between 'open interval' which matches the interval measured by VOT and 'aspiration' which is when there is aperiodic noise.

However, these intervals may not line up. It is possible to have both negative VOT as well as aspiration. For German stops, Jessen (1998) reported utterance initial voiced stop aspiration mean durations of 21 ms and additionally a mean of 23 ms of pre-voicing (i.e. -23 ms VOT). (The means may be somewhat misleading as a representation of the degree of voicing, however, as the subjects were reported to voice inconsistently.) Hindi uses both acoustic signals contrastively in a 2x2 paradigm (Kagaya and Hirose 1975). Other languages make a three-way contrast between aspirated, unaspirated, and pre-aspirated. Pre-aspirated stops have aspiration preceding closure. In Icelandic, pre-aspiration is similar at least in duration to an /h/ preceding the consonant and appears to be either an allophonic variant of the voiceless (i.e. aspirated) stops or the realization of voicelessness in geminate stops (Suh 2001 for a summary and analysis).

VOT's role in the voice contrast is generally thought of as a duration target: e.g. short-lag versus long-lag. But see below for Stevens and Klatt's (1974) musing that it may be tied to an acoustic, spectral target (Section 2.7).

Post-vocalic stops

For post-vocalic stops, the phonetic contrast involving the glottal signal is slightly different. The cessation of voicing after closure is more rapid in voiceless stops, so instead a Voice Termination Time (VTT) from the onset of closure to the termination of voice would be more appropriate.[1]

[1] I am not aware of the origin of this term but it seems to be generally used, and I do not know of a general study on VTT.

Intervocalic stops, being post pre-vocalic and post-vocalic, have three relevant measures: VTT, VOT, and duration of aspiration. However, intervocalic stops are often voiced throughout closure, with no cessation of voicing between the preceding and following vowel. In these cases, measures of VOT and VTT are meaningless: VOT would extend negatively back from the release to before the obstruent began itself, perhaps to the start of the utterance, while VTT would extend forward from the start of closure perhaps indefinitely. The rate of voicing during closure not surprisingly is highly variable. Edwards (1981) reported that 45% of medial voiced stops exhibited voicing throughout closure; Flege and Brown (1982) reported approximately 75%. Even [-voice] stops can be voiced throughout closure: 8% in Flege and Brown (1982).

Intervocalic stops and pharyngeal cavity expansion

Another related measurement is the amount of voicing during closure, without regard to whether the voicing is at the start (as in a VTT measurement) or at the end (as in a negative VOT measurement). (Discontinuous voicing during closure does not seem possible. If there is voicing, it is near the edges.) Edwards reported that the mean duration of the glottal signal in intervocalic voiced stops was 78 ms during closure with mean closure duration of 96 ms (that's 81% of closure duration). For voiceless stops the mean voicing duration was roughly 2.5 standard deviations lower, at 25 ms within a mean closure duration of 89 ms (that's 28%). (The measurement of longer closure duration for voiced plosives is unusual.) This take on voicing duration is then a fourth measurement of voicing for plosives.

Closely related to the glottal signal is pharyngeal cavity expansion. During stop closure with no nasal opening, subglottal and supralaryngeal air pressures begin to equalize during voicing. If the transglottal pressure difference is not maintained by subglottal articulators or oral cavity expansion, voicing will terminate. If the shape of the vocal tract were fixed, voicing would last for just 15 ms. Passive expansion of the vocal tract walls, keeping supralaryngeal pressure low, allows for 65 ms of voicing. (Fuchs 2005) Recall that Edwards (1981) reported that the mean duration of the glottal signal during closure in intervocalic voiced stops was 78 ms. Both “active” and “passive” gestures can increase the volume of the pharyngeal cavity: active gestures being greater activity in certain muscles while passive gestures being less activity or the inhibition of activity in other muscles. But in a study with three subjects, Bell-Berti (1975) found that the use of four gestures that increase pharyngeal cavity size, two active and two passive, was not consistent across subjects. All subjects used at least one active gesture, but not all the same one, and one subject did not use either passive gesture. From the fact that active gestures can be used to extend voicing but further that the use of active and passive gestures is mixed, Bell-Berti concluded that [lax] was a poor description of voicing. We should also take from this that there is probably no linguistic specification of any particular gesture — whether active or passive — in the mental representation of [voice]. Rather, these gestures are recruited by speakers in an ad hoc manner to instantiate a more abstract, perhaps acoustic target, e.g. to maintain a certain prescribed amount of voicing.

Fricatives

The difference between voiced and unvoiced fricatives is manifest (partly) in voicing duration during frication — but glottal pulsing is not robustly present throughout voiced fricatives just as it is not robustly present throughout closures of voiced stops. Stevens, Blumstein, Glicksman, Burton, and Kurowski (1992) measured the difference in voicing during frication. However, rather than measuring the raw duration within the period of frication during which there was glottal pulsing, Stevens et al. opted for a stricter measurement excluding voicing at an amplitude less than 10dB below the amplitude at the boundary between the fricative and the vowel (where it should peak). The authors did not motivate the use of a threshold on amplitude rather than relying only the presence of glottal pulsing at all — however, I can imagine one reason was that determining the precise start and end of glottal pulsing can be something of an art because of the neighboring nonperiodicity of the glottal signal. By their measure, voiceless fricatives had 4 to 26 ms of glottal vibration depending on context, while voiced fricatives had 29 to 58 ms. The mean frication durations were not given but were around 80–100 ms. Fricative durations generally vary too with the voice contrast, as discussed below, in such a way as to magnify the difference in durations of voicing relative to the total frication duration. Jessen (1998) reported for German fricatives around 10 ms (unvoiced) vs. 75 ms (voiced) of glottal pulsing, or around 7% vs. 85% of frication duration. As I noted impressionistically for stops, for fricatives voicing occurs, if anywhere, near the boundary with a vowel, rather than spontaneously in the middle of the segment, as one would expect given the high cost of initiating and terminating glottal vibration (Stevens et al. 1992).

The amount of glottal pulsing during frication depends on context. Based on electroglottographic data for productions of English /z/ and /s/, Smith (1997) reported the rate of voicing for sentence-final, word-final, and word-medial positions. In sentence-final position all /z/ tokens she collected had glottal pulsing for less than 25% of the duration of frication. In word-final position followed by a voiceless consonant /z/ tokens generally had between 25%-90% of the frication duration voiced. In word-medial position, the percentage of voicing ranged between 25%-100%. For /s/, fewer than 2% overall had voicing for more than 25% of the duration of frication, most of the remainder having no voicing at all. The amount of voicing in /z/ is also sensitive to voicing in a following fricative, with substantially less preceding a voiceless fricative (Stevens et al. 1992).

Schissel (2008)[2] compared the glottal pulsing differences in English utterance-final 'non-morphemic' /s,z/ as in 'house' (the noun /haʊs/ or the verb /haʊz/) to the plural and possessive morphemes '-s' which are generally analyzed as agreeing in voicing with the preceding segment. She found that neither non-morphemic /s/ nor morphemic '-s' following an unvoiced segment were ever, in her data, realized with glottal pulsing, and both non-morphemic /z/ and morphemic '-s' following a voiced segment were realized with voicing less than one-third of the time.

It is not entirely surprising that one would find that glottal pulsing is not always present during stop closures and frication intervals. Voicing is difficult to sustain in both cases because of the transglottal air pressure difference that must be maintained in spite of other factors. This is discussed in the next section.

In a transillumination study of German intervocalic (sometimes word-initial) stops and fricatives (Jessen 1998), maximum glottal opening during stop closure or frication distinguished the voiced and unvoiced forms: unvoiced obstruents had greater opening. (In the case of stops, the duration of glottal opening during closure also discriminated the voiced and unvoiced forms, unvoiced forms having 50% longer opening, but this might be a by-product of the total segment duration correlate to be discussed next.) Jessen believed that while a glottal opening appeared in both voiced and unvoiced obstruents, the opening in voiced obstruents was due to changes in intraoral air pressure as a result of the oral constriction, while the greater glottal opening in unvoiced obstruents was a result of an active gesture. Strong evidence was not presented for this position, however. A review of the glottis and its degree of opening can be found in Fuchs (2005).

[2] A class project for Phonetics I during which time I was the TA.

2.2 Obstruent duration

Voiceless obstruents have longer durations than voiced obstruents. For stops, the difference has been reported at around 60 ms, or one-to-two standard deviations, or a ratio of approximately 1.3-to-1, but studies have varied over whether the discriminating difference includes or excludes the burst and aspiration duration. Lisker (1957) included the release duration in measurements of intervocalic /p,b/ in trochaic minimal pairs and reported that durations for /b/ ranged from 65–90 ms and for /p/ from 90–140 ms. (Similar results were reported by Crystal and House (1988b); Edwards (1981).) Luce and Charles-Luce (1985) found similar results measuring up to the burst.

For fricatives and affricates a similar pattern obtains (Crystal and House 1988b; Stevens et al. 1992). Baum and Blumstein (1987) found in utterance-initial position unvoiced fricatives around 28 ms longer in frication duration than voiced fricatives, a ratio of unvoiced to voiced duration around 1.3. Despite the large difference, variability in duration was great and there was no clear boundary in duration times between the groups. The /s,z/ frication difference has also been found in French (Flege and Hillenbrand 1986), with a ratio greater than 2.0 word-finally.

Lisker (1957), Denes (1955), and Stevens et al. (1992) considered segment duration from the perceptual end, for word-medial stop closures, word-final frication durations, and frication durations in several contexts, respectively. Each showed that the unvoiced-voiced distinction was made by listeners at least in part based on the duration of the segment. Decreased segment duration lead to increased identification of segments as voiced, even if the segment has no glottal voicing. In Lisker's case, the perceptual judgments followed an s-shaped curve and were in agreement with the production measurements in terms of the location of the cross-over from voiced to unvoiced.

The duration pattern is in the direction we would expect given the added difficulty of maintaining voicing during an obstruent. Voiced stops and fricatives are both difficult, though for different reasons. As discussed above, once the transglottal pressure difference equalizes during stop closure voicing stops. Thus the longer the closure, the less likely voicing will continue throughout. In both stops and fricatives, voicing requires the maintenance of the right tension and adduction of the vocal cords and a transglottal air pressure difference. It is more difficult to maintain voicing in a fricative. Unlike in stops, supralaryngeal pressure can be kept low by allowing air to escape through the fricative's obstruction. But fricatives additionally require a pressure difference at the oral constriction, but there with the oral pressure higher. Coordination is required to hit the right targets at the two locations. See Ohala (1997) for a review of the factors involved in glottal voicing in obstruents.

De Jong and Zawaydeh (2002) mused, as discuss below in the section on vowel duration, that the presence of primary stress should amplify acoustic properties of segments that are linguistically specified. As opposed to the effect on vowel duration discussed below, the presence of stress on the syllable containing the obstruent has been reported to not magnify the voicing effect on duration (Crystal and House 1988b). To de Jong and Zawaydeh, this would indicate that closure duration is not linguistically specified and that the difference is a matter of, perhaps, physiology.

Could the duration difference be just a measurement bias? For plosives the termination of closure is marked reliably by a burst. The onset of closure is less reliably marked. At best there is rapidly decreasing intensity (e.g. Luce and Charles-Luce 1985), but persistence of glottal pulsing into closure might mitigate the intensity drop to make it appear as if the start of closure occurred later. This would create the illusion that voiced stops have shorter closure. (For instance, intensity decay time at the start of closure, measured as "the time needed for signal intensity, expressed in decibels, to drop from 90% of its peak value to 10% of its peak value", is 15–20 ms or 1.7 times longer before stimuli listeners identify as voiced stops, Hillenbrand, Ingrisano, Smith, and Flege 1984.) But since the closure duration difference does persist in whispered speech for /t,d/ and /s,z/ (Parnell, Amerman, and Wells 1977), the presence of voicing cannot be the cause of the duration difference. Frication duration differences are similarly found even between /s/ and partially and fully devoiced tokens of /z/, with about the same magnitude of duration difference in contexts that have greater devoicing (Smith 1997).

The duration of frication might still be a misleading measurement, Stevens et al. (1992) argued. While the frication duration varies according to the value of [voice], it seemed that the time from the initiation of the constriction gesture to its completion is constant across the two categories of [voice]. Stevens et al. measured this gesture intervocalically by tracking the F_1 transition in a preceding and following vowel. Though the duration of frication differed between /s/ at 108 ms and /z/ at 78 ms, the time from the start of transition in the preceding vowel to the end of transition in the following vowel was roughly the same for /s/ and /z/, at around 155 ms. In other words, the timing patterns of the supraglottal articulators were the same in /s/ and /z/, while the apparent difference in frication duration is due to a difference in timing of glottal gestures only. Because a different glottal position and transglottal pressure configuration are needed to initiate voicing, this all may mean that the frication duration difference is essentially a physiologically conditioned effect not under separate linguistic control from the specification of glottal voicing, and that the grammar in fact specifies equal durations for the two segments.

Evidence exists that a similar explanation cannot be made for stops based on an experiment that changed the glottal state of the stop without changing its underlying [voice] feature. Jansen (2004) measured regressive voicing assimilation (RVA), a coarticulatory process in English, in VC_1C_2 sequences. An example of RVA (based on Jansen 2004) is the different realization of /z/ in "Is {Bob, Pete} going?": [ɪzbɑb...] vs. [ɪspit...]. The voicing of /z/ seems to be neutralized to match that of the following segment. Jansen (2004, 2007) reported results for VCC sequences where the two consonants spanned a word boundary, e.g. in r-less British English 'brickwork depot' vs. 'brickwork tunnel' vs. 'Hamburg tenant' vs 'Hamburg dairy', etc. Jansen measured several acoustic properties of the velar stop. The effect of the voicing in C_2 on the duration of voicing during the closure of C_1 was near-categorical. While a /gr/ sequence showed 14 ms or 63% more closure voicing than a /kr/ sequence (which is the usual voicing difference), a /kz/ sequence showed 15 ms or 58% *more closure voicing in the /k/* than in the /g/ of a /gs/ sequence. On the other hand, stop closure duration was essentially not affected by the voicing feature on a following obstruent segment. The closure duration difference in /kr,gr/ was 14% and in /kz,gs/ 19% (/k/ greater in both cases). The glottal state cannot be said to be responsible for closure duration differences in stops.

Not unexpectedly then, the closure duration correlate of [voice] is reported to be not universal. In Danish, Hindi, Mandarin, and Swati Xhosa unvoiced stops have shorter closure (Jessen 2001). As it happens, Danish also has particularly long aspiration and it has been suggested that the closure duration correlate has been essentially sacrificed so as to highlight the aspiration contrast. In light of this cross-linguistic pattern, one would very much like to see whether these languages also lack a fricative duration correlate of [voice] (if they have the voice contrast for fricatives at all) — which would undermine a physiological account of frication duration.

To summarize, there is evidence on both sides of the question of whether obstruent duration is linguistically specified. In favor of a linguistic specification is that the pattern is not universal and that stop closure duration is not affected by the glottal state. On the other hand, the difference in frication duration seems to be determined solely by differences in the glottal state, no effect of stress has been found, and in either case a vague principle of "conserve energy" would explain why the more difficult, voiced obstruents would be shorter even if the duration difference itself were not linguistically specified.

2.3 Other correlates during the consonant

Release bursts are said to be shorter and weaker following voiced versus voiceless stops, in both aspirating (short lag/long lag VOT) and pre-voicing (pre-voicing/long lag VOT) languages (for references see Jansen 2004, page 52 and Fischer and Ohde 1990).

In the bilabial stop pair /b,p/, lip pressure has been reported to be greater for the unvoiced phoneme in German word-medial intervocalic position but no difference was reported for English (Jessen 1998:278 for a summary).

Maximum airflow during the frication also was different between /z/ and /s/, regardless of the amount of glottal pulsing, with even near-or-fully devoiced /z/ having less maximum airflow than an /s/ in the same context. This suggested that the glottal position was somewhere between the open state of unvoiced /s/ allowing greatest airflow and the constricted state of a fully voiced /z/ impeding airflow. (Smith 1997)

2.4 Preceding vowel duration

A cross-linguistic phenomenon has been observed in which the duration of a vowel is longer when preceding a voiced obstruent than when preceding a voiceless obstruent (House and Fairbanks 1953; Luce and Charles-Luce 1985; Crystal and House 1988a among many others, some cited below). It has been variously called the post-vocalic consonant voicing effect, the vowel length effect, extrinsic vowel duration, and pre-fortis clipping. To put the emphasis on the consonant, rather than the vowel, the phenomenon will be referred to here as the preceding vowel duration (PVD) difference or PVD effect.[3] The phenomenon is often described formally but descriptively as the distinction between /bæt/ versus /bæːd/ or V → [+long] / ____ [+voice] in the style of Chomsky and Halle (1968), though it is by no means a settled issue whether the phenomenon ought to be described in terms of phonological lengthening, or whether the PVD effect is even a single phenomenon.

The duration difference is quite large, at least in English. In Section 5.2 I present the results of a replication of various past work on measuring the PVD effect using minimal and near-minimal pairs in a word-list-reading experimental design. Representative of past work, vowel duration before [+voice] obstruents was found to be as large as 1.5 times greater in monosyllables and 1.3 times greater in trochaic bisyllables than the corresponding vowel duration in the other half of the (near-)minimal pair. Within individual minimal pairs, the tokens were easily separable: an average of 3 to 5 standard deviations separated [+voice] and [-voice] cases. But the effect is far reduced outside of word-final, pre-pausal context (Crystal and House 1988a).

[3] The existing terms each have their issues. "Fortis" is an antiquated term and "pre-fortis clipping" makes a commitment to the underlying process. "Vowel length" is a term used more often to refer to a phonological (i.e. binary) contrast rather than a continuous phonetic dimension. While "extrinsic vowel duration" has been used exclusively for this phenomenon, it is an odd term considering the number of extrinsic factors that affect vowel duration. And "post-vocalic consonant voicing effect" is unnecessarily verbose.

The PVD effect occurs within and across syllable boundaries (Chen 1970), for stops as well as fricatives (Umeda 1975; Stevens et al. 1992; Smith 1997). It affects not just an immediately preceding vowel but also the preceding vowel and any intervening sonorant (e.g. ‘sent’ vs. ‘send’, ‘false’ vs. ‘falls’, the vowel and sonorant both lengthening to roughly the same degree; Chen 1970; Crystal and House 1988a).

The PVD difference has been observed many, many times throughout the literature, especially in English. Chen (1970) is generally cited for his early, cross-linguistic comparison, although the sample size was dangerously small and the data were confounded by the position and context of the voiced/voiceless segment in each word (though it seems Chen got lucky). A number of factors have been shown to affect the PVD effect, at different levels of linguistic structure, but the results can be summarized that the PVD effect’s voiced–unvoiced difference is smaller when the vowel is in a context where it is expected for other reasons to have a shorter duration. This is Klatt’s (1973) “incompressibility” — that each successive shortening effect on a vowel has a diminished effect because of physical bounds on the speed of articulation. The voiced-unvoiced duration difference, in Klatt’s (1973) data, was 66 ms in a context where vowels were on average 165 ms in duration (the context was being in a monosyllabic word) but only 28 ms in a different context when the vowels were on average 117 ms (that’s in the first syllable of trochaic words). The effect of voicing on duration decreases after we apply the effect of position, one might say. But this incompressibility is so even when we consider the effects not as absolute differences but as multiplicative factors. Here I switch from viewing the PVD as a shortening, as Klatt did, but instead as a lengthening, as most research has done. In the first longer-vowels context voicing increased duration by 51 percent but in the second shorter-vowels context voicing increased duration by only 22 percent. In other words, the magnitude of the PVD effect is not constant either on an absolute time scale (i.e. milliseconds) or in terms of a proportion (1.51-to-1). (Also see Port 1981.)

Other work has looked at this variation in PVD ratio across contexts. The PVD ratio is much larger in pre-pausal position (a ratio of 1.5) than elsewhere (1.4 word-finally, 1.1 word-medially) (Umeda 1975, but also see Luce and Charles-Luce 1985; Crystal and House 1988a). The effect is also larger for stressed versus unstressed vowels (Davis and Summers 1989). It is also larger for vowels with greater intrinsic duration (Luce and Charles-Luce 1985) and in slow ‘tempo’ speech (1.18) than in fast ‘tempo’ speech (1.09) (Port 1981; Laeufer also notes work by Harris and Umeda in 1974). Other measurements of the PVD effect can be found in Summers (1987).

The PVD difference is also a perceptual cue. Denes (1955), looking at word-final /s,z/, found a strong effect of vowel duration on whether the final consonant was understood as voiced or unvoiced. He noted that the rate of perception as one category or the other was modeled well based on the ratio of the duration of the vowel to the consonant.

2.4.1 Comparison with other languages

The literature on the PVD effect has often portrayed it as a universal phenomenon but with a peculiar existence in English. The duration difference has been claimed to exist at least in Danish, Dutch, French, German, Hindi, Hungarian, Italian, Korean, Norwegian, Persian, Russian, and Spanish (see for references Kluender, Diehl, and Wright 1988), and of course English. Its apparent universality suggested early on that the PVD effect was a low-level possibly articulatory effect, something necessitated by other factors such as the airflow differences in voiced and unvoiced obstruents, and so not a part of linguistic competence.

This does not appear to be the case, however. The duration difference has been claimed to

English	1.51	Peterson & Lehiste (1960) as cited in Chen (1970)
English	1.45	House and Fairbanks (1953)
English	1.63	Chen (1970)
French	1.15	Chen (1970)
Russian	1.22	Chen (1970)
Korean	1.31	Chen (1970)
English	1.57	Zimmerman and Sapon (1958)
Spanish	1.17	Zimmerman and Sapon (1958)
English	1.23	Laeufer (1992)
French	1.15	Laeufer (1992)
Quebec French	1.32	Morasse (1995)
English	1.20	Flege and Port (1981)
Arabic	1.03	Flege and Port (1981)
Czech	1.02	Keating (1980)
Polish	1.0	Keating (1980)
Swedish	1.03	Elert (1964) as cited in Buder and Stoel-Gammon (2002)
German	1.11	cited in Fintoft (1961), as cited in Chen (1970)
German	1.20	Braunschweiler (1997), word medial
Catalan	1.17	Charles-Luce (1992)
Norwegian	1.23	Fintoft (1961) as cited in Chen (1970)
Hungarian	1.2	Jansen (2004) (long vowels long)

Table 2.2: Ratios of mean duration of vowels preceding voiced consonants to that before voiceless consonants in various languages reported in the literature. In some cases, including Polish and Czech, the mean of the individual ratios for minimal pairs is reported instead, which generally magnifies the value.

not be present in Polish and Czech (Keating 1980), based on pair-list readings of intervocalic /t,d/. (Non-final position is a position where the effect size is reduced, but in English word-list reading still readily observed, see section 5.2.) Similar claims have been made about Arabic (Flege and Port 1981; de Jong and Zawaydeh 2002) and Swedish (Buder and Stoel-Gammon 2002, based on Elert (1964)). The presence or absence of the PVD effect in a language cross-cuts whether the language has a phonemic vowel length contrast; for instance, Dutch, Hungarian, Swedish, Czech, and Arabic all have a vowel length contrast, while only Dutch and Hungarian are claimed to have a PVD difference. On the other side of the table, of languages without phonemic vowel length, English and French have a PVD effect while Polish does not.

A summary of findings of the PVD effect cross-linguistically is reported in Table 2.2. For each language, the ratio of the mean duration of vowels preceding voiced consonants to that before voiceless consonants is reported. English shows a much larger duration difference between voiced and unvoiced consonants than all other languages in which the PVD effect has been studied, at least at face value. Taking the results from Chen (1970) as a representative example (its problems notwithstanding), the ratio of the mean duration of vowels before voiced consonants to that before voiceless consonants in English is roughly 1.5 with an absolute difference between the mean durations of 92 ms, while in French, Russian, Korean, etc. the ratios and differences are 1.2–1.3 and 28–53 ms.

Comparisons across languages are fraught with problems. A ratio of 1.1 or less, such as in the reported cases of Arabic, Czech, and Swedish, might be a difference of a single pitch period, depending on vowel duration and F_0, which is quite a small difference to extrapolate much of anything. It may have more to do with intensity decay time, which is 15–20 ms greater before stops identified by speakers as voiced (Hillenbrand et al. 1984), see page 8, a difference far too small to explain the PVD effect in English but about the right size to explain the PVD effect in these other languages.[4] Chen's (1970) results have been suspected of overestimating the difference between English and the other languages (Kluender et al. 1988).

Zimmerman and Sapon (1958) very likely overestimated the effect in Spanish. They reported for the Latin American dialects of Spanish spoken by their subjects that intervocalic contrasts were between voiceless stops and voiced fricatives. If Spanish is like English where vowels are longer before fricatives than for stops (Umeda 1975), then the effect may have been largely due to not using true minimal pairs. I am not aware of any more recent studies of the PVD effect in Spanish. The results reported for Catalan, French, and German may have underestimated the effect. Charles-Luce (1992) for Catalan and Braunschweiler (1997) for German put the post-vocalic consonant in a non-final environment to escape word-final devoicing, which may reduce the PVD effect. (See page 17.). Laeufer's (1992) more detailed analysis of the PVD effect in French suggests that previous measurements of the PVD effect in French were also underestimated and that there is no great difference with English.

English and French vowels differ in many ways: French vowels are often shorter than their English counterparts; English final stops are longer and are often, unlike in French, unreleased and when they are released their releases are shorter than those in French; French pre-pausal voiced stops are often (74%) followed by a vocalic release (a voicing cue unavailable in English that may relieve the PVD effect of its communicative load); and varieties of French, unlike English, lack syllabic stress (Laeufer 1992). Controlling for some of these factors in the two languages, Laeufer found a PVD ratio in French of 1.42 and in English 1.6, much closer than past work found. Laeufer noted, however, that vowel durations were not consistent between the two languages (in one condition 150 ms and 209 ms, respectively), and this may account for the difference in the PVD between the two languages. Klatt's (1973) incompressibility says that durational changes will apply differently depending on the expected duration of the segment. When French vowels had similar durations to English vowels, which were the contexts of sentence-final position in French and sentence-medial position in English, the PVD ratios were much closer. (Also see Flege and Hillenbrand 1986 for similar results involving /s,z/.)

Laeufer's (1992) work is important in several respects. First, it highlighted language-specific details that make cross-linguistic comparisons susceptible to confounding factors. Second, in light of the fact that independent vowel duration factors interact with the PVD effect (Klatt 1973; Port 1981) and that when vowel durations were matched between English and French the PVD ratios were very similar, it is not entirely clear that English and French differ with respect to PVD at all.

[4] It seems useful here to mention Chen's (1970) explanation for the PVD effect in terms of the speed of closure formation. Because intraoral air pressure during consonant closure is higher for a voiceless consonant, since the build-up of air pressure extends into the lungs, whereas for a voiced consonant it extends only to the glottis, more effort is needed to form the closure in a voiceless consonant, he wrote. This greater effort may begin during the transition to closure, and the increased muscle effort may translate into increased velocity, a more rapid transition, and therefore a shorter vowel. Kluender et al. (1988), however, noted that later research has shown that closure velocity is not reliably greater in voiceless consonants, at least not in the contexts that show the PVD effect. A related explanation, given by Messum (2007), is based on the premise that aerodynamic effort is a limited resource and that voiceless stops and fricatives consume some of their preceding vowel's time because they require more effort. This hypothesis has not actually been tested, however.

One can compare the PVD effect to different types of vowel durational phenomena in other languages. The magnitude of the PVD effect in English is still smaller than duration differences associated with a phonological length contrast, such as in Dutch (see for a summary Dietrich 2006), for which van Bergem reported that long vowels were 1.8 times longer than short vowels. The PVD effect is outdone also by the Scottish Vowel Length Rule (Scobbie, Turk, and Hewlett 1999), which among other things involves lengthening by a factor of 1.8 specifically before voiced fricatives.

In summary, there are at least two categories of languages, if not three. English is the prototypical language exemplifying the PVD effect, and French may very well be included in the same category. Polish, Czech, Arabic, and Swedish have strong evidence for having no PVD difference (at least not beyond about a single pitch period if even that). As for the remaining languages, it is not clear. Where the effect has been probably overestimated, as in Spanish, the language might group with Polish etc. If languages remain, we may have a second type of PVD effect that is smaller than that of English and governed by a different mechanism entirely.

2.4.2 Interaction with other phenomena

When investigating whether a phenomenon is a matter of linguistic competence, rather than a physiological consequence of some other linguistic articulation, some have looked for whether the phenomenon interacts in interesting ways with other linguistic phenomena. In some cases, this is the same as asking whether the phenomenon is a productive process operating at the level of phonology or something more static below the level of the linguistic apparatus.

Stress

De Jong and Zawaydeh (2002, page 54) theorized that lexical stress could be a diagnostic for what aspects of a segment are linguistic and what aspects of a segment are not. “If some aspect of speech,” they wrote, “is part of the specification of a particular segmental phonemic entity, having that segment in a stressed syllable should be expressed by a more extreme instantiation of that aspect. However, if some aspect of speech is due to some other organizational or motor strategy, stress should not enhance its appearance in speech.” If only language was ever so straightforward. Let us not confuse de Jong and Zawaydeh’s hypothesis for a rigorously tested law, but we can take their points for what they are worth. It is also interesting to view this as a specific case of what has been reported elsewhere, that the PVD effect is greater in contexts in which vowel duration is expected to be greater for independent reasons.

De Jong (2004) contrasted vowel durations in the presence or absence of primary stress and in the context of either a voiced or voiceless following stop. His previous work (cited within) additionally included the contexts of stop versus fricative and singleton versus cluster coda. Vowels have been found to be shorter before voiceless stops (versus voiced stops), fricatives (versus stops), and cluster codas. The question was whether primary stress amplified the duration difference in each of those contexts. For instance: is the vowel duration difference between ‘bed’ and ‘bet’ (having primary stress) larger than the corresponding difference in ‘flower bed’ and ‘sports bet’ (having secondary stress), and in turn larger than that in ‘rabid’ and ‘rabbit’ (unstressed). In this case of post-vocalic voicing, yes, stress amplified the duration difference. (There was virtually no duration difference in the unstressed case anyway.) On the other hand, stress was not found to affect vowel duration differences when manner or coda size was a variable. The PVD effect, but not effects of clusters or manner of articulation, seemed to be a linguistic phenomenon.

De Jong and Zawaydeh (2002) applied this idea to vowel durations in Arabic. While Arabic has a similar use of stress as in English, they wrote, the vowel duration difference before voiced versus voiceless stops is considerably smaller than in English. Flege and Port (1981) found for Saudi Arabian Arabic and de Jong and Zawaydeh (2002) found for Ammani-Jordanian Arabic a mix of non-significant and significant duration differences. Although de Jong and Zawaydeh reported consistent patterns in the expected duration (that is, vowels longer before voiced consonants), the differences were on the order of 10 ms or a ratio of 1.1 throughout. This difference is far smaller than that observed in English and French, and roughly the difference of just one pitch period — small enough to wonder whether this is a reliable measurement in the first place. Not surprisingly then, the duration difference remained essentially nil whether or not the syllable containing the vowel had primary stress.

de Jong's conclusion has been that the PVD effect in English is a linguistic process while in Arabic it is not. I noted above that the voicing effect on stop closure duration was not amplified by stress (Crystal and House 1988b) — we might similarly conclude that this is evidence that the stop closure duration difference is like the Arabic vowel duration difference in not being a part of the linguistic specification.

Intervocalic flapping

The interaction of the PVD effect with English t/d-flapping is an instructive case because flapping neutralizes the voice contrast on the surface. Well known as the 'writer'/'rider' ambiguity, one must keep in mind that the ambiguity can impoverish the linguistic input so that even adults may not know the true (if there is such a thing) underlying form of infrequently-alternating words, such as 'giddy', which I innocently recently misspelled as "gitty".

There are three possibilities for how the PVD effect will come out in these words. The first is that the preceding vowel duration difference is based on whether the flap is pronounced with glottal pulsing or not. In other words, the PVD effect is a consequence of the articulatory gesture. Fox and Terbeek (1977) found no duration difference according to the glottal state, although the comparison was based on a very small sample size, in part because voiceless flaps are rare.

The second possibility is that the PVD effect depends on the lexical or underlying specification of [voice]. In that case we would expect a /raɪɾɚ/–/raɪːɾɚ/ contrast, despite the neutralization of other aspects of the /t-d/ phonemic contrast. If this were true, the PVD effect could be described as an early-ordered rule, relative to the rule for flapping. (I certainly do not intend to commit to an ordered-rule framework, but it is convenient for exposition.) The third possibility is that the PVD effect depends on the [voice] feature at the end of the phonological component of the grammar, after flapping neutralizes the difference between /t,d/. In that case, the PVD effect would produce no difference between the two words in the minimal pair.

Neither of the analyses above fit the observed data perfectly. Vowels preceding flaps continue to show a duration difference, but with a difference an order of magnitude smaller than usual. The original data is due to Fox and Terbeek (1977), probably of Chicago speakers. The facts were later investigated in New York City speakers by Huff (1980). Fox and Terbeek and Huff considered word-medial cases; Huff additionally considered monosyllabic words with final /t,d/ followed in the next word by an initial unstressed vowel, e.g. '{bite,bide} again' (both /baɪɾəgɛn/). Though these studies have been cited as answering the questions definitively, they had significant shortcomings: Fox and Terbeek lacked a control condition with non-flapped consonants and did not look across place of articulation, as a result confounding the potential effects of both flapping and syllable structure, and they surprisingly did not report any mean values. Huff's sample was relatively small and focused on a single dialect, and it also included short-a which is known to have allophonic or idiosyncratic variation depending on [voice] in the dialect Huff studied (e.g. Boberg and Strassel 2000).

Fox and Terbeek found a small but statistically significant difference in vowel duration preceding flaps depending on underlying voice. Huff also found a large and statistically significant difference in duration when the flap was word final, at a ratio of 1.2. This has not to my knowledge been replicated.

In many cases of flapping, morphological alternations are absent or infrequent, leaving language users without much evidence for the right underlying value of [voice] — as in my case of "giddy" versus "gitty". Some language users may have figured it out, and some not. Looking across individuals, the result of confusion is a reduction of the observed vowel duration difference — in the limiting case to no duration difference when there is total confusion over the underlying form. This actually fits the data quite well. Word-final alveolar stops are variably flapped, conditioned on the metrical structure in the word that follows. On the other hand, word-internal alveolar stops are either always flapped in that word or never flapped. Evidence for the underlying state of the internal flaps can only come from morphological derivations of the word, and it would make a great deal of sense if speakers were far more confused about the underlying state of flaps in word-medial position than in word-final position, where there should be little or no confusion. That lines up well with very small PVD differences for word medial flaps (Fox and Terbeek 1977) but normal-sized PVD differences for word-final flaps (Huff 1980).

If this is the correct explanation, we would expect all acoustic correlates of voice to be similarly reduced in the trochaic case as we take an average across speakers. Another correlate, discussed below, is the effect of [voice] on first formant frequency. Monophthongs have higher F_1 before voiceless consonants and the /ai/ diphthong has lower F_1 before voiceless consonants (Canadian Raising). Huff (1980) measured and reported formant values on the vowels preceding the word-final and word-medial flaps. In the monosyllabic case, where the speakers surely know the right underlying value of [voice], F_1 was around 200 Hz higher for /æ/ and 100 Hz lower for /aɪ/ before flapped /t/ than flapped /d/. These formant differences are consistent both in direction and magnitude with what is found before voiced and unvoiced consonants in non-flapping environments, as discussed below. In the trochaic case where speakers may not know the underlying value of [voice], the effect size did not appear to be appreciably reduced, meaning that the speakers also had the correct underlying [voice] value set in these cases. If this were the general case, the reduced effect on duration could only have been a result of the flapping process and not erroneous values in the speakers' lexicons. Unfortunately, Huff's results do not agree with the results of my own experiment reported in Section 5.2, which show that F_1 is neutralized by flapping as well. In other words, there is no reason yet to believe speakers do have the correct underlying form.

No standard model of phonology can account for these facts. If the PVD effect is essentially an early-ordered rule before flapping, then the reduced (and essentially extinguished) duration difference before flaps is unexpected. If the PVD effect is essentially a late-ordered rule after flapping, then we would expect the duration difference to be completely neutralized. And yet, a statistically significant difference remains. Two phenomena have been proposed in the literature: 'incomplete neutralization' (see e.g. Port and O'Dell 1986) and an effect of orthography (Warner, Good, Jongman, and Sereno 2006). But I do not consider these further.

Devoicing and voiceless speech

Devoicing provides another case of dissociation between vowel duration and surface voicing. One case of devoicing in standard English appears to be lenition in the (word-final) /s,z/ contrast. As noted several times above, the presence or absence of the glottal signal does not appear to affect other aspects of the voice contrast. That is, so-called voiceless /z/'s continue to be paired with longer preceding vowels while they have less frication duration and intensity (Smith 1997).

This contrasts with the results of a preliminary study (Schissel 2008) that showed that the /s,z/ alternation of the morphological suffix '-s' show preceding vowel duration differences. This alternation is considered to be phonological rather than phonetic, and along with the data regarding flapping suggests that the PVD effect operates late in the phonological system but still within the phonological system.

Many other languages have devoicing as an active phonological rule, and one might expect those cases to operate similarly to the '-s' voicing alternation or to the neutralization of flapping. In German syllable-final stops are devoiced, purportedly neutralizing the distinction between final stops as in 'Bund/bunt' (Fuchs 2005, page 24). Nevertheless, several measures of the phonetic realization of these stops continue to show a small difference between the voiced and voiceless stops despite the claims of neutralization: the preceding syllable nucleus is 15 ms longer before an underlying voiced versus voiceless consonant (making for a ratio of 1.09), the duration of glottal pulsing into closure is 5 ms longer, release aspiration is 15 ms shorter, and closure duration is marginally shorter (Port and O'Dell 1986). (Unfortunately Port and O'Dell (1986) did not include non-'neutralized' stops as a control, but we can compare the acoustic differences to those found in non-neutralized, word-medial context by Braunschweiler (1997), who found a vowel duration ratio of 1.20. As we know, the PVD difference in word-medial position underestimates the PVD difference that would be found word-finally (Klatt 1973), if it weren't for devoicing, and so Port and O'Dell's results do show a clear reduction of the PVD effect in devoiced context.) This case of devoicing appears to be very similar to what is found in t/d-flapping in English. (That Port and O'Dell's results were low might be attributable to the same question of inferring the underlying form of rare flapped words discussed above: some words in their corpus were rare and one lacked morphological evidence for the underlying voicing feature. If the speakers did not know the underlying form, a mean value across speakers will underestimate the true effect.)

Additionally, the PVD effect persists in whispered speech at the same ratio as in normally phonated speech (a ratio as high as 1.8 in Sharf 1964), and laryngectomized patients who phonate using the esophagus and need neither laryngeal nor pulmonary articulation also exhibit a very high PVD ratio: 1.6 ratio in normal subjects, 1.7 in esophageal speech, the difference likely due to slower speaking rate in esophageal patients (Gandour, Weinberg, and Rutkowsky 1980).

Voicing assimilation

The final case of interaction noted here is that between the PVD effect and regressive voicing assimilation, which provides another example of the dissociation between the PVD effect and surface voicing. In Section 2.2 I reviewed Jansen (2004, 2007) which showed that RVA neutralizes the difference in the glottal state during stop closure but has no meaningful effect on stop closure duration. Similarly, there is essentially no effect of RVA on preceding vowel duration. This is in agreement with the coarticulatory explanation of RVA and explaining the PVD effect as a phonological process, rather than one based on the phonetic realization of the consonant. It also contrasts with RVA in Hungarian, thought to be a phonological rather than coarticulatory process, which does reduce or neutralize the preceding vowel duration difference (Jansen 2004). This does not indicate much about the PVD effect in Hungarian since RVA neutralizes voicing at both a phonological and phonetic level.

2.4.3 Phonological status

The evidence presented above regarding the interaction of the PVD effect with voicing lenition, whispering, and voicing assimilation strongly indicates that the PVD effect is not simply conditioned on surface voicing. Many sorts of reductionist explanations have failed (as in footnote 4). Instead, it seems to be a phonological process. But, even at the level of phonology there is still a wide range of ways the phenomenon might be encoded.

Because stop closure duration also correlates with [voice] — *shorter* for voiced consonants — it has been proposed that the duration of the vowel varies inversely with the duration of the stop in order to maintain constant syllable duration (Chen 1970 cited Kozhevnikov and Christovich (1967:107) and B. Lindblom (1967:21) for the hypotheses of compensatory temporal lengthening.) This hypothesis is likely to at least be a factor. Vowels before Italian geminate consonants are shorter than they would be before non-geminate consonants (Esposito and Benedetto 1999), and this effect is apparently exceedingly common across languages (though not universal, see Kluender et al. 1988, page 161 for references). But the compensation between vowel and consonant duration is not exact on an absolute scale. Italian geminates are roughly twice as long, around 90 ms longer, than their non-geminate counterparts, but preceding vowels are roughly 25% or 40 ms shorter before geminate consonants (Esposito and Benedetto 1999). Kluender et al. (1988) reported other work on English and Swedish indicating that while there may be such a thing as compensatory vowel duration adjustments made inversely to other durations in the syllable, the compensation is neither total nor large enough to explain the PVD effect. (Also see van Santen and Shih 2000 and Braunschweiler 1997 for corroborating evidence in German.)

If not the sum of the durations of the vowel and consonant, then perhaps it is the ratio of their durations that is held constant. I noted above that Denes (1955) found that rate of perception as [+voice] was modeled well based on the ratio of the duration of the vowel to the consonant. Luce and Charles-Luce (1985) investigated this from the production end and came to a negative conclusion regarding the ratio: the absolute duration of the vowel was more reliably able to distinguish the two voice categories than the ratio (as he put it, though with a somewhat questionable mathematical approach relying on p-values to rate the reliability of the measure).

The above failed explanations were in terms of constancy. Kluender et al. (1988, page 156) made a proposal in terms of exaggeration: the PVD effect exists because speakers “select acoustic cues that have mutually reinforcing auditory effects. Talkers signal phonetic contrasts using a ‘conspiracy’ of cues that enhances the perceptual distinctiveness of features and segments.” In other words, the effect is deliberate, in the sense that it is a part of the linguistic knowledge of the language, and not physiological. In the case at hand, the PVD effect exaggerates the closure duration cue. By shortening vowels before voiceless consonants, the increased closure duration of voiceless consonants is perceived to be longer than it otherwise would be.

Kluender et al. further hypothesized that auditory enhancement would be unnecessary both in the case where the duration difference is so great that vowel duration changes would not increase its perceptual salience and in the case where no consonant duration difference exists to be enhanced. For the first case, that a language with a very large consonant duration difference will not need an enhancing vowel cue, Kluender et al. point to Turkish. Turkish has consonant gemination, and with a quite large duration difference between simple and geminate consonants at a ratio of 3:1. But it lacks any vowel duration compensation for gemination. This contrasts with Italian in which, as noted above, vowels are shorter before geminate consonants (Esposito and Benedetto 1999).

For the second case, Kluender et al. note Arabic, the only language they knew to lack a closure duration correlate to voice (Flege and Port 1981). In line with their hypothesis, the vowel duration difference in Arabic is very small, if it exists at all (Flege and Port 1981; de Jong and Zawaydeh 2002).

Kluender et al.'s story though enticing is still, however, not an *explanation*. It explains why certain gestures ought to be recruited, but it does not say that the gestures *must* be recruited or that they cannot be recruited in other cases. And, the facts seem to show that these sorts of counterexamples exist. Polish has a closure duration difference quite comparable to English but lacks a PVD difference (Keating 1980). And there may be cases where the closure duration difference is reversed (i.e. voiced consonants longer) but the PVD effect is not. According to references cited by Jessen (2001), Danish, Hindi, Mandarin, and Swati Xhosa have a reversed closure duration correlate, but as noted above, Danish and Hindi are also claimed to have a PVD difference. (One would need to read the work on Danish and Hindi to know whether this actually stands against Kluender et al.'s view.) If these are not inconsistent with Kluender et al.'s hypothesis, then they merely have punted the hard question: of the languages with a closure duration contrast, why should some languages recruit vowel duration but not others? The mutually reinforcing hypothesis also does not address why in English the PVD effect appears to be greater than in other languages that also show a PVD difference (if such differences are real; Kluender et al. denied the difference). This angle has been insightful, but in the end may not provide anything testable.

Where does this leave the PVD effect? As far as anyone could be able to tell, the PVD effect in English is a phonological process separate from other phonological aspects of [voice].

2.5 Fundamental frequency in following vowels

Fundamental frequency is slightly lower in vowels following voiced stops. For stressed vowels surrounded by the same stop on either side, House and Fairbanks (1953) found just a 4 Hz difference (a ratio of 1.04), tentatively attributing it to a lower intrinsic or natural fundamental frequency for glottal pulsing during voiced stops. Jessen (1998) is a relatively recent and through examination of F_0 perturbation following word-initial obstruents in German. In the first five pitch periods following stop release, or for fricatives essentially starting with the onset of formant structure, F_0 after voiced obstruents was reported consistently lower on the order of 10–20 Hz for at least 75–150 ms into the vowel. And although there had been claims that the voicing difference was the difference between a falling and a rising contour, Ohde's (1984) study of word-final stressed syllables in English found that F_0 tended to decrease in the first five pitch periods in both voicing categories. — F_0 simply seems to start off higher for unvoiced consonants. (Also Edwards 1981, and Castleman and Diehl 1996 for references to other studies including perception studies.)

But when actual aspiration is considered separately from the phonological contrast, a different pattern emerges — indicating there are actually two separate F_0 perturbation effects. Ohde (1984) performed a three-way comparison of English speakers' 1) [+voice] stops (e.g. in nonsense word /həˈbib/), 2) [-voice] stops in a position where they are aspirated (e.g. in /həˈp$^{(h)}$ip/), and 3) [-voice] stops in onset clusters following an 's' where they are unaspirated (e.g. /həˈspip/). The difference in

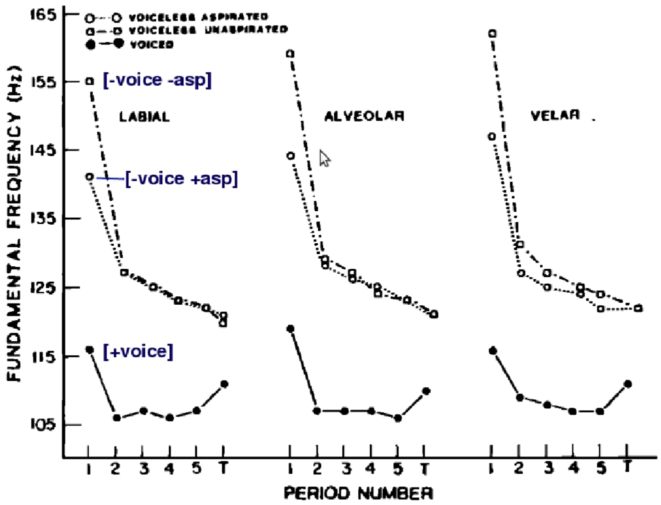


Figure 2.1: Ohde (1984): “Average F_0 from voicing onset to the vowel target (period T) for voiceless aspirated, voiceless unaspirated, and voiced stops according to place of articulation.” (Additional labels added for clarity.)

F_0 between voiced stops (1) and voiceless unaspirated stops (3) was the greatest, with a long lasting 10–20 Hz difference with voiced stops lower. However, the comparison of voiceless unaspirated stops (3) to voiceless aspirated stops (2) showed a 10–15 Hz difference in the first pitch period only, but with *aspirated* stops lower. This is unexpected since aspiration is always associated with the voiceless category in languages that make a binary [voice] distinction with aspiration, and so we should expect aspiration to exaggerate the difference rather than reduce it. I have included Ohde’s figure as Figure 2.1. Kingston and Diehl (1994) reported more variable results, with voiceless aspirated stops actually having a much higher F_0 than the voiceless unaspirated stops.

In Hindi (Kagaya and Hirose 1975 and references in Jessen 1998, page 329), which makes a four-way contrast varying both voice and aspiration, F_0 is 34–58 Hz greater for voiceless stops compared to their corresponding voiceless category (that is, keeping aspiration the same), and is 10–34 Hz lower for the aspirated stops compared to their corresponding unaspirated category. See Table 2.3. In Madurese, which makes a three-way contrast as in Thai (and in parallel to the three English cases above) aspirated stops similarly lowered the F_0 of the following vowel by 10–40 Hz compared to the unaspirated voiceless stops (Cohn and Lockwood 1994 as cited by Downing and Gick 2001, who note that work on Thai has yielded mixed results). The patterns in English, Hindi, and Madurese agree that the pattern of a higher F_0 for the voiceless class in the so-called aspiration languages has little to do with aspiration. (Jessen 1998, page 109 questions the validity of the Hindi data on the grounds that it may be due to a confounding effect of breathy phonation — although the same potential confound was found to not be a factor in F_0 perturbation in German.)

Is F_0 perturbation a matter of performance or competence? Ohde (1984) and Jessen (1998) attempted to reduce F_0 perturbation to a physiological consequence of a difference in vocal cord tension. High tension in the vocal cords to prevent voicing in voiceless obstruents affects the voicing that follows: tension reduces the length of the glottis, thereby increasing its natural frequency. (This appeared to be supported by measurements of the cricothyroid muscle, which affects horizontal vocal cord tension, in work on English and Dutch, among other physiological measurements.)

	Voiced	Voiceless
Unaspirated	154	188
Aspirated	120	178

Table 2.3: F_0 perturbation in the four-way contrast of Hindi: Mean fundamental frequency in Hz in a vowel following a stop. F_0 was measured following the release of voiced stops and following the start of voicing for voiceless stops (Kagaya and Hirose 1975).

Stevens (1998, page 466), however, computed based on a mathematical model of the glottis that the effect on F_0 should be on the order of 5–7 percent, much less than the 10–15 percent difference that is observed (according to Stevens). Vocal cord tension may be a contributing factor, but the rest of the effect remains unexplained. Ishihara (1998), for instance, reported that F_0 perturbation in Japanese and English persisted whether or not the voiced half of the minimal pair had pre-voicing. (Jessen 1998 reported something similar for German.) Granted, there may be vocal cord tension differences even if they do not result in pre-voicing, but I think some difference would be expected. It seems then that F_0 perturbation is linked to neither aspiration (as explained above) or glottal vibration, which leaves few options besides it being a part of the phonological specification of [voice] in much the same manner than glottal vibration itself is.

Other aerodynamic explanations had been proposed for F_0 perturbation (see citations in Ohde 1984 and Jessen 1998), but they do not fit the data. The aerodynamic effect, based on higher airflow causing a Bernoulli effect on the glottis, is expected only in the first 10–15 ms of the following vowel. This would be useful for explaining the effect due to aspiration, except that they predict higher F_0 with aspiration, contrary to fact. Jessen (1998, page 109) noted that voiceless obstruents tend to be followed by more breathy voice quality than voiced obstruents, in German, English, and other languages (see references within). This might affect F_0. Ohde (1984) noted that the height of the larynx also varies between the voicing categories and could be the source of another explanation. For a survey of explanations, see Kingston and Diehl (1994).

In fact, neither a tension or an aerodynamic explanation could be the sole explanation. Jessen (2001) noted that the same perturbation effect is found in languages that make a binary voicing distinction in initial position based primarily on aspiration including English, Cantonese, Mandarin, and Danish, or based primarily on pre-voicing such as French and Japanese. And in Tamil, in which voice is not contrastive but predictable from gemination, no perturbation effect is found (Kingston and Diehl 1994), and thus F_0 perturbation could not be a physiological consequence of any other aspect of [voice] used by Tamil.

There is also some literature on the effect of voice on the fundamental frequency preceding the consonant. For this context, more work has been done in perceptual studies than production studies. Codas are perceived more often as voiced when the fundamental frequency steady state or offset is lowered in the preceding vowel (Castleman and Diehl 1996). In CVC syllables, a mere 15 Hz increase in F_0 from 95 to 110 increases the rate of voiceless perception by roughly 20 percentage points. In VCV sequences, the effect is much smaller, and is smaller than the perceptual effect of varying F_0 at the onset of the second vowel. Hawkins and Nguyen (2004) reported from acoustic measurements of production data no F_0 difference at the onset and mid-point of the vowel (at least not greater than 3 Hz, which is hardly perceivable), and that F_0 was approximately 20 Hz higher at the vowel end before a voiced consonant. However, they attributed the difference at the vowel end to the difficulty of measuring fundamental frequency in the vicinity of glottalization (of voiceless codas). Looking at /l/ onsets, Hawkins and Nguyen (2004) reported no effect on onset fundamental frequency of coda voicing, but note their earlier work that showed a perceptual effect on coda voice when onset fundamental frequency is varied (lower F_0 is, again, correlated with voice).

2.6 Formant structure of preceding vowels

The first formant of vowels has also been seen to vary with the voicing of the following consonant. In monophthongs, F_1 lowers before voiced consonants. In the diphthong /aɪ/ in some dialects F_1 is greater. These patterns have been observed both in the acoustic phonetics literature and in the sociophonetics literature, which tends to treat it as a sociophonetic variable subject to dialectal variation.

2.6.1 Monophthongs

In a monophthong preceding a stop, a relatively small-scale effect on F_1 appears throughout the duration of the vowel (Summers 1987 but see also Moreton 2004 and references therein, Hawkins and Nguyen 2004) and in an unexpected direction. F_1 is lower in the voiced context by around 10–20 Hz at the onset of the vowel, 35–45 Hz during its steady-state, and 90–140 Hz at the onset of the obstruent (these were /a/ and /ae/ vowels with mean frequencies around 775 Hz, so this is up to nearly a 20% difference at the onset of the obstruent). These findings may be unexpected if we were to make a prediction based on what we know about vowel duration. The change is opposite to what would result from hyper/hypoarticulation due to the longer/shorter duration of vowels in voiced/unvoiced context (Moreton 2004). The F_1 curves in Summers (1987) had a ∩ shape, and so hypoarticulation in the unvoiced case would be a lower steady-state F_1. But this is the opposite of what was found. See Section 5.2 for additional data.

A similar pattern was found in vowels preceding the strident fricatives (Stevens et al. 1992). In the roughly 30 ms before the vowel-consonant boundary, F_1 decreased by roughly 400 Hz in the voiced case, but 0–200 Hz in the voiceless case — making for a 200–400 Hz difference between voiced and voiceless cases. Stevens et al. explained the difference on purely physiological grounds: Because voicing is cut off earlier in the voiceless case, less of the F_1 transition can be seen. If it were not for the change in glottal state, we would see that the formant transition was the same in both voiced and voiceless cases. But as noted above, F_1 differences exist not only at the vowel offset but during the earlier steady state as well.

The correlation with F_1 has been partially supported by perception research as well. At least, F_1 offset has been shown to have a perceptual effect. The lower it is at the vowel-consonant boundary, the more likely subjects are to identify the consonant as a voiced consonant. However, in one case (Fischer and Ohde 1990) the choice in the subjects' minds seemed to be not between a voiced or voiceless consonant, but between a voiced consonant and a vowel — putting into question the experimental design. (Of note, Fischer and Ohde found that the effect on perception was greater for vowels with higher F_1 steady states, the mid and low vowels, which have more room to fall in the transition.)

Three experimental angles can be brought to bear on the phonological status of the effect on F_1. The first angle is whether flapping neutralizes the effect on F_1, as discussed above in Section 2.4.2. The results of Huff (1980) indicated that the F_1 difference was not neutralized. If this were the case, the effect would necessarily be phonological since it would rely on the underlying state of [voice] and not the surface realization where voice is neutralized. But, my own results reported in Section 5.2 indicate otherwise, that F_1 is indeed neutralized by flapping. While this does not rule out a late-ordered rule (i.e. bled by flapping), it does not mean that the effect is necessarily phonological.

The second angle comes from Jansen (2004, 2007) who measured the effect of regressive voicing assimilation (RVA) on various acoustic properties. Recall that RVA occurs in a VC_1C_2 context where the [voice] feature of C_2 affects the realization of C_1. (In Jansen's experiments, the vowel was always the long central mid open vowel /ɜː/ of British English.) As noted earlier, the duration of the glottal signal during C_1's closure is nearly neutralized if an immediately following obstruent has a conflicting value of [voice], but there is essentially no effect on closure duration or preceding vowel duration. Jansen (2004, 2007) also measured preceding vowel F_1 at 10 ms before the onset of the (first) obstruent. When C_2 was the control consonant /r/, F_1 was lower by 26 Hz in the voiced case — matching the well-known effect for monophthongs described above. When the first obstruent was followed by /t/, the F_1 difference was completely eliminated, and it was considerably reduced when the obstruent was followed by /d/ and /z/ as well. F_1 thus patterns like glottal signal duration but not either of the segmental duration correlates. Since RVA in English is believed to be merely coarticulation of the glottal signal, then this indicates the F_1 effect is a physiological consequence of the glottal state.

I discuss the third angle in the next section.

If this is right, then there is still a second phonological effect that affects F_1 in the same way. There is a limited and dialectally varying phonological effect called short-a tensing. The triggering environment differs from dialect to dialect, but [voice] is often relevant. The effect is a raising and peripheralization of the short-a nucleus from roughly [æ] (lax) to [ɛə] (tense). For instance, in New York City short-a is tensed before voiced /b,d/ but not voiceless /p,t/, and before voiceless /f,s/ but not before voiced /v,z/ (Boberg and Strassel 2000). Raising corresponds to a lower F_1, which for New York City stops matches the general pattern of the effect of [voice] on F_1 discussed above. It would be interesting to test the effect of RVA on voice-conditioned F_1 in a dialect with short-a tensing.

There may also be an effect on F_2. F_2 is higher before voiced consonants in high-front, low-front, and low-back/central vowels (but not back-rounded vowels), though the statistical significance reported was weak (Hawkins and Nguyen 2004).

2.6.2 Diphthongs

Other things being equal, we might predict the effect of post-vocalic [voice] on diphthongs based merely on the elongation of the vowel preceding voiced consonants. In a comparison of diphthongs across speaking rate, Gay (1968) found that what remained constant across vowel duration in the up-gliding diphthongs /ʉɪ,aɪ,au,eɪ,ou/ was the rate of change of F_2 — as opposed to the offset target frequency. "If the speaker rate is slow, the target is reached and the gesture is completed; if the speaker rate is fast, the movement, while on course, is cut off before reaching the final target" (p1573). If formants are affected the same way by [voice], then we would expect peripheralization in [+voice] context and centralization in [-voice] context. But this pattern did not hold for the voice contrast, despite a duration difference. That is, although vowels preceding voiced consonants were longer, the effect on F_2 was different. Instead of a longer glide, a longer steady-state was generally found.

Though Gay did not find an effect of final consonant voicing on off-glide formants, Moreton (2004) found in the front- and up-gliding diphthongs (/aɪ,ɔɪ,eɪ/) peripheralization before a voiceless stop: lower F_1 by roughly 10–25% and higher F_2 by roughly 3–15%. (Similar but smaller effects were found for F_1 at the nuclear target.) This pattern for F_1 is in the opposite direction of what has been found for the low monophthongs, leading Moreton (2004) to propose a unified explanation for the lowering of F_1 in monophthongs and the raising of F_1 in high-off-glide diphthongs. He proposed a "Pre-Voiceless Hyperarticulation" hypothesis: low-monophthongs should become lower and high-diphthong off-glides should become higher. But what of high monophthongs and in-gliding diphthongs? Moreton's hypothesis has yet to be tested on these crucial cases (see Section 5.2.3). Finally, Moreton did not give a reason for what the underlying mechanism is. As discussed above, the evidence points to the F_1 effect in monophthongs being physiological, meaning a consequence of the glottal state or some other aspect of [voice] articulation and not a matter of linguistic competence. Can something physiological have a hyperarticulatory effect? Perhaps it could, but the details need to be spelled out.

Besides the effects observed by Moreton, there is also a large-scale effect called Canadian Raising affecting the diphthongs /aɪ,au/. Canadian Raising is found (or at least was found several decades ago) in English speakers in Ontario as well as the "inland northern" United States from western New England across the Great Lakes (and at the time in Charleston as well, but not anymore, Joe Fruehwald, p.c.). For speakers in these dialects, the nuclei of the diphthong is raised before voiceless consonants roughly to [ʌɪ,ʌu] (e.g. [raɪd] versus [rʌɪt]). Canadian Raising is clearly not a physiological consequence of some other aspect of voice. For one, it does not occur in many dialects of English that are, as far as anyone has observed, otherwise identical in the relevant ways. It is also in the opposite direction of the effect of [voice] on monophthongs. Whereas F_1 is lower for monophthongs in a voiced context by 10–50 Hz in the first half of a monophthong (Summers 1987), it is higher on the order of 100 Hz during the nucleus of the diphthong /ai/. Also, as noted above, whereas speaking rate affects the off-glide target of the diphthong but not the nucleus's target (Gay 1968), it is the nucleus that is primarily changed in Canadian Raising.

Canadian Raising also has many exceptions. It does not occur across morphological boundaries ([flaɪ-swaɾɚ] 'fly swatter', but not [flʌɪ-swaɾɚ]) or in certain configurations of stress ([daɪˈkaɾʌmi] 'dichotomy'). Still, there appears to be lexical exceptions ([tʌɪgɚ] 'tiger'), idiolectal variation, unusual phonemic contrasts ([aɪdl̩] 'idol' vs. [ʌɪdl̩] 'idle'), and variability in the application of raising before segments with nondistinctive voice such as /r/, all of which complicate the picture and suggest the presence a phonemic split rather than a rule (Joos 1942; Chambers 1973; Vance 1987). Still, Idsardi (2006) provides examples of vowel quality changes in semi-productive morphological processes ([dəsaɪd] 'decide' vs. [dəsʌɪsəv] 'decisive') and productive syntactic processes ([lʌɪ-to] 'lie to' vs. [laɪ-əbaut] 'lie about'), meaning Canadian Raising is a productive rule even if overridden in some lexical entries.

Joos (1942) reported two Canadian dialects, one in which flapping of intervocalic /t,d/ neutralized raising ([raɪɾɚ] for both 'writer' and 'rider') and one in which the vowel quality contrast remained at least for some words ([raɪɾɚ] vs. /[rʌɪɾɚ]). Only the latter dialect appears to remain (Chambers 1973; Vance 1987).

Both the monophthong and the diphthong patterns are found in several other languages besides English, but not the patterns in reverse. That is, raising before voiceless consonants in monophthongs and before voiced consonants in diphthongs is not found, or very rare (Moreton 2004).

2.7 Formant structure of following vowels

In a stressed vowel following an obstruent, a rapid rise in F_1 of several hundred hertz can also be a cue to voicedness (Stevens and Klatt 1974). While it is a perceptual cue, it may not be an independent gesture on the production side. The difference between the voiced and voiceless cases is that the formant transitions appear truncated (on the starting side) in the case of voiceless obstruents. Aspiration perhaps masks post-obstruent transitions and so the first measurable values of F_1 may be at a different point in the articulation, although the articulation is otherwise the same. This has been called F_1 cut-back.

Stevens and Klatt (1974) nevertheless suggested two ways that F_1 cut-back may still have a role in linguistic competence after all. First, the linguistic difference may not be an active gesture to control F_1, but instead it might be an acoustic target: that there shall be no rapid spectral changes following voiceless releases. The duration of aspiration may then be recruited to achieve this target. The longer VOTs for voiceless velar stops versus dentals, and dentals versus labials, may be explained by the fact that longer aspiration is needed to mask longer transitions. What Stevens and Klatt suggested then is that aspiration's role in [voice] is not a duration target such as "long-lag" but rather an acoustic, spectral target. That means that F_1 cut-back could be rather the flip-side of VOT.

For a second hypothesis for how F_1 cut-back may relate to a linguistic target, they note the difference in duration of [r] in [br] versus [pr] onset clusters. In [pr], [r] has a greater duration, and Stevens and Klatt proposed that this is a strategy to delay the rapid spectral changes from the [r] to the following vowel which might otherwise be misinterpreted as a cue for voicing of [p]. I have not seen any study validating whether there is a consistent duration effect in onset clusters such as these.

2.8 Other correlates during surrounding vowels

Summers (1987) reported greater jaw lowering in a vowel preceding a voiceless obstruent, meaning the jaw appeared to be doing more work despite a reduction in the total duration of the vowel. At the end of the vowel, the final jaw position did not vary with [voice] but the duration of the rising gesture was longer before a voiced obstruent. A longer duration is expected on the grounds that overall vowel duration is longer in this context, but also unexpected because the jaw has greater ground to cover before a voiceless obstruent. As a result, the jaw motion in the voiceless context is more rapid. (Lisker 1957 noted a similar pattern in formant transitions.)

Lisker (1957) noted that, for pre-vocalic /p/ versus /b/, the overall intensity of phonation is greater or rises more rapidly following /p/. If true, this is another acoustic correlate of [voice]. However, we can easily imagine that this difference follows directly from the different transglottal pressure configurations due to differences in VOT.

2.9 Syllable onset differences

Hawkins and Nguyen (2004) found that coda voicing can affect the realization of onset /l/, and probably onset fricatives as well. The duration of an onset /l/ is longer preceding a voiced coda than a voiceless coda, similar to the duration difference found for the syllable nucleus, the PVD effect. However, it is a much smaller effect. The mean duration ratio for /l/ was just 1.05 or 4 ms, which is approximately a single pitch period, though in particular instances higher ratios such as 1.18 were found. The small difference is roughly the magnitude of the PVD effect in languages besides English and French, where a physiological explanation is more likely than a specific duration specification in the grammar. Hawkins and Nguyen also found lower F_2 (by 125 Hz or 7 percent) and lower spectral center of gravity during onset /l/ in voiced context, something that might be explained by, for instance, anticipatory changes to oral cavity volume. But articulatory data is lacking. Hawkins and Nguyen preferred the interpretation that voice is manifest as a syllable-wide attribute, a contrast of "somber" (for voiced) versus "bright" (for voiceless) in a non-segmental model of acoustics or perception.

2.10 Summary

The feature [voice] is associated with many acoustic changes during the consonant itself, in preceding and following vowels, and even as far back as a preceding onset. This chapter attempted to survey the entire literature of the acoustics of the production of the [voice] contrast in adult English speakers.

Some aspects of this contrast are primary aspects of [voice] itself — targets that are a part of linguistic competence. Certainly the state of the glottis is a part of the [voice] target, manifesting variously in voice onset time, voice termination time, the probability that voicing occurs in closure, etc., depending on context. Other acoustic correlates of [voice] with evidence pointing in the direction of linguistic specification is that closure duration is longer in voiceless stops than voiced stops and the PVD effect that vowels are longer before voiced consonants. This is at least the case in English and French where the observed PVD ratios are relatively high. Some languages, as noted earlier, have no PVD difference. And some have been measured to fall somewhere in between, though better comparisons between languages are needed. F_0 perturbation due to voice is probably also phonologically based, with no complete phonetic explanation having been found and that the effect occurs in languages what make a voicing distinction either based primarily on the glottal signal or based primarily on aspiration.

Other measurements seem to lead in the direction of linguistic performance. The duration difference in frication correlated with [voice] seemed to be merely an interaction with the glottal state, and that the articulatory gestures are otherwise the same — contrasting with the correlate of closure duration for stops. Likewise, F_1 on the following vowel may be "cut-back", meaning the articulatory gesture is the same but aspiration covers up the formant transition. The effect on the F_1 of monophthongs also is most likely a physiological effect, especially in light of its neutralization by regressive voicing assimilation.

Chapter 3

The Acquisition of [voice]

Voice is composed of many elements. As discussed in the preceding chapter, several acoustic correlates of [voice] seem to be a part of linguistic competence. These correlates of [voice] are free to vary from language to language and their use (or absence) in any given language must be learned by infants. The state of the glottis in terms of VOT and VTT, closure duration, preceding vowel duration, and fundamental frequency following the consonant are correlates of voice that seem to be a part of linguistic competence. These contrast with the non-linguistic aspects of [voice] which result from physiological dependencies on other aspects of [voice]. Frication duration and preceding vowel F_1 were thought to be such correlates on the basis that their correlation with [voice] could be attributed to changes in the glottal state.

Many techniques have been used to explore what children learn about [voice] during development, including observing phone inventories, measuring acoustic properties in production, and infant discrimination/perception tasks. But the phonological side of [voice] is only recently starting to receive attention, e.g. by Dietrich, Swingley, and Werker (2007) and van der Feest (2007).

In this chapter the reader will first be introduced to some basic facts about language acquisition as it relates to the development of the [voice] contrast, and especially the preceding vowel duration (PVD) effect. Although much past work has already been done on this subject showing that infants exhibit a PVD difference as early as age 2 years, several important questions remain. For one, no reliable developmental trend has been observed and replicated in infants aged 1;6–4. That is, research so far has equivocated on whether infants develop an adult-like PVD effect during the course of language acquisition by starting with no PVD difference and increasing the duration of vowels before voiced stops over time, by decreasing the duration of vowels before voiceless stops over time, or whether the PVD effect is in fact the unmarked state in universal grammar.

Secondly, we also ought to not take it for granted that infants exhibiting the PVD effect do so because they have learned the same phonological process that adults are thought to use. This chapter will also consider alternate hypotheses that could explain the PVD effect pattern in infants. Evidence against each possibility will be presented.

3.1 Gross development of [voice] production

Meaningful speech with adult-like phone production begins by roughly 15 months of age normally (Stoel-Gammon 1985). The consonant phone inventory, as recognized by transcribers, grows over time. In Stoel-Gammon (1985), an inventory at an age meant the phones present in at least half of

	t	s	d	k	z	b	p	g	v	f
initial	16	17	12	12	0	14	9	10	1	8
medial	22	10	12	19	4	10	8	5	6	5
final	35	15	14	10	15	0	4	1	4	1
all	26	15	13	12	8	7	6	5	4	4

Table 3.1: Phone frequencies in adult infant-directed speech, as a percent of occurrence in each context. (Rows sum to 100, modulo rounding.)

the sampled children's inventories in hour-long recording sessions. The number of children varied across ages, from 7 to 33. Observed first in word-initial position are unaspirated (i.e. voiced) stops by 15 months, followed by aspirated stops at 21 months, followed by the (contrastively) voiceless fricatives at 24 months. (Voiced fricatives are considerably more difficult to produce because of the coordination of two constriction gestures and did not occur within 24 months.) The word-final phone inventory is at each age much smaller, and the order of occurrence of phone types may be different; however, this may simply reflect distributional differences in the language environment rather than something about language development. There, voiceless stops appeared first at 18 months, followed by voiceless fricatives at 24 months. (Voiced stops and fricatives did not occur in this position by 24 months.) In a study later on of infants from 11–18 months, voiced and voiceless stops and fricatives were found in both word initial, medial (intervocalic), and final position (Stoel-Gammon 2002). Still, perceptual abilities are continuing to become more adult-like even into the second decade of life (Hazan and Barrett 2000).

There are a number of limitations to approaching language development by measuring phonetic inventories. For one, phones are not equally frequent in the child's linguistic environment. An infrequently observed phone doesn't indicate necessarily that it is less fluent for the child, when in fact it may just be that it occurs only in infrequent words. A maximum of 50 tokens were drawn per child per recording session in Stoel-Gammon (1985), and in both Stoel-Gammon (1985) and Stoel-Gammon (2002) no effort was made to compare phone frequencies in the children to their frequencies in infant-directed speech. It would be premature to take the absence of phones from the inventories as evidence that the child could not actually produce the phone.

We can grossly estimate expected phone frequencies in child speech by looking at infant-directed speech, where the absence of a phone ought to bear on its frequency of use and not, as in infant speech, its difficulty to be produced. The Providence corpus (Demuth et al. 2006) contains orthographic transcriptions of infant-directed speech to six children aged 1–4 years and provided the necessary data to test the hypothesis that phone frequencies vary substantially. Each word in the orthographic transcription was matched against its first entry in the Carnegie Mellon University Pronouncing Dictionary. The consonant phones were separated according to whether they occurred word initially (including in a cluster), medially, or finally (including in a cluster). One million /p,t,k,b,d,g,s,f,z,v/ tokens occurred in the corpus. Their relative distribution is given in Table 3.1. As the figure shows, the phones are not equally frequent. The labiodental fricatives are as much as 35 times less frequent than /t/ in word-final position. /s,z/ are relatively common, but not in all positions. The voiced fricatives were the rarest segments in word initial position (occurring a total of 1 time), which may explain why they were not observed in word initial position in Stoel-Gammon (1985). What we ought to conclude from this table is that observed phone inventories especially in small samples are likely to underestimate the child's linguistic abilities.

Phonetic transcription alone also only tells us what articulatory gestures the infant has to work with and not anything about its interface with phonology. If the child produces [b] and [p] in free alternation for both /b/ and /p/ phonemes, an analysis of a phonetic inventory will reveal two phones but fail to indicate that the child has not learned either the phonetics of underlying [voice] or the [voice] feature values for the words in his lexicon. Comparing the child's phonetic output with the *adult* target shows whether the child is learning adult phonological categories. Snow (1997) measured VOT on word-initial stops and found a difference between voiced and unvoiced stops (that is, voiced or unvoiced according to their feature value in adult phonology) starting at least by age 1;6, the earliest age considered. Children at that age produced 41 ms more aspiration in voiceless stops, and 73 ms at age 1;9. (In adults the difference is roughly 60 ms.) This shows that not only is the child forming two categories, but that the child is learning by at least age 1;6 the right [voice] (or [asp]) feature value for the initial stop phonemes in the words in his lexicon.

3.2 Development of the preceding vowel duration correlate

As in adults, much past work has been reported on the PVD effect in infants. While past research has established that the duration difference is exhibited by infants, there has been no consensus on the presence of a developmental trend.

The earliest work on the acquisition of the PVD effect appears to be in an unpublished doctoral dissertation by M.A. Naeser in 1970 at the University of Wisconsin, which Krause (1982) reported as finding a vowel length difference before voiced and voiceless consonants in the spontaneous and imitated speech of age-1;9 (and up) children. Both experimental and corpus studies since then have confirmed the conclusion. Taking the results in reverse order by subjects' age, DiSimoni (1974) found a vowel duration difference at age 9, and DiSimoni (1974) and Krause (1982) both found a duration difference at age 6. Baran and Seymour (1976) reported a PVD difference at age 5 (in African American Vernacular English speakers). Both DiSimoni (1974) and Krause (1982) found a durational difference at ages 3–4, with ratios of about 1.8. At age 2;6 and 2;0, PVD ratios were reported of roughly 1.7 and 1.4, respectively (Buder and Stoel-Gammon 2002). Ko (2007) reported a difference before the age of two, based on a corpus study.

Three papers have reported on a developmental trend. Both DiSimoni (1974) and Krause (1982) found that the durations of vowels before unvoiced consonants was relatively stable from age 3 to 9 (DiSimoni) or 3 to "adult"-age (Krause). But DiSimoni reported an increase in duration of vowels before voiced consonants with vowels before voiceless consonants stable, and so an exaggeration of the PVD effect, and Krause a decrease for voiced consonants producing a decline in the PVD difference with age. Ko (2007) similarly looked at the duration over time for the vowel in four frequent words in two children from age 1;6 to 4 and found no reliable pattern. A confounding factor is the developmental trend of overall vowel duration. Lee, Potamianos, and Narayanan (1999) reported vowel duration decreases by 25% from 5 to 7 years of age and roughly a 3-4ms decrease per year until age 15. In light of this, the findings of roughly constant vowel duration preceding unvoiced consonants is quite a surprise. Not only that, but we know that the PVD effect decreases as vowel duration decrease (i.e. incompressibility, see Section 2.4), and so a stable mean PVD ratio during a time when overall vowel duration is decreasing could, in fact, indicate learning. From these studies, the developmental trend of the PVD effect is so far not clear.

What kind of developmental trend do we expect? Or, what is the unmarked or native state? Could children actually start with a PVD difference and lose it over time in languages that do not have a PVD difference (Buder and Stoel-Gammon 2002)?

To test this hypothesis, Buder and Stoel-Gammon (2002) compared English and Swedish 24- and 30-month-old infants in their productions of vowel duration before voiced and unvoiced stops in an elicitation task. Swedish has a phonemic vowel length contrast and essentially no PVD effect (a ratio of 1.03). Buder and Stoel-Gammon predicted that the PVD ratio in Swedish children would resemble that of English children at first and then decrease with age as vowel duration became increasingly used to signal phonemic length — unlearning the PVD effect. The results were compatible with the prediction, with English and Swedish children at 24 months showing PVD ratios of approximately 1.4 and 1.7, respectively, and at 30 months 1.7 and 1.1, respectively. A ratio of 1.7 in Swedish infants is remarkable, considering the lack of supporting evidence for a difference this large in Swedish adult speech. (The PVD difference is probably not exaggerated in infant-directed speech either. The motherese of Dutch, which like Swedish has a phonemic vowel length contrast, lacks the sing-song-ish vowel elongation that is found in e.g. English motherese, and this is thought to be because it could change the meaning of the utterance, Dietrich et al. 2007. It stands to reason that Swedish would similarly limit duration changes of vowels in motherese.). If the time trend were true, it might support the hypothesis that the PVD effect is the unmarked option, that infants start *with* a PVD difference. In infants learning languages which exhibit the effect, including English, the PVD effect would be stable over time, while it would be attenuated in other languages. We should not be surprised, then, that no consensus on a developmental trend has been found for the PVD effect in English.

But from just two points in time for the Swedish children, it is hard to read much into these results. Durations are highly variable and we saw from DiSimoni, Krause, and Ko that three papers can produce all logically possible results.

3.3 Grammatical status of the PVD effect

Although we may see vowel duration differences in infants that resemble the adult pattern, we cannot take for granted that this is due to the same phonological process in infants. There are several reasons why a child might have a surface PVD difference:

1. Adult-Like: The process may be phonological in the adult-like way, where the duration of a vowel depends on the voicing of the following consonant in a productive process.

2. Phonemic Vowel Length: The duration difference may be phonological, but stored as a property of the vowel rather than the consonant. In other words, the child may employ a phonemic vowel length contrast and has “learned” a distribution of short versus long vowels that only coincidentally correlates with post-vocalic voicing.

3. Performance: The duration difference can be explained by a physiological process related to some (other) articulatory correlate of [voice].

4. Epiphenomenal: The duration difference is epiphenomenal. Although the infant’s vowels exhibit the expected pattern, the difference is due to other factors, for instance having a lexicon in which, coincidentally, vowels with longer intrinsic duration (i.e. the so-called tense vowels and diphthongs) occur more often preceding a voiced coda. Alternatively, voiced codas may occur more often in words more prone to being lengthened due to being focused.

We can be asking the same type of question of any other acoustic correlate of [voice], whether it be formants, the glottal signal, or consonant release. And we should be prepared for there being a different answer at different points during language development.

We would like to simply test for whether the PVD effect is a productive process in infants. Child elicitation studies are difficult to carry out, and it would especially difficult to ask children to speak novel words without also giving them a hint as to how they should be spoken. For many reasons an elicitation study was ruled out for this dissertation, which instead is based on corpus data.

Some of the hypotheses about the PVD effect can be more easily ruled out than others. One of the hypotheses is that vowels with shorter intrinsic duration, such as the lax /ɪ/, tend to pair with voiceless consonants more often than the longer or tense vowels. As a result, we would observe a duration difference correlated with [voice], even though the correlation was entirely by chance. But it is not the case that shorter vowels tend to be paired more often with a following voiceless consonant than longer vowels. In both the infant-directed adult speech and the child speech in the 'Lily' portion of the Providence corpus (Demuth et al. 2006), the correlation is strongly the reverse.[1] To test this, the log-ratio of the number of occurrences of each vowel type preceding voiced consonants to the number before voiceless consonants was computed (all with primary stress). This was compared to the rank ordering of the vowel when they are ordered by their intrinsic vowel duration (at least as given by Umeda 1975), i.e. 0 for /ɪ/, 1 for /ɛ/, and so on. For adult speech, the correlation is -.76 ($N = 124,021; p = .01$), for child speech -.44 ($N = 27,767; p > .2$). The negative correlations indicate that longer vowels tend to occur more often than shorter vowels with voiceless consonants. For instance, /aɪ/ occurs nearly four times as often with a voiceless consonant than a voiced consonant in the infant-directed speech and is one of the longest vowels, whereas /ɪ/ occurs twice as frequently with voiced consonants than /aɪ/ but is one of the shortest vowels. If anything, this empirical distribution of phones would obscure rather than create a vowel duration pattern like the PVD effect. From this we can rule out one type of epiphenomenal explanation for the PVD pattern in infants.

It is well known that infants 'lose' their ability to make discriminations on non-native acoustic dimensions. For instance, while English- and Hindi-learning infants at 6–8 months both can discriminate dental from retroflex stops, only the Hindi-learning infants continued to do so at 10–12 months. Or while Spanish and Catalan infants can discriminate a Catalan-specific vowel contrast at 4 months, only the Catalan-learning infants can at 8 months. (See Mugitani, Pons, Fais, Dietrich, Werker, and Amano 2009 for a summary.) The loss of the ability to discriminate non-native contrasts is certainly a net-gain for the child: the child will be at a disadvantage if he tries to find meaning for distinctions where none exists. Losing non-native contrasts is just another way to say forming native categories.

Perceptual abilities are found quite early. As for discriminating values of [voice], a categorical VOT distinction in perception in English is possible at least as early as 1–4 months, when infants have been shown to recover from habituation when presented with (short-lag) voiced or (long-lag) voiceless stops that differ by only a 20ms change in VOT (Eimas, Siqueland, Jusczyk, and Vigorito 1971). Similarly, infants aged 2–2.5 months were shown to discriminate /s/ from /z/ and /t/ from /d/ in word-final position in a high-amplitude sucking dishabituation design (Eilers 1977). Though Eilers (1977) pursued the question, it was not entirely clear exactly which acoustic cues the infants were using. When presented with [at] vs [aːt] stimuli, which adult speakers were reported to identify

[1] As opposed to the acoustic analysis reported in Chapter 6, this counts all tokens in the entire corpus. Dictionary definitions were used to categorize consonants as voiced or voiceless.

as /at/ and /ad/, the infants in the study did not notice a difference. The infants must have been using other factors, such as voice termination time or release burst voicing.

Ko, Soderstrom, and Morgan (2009) returned to this question in a looking-time task with infants aged 8 and 14 months. They were played potentially familiar monosyllabic words (bag, back, cub, cup, pig, and pick) with digitally altered vowel duration. In the mismatch condition, the vowels preceding voiced codas were shortened by half and vowels preceding voiceless codas were lengthened by 60 percent. (In the match condition, vowels were first mismatched, and then digitally altered to return to their original duration — to ensure that all stimuli were similarly digitally altered.) The stimuli were read in infant-directed speech style and with a "strong coda release" so that F_0 and formant transition cues were available to the infant to potentially conflict with the vowel duration cue. Infants looked at a light located at a sound source which played the stimuli; if the infant detected a mismatch, looking times would be different in the match and mismatch conditions. Of the eight conditions in this age × voicing × vowel duration ($2 \times 2 \times 2$) design, the only difference found was that the short vowel with voiced coda conditions in the 14-month-olds stood apart from the rest with a shorter looking time.

If infants have no linguistic knowledge about the relationship between vowel duration and postvocalic consonant voicing, the 14-month-old infants should not have been able to have a different behavior in match and mismatch conditions. (If the looking time difference was caused by some inherent acoustic novelty of the mismatch, then the 8-month-olds would probably have been expected to show the same pattern, which they did not.) At least in so far as perception goes, and if we accept the conclusions presented in the study, some linguistic knowledge about the PVD effect must be available to the 14-month-olds and is acquired after 8 months.

Other work has agreed, albeit weakly, that the PVD effect in infants results from linguistic knowledge and not a physiological process. Children often omit word-final stops: In a study of three children age 3;10-7;6 with clinical problems that likely exacerbated their rate of word-final stop deletion to the rates ranging from 0 to 97%, all three children showed the PVD effect in utterance-final context when the consonant itself was deleted. The older children had PVD ratios of 1.3 and 1.6 (while the youngest and most impaired child's ratio was 1.1) (Weismer, Dinnsen, and Elbert 1981). Still, the fact that a stop appeared to be deleted does not mean that no aspect of its articulation was performed. The vocal cords might abduct even for a voiced stop in which closure is not achieved, and this could, in turn, perhaps explain other acoustic differences of [voice]. Also, of course, we would prefer to have production data from clinically normal children if possible.

The results for Ko et al.'s (2009) 14-month-olds are difficult to interpret, however. This was a two-by-two design: infants were given short or long vowels and voiced or voiceless codas (meaning the acoustic cues besides vowel duration). The hypothesis was that the mismatch conditions (short vowel with voiced coda; long vowel with voiceless coda) would have longer looking times compared to the match conditions (short vowel with voiceless coda, long vowel with voiced coda). But this was not entirely born out, as described above. The interpretation of the results depends on how the conditions are grouped. Ko et al. believed that the asymmetry in the data was about vowel duration: only in the short duration category did infants notice a mismatch (i.e. incorrect coda voicing); looking times between the match and mismatch were similar in the long-duration group. They reasoned that infants are less sensitive to mismatches with long-duration vowels because phrase-final lengthening and vowel elongation in infant-directed speech make lengthening not an unusual occurrence, and so not worth discriminating. But Ko et al. could have equally parsed the results in two other ways. By grouping the categories the other way, only in the voiceless conditions was a mismatch (i.e. long vowel) of interest to the infant; in the voiced conditions no looking time difference was found when the vowel was shortened. In other words, infants were interested when the vowel *was too long*.[2] Or in the third perspective, only in the match conditions was there a looking time difference. That is, when vowel duration matches other cues infants look longer at longer vowels (after all, there is more of it to look at it), but when there is a mismatch infants always look a long time (which was the expected outcome in any case). Certainly something anomalous and interesting is happening, but it is impossible to know exactly what.

While something like native categories begin to form well before the end of the first year of life, this is not the same as storing a phonological feature value in a lexical entry. For instance, according to van der Feest (2007), Dutch-learning infants at age 1;8 cannot detect when an image of a word they know is paired with a mispronunciation made by a [voice] mismatch in initial position, i.e. a picture of a cat paired with the auditory stimulus [pus] (correct) or [bus] (mismatch). This was tested with a split-screen preferential looking paradigm, with the screen split between the target image and a distractor. By age 2 the infants have learned the word discrimination task, at least partially: Looking times at the target and reaction times were different when a voiceless word was mispronounced voiced, but not in the other direction. A similar effect was found for English-learning infants aged 14–21 months in Swingley (2009) with minimal pairs such as ‘boat’ versus ‘poat’ (mispronounced in onset) versus ‘boad’ (mispronounced in coda), and ‘cup’ versus ‘gup’ (onset) versus ‘cub’ (coda). There the infants noticed a difference in both onset and coda, based on looking times to a target and distractor. Apparently, according to van der Feest, Dutch-learning infants are able to make the perceptual distinction when given the auditory stimuli alone. Thus infants seem to recognize that the tasks are different, in particular that the task with images is about word learning and not about acoustic discrimination. It seems the ability to make a perceptual distinction does not translate immediately into the ability to associate the distinction with a word in the lexicon and to use that knowledge in this type of task.

It is actually surprising then that the PVD effect would be observed in infants earlier than age 2 when van der Feest’s (2007) results would suggest that the specification of [voice] in lexical entries is only beginning to be made at this time. This exactly highlights the contrast between studying phonetic inventories and studying infant phonological knowledge. An infant at, say, age 1;5 would very well be expected to produce both “voiced” and “voiceless” stops. But at the same time, he would be expected to not be able to associate the right value of this feature with lexical entries. The two segments would have to be in essentially free alternation.

Dietrich, Swingley, and Werker (2007) showed in a habituation task that English-learning children at age 1;6 do not notice a near doubling or halving of the duration of a vowel in either a foreign (Dutch) or English nonsense word when paired with its image — although they can make the discrimination when there is no pairing with an image (Mugitani et al. 2009). On the other hand, Dutch-learning children do notice the vowel duration mismatch (they dishabituate). Dutch is a language with a phonemic vowel length contrast, unlike English, and so vowel length must eventually be stored in the lexicon. It follows from the interpretation of van der Feest (2007) that English-learning children are correctly not entering vowel duration information into their lexicon

[2] This may not be about vowel duration, however. In van der Feest (2007), a study of Dutch-learning infants’ perception of syllable-initial voice, the same direction of asymmetry was found. Mismatches were detected with voiceless onsets only, and not voiced onsets. But I keep this as merely a footnote because mismatch meant something different in van der Feest (2007): Rather than a mismatch between two acoustic dimensions, it was a mismatch between the acoustic signal and the infants’ knowledge of the correct pronunciation of the word indicated visually. Still it is interesting that the same direction of asymmetry was found.

while Dutch-learning children are. As Mugitani et al. concluded, English-learning infants aged 1;6 do not have a phonemic vowel length contrast. This rules out one possible explanation for the PVD effect in infants.

Below is a summary of the evidence against each possible explanation of the PVD effect in infants besides an active adult-like phonological process:

1. Phonemic Vowel Length: English-learning children will not notice a change in vowel duration when paired with the image of the word it is in. (Dietrich et al. 2007; Mugitani et al. 2009).

2. Performance: The PVD difference persists in tokens with omitted stops (Weismer et al. 1981) and infants notice some vowel duration–coda voicing mismatches (Ko et al. 2009).

3. Epiphenomenon: The distribution of vowel types in adult infant-directed speech rules out this hypothesis.

3.4 Discussion

The review of the literature in this section directs us to a picture of [voice] that emerges in three stages. In the first stage in the first few months after birth, acoustic dimensions that take part in [voice] become perceptually accessible, including VOT and only later preceding vowel duration. During the second stage around 8–14 months, associations between phonetic cues to [voice] begin to form, as seen by infants' ability to notice a mismatch between vowel duration and the other cues to post-vocalic voicing. But at this stage, the infants are not storing this information into their lexical entries. Only between 20 and 24 months does it appear that voicing information is stored in lexical entries.

It is still not known exactly what information about [voice] is stored in linguistic knowledge and lexical entries and when that begins to happen. In van der Feest's (2007) study of 20- and 24-month-old Dutch infants, a mispronunciation involved two levels of representation: the underlying [voice] feature and the surface acoustics and articulation of pre-voicing (or lack thereof). From this study alone, we cannot infer which level of representation was the basis of the infants' discrimination at 24 months. Likewise in Ko et al. (2009), we cannot be sure whether the 14-month-old infants' behavior was due to having trouble choosing a value for [voice] in the face of incompatible acoustic evidence (preceding vowel duration versus other cues) or whether the infants merely noticed a statistical anomaly in the acoustic signal.

The inquiry into what is happening in the minds of language users is a program of study rather than a single experiment. And when it comes to infants, for whom we must give up the premises that they are competent speakers and that their grammar is stable, finding answers becomes even more challenging. The direction forward is to push deeper into the richness of [voice] in infants' grammars. Some of the evidence presented in this chapter provided evidence that the PVD effect in particular is a phonological phenomenon in infants, and if true would mean that van der Feest (2007) and Ko et al. (2009) did indeed reveal facts about phonological knowledge. But nothing is known for sure and additional evidence, to be presented in Chapter 6, will contribute more perspective to these questions.

Chapter 4

The Phonetic Implementation of Features

In the preceding chapters I explored the acoustic correlates of [voice] and discussed the evidence for and against each correlate being a part of linguistic competence — meaning, whether the correlate is encoded mentally. The next question is, if a correlate *is* mentally encoded, *how* is it mentally encoded? This is an important question when investigating language acquisition because it shifts the focus from merely an observation of the infant's behavioral changes to an insight into the infant's mental representations.

Feature theory has generally carried the heavy weight of classifying phonemes into groups according to how they participate in phonological rules (Chomsky and Halle 1968) as well as how phonemes are realized at the level of phonetics. The [voice] feature separates phonemes into two groups, the [+voice] "voiced" phonemes and the [-voice] "voiceless" phonemes. At the level of phonology, this distinction plays a role in such rules as voice agreement in the '-s' morpheme in English ('cats' [s] versus 'dogs' [z]) and final devoicing in languages including German. If a process like these generalizes, that is, if it is productive, it is because it is encoded not as acting on a fixed list of words (cat, dog) or phonemes (/t/, /g/) but because it picks out its conditioning environment based on an abstract feature specification (e.g. [-continuant, +voice]). At the level of phonetics, there are generalizations as well — ones we believe are related to the [voice] feature somewhere within the phonetic or phonological subsystems of grammar. This is that all of the [+voice] phonemes share the same properties when it comes to the glottal state (VOT, etc.), aspiration, stop closure duration, and preceding vowel duration (PVD). But whereas phonological phenomena such as ordered rules or constraints are investigated within a formal framework, considerably less attention has been paid to the formal system that connects features to their phonetic counterparts.

A Compositional Phonetics

Anyone working in phonetics and phonology will have some model in mind for how the two parts of the linguistic system connect, and by and large there is a lot that most will agree on about the nature of the interface. For one, most would say the phonetic exponence of a phonological segment is what I would call "compositional", to borrow the term from semantics. A "compositional semantics" is one in which the meaning of a term is derived from the term's parts and how they are put together (and from nothing else). Here, the phonetic exponence of a phoneme is determined by its featural specification (its parts) and nothing else. For instance, once we have identified the acoustic (or perhaps articulatory) signatures of the phonological features [+voice] and [-voice], [+strident] and [-strident], [+high], [+back], etc., we could put the pieces together to predict the phonetics of any segment in the language's inventory.

This idea goes back to Jakobson and Halle (1956), to whom we owe much of modern feature theory. They wrote:

> The speaker has learned to make sound-producing movements in such a way that the distinctive features are present in the sound waves, and the listener has learned to extract them from these waves. (p8)

Unfortunately there is no getting around that phonetics is not strictly compositional. It is readily observed that despite the use of a small set of binary features to represent the vowel inventory, the acoustic vowel space is not organized in a discrete grid pattern (even with a rotation of the vowel space). In other words, the formant targets of vowels cannot be entirely modeled based on their feature specifications alone: each vowel phoneme misses the target we would expect for it given its featural specification alone. Similarly, VOT values in voiceless stops vary across place of articulation: velar $\gg$ dental, labial (Lisker and Abramson 1964a). This variation cannot be accounted for by the featural specification either. (Except recall Stevens and Klatt 1974's proposal that this is due to a constant acoustic target, discussed in Section 2.7.) And that is just within a language. Precise VOT values vary from language to language (Lisker and Abramson 1964a), the formant targets of /i/ vary greatly depending on the number of vowels in the language (Lindau and Ladefoged 1986), and the timing of the releases in ejectives varies from language to language (Browman and Goldstein 1986). It can't be that these features are universal down to the phonetics and also that the phonetics-phonology interface is strictly compositional. One must be wrong.

But if it weren't true at all that phonetics is compositional, that would make each phoneme's exponence no more than an idiom. It would be a coincidence that phonemes that shared feature values also shared phonetic properties. The whole of Chapter 2 would be a mystery. Not only would we ask why /p,t,k/ all contrast similarly with their cognates /b,d,g/, but why should there be cognates at all?

This conundrum arises not just from how features are put together but also, as I discuss next, from what the inventory of features looks like and how phonetically rich features are allowed to be.

Narrow versus Algebraic Features

Starting with Jakobson and Halle and continuing even through today there has been a line of inquiry into what is *the* phonetic correspondence of each feature known to phonologists. They wrote, "The sameness of a distinctive feature through all of its variable implementations is now objectively demonstrable" (p14). (Whether it is an acoustic property or an articulatory property that is invariant has been another subject of debate (see Clements and Hallé 2010 for a summary), but the difference is not important here.)

There is one major problem, and that is that after at least a half century of debate on the subject there is little sign of a consensus on what the phonetics of features actually are. This is not exactly surprising. Parker (1977) observed early on that of the many acoustic correlates of [voice], not even one seems to occur in every context in which a consonant can appear. [voice] in initial stops is distinguished primarily by VOT, a context in which closure duration and preceding vowel duration are not applicable. Final stops may be distinguished by preceding vowel duration, but there is no VOT or closure duration to be measured in this context. Parker wrote, “[E]ach of these acoustic cues must of necessity be mutually exclusive with at least one of the others,” meaning there is always one cue that doesn’t occur with the others.

The complex presentation of acoustic correlates of features leads to a perpetual debate (or refinement) over such issues as whether voicing in stops of any given languae is a [voice] distinction (generally tied to glottal vibration) or a [tense/lax] distinction (generally tied to some abstract notion of force) (Jessen 1998, 2001), or if voicing in fricatives is distinguished by [voice] or the feature [spread glottis] (generally tied to aspiration) (Vaux 1998).

Jessen (1998, 2001) followed Jakobson and Halle’s (1956) position, what van Rooy and Wissing (2001) called the “narrow” view of features, that each feature names a particular acoustic/articulatory distinction that is consistent across contexts. Jessen asked which of [tense/lax] or [±voice] is more appropriate for German and similar languages including English, and what is the phonetic exponence of these features that is common across contexts? He decided on [tense/lax] for German, with this sort of contrast realized primarily as a duration difference of the segment, especially in terms of aspiration. [tense] segments should have some common longer-duration property. Languages which employ [±voice] have a difference in the presence of glottal vibration during the segment, i.e. a pre-voicing/short-lag VOT difference, and not a short-lag/long-lag difference which is a matter of aspiration and thus an implementation of [tense/lax].

Jessen still ran into the problem of defining what the singular phonetic property of [tense/lax] is. In Jessen (1998) he vacillated between whether [tense/lax] is primarily a matter of aspiration (in stops) or total phone duration (in fricatives, page 279), and also allows a language to employ [tense/lax] to make the voicing contrast primarily as a difference in aspiration duration, closure duration, total duration, and/or preceding vowel duration, and not aspiration duration necessarily. He noted that Danish, Hindi, Mandarin, and Swati Xhosa’s [tense] stops are aspirated but have a shorter closure duration than their unaspirated counterparts (opposite to the usual pattern, see Chapter 2). [tense] cannot both refer to an overall or nonspecific longer duration requirement and also appear in languages in which [tense] segments are in some way shorter. He also expanded the feature to other uses, allowing it to be responsible for gemination contrasts. Given all of these possible meanings of [tense/lax], Jessen’s position is really not narrow at all.

The opposite of narrow features are “algebraic” features, with “the maximal estrangement between phoneme and sound” (Jakobson and Halle 1956, p. 15). On the algebraic side was Keating (1980, 1984) who proposed that the phonological feature could have different phonetic representations from language to language:

> [A]t one level of feature representation, all two-category voicing contrasts should be represented as [±voice], regardless of their phonetic specifications. The intention of this proposal is that the phonological rules which refer to voicing be equivalent across languages at a higher level than the phonetic. Thus voicing assimilation and devoicing processes will be described similarly across languages, regardless of the actual phonetic contrasts involved. (Keating 1980, p233)

> [T]he occurrence of a phonological rule in a language should not depend on, or be correlated with, the phonetic details of the language. (Keating 1984, p292)

Voice-related phenomena that are found both in languages with an apparent (narrow) [voice] contrast and those with an apparent [tense/lax] contrast give support to the separation of phonetics from phonology. Keating (1984) cites the cases of the PVD effect and F$_0$ perturbation as such phenomena. As discussed in Chapter 2, F$_0$ perturbation had the same pattern in languages with a single voice-like contrast regardless of whether the contrast was primarily voicing or aspiration. Also see Kingston and Diehl (1994).

But here it gets murky because an author can simultaneously claim that a feature is phonetically abstract, but only up to a point. Keating (1980) mapped [voice] onto VOT specifically. The abstraction was not that [voice] could map to any phonetic dimension or combination of dimensions, but that [voice] maps to particular points on the VOT spectrum: pre-voicing, short-lag, or long-lag.[1]

Vaux (1998) debated the featural specification of voiced versus voiceless fricatives. While he leaned on articulatory evidence, he intended to make a phonological point: "It should be noted that the theory presented here is primarily a theory of phonological rather than phonetic representations" (p509). And Vaux's most interesting points have a distinctly phonological appearance. In support of the use of the aspiration-feature [spread glottis] for fricatives is that "voicing" in fricatives patterns in some cases with the presence of aspiration in stops in what can only be a phonological rule. The first interesting case that I will summarize occurs in the Armenian dialect of New Julfa, which has a four-way stop contrast parallel to that of Hindi (i.e. voicing and aspiration both contrastive). There is a certain 'kə-' prefix in which /k/ is aspirated just in those cases when the prefix is followed by either an aspirated stop or a voiceless fricative (/k+g^h-o-m/ → [g^{h}əg^{h}om], /k- savor-ie-m/ → [k^{h}əsavoriem]) but not a vowel, unaspirated stop, or voiced fricative. The second interesting case is similar: In the Seville dialect of Spanish, /s/ debuccalizes in coda position to /h/, but it also causes a following stop to become aspirated (/los padres/ → [loh p^{h}aðreh]). Given the pairing of fricative voicelessness with aspiration, Vaux argues that they are governed by the same feature. The articulatory differences between aspiration and frication and the distance in each example between the triggered aspiration and its conditioning environment make these easily understood as phonological rules rather than phonetic, e.g. coarticulatory, processes. Vaux's style of analysis is precisely the type of "algebraic" detached-from-phonetics approach to phonological features that Jakobson and Halle (1956) sided against. From Vaux's arguments, nothing about the actual articulation or acoustics of fricatives is deduced (nor was that intended).

An algebraic linguist needs two types of features. On the one hand, there are phonological features. These are presumed a part of the grammar and are used to model categorical phonological processes. They are the features of Chomsky and Halle (1968). On the other hand, there are phonetic features, which are terms to identify acoustic or articulatory properties of speech sounds without regard for any mental representation of those differences. Phonetic features are points on in the continuous, multi-dimensional space of acoustics or articulation. They are used in the same way we use terms like "voiceless unaspirated [p]" and "voiced unaspirated [b]", which may not even have any particular grammatical significance in a language (i.e. when they are not phonemes). Having a distinction between phonetic and phonological features is relevant and useful for conducting coherent discourse on the phonetics-phonology interface.

Unfortunately the notation of features is confusing, since [+voice] can be used in the narrow sense to refer ambiguously to both the phonological and phonetic aspects of a phoneme. Keating (1984) used curly braces for something like what I am calling phonetic features. She wrote of {voiced}, {vl.unasp.}, and {vl.asp.}, which divide the acoustic space differently than the conventional features [±voice] and [±asp] do. Changing feature notation would not be easy. Since brackets are used to discriminate phones from phonemes (i.e. [b] versus /b/), standard feature notation such

[1] This is perhaps a narrow reading of Keating (1980).

as [+voice] ought to refer to a narrow or phonetic feature only. This square-bracketed "[±voice]" should describe the difference between the phones [b] and [p] (but not [p] and [p^h] which would be distinguished by [±asp]). If future phonologists want to avoid ambiguity, perhaps a slash-bracketed "/±voice/" should be used describe the difference between the phonemes /b/ and /p/ (regardless of whether they are English [p] and [p^h] or Dutch [b] and [p]). (For the sake of clarity, however, I continued to use [voice] throughout to refer to the phonological feature only.)

Redundant Features Undergeneralize

Stevens, Keyser, and Kawasaki (1986) divided features even further. They split phonological features into distinctive features — those that signal differences between words — and redundant features, which do not. Redundant features provide the listener with added cues that might be useful in the face of noise. For instance, [+round] enhances the acoustic properties of [+back] because both lower F_2. This, they say, explains why English back [u] is rounded and front [i] is unrounded. In these cases, [round] is recruited as a redundant feature. Similarly, [back] enhances [distributed] because "[a] fronted tongue-body presumably provides a favorable posture from which the apico-alveolar construction can be achieved, whereas a backed tongue-body position provides a posture that favors formation of the dental or interdental construction" (436), explaining a correlation between these features in Malayalam.

Stevens et al. also addressed the correlates of [voice] in terms of redundant features. A nasal murmur at the start of stop closure contributes low-frequency energy, which supports the perception of glottal vibration (which is also low-frequency energy). In the Tokyo dialect of Japanese, they report, [k] and [g] sometimes contrast intervocalically as [k] and [ŋ]. In other words, [nasal] can be recruited into the featural specification of voiced stops to enhance their acoustic contrastiveness. How far can redundant features take us in explaining the ensemble of correlates that make up [voice]? In this model, all of the grammatically-stored correlates of [voice] are encoded as features. Let us say that English voicing uses [tense/lax] distinctively, which accounts for the duration of aspiration. Then separate, redundant features must be employed to account for pharyngeal cavity size gestures, closure duration, F_0 perturbation, and the PVD effect. And this is not to speak of how the realization of [voice] varies by context.

This model of the exponence of [voice] misses the point of a phonological voice feature, however. First, consider what the specification of /p/ is in a redundant feature model. Like all phonemes, it is a bundle of features: [+labial, -voice, +tense, -PVD, +F_0 perturbation, etc.]. Redundant features are recruited into a segment's bundle of features as needed. A segment is free to have whichever redundant features makes sense for it. For instance, [+back] vowels are not *necessarily* rounded. Only the non-low [+back] vowels are rounded. As far as the grammar is concerned, the relationship between [back] and [round] is accidental. It might be the unmarked state that they co-occur, it might be a historical development, but the grammar does not say that they must co-occur. This is problematic for [voice]. Under a redundant feature model, it would be an accident that *all* contrastively voiced segments make use of the *same* correlates of [voice], when each segment is free to choose its own redundant features. /p/ recruits a redundant feature for F_0 perturbation, and so does /t/, and so does /k/, etc. The redudant feature model is missing a generalization (it undergeneralizes, see Section 7.1.4). Since [voice] and its acoustic correlates make a regular paradigm, a model that leaves the relationship between [voice] and its components unspecified is not acceptable — at least not until the paradigm is shown to be wrong. What is needed is a model in which multiple acoustic properties are stored within a single feature without the requirement that they all be present in all contexts.

Only if a feature can be more abstract than a single acoustic/articulatory dimension can a complete account of [voice] be modeled within the grammar. Stevens et al. (1986) introduced the notion of a "cover feature" that combines several low-level features. Jessen (1998) explored the possibility that a feature has a basic (e.g. primary) correlate that is essentially invariant across contexts but may also be associated with other secondary correlates that are potentially context-dependent. In the case of [tense/lax], the basic correlate is the presence of aspiration for [tense] with secondary correlates including preceding vowel duration, etc. Cover features and Jessen's features differs from the redundant feature model by allowing [voice] to subsume other aspects of acoustics and articulation, and explaining why the correlation applies necessarily to all segments marked for [voice]. In a model using cover features it might be that "[-voice] = [+spread.glottis +tense, -PVD, +F_0 perturbation]" and this is a part of the grammar, and /p/ simply is [+labial, -voice]. The correlates of voice *necessarily* carry from one phoneme to the next, as the data seems to indicate. Note that this isn't a departure from the narrow interpretation of features. A feature of this sort still has a particular phonetic exponence across contexts, but the exponence involves several acoustic dimensions.

Where is the PVD in this model?

One of the dimensions in particular should raise a red flag. If the PVD effect is treated as a component of [voice], then the model allows a feature to encode something about the phonetic realization of the *preceding* segment. This model might violate our sense of the linguistic system as operating segment-by-segment from left to right, or our view of phonetics as being compositional. That is not very appealing, and it is certainly not very narrow. But it would not be the first complex phenomenon to be proposed to be encoded within a feature. Van Rooy and Wissing (2001) proposed that even Optimality Theory constraint rankings could be encoded as part of a feature. Observing (or claiming) that regressive voicing assimilation of the regular phonological type found in many languages besides English is found in a language whenever the voicing contrast in the language is made by narrow [voice] (versus [tense/lax]), they propose that RVA is not a separate phenomenon from narrow [voice]. "[RVA] is an inherent consequence, even property, of the distinctive feature [voice]," they wrote.

But this is not the only way to think about what is going on. Browman and Goldstein (1986) proposed an interesting account of the PVD effect within the framework of Articulatory Phonology. Rather than thinking of the duration of the vowel as a parameter which is affected by the following [voice] feature, they proposed that the duration is a consequence of the relative timing of the gestures involved in the two segments. It is relative timing, in terms of a phase relation, that is the parameter affected by [voice]. "Phase" in Articulatory Phonology is used like that in waves but means something like the scheduled time of an articulatory event. "[T]he phase angle for /b/ relative to the vowel is somewhat greater than it is for /p/ (approximately 205 degrees vs. 180 degrees)" (p246), they wrote. If the consonant starts later, then the vowel will come out longer. Browman and Goldstein's account of the PVD effect is that it is actually just a property of the post-vocalic consonant, which is a handy point of view since it allows the PVD effect to be accounted for not with a rule that involves both the vowel whose duration changes and the following consonant which houses the [voice] specification, but with the consonant alone. This exemplifies one of Articulatory Phonology's premises: phonetics appears more discrete when we view it along the right dimensions.

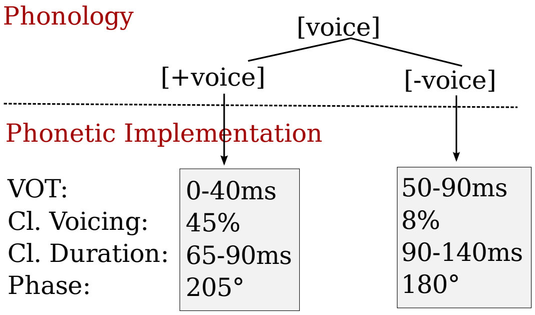


Figure 4.1: A sketch of the phonetic implementation of [voice] in English, including the dimensions VOT, closure voicing, and closure duration (see Chapter 2), and phase (see this section).

This is a concrete implementation of the compensatory lengthening/shortening account of the PVD effect discussed in Section 2.4.3. Unfortunately, evidence indicated that the PVD effect could not be fully described by a compensatory adjustment.

Finally, must redundant features be *features*? In the specification of [-voice] above — "[-voice] = [+spread.glottis +tense, -PVD, +F_0 perturbation]" — we are still forced to define the exponence of each of the redundant features if we cling to a compositional phonetics. But what if the [+spread.glottis] in aspirated stops differs in exponence from the [+spread.glottis] in voiceless fricatives (which could be true in the New Julfa case described above), or different from voiceless sonorants? Let's say this were true, then we perhaps might as well get rid of the intermediate level of abstraction for the redundant features. Instead, [-voice] would specify its exponence in continuous, phonetic space directly without reference to the values of other features: [-voice] = [VOT=40ms, closure=75ms, etc.]. This is sketched in Figure 4.1.

Earlier I noted that Bell-Berti's (1975) data indicated pharyngeal gestures are recruited by speakers in an ad hoc manner to achieve a more abstract target, such as to maintain a certain prescribed amount of voicing. Where does this fit in the model? On the one hand, ad hoc recruitment of gestures looks a lot like redundant features and it is exactly parallel to Kluender et al.'s (1988) explanation of the PVD effect (which I rejected) that it is recruited to support another aspect of the [±voice] difference. Do the pharyngeal gestures undermine a model that denies undergeneralization? No. The ad hoc nature of pharyngeal cavity articulators was on a speaker-by-speaker level, not a phoneme-by-phoneme level. That is, if there were secondary features here they would be the specifications of different particular muscles for different phonemes. But that is not what was seen. There could still be, in principle, a fully generalized specification across phonemes in the [voice] paradigm regarding the maintenance of the glottal signal — even if that specification may vary slightly from speaker to speaker in terms of the muscles used.

Final Thoughts

At this point it is hard to say more about the mental representations of the phonetic aspects of features without adopting a fuller picture of the phonetics-phonology interface and the phonetic subsystem of the grammar (such as that provided by Articulatory Phonology). Any model that accepts an at least partially compositional view of phonetics must deal with the point of this chapter, namely that features must specify their phonetic implementation, which may be along numerous dimensions. But that is not the end of this chapter, since every model should make predictions.

The crucial prediction of this model is the following: so goes [voice], so goes its correlates. As opposed to a redundant feature model, this model predicts speakers will carry their correlates of [voice] to new phonemes, such as in L2 learning. Whether this actually occurs is not known since testing this is not readily possible without a rich model of L2 acquisition and phonological transfer. Flege and Hillenbrand (1986) and previous work by Flege tested native English, French, Swedish, Finnish, and Arabic speakers' perceptions of the English /s/-/z/ contrast by altering vowel and consonant duration along a continuum. This is the sort of experiment that might bear on the question. (Arabic [voice] does not involve a duration difference of either the preceding vowel or the consonant; Swedish and Finnish lack a /z/ phoneme; and French is similar to English.) The case of Swedish and Finnish speakers is the most relevant since they are learning a new phoneme, /z/. Would they apply their (L1) knowledge of [voice] to the new phoneme? First, what does [voice] look like in their native language? As indicated in Table 2.2, Swedish probably does not have a PVD difference, or not one nearly as large as in English. Nevertheless, even the Swedish speakers inexperienced in English made use of vowel length in making perception judgments. This is surprising at first, but is easily explained either in terms of these speakers reinterpreting the vowel length difference as the phonemic length contrast found in their native language (as suggested in the paper), or alternatively that they are simply very good learners. Neither Swedish nor Finnish nor inexperienced Arabic speakers showed sensitivity to frication duration. Since Arabic does not make use of a frication duration cue in any case, that result is not relevant here, and I am not aware of whether that correlate is used in either Swedish or Finnish (though I suspect not if they have a phonemic length contrast on consonants). What we can take away from Flege and Hillenbrand (1986) is that answering questions of this sort is not easy and requires choosing the right languages to control for many factors.

Finally there is the question of how this model relates to first language acquisition. In a model such as this, the acquisition of [voice] is not just a matter of mapping phonemes to phones but acquiring the generalization of the acoustic properties across phones. The full acquisition of [voice] requires solving several subproblems: storing the correct feature values in lexical entries, determining which phonetic dimensions are relevant to each value of the feature, setting the continuous-valued parameters, and perhaps mediating that by segmental context. Acquisition might follow a number of stages: encoding phonemes with a large number of narrow features, generalizing (and possibly overgeneralizing) the co-occurrence of narrow features by replacing them with new algebraic features, and finally learning the idiosyncratic differences that do exist between phonemes. Just as in Chapter 3, the direction forward is to start looking inside [voice].

Chapter 5

Experiments 1–2: Adult Speech

What infants learn is a function of their environment. Two experiments are reported in this chapter which fill in gaps in our knowledge of the preceding vowel duration (PVD) effect in adult speech.

Although the PVD effect has been reported widely in studies of English, no comprehensive investigation into dialectal variation of the phenomenon had previously been reported. Dialectal variation is an important factor for an infant study so that we can be reasonably sure the infants being studied are learning a language (i.e. dialect) in which the PVD effect in fact occurs. The first experiment in this chapter was a corpus study of vowel duration in the Atlas of North American English (Labov, Ash, and Boberg 2006). The study found that the PVD effect appeared in every dialect region of North America covered by the atlas. Though the magnitude of the PVD effect varied from dialect region to dialect region, the variation was not so great as to put into question that all infants exposed to one of these dialects will be in the environment of the PVD effect.

The second experiment described in this chapter was a production study of adult speech. Several important questions remained about the PVD effect that had not been directly addressed in the literature. The study was carried out to establish the effect of syllable structure and flapping on the magnitude of the PVD effect, the effect on F_1 of the preceding vowel, and to test directly whether the PVD effect is a productive process by using nonsense words.

5.1 Experiment 1: English Dialectal Variation

The following was conducted through joint work with Keelan Evanini and was first reported in Tauberer and Evanini (2009).

The Atlas of North American English (Labov et al. 2006), henceforth ANAE, was used to measure any dialectal variation in the PVD effect across the dialects of the atlas. Wide variation in the PVD effect has been found across languages, and while the PVD effect had already been investigated specifically in several dialects of American English, Evanini and I reported the first comprehensive survey across North American English.

5.1.1 Previous work

Two past studies have investigated dialectal variation of PVD within American English.

Veatch (1991) conducted what we ought to consider a preliminary study using only a handful of speakers spread across four dialects of American English: speech in Alabama, Chicago white speech, Jamaican Creole, and Los Angeles Chicano speech. His results were fairly inconclusive. For Alabama, Chicago white, and Jamaican Creole, one speaker from each dialect showed a weak but significant duration difference between vowels before voiced and voiceless consonants (ratios ranged from 0.93–1.16). The Los Angeles Chicano speaker had a statistically insignificant difference. This of course does not provide any support that Los Angeles Chicano speech lacked the PVD effect, but it left open the possibility.

Jacewicz, Fox, and Salmmons (2007) collected a more rigorous corpus from 18 speakers in each of Madison, Wisconsin; Columbus, Ohio; and western North Carolina. Each dialect showed a duration difference by voicing, but an interaction by dialect was found indicating that the PVD effect was not the same in all dialects. However, other factors such as overall differences in mean vowel durations by dialect would impact the PVD effect and might indicate, statistically, an interaction by voicing and dialect where the effect is actually unrelated to the PVD effect.

While not a comparison across dialects, Baran and Seymour (1976) measured the PVD effect in 5-year-old speakers of African American Vernacular English (AAVE). Baran and Seymour thought AAVE an interesting dialect to study in this case because of its supposed devoicing of final consonants, although the devoicing of final consonants is by no means unique to AAVE among English dialects (e.g. devoicing of fricatives is common generally, Smith 1997). Their data, collected from six minimal pairs spoken by 20 children, showed a PVD ratio of 1.4. This result contributes to a picture that the PVD effect is universal in dialects of American English, but it also is an interesting case for showing PVD dissociated from other aspects of [voice] including phonetic voicing itself.[1]

5.1.2 Methodology

Description of the ANAE corpus

The ANAE was collected by Labov et al. (2006) in an attempt to provide a comprehensive view of the current sound changes in progress in all of the major dialect regions of North America. Previous corpora that contained dialect variation were not adequate for dialectological purposes since they did not control well for the geographic background of the speakers and did not provide a broad sample of cities and subregions from within the major dialect regions.

The ANAE interviews were conducted over the telephone so that speakers from all regions could be accessed efficiently. The sampling methods ensured that the corpus contains more accurate and more fine-grained dialect information than existing corpora: at least two speakers were selected randomly from every city in North America with at least 50,000 inhabitants, and only speakers who had lived their entire lives in that city were chosen. In total, 762 speakers were interviewed for the ANAE. Of these, a subset of 439 were selected for detailed acoustic analysis by the ANAE authors. Interviews were a mix of spontaneous speech, minimal pair tests, and other elicitation methods, though the speech style was not consistently recorded and so could not be controlled for below.

Table 5.1 provides the dialect region affiliation of the speakers in the portion of the corpus used in this study. In addition to dialect region, we also investigated phenomena at the more specific dialect level (as defined in Labov et al. 2006) and report results for the Boston and Maine speakers, both part of the Eastern New England dialect region.

[1] Baran and Seymour did not report anything about the acoustics of the supposedly devoiced segments, leaving open the question of whether they were actually devoiced and whether the devoicing differs from the speech of 5-year-olds in other dialects, e.g. whether this is a phonetic or phonological process.

Dialect Region	Speakers
North	124
South (Region)	76
Midland	63
West	41
Canada	28
Western PA	13
Mid-Atlantic	12
Eastern New England (ENE)	10
Southeast	10
New York City (NYC)	5
Total	*382*
Dialect	Speakers
Boston	5
Maine	2
Inland North	61
Pittsburgh	6
South (Dialect)	58

Table 5.1: Counts of speakers used in this study by dialect region or dialect in the ANAE corpus.

In total, ca. 300 vowels (always bearing primary lexical and phrasal stress) were analyzed in the ANAE corpus for each of the 439 speakers, yielding 133,723 tokens. We omitted from analysis the 26 speakers labeled as coming from dialect "T" in the ANAE corpus, since these speakers do not form a coherent dialect region. Also, a small portion of the audio files had to be excluded due to inadequacies of the corpus from a speech processing perspective. We also only considered word-final syllables, as the position within a word has a large effect on duration. In total, the results presented below are based on an analysis of 34,439 tokens from 382 speakers.

Duration measurement and normalization

In order to obtain duration measurements for the vowels in each of the tokens, the corpus was processed with the forced alignment system described in Yuan and Liberman (2008). The system is based on monophone hidden Markov models with 32 gaussian mixture components on 39 PLP coefficients trained on 25.5 hours of speech from the SCOTUS corpus.

We normalized durations not by speaker as is usually done but in order to minimize potentially confounding contextual effects not of interest to our study. A correlation in the corpus between vowel quality and the voicing of the following consonant, for example, would undermine the interpretation of the results; factoring out variables not of interest also reduces the variation in the remaining data and increased our ability to make statistical inferences. We normalized durations by fitting a linear model to the log-duration data and then subtracting from each duration measurement the predicted components due to the unwanted factors. The model contained vowel identity and post-vocalic place and manner of articulation as predictors.

The log-duration model we chose treated each factor as having a multiplicative effect on duration, rather than an effect in absolute terms (i.e. seconds). It is somewhere between a simple linear model and the more complex model proposed by Klatt (1973) based on the notion of "incompressibility" — that each successive shortening effect on a vowel has a diminished effect because of physical bounds on the speed of articulation. Because the duration phenomena do not combine precisely multiplicatively (see Section 2.4), some confounding correlations no doubt remain after this process.

5.1.3 Results

Vowel duration by region

Table 5.2 reports mean duration in word-final syllables by dialect region. The region with the shortest vowels was New York City at 133 ms, while the South and Southeast regions had the longest mean durations at 156 ms and 159 ms, respectively. The differences between the South and the two regions ranked directly below it, the Midland and the West, were not significant, but to the next region, Western PA, the difference was significant (Tukey post-hoc $p < .01$).

Region	Duration	Region	Duration
NYC	133	Western PA	150
ENE	140	West	153
Canada	142	Midland	154
Mid-Atlantic	146	South	156
North	149	Southeast	159

Table 5.2: Mean durations (ms) of vowels in word-final syllables by dialect region.

At first glance, Table 5.2 might seem to underlie the commonly held perception that Southerners speak with a slower overall speaking rate than other regions. However, we found no such regional difference in a large corpus of spontaneous speech containing regional variation (Cieri, Miller, and Walker 2004): the mean speaking rates for 445 Northern and 1,421 Southern speakers from the Fisher corpus are both 193 words per minute (it is impossible to calculate speaking rate in the ANAE corpus, since the interviews were not transcribed). Speaking rate, then, does not explain the vowel duration difference observed. We gave an alternate explanation in Tauberer and Evanini (2009).

Vowel Length Effect

Table 5.3 reports the vowel duration ratio (mean duration before voiced obstruents to that before voiceless obstruents) in each of the dialect regions for vowels preceding a stop, fricative, or affricate. The dialect regions show similar duration ratios in the range of 1.15–1.26. A duration ratio around 1.2 is what would be expected given the nature and mix of the speech tasks in the corpus. Durations were normalized as described above. The standard errors suggest that there are dialect differences, such as between the South (ratio 1.18, s.e. 0.02) and the Southeast (ratio 1.24, s.e. 0.02) and Western Pennsylvania (ratio 1.26, s.e. 0.04).

The only outliers among the dialects were the Boston and Maine dialects, with ratios of 1.34 (s.e. 0.04) and 1.00 (s.e. 0.10), respectively. (These dialects also had a very small number of tokens applicable for analysis in this section — 356 and 175 — from just six and two speakers, respectively.) The Maine dialect's duration ratio at 1.00 is considerably less than what has been found in any comparable study of English, but a regression analysis of normalized log-durations showed the interaction between voicing and membership in the Maine dialect to be nonsignificant. The Boston dialect differed significantly from the rest ($p < .02$).

Region	Ratio	(SE)	Region	Ratio	(SE)
NYC	1.15	(.06)	North	1.23	(.02)
South	1.18	(.02)	ENE	1.24	(.06)
Canada	1.18	(.03)	Southeast	1.24	(.02)
West	1.20	(.03)	Mid-Atlantic	1.23	(.06)
Midland	1.20	(.03)	Western PA	1.26	(.04)
Dialect	Ratio	(SE)	Dialect	Ratio	(SE)
Maine	1.00	(.10)	Boston	1.34	(.04)

Table 5.3: Post-vocalic voicing duration effect as a ratio (pre-voiced vowel duration to pre-voiceless duration) for the dialect regions and two dialects, and standard errors of the mean.

The values reported in Table 5.3 are the means of the PVD ratios computed for each speaker. This differs from what was reported in Tauberer and Evanini (2009) which was the speaker-pooled PVD ratio. The choice here to compute means by speaker first allowed for standard errors to be computed.

Some of the dialect differences can perhaps be attributable to overall vowel duration differences. The dialect region with the shortest vowels, NYC, also had the smallest voiced–unvoiced ratio, as expected based on incompressibility (Klatt 1973). On the other hand, the difference between the South and Southeast cannot be explained based on their overall vowel durations because their mean durations were similar.

In summary, the PVD effect appears across dialects of American English, with some dialectal variation in magnitude and with outliers warranting further study.

5.2 Experiment 2: Adult Lab Speech

This section describes a production experiment carried out to establish the effect of syllable structure and flapping on the magnitude of the PVD effect and the effect on F_1 of the preceding vowel. It also addresses two other aspects of the PVD effect in light of the future research direction in the acquisition of the PVD effect: the effect of speaking rate and whether the PVD effect is a productive process.

The common wisdom regarding the PVD effect and flapping had been that flapping induces incomplete neutralization. This was supported by Fox and Terbeek (1977) and Huff (1980), as discussed in Section 2.4.2. However, while both studies reported PVD differences (much smaller than what is usually found in English), neither study included a control group. Two problems arose in interpreting the results. First, because the PVD effect is highly variable depending on speaking task and phonological context, without a comparison to same-task data without a flap it was premature to make a strong statement that flapping in fact decreases the PVD effect. Second, because flapping occurs in a particular stress context, it had not been shown that the incomplete neutralization (descriptively speaking) has to do with the flapping or the context. If might have been that the PVD effect was reduced because the post-vocalic consonant was not word-final (Klatt 1973) or because it syllabified with the vowel to its right, and not because it was a /t,d/ turned into a flap. An effect of syllable structure would be quite reasonable because in syllable-initial position other [voice] cues such as aspiration become available, lessening the communicative load of PVD. I answer these questions below.

Secondarily, preliminary evidence from Huff (1980) suggested that the effect on preceding vowel F_1 due to [voice] persisted despite a flapped /t,d/. Normally F_1 of a monophthong is lower preceding a voiced consonant on the order of 10–100 Hz (see Section 2.6). If it were true that flapping does not neutralize this effect, it would indicate that the effect on F_1 was phonological rather than phonetic. Additionally, it would show that speakers had knowledge of the underlying form, eliminating the lack-of-knowledge hypothesis for the incomplete neutralization of the PVD effect discussed earlier (see page 16). If the F_1 difference is neutralized by flapping, then one of the two hypotheses is likely wrong.

The high variability of speaking rate in the developing child has prevented measures of the PVD difference in children from being properly compared with other measures at different ages. Because of the importance of understanding the interaction of speaking rate and the PVD effect, I sought to replicate some of the work of Port (1981), who showed that the PVD effect is larger in slow 'tempo' speech than in fast 'tempo' speech. I initially intended to use the results of manipulating this variable to understand whether changes in vowel durations in children are a result of changing speaking rate or the development of the PVD effect.

Taking the tentative assumption that the PVD effect is not based in physiology, the last aspect of this experiment was to measure PVD ratios in minimal pairs of nonsense words. Though it was fully expected that the PVD difference would show up *productively*, an effect of processing an unfamiliar word may nevertheless affect the magnitude of the PVD effect. Having the results of this task at hand is an important cornerstone before moving on to infant speech in Chapter 6.

5.2.1 Methods

Five university students, all native speakers of English with no prior knowledge of the purpose of the task, participated in the experiment, three associated with the linguistics program at the University. The participants were compensated $15 for their time. The participants were asked to read a list of sentences provided to them, which they had not seen before. Recordings were made in the University's linguistic department's Phonetics Lab's sound booth at 44,000 Hz.

Each sentence was a frame sentence of the form "Say ____ for me." containing a target word. Target words contained a VC sequence either in a monosyllabic word, in the first syllable of a trochaic disyllabic word (the 'tautosyllabic' condition), or crossing a syllable boundary in a trochaic disyllabic word (the 'heterosyllabic' condition). The words came in minimal or near-minimal pairs, differing in the voicing of C and potentially in the segmental content after C (only when no minimal pair could be found). They were drawn from a list of 163 real English and novel nonsense words listed in Appendix A. Vowels were mixed and included both monophthongs and diphthongs. C ranged over the six stop consonants /p,b,t,d,k,g/. It was expected that flapping would occur for /t,d/ in the heterosyllabic target words and nowhere else. We will call these 'flap target words'. Some words given to the participants were discarded due to experimenter error. Sample target words were:

thought, thawed (monosyllabic); crapshoot, crabmeat (tautosyllabic); seater, cedar (heterosyllabic)

nonsense words: chack, chag (monosyllabic); geetmonk, geedmonk (tautosyllabic); nuckist, nugist (heterosyllabic)

The list included words with the short-a vowel which is known to have allophonic or idiosyncratic variation depending on [voice] (e.g. Boberg and Strassel 2000) in some dialects. Several of the five participants reported growing up in a city which is part of a dialect region in which short-a tensing occurs, but all participants may very well have had some version of this effect.

The reading task was divided into five blocks. In each block, subjects were given each of the sentences to read three times, presented in a random order (although subject confusion resulting from double-sided instructions resulted in some words being read fewer or more times). Each block of the task presented the subjects with either the real words or the nonsense words and participants were instructed to speak in either a normal, slow, or fast speaking rate. Prior to the recording subjects were informally told what a fast or slow speaking rate might sound like, with spoken examples from the experimenter (me; but not using any of the target words). The sequence of experimental blocks is given in Figure 5.1. Note that the nonsense words were only recorded at the normal speaking rate, and the second iteration of nonsense words in block 4 was a filler to allow the participants to find their normal speaking rate again between switching from slow to fast. The recordings of block 4 were discarded. Excluding block 4, an average of 967 tokens were recorded per participant. Average recording time was approximately 45 minutes per speaker.

Block	Word List	Speaking Rate
1.	Real	Normal
2.	Fake	Normal
3.	Real	Slow
4.	Fake	Normal (filler)
5.	Real	Fast

Figure 5.1: Experimental design for Experiment A.

Recordings were passed through the Phonetics Lab's forced alignment toolkit (Yuan and Liberman 2008). The forced aligner time-aligns phone boundaries according to an orthographic transcript and pronunciation dictionary. The system is based on monophone hidden Markov models with 32 gaussian mixture components on 39 PLP coefficients trained on 25.5 hours of speech from the SCOTUS corpus. It is based on the HTK software toolkit.

As the participants were asked to read from a list, a transcript was almost readily available — almost, owing to the fact that most participants strayed from the instructions somewhere. Phone boundaries for the target vowels were then manually corrected based on spectrograms, with vowels located by clear formant structure and a fast intensity rise and decline at the start and end. Resulting from poor design choices by the experimenter, several target words contained sonorant consonants preceding the target vowels making it difficult (if not impossible) to determine the boundary between the preceding consonant and the vowel. In these cases, the boundary was often left where the forced aligner located it.

First formant frequency of the preceding vowel was also measured automatically using Praat (Boersma 2001), with a time step of 5 ms and other standard options. The maximum value was found for the measurements within the bounds of the vowel. No hand correction was performed.

5.2.2 Results

Vowel Duration

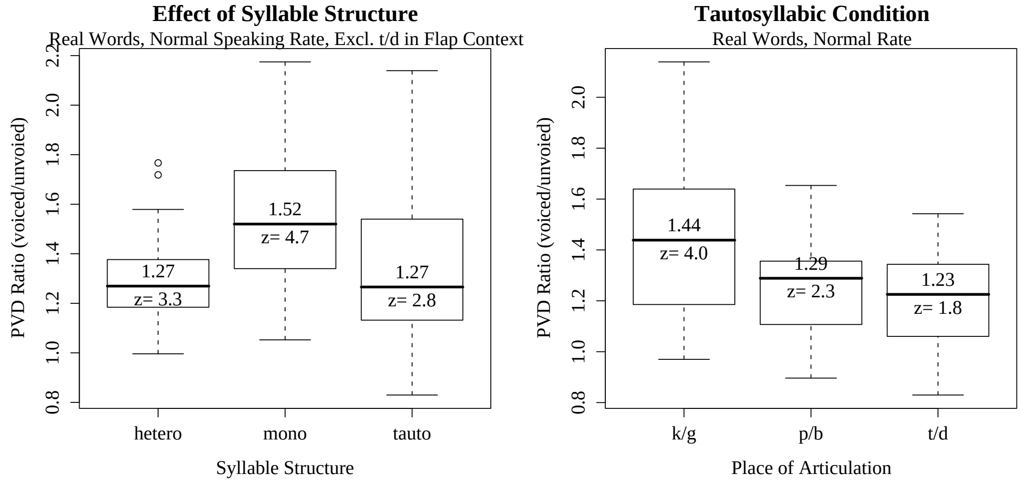
For each speaker and separately for each block, the mean duration of each vowel was computed over the roughly three utterances of each word. A PVD ratio was then computed for each minimal pair (still separately by speaker and block) by taking the ratio of the two previously computed mean durations (the mean vowel duration before voiced stops divided by the mean vowel duration before voiceless stops). Only minimal pairs for which the speaker spoke each half of the pair at least twice were included.

Considering the real, monosyllabic words at a normal speaking rate, the median PVD ratio ranged among the speakers from 1.25 to 1.70 (a one-way ANOVA indicated a highly significant main effect for speaker, $p < .001$). This is likely a result of variability in speaking rate and flap production. While the ratio indicates the magnitude of the difference on the whole, it does not indicate the separability of the two groups. High variance in vowel duration could mean that the distributions of durations greatly overlap and so a listener could not reliably determine the category from the token, even if the means are far apart. To measure the separability of the distributions, the number of standard deviations between the mean voiced and mean unvoiced durations were computed. This is called the separability z-score below. (Because variance was seen to increase with duration, and the unvoiced and voiced categories differ in duration, durations were considered on a log scale so that the unvoiced and voiced categories would have a similar variance. See Rosen 2005.) As with the ratio computation, a separability z-score was computed once for each speaker, block, minimal pair triplet. Mean separability z-scores ranged from 2.4 to 7.0 standard deviations.

The first question was whether the apparent reduction in PVD ratio seen for flapping (Fox and Terbeek 1977; Huff 1980) was in fact true, and whether it was due to flapping or a syllable boundary separating the vowel from the post-vocalic consonant. Indeed, in flap target words the PVD ratio is smaller. The mean PVD ratio for real disyllable words at a normal speaking rate, excluding the flap target words, was 1.29 (or a mean absolute duration difference of 25 ms). For the flap target words of these, the mean ratio was 1.07 (a mean absolute duration difference of 9.6 ms). Though the ratio was smaller, it still represented a statistically significant difference between vowels preceding voiced or voiceless consonants (paired t-test, $p < .02$). However, at least some and perhaps many of the [t,d] segments were not realized as flaps — this was not checked comprehensively. We cannot be sure that the reason the neutralization appears incomplete here was not due to the fact that the rule of flapping was not uniformly applied by the speakers in the first place.

This PVD difference in flap target words was not due to the syllable boundary between the vowel and consonant. Excluding flap target words, the PVD ratios in the tautosyllabic and heterosyllabic groups (real words at normal speaking rate) were virtually the same, 1.23–1.24. Nor was the effect due only to place of articulation: in the tautosyllabic group there was no significant main effect of place of articulation. (See Figure 5.2.) Figure 5.3 shows the PVD ratios by place of articulation in the heterosyllabic group only, showing the /t,d/ pairs to be different. (The difference was highly significant, $p < .001$; Speaker 1 was excluded as she was impressionistically deemed to not consistently flap.)

The second question was what the effect of speaking rate is on the PVD effect. Recall that speakers were asked to speak at normal, slow, and fast speeds, and given brief examples, but their actual tempo was otherwise not enforced by the experimental design. Whether the speakers adjusted their tempo according to the instructions was determined by measuring mean vowel duration (the



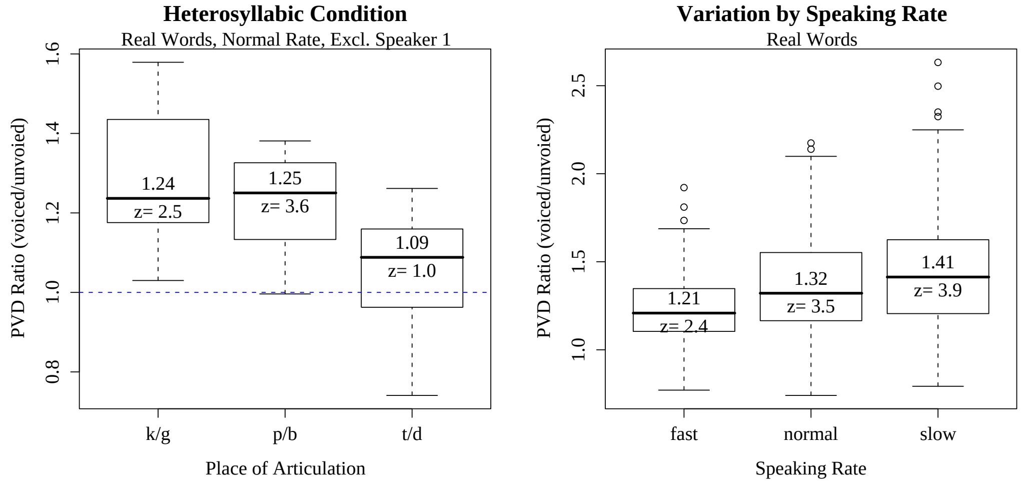
Effect of Syllable Structure

Real Words, Normal Speaking Rate, Excl. t/d in Flap Context

Tautosyllabic Condition

Real Words, Normal Rate

Figure 5.2: PVD ratios. Data points summarized by the boxes are the voiced/unvoiced ratio for each minimal pair separate by speaker. Left: By syllable structure, excluding flap target words in the heterosyllabic group. Right: By place of articulation in the tautosyllabic group. Median values are indicated above the median line in the boxes. Below the median line, the mean separability z-score is reported.



Heterosyllabic Condition

Real Words, Normal Rate, Excl. Speaker 1

Variation by Speaking Rate

Real Words

Figure 5.3: Left: PVD ratio by place of articulation in the heterosyllabic group. Speaker 1 was excluded from this graph since she (impressionistically) failed to flap more often than other participants. Right: PVD ratio by speaking rate.

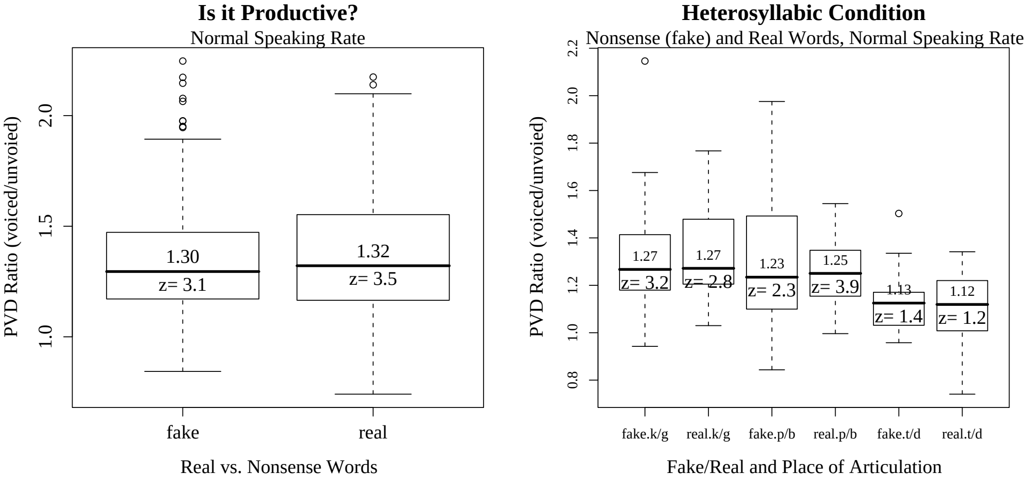


Figure 5.4: PVD ratios: Left: Real and nonsense words at a normal speaking rate. Right: Real and nonsense words in the heterosyllabic group, by place of articulation.

vowels in the relevant VC pairs) and the duration of the frame sentence minus the duration of the vowel. Both measurements show speakers were able to adjust their speaking rate, although the distributions for the three speaking rates were considerably overlapping. The PVD ratio also varied by speaking rate in the expected manner, from 1.19 in the fast condition to 1.35 in the slow condition (Figure 5.3). The difference was highly significant (one-way ANOVA, $p < .0001$).

The last question was whether the phenomenon is productive in nonsense words. It appears to be so. The PVD ratio was nearly identical for real and nonsense words (Figure 5.4). (Participants were free to choose their own pronunciations for the nonsense words and frequently chose different vowels for different halves of a minimal pair. This leaves open the possibility that phonotactic regularities may make it appear that a PVD difference exists, if participants tended to choose vowels with a longer inherent duration in pre-voiced-stop context. But based on the survey of adult infant-directed speech, in Section 3.3, this is unlikely to be the case.)

Vowel Quality

Two analyses were made regarding vowel quality. In the first comparison of F_1 between the voiced and voiceless halves of the minimal pairs, only real words in the heterosyllabic condition with monophthong vowels were considered. This is the condition in which pre-voice F_1 is predicted to be lower. Outside of a flapping context (i.e. place of articulation was velar or labial: backing/bagging, chucking/chugging, flocking/flogging, sopping/sobbing, seaport/seabed), F_1 was on average 68 Hz lower before the voiced consonant (N=73, $p < 0.0002$). (The fake words showed an even larger difference.) In a flapping context (i.e. alveolar place of articulation: seater/cedar, catty/caddy, petal/pedal), the difference was not significant for the real words (N=43), fake words, or all words combined. See Figure 5.5.

The second analysis was of the diphthong /aɪ/ which undergoes Canadian Raising in the voiceless context (the lowering of F_1, see Section Section 2.6). The question here was whether Canadian Raising was neutralized by flapping. No real-word tokens were included with the diphthong /aɪ/

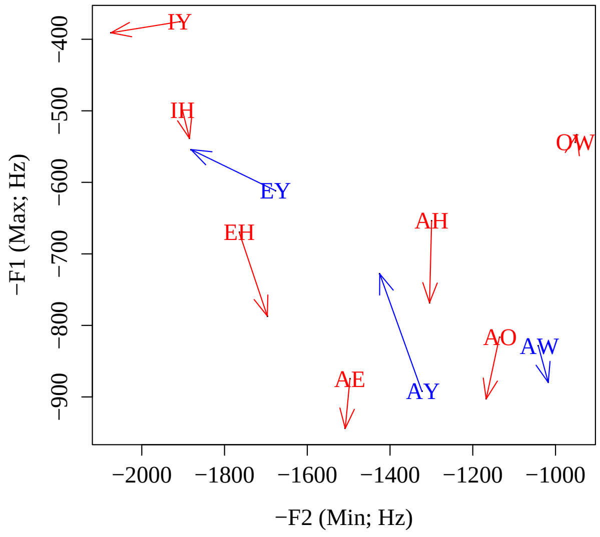


Figure 5.5: Effect of post-vocalic voicing on vowel quality. Vowel names are located at the mean max-F_1/min-F_2 coordinate for tokens before voiced consonants. Arrows indicate the corresponding vowel quality before voiceless consonants. (Arrow lengths are computed as the mean of the within-speaker, within-minimal-pair differences, rather than pointing to the mean vowel target.) Monophthongs are in red, diphthongs in blue. As opposed to the results reported in the text in this section, all collected tokens are included in this graph except those in the flapping context.

followed by an alveolar stop, but the made-up words "jiteing/jiding" were included. This pair was intended to be spoken with the /aɪ/ diphthong, but since participants were only given the words in orthographic form they often but not always used the intended vowel. Participants produced tokens following the pattern of Canadian Raising, with F_1 141 Hz lower, on average, for the voiceless halves of this minimal pairs (N=5, $p < 0.007$).

5.2.3 Discussion

Vowel Duration

To summarize the results so far, past work has been replicated with a proper control condition. The PVD ratio was previously found by Klatt (1973) to be greater in monosyllables (1.5) than in the first, stressed syllable in disyllabic words (1.3). Similar ratios were found here: 1.41 and 1.24, respectively. Fox and Terbeek (1977) and Huff (1980) previously reported small magnitudes for flaps. No significant effect was found for adding a syllable boundary between the vowel and consonant (tauto- vs. heterosyllabic conditions) or varying place of articulation alone, but the interaction to produce flapping did reduce the PVD difference. An effect of speaking rate was replicated, with increased speaking rate correlated with decreased PVD ratio. Finally, the PVD effect appears to be productive, assuming it is not physiologically based.

With regard to flapping, the experiment here as well as that of Fox and Terbeek (1977) crucially relied on the assumption that speakers have the underlying forms for the stops that we think they do. For the flapped pairs here including ‘petal’/‘pedal’, it is not a trivial assumption that the speakers have an underlying /t/ and /d/ when morphological alternations that reveal the underlying form of these words are exceedingly rare. (On the other hand, the participants in this study were reading from a printed word list, which provides an orthographic hint.) Without a distinction in underlying form, no vowel duration or F_1 difference would be expected — and while some vowel duration difference persisted in the flap target words, no difference was found here for F_1. On the other hand, Canadian Raising was observed to persist in a flapped nonsense word, which necessarily has less evidence for its underlying form than a real word. As a result, we must conclude that orthographic information is enough to provide the speaker with an underlying value for [voice] and that the neutralization of the PVD and F_1 effects was due to the neutralization of [voice] in the flapped segment. (This is as opposed to what was noted earlier, that Huff’s (1980) results indicated that F_1 was not neutralized.)[2]

Vowel Quality

Figure 5.5 showed the diverging effect on vowel quality of post-vocalic voicing: 1) the mid and low monophthongs are lowered in voiceless context, 2) the high vowels are lowered to a much smaller extent, and 3) the front- and up-gliding diphthongs are raised in voiceless context. (This leaves out the AW diphthong which stands apart.) Points 1 and 3 are not new findings. As discussed in Section 2.6, Summers (1987) reported that F_1 of /a/ and /ae/ were higher (i.e. the vowel was lowered) in the voiceless context by around 10–20 Hz at the onset of the vowel, 35–45 Hz during its steady-state, and 90–140 Hz at the onset of the obstruent. Moreton (2004) found in the front- and up-gliding diphthongs (/aɪ,ɔɪ,eɪ/) lower F_1 (i.e. the vowel was raised) by roughly 10-25% before a voiced stop. (Similar but smaller effects were found for F_1 at the nuclear target.) This lead Moreton (2004) to propose a “Pre-Voiceless Hyperarticulation” hypothesis, that monophthongs and diphthong offglides are peripheralized before a voiceless stop.

Data on high monophthongs and in-gliding diphthongs had not been studied, leaving out two crucial cases from Moreton’s predictions. In this experiment the data for the high/front vowels /i,ɪ/ incidentally bears on the picture Moreton put forward. The effect of [voice] on F_1 of these vowels is much less. They continue to be lowered before a voiced stop (contra Moreton’s prediction), but they are lowered less than all of the mid and low monophthongs. They are, at least here, clearly not peripheralized in the height dimension as Moreton’s (2004) hyperarticulation hypothesis would predict. This experiment was not actually carried out with the intention of testing Moreton’s hypothesis, unfortunately, or other high vowels would have been included. In-gliding diphthongs are harder to find and more so in a context that precedes contrastive voicing (that is, not before a rhotic as in the in-gliding diphthong of r-less ‘beer’).

[2] If this is not convincing, then a possibly better test for the interaction between [voice] and flapping was that used by Huff (1980), looking at flapping across word boundaries with a word-final [t,d] that flaps before a word-initial unstressed vowel. In these cases, flapping is relatively rare and the underlying form of the flap should be known to the speaker. Huff found a fairly large PVD ratio for vowels before flaps, 1.2, though we do not know what kind of confidence interval to apply and the number of tokens considered by Huff was relatively small. If the data is to be believed, then it would be evidence the other way: that when underlying forms are known to the speaker, then the PVD pattern persists despite [voice] neutralization.

Chapter 6

Experiment 3: Infant Speech

Research on [voice] in infants in the past could be put roughly into two groups. On the one hand, there have been studies of production and perception of the [voice] contrast at the level of swapping one segment for its voice cognate. Mispronunciation tasks have been used in this case. In van der Feest (2007), Dutch-learning 24-month-olds were presented with a word (possibly mispronounced), its picture, and a distractor picture. The words had contrastive voicing in the onset and were mispronounced by changing a voiced segment to a voiceless segment, or vice versa. The infants showed a mispronunciation effect (longer time to shift gaze to the correct picture and more time spent looking at the distractor) when a voiceless-onset word was pronounced voiced (but not when a voiced-onset word was pronounced voiceless; no mispronunciation effect was found, either for voiceless or voiced onset words, for 20-month-olds; in adults, who showed a mispronunciation effect, no voiced-voiceless asymmetry was found). A similar effect was found for English-learning infants aged 14–21 months in Swingley (2009) with minimal pairs such as 'boat' versus 'poat' (mispronounced in onset) versus 'bode' (mispronounced in coda), and 'cup' versus 'gup' (onset) versus 'cub' (coda). Stoel-Gammon's (1985) survey of infants' segmental inventory in production is another example of this line of research. What van der Feest, Swingley, and Stoel-Gammon's work have in common is that the acoustic details of the contrast were not the primary focus.

The other direction of research on [voice] has been interested in particular phonetic dimensions of the feature: VOT, preceding vowel duration (PVD), etc. Infants are known to make a categorical perceptual VOT distinction around 1–4 months (Eimas et al. 1971). In a looking-time task with infants aged 8 and 14 months, Ko et al. (2009) presented the infants with words with digitally altered vowel duration. In the mismatch condition, the vowels preceding voiced codas were shortened by half and vowels preceding voiceless codas were lengthened by 60 percent. The 14-month-olds (but not the 8-month-olds) showed a mispronunciation effect: they looked longer at voiceless codas with long vowels (i.e. the mismatch) than voiceless codas with short vowels. (An asymmetry was found here too: no mispronunciation effect was found for voiced codas.) On the production side of the PVD effect, the reader has already been referred to DiSimoni (1974), Krause (1982), Buder and Stoel-Gammon (2002), and Ko's (2007) corpus study in Chapter 3.

Less common is the investigation into the relationships between multiple acoustic components of a single feature, such as in cue trading. In cue trading, the enhancement of one acoustic correlate of a feature makes up for the absence of another. For instance, Simon and Fourcin (1978) brought cue-trading to infant speech perception. They varied two correlates of [voice], VOT and F_1 onset in the following vowel, and found that the use of F_1 develops between ages four to six years.

This contrasted with the development of a categorical perception of [voice] based on VOT which develops by year four, with the voiced-voiceless boundary shifting from 30 ms at that time to 25 ms at age 11–12 years. Still while some things are known about the components of [voice] in language development, it is mostly on the perceptual end of language use and, further, on word-initial stops.

This chapter presents the results of a new longitudinal corpus study of the production of [voice] by American English-learning infants looking across several acoustic dimensions of the [voice] feature in post-vocalic stops and fricatives. The goal of this study was to examine the individual acoustic correlates of [voice] as a part of the larger whole. Some of the questions to be answered include: Do the individual correlates develop at the same time? Do infants use each correlate independently or are they used in an all-or-nothing pattern? Additionally, I sought in particular to add an additional angle to the question of the developmental trend of the PVD effect, which so far has had conflicting accounts in past work.

While the PVD component of the present study is similar to Ko (2007) in a number of ways (in fact our data comes from the same corpus), the study here differs in several important respects. Ko first drew 364 tokens (136 voiced, 228 voiceless) participating in six minimal or near-minimal pairs such as ‘duck’/‘bug’ (each minimal pair came from one child, with a total of four children examined). Each minimal pair showed a reliable vowel duration difference on its own. In the second part of her paper, Ko selected four words (totaling 343 tokens, controlled for prosodic positions) and for each word computed a correlation between vowel duration and age. No reliable developmental trend was found. In the study presented in this chapter, a much larger and less balanced sample of tokens was drawn but a more complex multi-variate analysis was used to detect the presence of a vowel duration difference and an interaction with age.

A corpus study has its advantages and disadvantages compared to an experimental study. Experiments such as elicitation tasks, especially with infants, are time-consuming to carry out with the result being that only a relatively small number of tokens can be collected. Collecting longitudinal data is even more difficult. A large corpus of longitudinal data simplifies the step of data collection, leaving more time for analysis.

Of course, the inability to control a variable experimentally always brings its challenges. Correlations between [voice] and other variables in the corpus, due to both randomly sampling only a small set of child speech as well as to patterns present in the adult language, must be addressed during data analysis. An abundance of high vowels paired with voiceless consonants may exaggerate the effect of voicing on duration and may hide the effect of voicing on F_1. High vowels are naturally shorter than low vowels, and have lower F_1. This is addressed below by a) including these types of factors in the statistical models, and b) by using a new normalization technique before displaying the data points visually.

Another confound is that some words may be more apt to be said in an excited state than others. “Daddy” tokens were excluded precisely for this reason, but others such as “puppy” and “doggie” remained in the data and raise this question. Prosodic focus is associated with changes in duration and pitch, which again may confound correlations with [voice]. It is impossible in a non-experimental design to account entirely for this possibility. Even accounting for it partially would be very time consuming. One might ask adult listeners to rate the excitability of each token and then to include that in a regression model of duration and other acoustic cues. I did not pursue this route. A confound to the age variable is the changing vocabulary of the children over time. This is a confound if the words have idiosyncratic acoustic properties, either because some words are more excitable or have other peculiar duration, spectral, or other properties of interest. These confounds could not be avoided.

6.1 Hypotheses

The audio recordings of two children from a longitudinal corpus were analyzed for vowel duration, F_1, consonant duration, and closure voicing intensity in VC tokens. Several hypotheses were made.

A vowel duration effect of consonant [voice] has been shown to exist in infants several times. Buder and Stoel-Gammon (2002) and Ko (2007) found a difference with 2-year-old and younger infants. Buder and Stoel-Gammon (2002), however, suggested that the PVD effect is the unmarked option and that infants learning languages without a PVD difference would have to unlearn it, whereas infants learning languages like English will show no developmental change. Reports of a developmental trend of the PVD effect in English-learning infants have been contradictory so far, but contra Buder and Stoel-Gammon the PVD effect seems an unlikely candidate for something stored in a Universal Grammar. Thus a positive developmental trend is predicted.

Furthermore, to substantiate that the PVD effect is an independent linguistic process, a statistical model will be used to determine whether preceding vowel duration is predicted better by the phonological feature or the other phonetic aspects of the utterance. If the PVD effect is merely a by-product of the glottal state, then a positive correlation should be seen between closure voicing intensity and vowel duration, for instance. Or if the PVD effect is a consequence of compensating for changes in consonant duration, then a negative correlation should be seen between consonant duration and vowel duration. The reason for considering the [+voice] and [-voice] tokens separately is that when pooled these correlations are sure to arise, because each of these acoustic dimensions of [voice] is correlated to [voice], and so is correlated with all the rest. On the other hand, it is predicted that the PVD effect is an independent process that will not be correlated with other aspects of [voice].

A main effect of age on first formant frequency is expected: as the vocal tract length of the child increases due to normal development, the formant frequencies, as the resonances of the vocal tract, will decrease. Fitch and Giedd (1999) reported mean vocal tract length of 9.92 cm for children 2–4 years old and 10.54 cm for children 5–6 years old, based on MRI scans. Using the formula for first formant frequency of a neutral vowel $F_1 = c/4L$ where c is the speed of sound (35,000 cm/sec) and L is the length of the vocal tract (Johnson 2003), the expected F_1's at these two age ranges is 882 and 830 Hz respectively, or a decrease of 1.7 Hz per month during the roughly 1.5-year period.

Nothing to my knowledge is known about F_1's role in [voice] in infants. Since F_1's role in [voice] in adults appears to be an automatic, physiological consequence of some other aspect of [voice], most likely the glottal state, a positive correlation is expected between the glottal state and preceding F_1.

As with F_1, little if anything is known about consonant duration in infant speech. Since the PVD effect has been found in infant speech, it is hypothesized that consonant duration in infant speech will also vary according to [voice] as it does in adults.

6.2 Methods

6.2.1 Corpus

The corpus of Demuth et al. (2006) served as the basis of this dissertation. It was used originally, in that paper, to study the acquisition of word-minimality effects and the hypothesis that children would lengthen vowels before deleted coda consonants to retain the bimoraicity of the word. Here I have found a new use for the data to investigate [voice].

The CHILDES database manual describes the corpus as follows:

> The corpus contains longitudinal audio/video recordings of 6 monolingual English-speaking children's language development from 1–3 years during spontaneous interactions with their parents (usually their mothers) at home . . . The total corpus consists of 364 hours of speech.

Of the participants, I selected 'Alex' and 'Lily' for further annotation and analysis. Lily was also a subject in Ko (2007), who used the same corpus. Alex was recorded for one hour every two weeks beginning at age 1;9, the onset of first words, and ending at age 3;5. A total of 16,460 vowels[1] were recorded from Alex. Lily was recorded from age 1;1 to 2;0 every two weeks, from age 2;0-3;0 every week, and from age 3;0-4;0 monthly, for a total of 30,646 vowels.

Of those vowels, only the stress-bearing vowels preceding a coda consonant with distinctive voice (i.e. stops and fricatives but not nasals or liquids) were used.

While the corpus is unique in having longitudinal recordings for several children, it is not known for its sound quality. Despite having a high sampling rate, the acoustic environment was the children's homes and not a speech recording lab. Background noise (hums, toys, telephones, outdoor noises) was common, sometimes drowning out the entire linguistic signal. However, often the noises were inconsequential for the acoustic analysis. Occasionally the child was too far away from the microphone. The recording environments also seemed to have soft echoes (though having no experience with the acoustic analysis of echoes I was not entirely sure), which I attempted to ignore. On rare occasions the adult and child speech overlapped. While this prevented some of the recordings from being used in the acoustic analysis described below, a substantial amount (surely well more than half of the corpus) was usable. Durations, at least, are easily measured even in quite poor signals.

6.2.2 Data preparation

Sample selection

Several measures were desired for analysis. For the (relevant) vowels, duration and spectral measures were desired, as well as phonemic identity based on the adult lexical entry (in other words, what vowel it was). For the consonants, duration and the presence of glottal voicing were desired, and their phonemic identity (place & manner of articulation and voicing feature) based also on the adult lexical entry.

While the corpus already included an orthographic and phonetic transcription of the child speech, as well as time-stamps at speaker changes, time-alignments of the vowels and following consonants

[1] This is a count of syllable nuclei, as opposed to the count of vowels in the corpus's Unicode transcription, which encoded diphthongs as two vowel characters.

of interest were needed in order to determine phone durations and spectral qualities. Likewise, neither the orthographic nor phonetic transcription gave the phonemic detail needed.

Rather than annotating the whole corpus with time alignments at the phone level, a small subset of relevant VC pairs plus their surrounding segments were chosen according to the following criteria:

1. The vowel had primary or secondary stress.
2. The post-vocalic consonant was one of /p, t, k, s, f, b, d, g, z, v/.
3. There was no utterance or word break between the segments. Utterance breaks were annotated manually, generally at the locations where breaths were taken in a pause or at the boundaries of what seemed like prosodic phrases.
4. The word containing the VC pair was not in a stop list constructed to exclude function words that are often short and stressless (despite a dictionary entry that may specify otherwise) or other words commonly produced with unusual prosody. The stop list comprised: at, but, daddy, did, does, get, got is, it, it's not, out, pop, that, that's, this, up, who's, yes, yup.
5. A preceding segment was not /r, l, w/. It is difficult and sometimes impossible to determine a boundary between these segments and a vowel.
6. Tokens were also rejected if the child was singing, whispering, yelling them, had question-intonation, if other segments in the word had a significant speech errors such as deletion, or if the audio quality was too poor.

In order to screen the large corpus for the few thousand relevant tokens, it was necessary to be precise about each of these items and then develop a computer program to implement the checklist. For (2), I am talking about the phonemic identity of the consonants, not features at the level of phonetics. The reason for this is that I am investigating [voice] as a phonological phenomenon, encompassing many different aspects of acoustic phonetics. To classify the consonants according to their phonetic voicing (glottal signal) would miss the point. The children also commonly misarticulated consonants, and there was some transcription error, so I also decided that the identification of the consonant phoneme (e.g. as a /t/ versus a /d/ or /p/) should be based on the canonical phonemic identity as would be given in a pronunciation dictionary. Similarly, for (1), the stress of the vowel, one must look at canonical stress in the adult form of the word to avoid making thousands of impressionistic judgments by how it was realized in the child's speech.

That said, infant speech is particularly prone to error. Rather than rejecting too many tokens, speech errors were generally allowed so long as the segments had any exponence at all. Vowel quality, as uttered, was ignored. Some 25% of the (relevant) vowel tokens had a mismatch in vowel quality between the canonical vowel identity and the identity as phonetically transcribed in the original corpus. Likewise, the phonetic features of the consonant as uttered were largely ignored. There was nearly a 30% error rate on consonants, with errors in place and manner. Some were uttered as sonorants, but where they were still seemingly an attempt at the consonant they were retained for analysis.

All of these requirements necessitated that the canonical phonemic transcription of each child utterance be computed automatically. I implemented a Python program to do this. The approach was to use the Carnegie Mellon University Pronouncing Dictionary (cmudict) to get the possible canonical phonemic transcriptions of each utterance based on the orthographic transcription already present in the corpus. The canonical phonemic transcriptions were used to identify utterances containing relevant tokens.[2]

Because there were still far more tokens than I could go over myself, I prioritized the tokens so they would be maximally useful to the study. The best tokens would be those that occur in near-minimal pairs, so that the sample is as close to balanced as possible. There were a few actual minimal pairs (occurring in the sample were back ∼ bag, eggs ∼ x /eks/, have ∼ half, peas ∼ piece, pick ∼ pig, and the near-hit toast ∼ toes). And although there were ten times as many word types not part of a minimal pair in my final sample, I had prioritized it so that many at least would occur in a near-minimal pair where the vowel and following consonant's place and manner are the same, but the rest of the words may be different. /ɪk,ɪg/ and /ʌb,ʌp/ words, for instance, were chosen over words with less common vowel–consonant bigrams. The most frequent words are included in the results section.

The phonemic identity of each vowel was also identified, according to cmudict.

Tokens were also labeled with whether or not they were "pre-pausal", using a slightly peculiar definition of pre-pausal. While pre-pausal lengthening has been shown at times to be confined to the final syllable before the pause (Wightman, Shattuck-Hunagel, Ostendorf, and Price 1992), the data here showed that vowels separated from the end of an utterance by only unstressed syllables were also lengthened. For the purposes of analysis, "pre-pausal" is defined as the rightmost stressed syllable within an utterance.

Forced-alignment

Thinking ahead to having to manually annotate the start and ends of the relevant vowels and consonants, I performed automated forced alignment on the entire corpus for Alex and Lily. Forced alignment would, at the least, provide a first approximation to the final alignments. The Penn Phonetics Lab Forced Alignment Toolkit (P2FA, Yuan and Liberman 2008) was used. P2FA is a Hidden Markov Model-based tool based on HTK (Young, Evermann, Gales, Hain, Kershaw, Moore, Odell, Ollason, Povey, Valtchev, and Woodland 2006), using multiple gaussian mixture monophone models on 39 PLP coefficients. Forced alignment is usually used by providing an orthographic transcript and a recording, plus a pronunciation dictionary such as cmudict. Essentially, the transcript is converted into a sequence of phonemes from which is constructed the HMM model. Since child speech differs substantially from canonical pronunciations of words, instead of providing an orthographic transcript a phonetic transcription was used directly instead. That is, the transcript was a list of monophones in the acoustic model, and no intermediate mapping from words to phones was needed.[3] In addition, the alignment process was sped up by breaking down the corpus into utterance-sized chunks based on the utterance-level time stamps already present in the corpus. (The average utterance length was around a few words.)

Knowing from preliminary work that P2FA's adult speech acoustic model based on Supreme Court justice recordings did not work very well on infant speech, I opted to train a new model from scratch using Lily's speech. Infant speech has higher fundamental and formant frequencies, a greater range in frequency space, and far more variability in the articulation of frication noise — among perhaps many other differences that a PLP-based forced aligner might be sensitive to.

[2] Although cmudict is called a pronunciation dictionary, dictionary entries are concatenations of phonemes, not phones. Flaps, for instance, are encoded as either T or D, according to the underlying form.

[3] In practice this was accomplished with a trivial map from phones to themselves.

Creating a new acoustic model required converting the Unicode phonetic transcription into a format suitable for HTK, the HMM toolkit behind P2FA. One- and two-character (i.e. diphthong) vowels were converted into ASCII labels based on cmudict's convention (AA, EY, etc.) plus a binary stress annotation based on the transcript. Three steps of reestimation were used, followed by revising the silence model according to the HTK tutorial, a second gaussian mixture component was added, followed by two more steps of reestimation. Because /t/ and /d/ were inconsistently transcribed as flaps (and so the acoustic model included an HMM for flaps), P2FA was allowed to choose either [d] or [ɾ] for a [d] in the transcript by adding a second mapping to the dictionary from [d] to [ɾ]. A forced alignment step was run to take new guesses about whether any [d] was really a [d] or a [ɾ]. Two more steps of reestimation were performed on the revised transcript, followed by adding a third mixture component, followed by two more steps of reestimation. Finally, the forced alignment step was performed to get the time alignments.

The model trained on Lily's speech was used both for Lily and Alex's forced alignment.

I have not quantified the accuracy of the forced alignment, but I can give a general impression. Virtually all boundaries had to be moved at least a little. Most were not further from the correct location by more than a segment or two. But when there was significant background noise in the signal, especially speech of the mother[4], the forced alignment boundaries were often hopelessly far away. A common interesting case is when the mother or child repeated what the other had said within the time bounds of the utterance, which was most likely to (quite legitimately) confuse the aligner.

6.2.3 Acoustic analysis

Segmentation and duration

The vowel and consonant tokens identified above were analyzed on several dimensions. The first dimension of interest was vowel duration. All acoustic measurements involve some operational simplification from the abstract linguistic concepts that are actually the target of research, and measuring vowel duration is problematic because there are often no hard boundaries between a vowel and its surrounding segments. The easiest case is the boundary between a unaspirated stop and a following vowel: it is common to say the vowel starts after the termination of the release burst, which is almost always easily visible on a spectrogram and is very short-lived.

Going from a vowel into stop closure the boundary is much less clear. Multiple articulations are occurring simultaneously and at different rates. The closure itself takes some time to form, during which the formants begin to slowly disappear as they move to their loci for the stop's place of articulation. This along with glottal abduction or adduction depending on the [voice] of the consonant reduces the overall intensity of the speech signal. And the glottal signal can continue into or throughout the stop whether or not the stop is phonologically voiced. Of all of these signals, I used a sharp drop in or disappearance of formant intensity to mark the end of the vowel, where applicable. For Lily, I relied on the spectrogram only — since the spectrogram is based on a sliding window it has a limited time resolution. For Alex I primarily made use of the waveform by looking for the characteristic shapes of the high frequency formant energy (i.e. higher than F_0).

Other types of boundaries were harder. If the stop is realized as a flap (as can happen in intervocalic position, and not just for alveolar stops), it is a continuous gesture with no obvious start or

[4] The corpus does have audio in two channels, for the mother and child, but because they often sit close together there is significant overlap.

end or boundary between the vowel before and the vowel after the flap. In these cases, I was forced to take a best guess.

The boundary between vowels and fricatives (and the occasional stop misarticulated as a fricative) was easier to formulate though not always easier to measure. Here I looked for the start or end of frication, if it was visible at all. Sometimes there were gaps between the end of the formants and the start of frication, and these counted toward the duration of the vowel. Intervocalic fricatives also sometimes appeared more like approximants with a continuous rather than discrete gesture, and best-guess boundaries had to be placed.

The duration measurement for consonants included the total duration of closure, release burst, and aspiration if any for stops, or the duration of frication for fricatives. Durations of consonants were excluded from analysis in a number of context. Their durations in utterance-final position are unreliable because of the variability of release bursts. Because many of the acoustic properties of consonants are affected by the context following the consonant, these properties were only taken in *intervocalic* context which was thought to prevent extraneous effects such as regressive voicing assimilation. Intervocalic context included when the sonorants R, W, L, Y, M, N followed the consonant, besides vowels.

Closure Voicing intensity

The intervocalic stops were also annotated for degree of voicing during closure. There is no obvious acoustical measure for this. The glottal signal during closure can vary along a number of dimensions all relevant to considering whether the use of the glottal pulse is correlated (and how much) with the phonological voice feature. These acoustic measurements include voice termination time, voice onset time (if negative), and intensity. Is a low-intensity glottal pulse throughout closure more voiced than a high-intensity glottal pulse that terminates quickly? Ideally, many acoustic measurements could have been taken. Instead, a qualitative annotation and a quantitative acoustic measurement were made. The qualitative annotation was a choice between no voicing, partial voicing, full voicing, and vocalized, meaning the segment appeared to have no closure. This decision was made by visual inspection of the waveform, spectrogram, and pitch track. Voicing that comes to a quick end was marked as no voicing. A similar measurement was made for fricatives.

A quantitative measurement based on RMS intensity during a stop's closure was also used. It was computed as the RMS intensity of the signal during the first 100 ms or first 85 percent of the duration of the stop, whichever was shorter, divided by the RMS intensity in the last 100 ms of the preceding vowel. By using this ratio the measure is robust to overall changes in speaking intensity. Both the vowel and stop must have been 50 ms long or more, in order to get a reliable measure of glottal signal intensity. Because the intervals for stops were annotated as including burst and aspiration, the final 15 percent of the consonant was excluded in the hopes of removing these signals that would influence the RMS intensity. No corresponding measure was made for fricatives as the intensity of a fricative may be due as much or more to the frication noise as to the voice bar.

First formant frequency

First formant frequency and low-band spectral center of gravity of the monophthongs only were measured using Praat's automated methods (Boersma 2001). Measurements were taken at the mid-point of the vowel and shortly before offset. No substantial differences were found among these four measurements, and so we report results for the first format frequency at the mid-point only.

6.3 Analysis

6.3.1 Duration modeling

In order to make any inferences about the data there must be a model. A simple and plausible model for duration is a multiplicative model, i.e. $D = D_0 \cdot k_1 \cdot k_2 \cdot \ldots$. That is, the duration of a vowel starts with some base or intrinsic duration D_0 (Umeda 1975), and then other multiplicative factors are applied depending on context. If k_1 is the factor for the context of a following voiced consonant, then k_1 approximately equals 1.3 in adult English. If k_2 is the factor for pre-pausal context, it might equal, say, 2.0. The duration of any pre-voiced pre-pausal vowel would be predicted to be $D = D_0 \cdot 1.3 \cdot 2.0 = D_0 \cdot 2.6$. Only with a model can we test whether any of the parameters have interesting values, given the data, by fitting the model to the data.

We don't know what the underlying processes are that govern segment duration, unfortunately, so in choosing a model one is forced to balance simplicity, linguistic plausibility, and goodness of fit. The multiplicative model is simple and linguistically plausible, but it does not fit everything we know about segment duration. Klatt (1973) observed what he called "incompressibility", that phenomena such as the PVD effect have smaller effects, measured as a percent change, in word-internal position than in word-final position (see Section 2.4). The idea of incompressibility is that when a word-internal shortening rule applies first, the PVD effect when viewed as vowel shortening has less room to function because the vowel is approaching a lower bound on duration due to articulatory limits.

In Klatt's model, phenomena such as the PVD effect and pre-pausal lengthening are ordered rules applied successively starting with the intrinsic vowel duration D_0. Each rule has the form $D_i = k_i(D_{i-1} - mD_0) + mD_0$.[5] m is a factor, suggested by Klatt as 0.45, that gives the absolute minimum duration of the vowel when multiplied by its intrinsic duration, D_0. Taking the same factors as above and assuming D_0 = 100 ms and $m = 0.45$, then the duration of a pre-voiced vowel is computed as $D_1 = 1.3(100 - 45) + 45 = 116.5$ ms. If it is also pre-pausal, the duration is computed as $D_2 = 2.0(116.5 - 45) + 45 = 188$ ms. (Note that if m were zero, it would be the same as the multiplicative model.) This model is linguistically plausible and explains the known data better than a multiplicative model, but it is not quite as simple as the multiplicative model (it is harder to estimate the parameters; more on that momentarily). The recursive formulation that Klatt provided can be expanded into a formula which resembles the simple multiplicative model:

$$D = (D_0 - mD_0) \cdot k_1 \cdot k_2 \cdot \ldots + mD_0$$

Van Santen (1994) proposed a generalization of this to sums-of-products models. Such models are, as the name suggests, sums of products of the form $\Sigma_i \Pi_j k_{ij}$. The Klatt model is a particular instance of this that is the sum of two product terms: $(D_0 - mD_0) \cdot k_1 \cdot k_2 \cdot \ldots$ and mD_0. A generalized sums-of-products model allows for many more possible ways that the factors can combine, leading to a much larger number of parameters to estimate. For a model involving the same two factors as above, the multiplicative and Klatt models have just a single form. But the sums-of-products model has five possibilities: $k_1 + k_2$, $k_1 \cdot k_2$, $k_1 + k_1 \cdot k_2$, $k_2 + k_1 \cdot k_2$, and $k_1 + k_2 + k_1 \cdot k_2$. Or this can be thought of as a single model with twice the number of parameters:

$$D = k_1 + k_2 + k_1' \cdot k_2'$$

where k_1 and k_1' are separate additive and multiplicative effects of the PVD effect, and k_2 and k_2' are additive and multiplicative components of pre-pausal lengthening. While additional parameters can

[5] $m \cdot D_0$ is Klatt's $D_{\min}$.

always provide a better fit to data, more data and less noisy data is needed to judge the reliability of the fit (i.e. in terms of a confidence interval) and over-fitting must be avoided. Sums-of-products models are the most complex model of the three discussed so far. They are also less linguistically plausible because of the fast growth in the number of ways phenomena can interact as the number of phenomena increases. But in virtue of having more parameters they will be able to explain more of the variation in any given data set.

Finally, another model of phone duration was proposed (but ultimately rejected) by van Santen and Shih (2000), where phone duration is dependent on a separately computed syllable duration only, and not on its own context. Syllable duration would be computed first based on factors such as the number of phonemes in the syllable, the position of the syllable in the utterance, and stress. The syllable duration is then split across the segments within the syllable taking into account the intrinsic duration of each segment and its inherent elasticity, e.g. that some segments vary in duration more than others, and perhaps its position within the syllable. Although van Santen and Shih were mathematically precise about how to distribute duration within a syllable, they did not offer a formal model of how context determines syllable duration — that would need to be specified before the model could be used in practice.

Although many decades old by now, Klatt's model still provides the best middle-ground when weighing simplicity, plausibility, and goodness of fit.[6] The initial difficulty with applying Klatt's model is that given a corpus of vowel duration measurements and labels, it is not as simple to estimate the model parameters D_0 and k_i as in the simple multiplicative model. The multiplicative model has the advantage that it can be transformed into a linear model by taking the logarithm of both sides, and then the parameters can be estimated using simple linear regression. Assuming all of the labels are binary (voiced/unvoiced, pre-pausal or not), the form of the linearized model is derived as:

$$D = D_0(k_1)^{x_1}(k_2)^{x_2}\ldots \qquad \text{(multiplicative form)}$$

$$log(D) = log(D_0) + log(k_1)x_1 + log(k_2)x_2 \cdot \ldots \qquad \text{(linearized form)}$$

where D and x_i are values from the data, $x_i = 0$ if the context is not met or 1 if the context is met. In the multiplicative form, a value of 0 for x_i means $(k_i)^{x_i}$ will be 1, so essentially the factor has no effect. If x_i is 1, then $(k_i)^{x_i} = k_i$ and the factor is included. The same logic holds after the model is linearized.

The parameters of Klatt's model cannot be estimated using linear regression, but maximum likelihood estimation can be used instead. Under this approach, the goal is to find the parameter settings that are most likely given the data at hand. This is accomplished by writing a likelihood function — given a set of data, how likely is a certain setting of parameters — and then maximizing that likelihood using a generic maximization algorithm. The result is a good fit of the data.

The likelihood function can be made based on an assumption of what distribution the residuals are drawn from (that is, subtracting the values predicted by the model from the observed values). For any setting of the model parameters, we can say the likelihood of those parameters being correct is the probability of seeing the corresponding residuals. If we make the assumption that the residuals are distributed according to a normal distribution (with mean zero and some unknown standard deviation also to be estimated), the likelihood of a parameter setting is the product of the values of the normal distribution density function at the residuals (i.e. the probability of seeing each residual). The likelihood function is then maximized by searching for optimal settings of the parameters using

[6] See van Santen (1994) for a comparison of other models, including one based on decision tree classification. It is beyond the scope of this dissertation to exhaustively justify the choice of model.

any general maximization algorithm. In this case I use the "nlm" function in R. Standard errors for the parameters are produced through this process as well, which allows us to judge the reliability of the estimate. (For the theory behind maximum likelihood estimation, see Fergusun 1996.)

Several existing R functions have been written for maximum likelihood estimation. However, I found them cumbersome for testing many models and examining the output. As a result, I constructed a new R function for performing MLE. The code is given in Appendix B.

The use of maximum likelihood estimation will become clearer in the results section when it is fit to the vowel duration data.

6.3.2 Modeling first formant frequency

As in the case of segment duration, it is also important in the case of formant frequencies to choose an appropriate model for analysis. We have already seen work suggestive of a model of diphthong F_2 in Section 2.6.2: Gay (1968) found that in up-gliding diphthongs of various durations the slope of F_2 remained constant, so that longer diphthongs have higher offsets. That suggests a model in which slope is a parameter. But the effect on F_2 due to [+voice] was different. Instead of a longer glide, a longer steady-state was generally found. Moreton (2004) proposed peripheralization before voiceless stops (lower vowels become lower, higher vowels become higher). One would still have to formalize that mathematically. But it didn't entirely bear out in the data presented in Section 5.2: all of the monophthongs were lowered before voiceless consonants, though to different degrees. At least as the interaction between F_1 and voicing goes, we have several hypotheses but few confirmations from data.

Without any standard models to choose from, a linear model using F_1 frequency in Hertz will be sufficient for testing the effect on F_1 of [voice] and other factors. However, rather than fitting a standard fixed-effects model to the data, mixed (i.e. fixed plus random) effects models will be used instead. The advantage of including random effects is that the effects of the different vowel qualities and place/manners of articulation of the consonant can be treated as being drawn from a normal distribution which improves their estimation by looking across the values of each factor.

6.4 Results

6.4.1 Data overview

A total of 506 VC tokens were initially collected for the child Alex and 595 for the child Lily. Histograms in Figure 6.1 show the distribution of the collected tokens by age (from 14–42 months) and vowel quality. Some aspects of the distributions reflect the underlying nature of infant speech: children speak few coherent words before around age 18 months. Other aspects of the distribution reflect the contents of the corpus (fewer recordings were made after age 3) and the selection of tokens I made here: attempting to flatten out the distribution, and avoiding vowels without appearing in a fair number of both voiced and voiceless tokens.

For some vowels, there was an overwhelming imbalance in that they would be followed predominantly by either a voiced or voiceless consonant. Tokens with these vowels were removed leaving AA, AE, AH, AO, EH, EY, IH, IY, OW[7]. Additionally, because there were so few tokens

[7] cmudict conventions are used because the identity of the vowel in each word was determined by referencing the cmudict pronunciation.

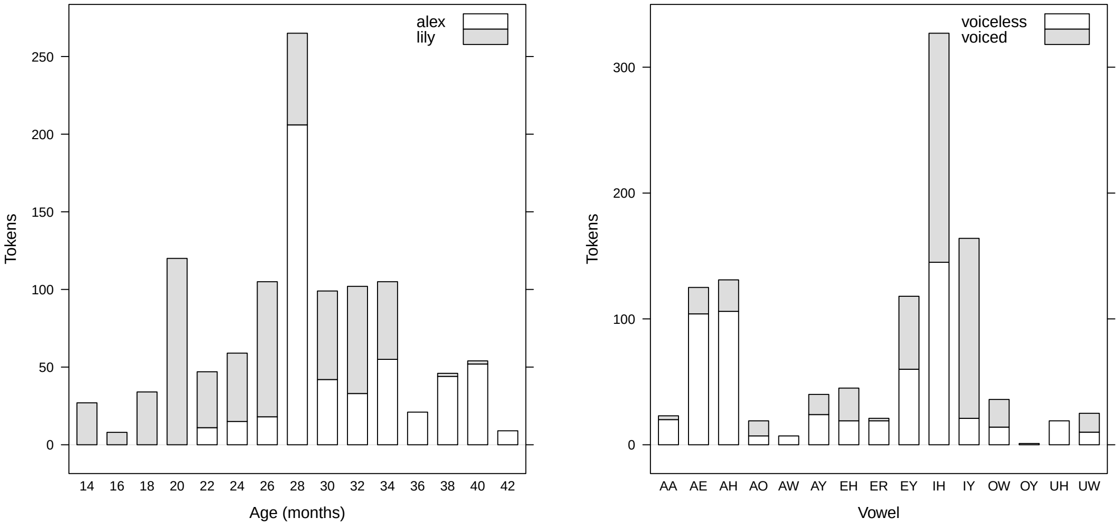


Figure 6.1: Left: Distribution of tokens across the ages in which the children were recorded. Right: Frequency of vowels, split by whether they preceded a voiced or voiceless consonant.

from Lily after age 34 months, these additional tokens were reassigned to age 34 months so they could be usefully binned in the figures. Finally, outliers were removed. This left 443 tokens for Alex and 517 tokens for Lily (960 in all).

161 distinct words (types) were collected in the data and remained after outliers were removed. The most frequent words are listed in Table 6.1. The 15 most frequent words accounted for 54 percent of the corpus. There was no obvious trend that the most exciting (e.g. animate) words were the longest, although this certainly could have been a factor.

Consonants were not evenly distributed across age, and some were rare in the corpus altogether. For Alex, /d,s,f/ were quite uncommon, and /p,b/ were found mostly through only age 28 months. I return to this in section 6.4.4. Although several phonemes were uncommon for both children, these cannot always be attributed to the child's linguistic environment. Figure 6.2 shows the percent difference between the relative frequency of each consonant phoneme in the child's speech compared to the child's parent's speech, in intervocalic context only, based on a count of the tokens throughout the entire corpus for each child. For instance, Lily used labiodental fricatives about .75 percentage points less (actually about a 20 percent difference) than her parent in the corpus. More on this later.

6.4.2 Preceding vowel duration

A number of factors affect the duration of a vowel in adults, including speaking rate, position in the utterance, and post-vocalic voicing. The goal of this section was to determine whether the post-vocalic voicing effect (PVD) can be found in the children in the corpus and, if so, whether there is a developmental trend.

The first thing to do is to plot the data. Figures 6.3–6.4 shows vowel durations by age, post-vocalic voicing, and child. Figure 6.3 shows vowels in pre-pausal position and Figure 6.4 shows vowels in non-pre-pausal position. The figures show a fairly robust effect of voicing across the age

	voiced	frequency	mean.duration
big	*	147	0.190
puppy		63	0.142
he’s	*	45	0.138
baby	*	42	0.202
hat		38	0.139
these	*	37	0.246
six		23	0.167
eight		17	0.254
pick		17	0.117
cheese	*	16	0.254
fix		16	0.119
apple		15	0.126
she’s	*	15	0.175
stick		15	0.161
chicken		14	0.193

Table 6.1: Fifteen most frequent words in the corpus and their mean duration, ordered by frequency. Voiced types are marked with an asterisk for convenience.

ranges. I will return to quantifying the PVD effect and assessing whether there is a developmental pattern shortly. The problem with interpreting Figures 6.3–6.4 for a correlation between the PVD effect and age is the same problem that we try to resolve with regression: when multiple correlated variables are in play, a simple correlation between any two will be confounded by effects of the others. And because this was an observational study with no control over the distribution of tokens, care must be taken to make sure there were no unforeseen systematic patterns. So, for one, age is correlated with several aspects of language development which are in turn correlated with the acoustic correlates of [voice]. A simple plot or a correlation coefficient of just age and vowel duration may be confounded by other effects.

Here is one way this is a problem. The greatest effect on vowel duration was whether the vowel was pre-pausal. Mean utterance length increases with age, and so as age increases vowel duration should decrease because more tokens are not pre-pausal. We also know that the PVD effect has less of an effect in contexts where vowels are shorter. Thus, other things being equal, we expect the observed PVD ratio to decrease with age on account of the change in utterance length, and not because of a change in the PVD effect itself. This confound must be taken into account when examining unbalanced data.

In Alex’s case, multi-word utterances began at around age 2;5 and in Lily’s case around age 2;1. Within six months following the onset of these utterances, tokens were roughly evenly split between pre-pausal and non-pre-pausal. A plot of this is given in Figure 6.5.[8]

To estimate the effect of context on vowel duration, I fit a Klatt model to the vowel duration data using the R function I described in Section 6.3.1. The model here has two parameters: an effect of being pre-pausal (k_1) and an effect of the following consonant being voiced (k_2). The model is

[8] I don’t intend the reader to generalize this. Recall that the distribution of tokens in the data is based not only on the children’s free expression but also on the priorities I used in selecting tokens to analyze.

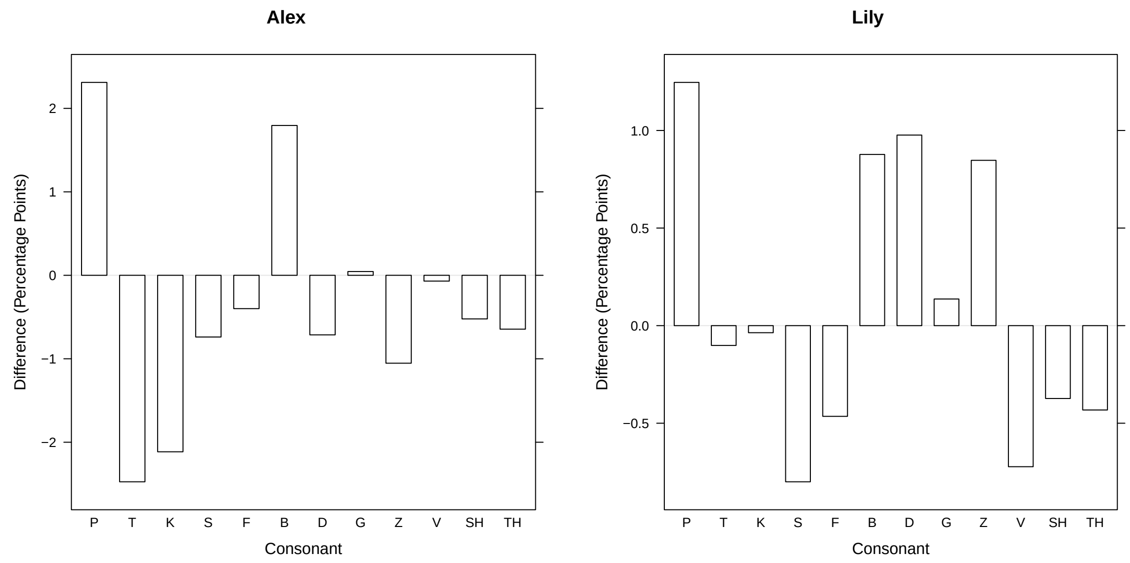


Figure 6.2: Comparison of the relative frequencies of intervocalic consonant phonemes in child speech to the speech of the child's parent, based on the entire corpus (that is, not just the tokens which I annotated). Each bar is the relative frequency of the phoneme (out of the 12 phonemes shown) for the child minus that for the adult. (Alex: N=15644, Alex's parent: N=61228, Lily: N=33622, Lily's parent: N=165506)

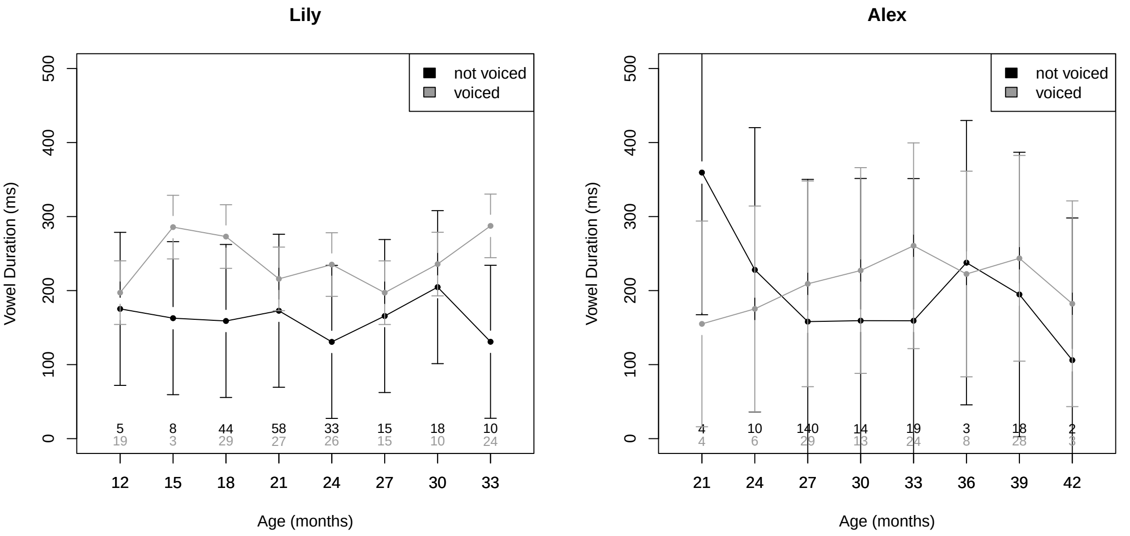


Figure 6.3: Raw vowel duration by age, post-vocalic voicing, and child in pre-pausal position. Here and elsewhere, 95% confidence intervals are shown.

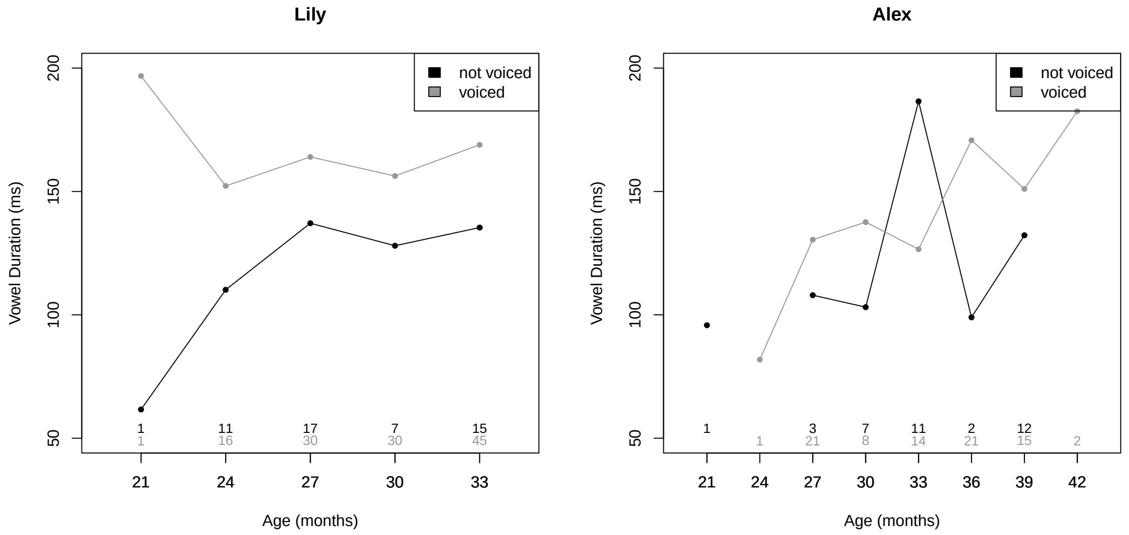


Figure 6.4: Raw vowel duration by age, post-vocalic voicing, and child in non-pre-pausal position.

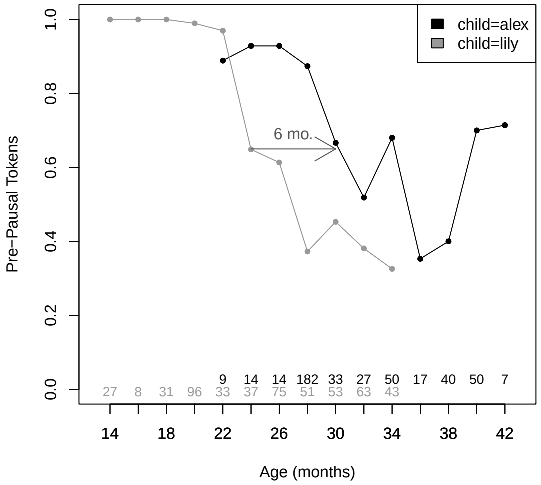


Figure 6.5: Relative frequency of tokens at each age which were not pre-pausal, according to the definition given in the text.

thus $(D_0 - mD_0) \cdot k_1^{x_0} \cdot k_2^{x_1} + mD_0$, where x_0 is 1 if the token is pre-pausal, zero otherwise, and x_1 is 1 if the token is before a voiced consonant, zero otherwise. Estimating Klatt's value for m proved problematic for the solution technique (the resulting standard errors were very high), so I have assumed it to be 0.45 throughout, following Klatt.

The model is fit with the following R command after loading the MLE function given in the appendix:

```
> est <- mle(data, data$vowel_duration, function(data, D0 = 0.1,
+     k1 = 1, k2 = 1) {
+     m = 0.45
+     (D0 - m * D0) * k1^data$prepausal * k2^data$voiced + m *
+         D0
+ }, nullhyp = list(D0 = NA))
```

```
    estimate stderr  null pval
D0  0.121    0.00439 NA   NA
k1  1.70     0.083   1    2.12e-17
k2  1.58     0.0622  1    1.95e-20
σ_ε 0.081    0.00185 NA   NA
r^2  0.157
```

The result of the maximum likelihood estimation shows that the baseline duration of vowels (regardless of quality) is 121ms, that the pre-pausal lengthening factor is approximated at 1.7, and the PVD effect lengthening factor at 1.6. These last two are highly statistically significantly different from 1.0. (See the appendix for an explanation of σ_ϵ.)

The results are comparable to what is found using a multiplicative model, solved with linear regression:

```
Call:
lm(formula = log(vowel_duration) ~ prepausal + voiced, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-1.464867 -0.272620 -0.008802  0.268704  1.187246

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.20934    0.03279  -67.37   <2e-16 ***
prepausalTRUE  0.32664    0.03187   10.25   <2e-16 ***
voicedTRUE     0.31436    0.02930   10.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4361 on 957 degrees of freedom
Multiple R-squared: 0.1528,       Adjusted R-squared: 0.151
F-statistic: 86.31 on 2 and 957 DF,  p-value: < 2.2e-16
```

Note that the coefficients must be transformed out of log-space using the exponential function to interpret them as a ratio: for pre-pausal lengthening this yields 1.39, for the PVD effect 1.37. Factor estimates from a Klatt model cannot be directly compared to the corresponding estimates from a multiplicative model though. Klatt's model takes into account the nonlinear way in which the factors combine. For a comparison, one must compute the ratio predicted by the Klatt model for a single factor holding all other factors constant (for instance by omitting them):

$\frac{(D_0-mD_0)\cdot k+mD_0}{(D_0-mD_0)+mD_0} = m+(1-m)k$

Given $m = 0.45$, the vowel duration ratios predicted by the Klatt model are 1.39 for pre-pausal lengthening and 1.32 for the PVD effect. These are similar to the ratios computed by the linear regression.

There are other confounding variables besides utterance length. The distribution of vowel quality changes over time, and vowel quality influences vowel duration because each vowel has a slightly different intrinsic duration. Place and manner of articulation of the following consonant has been shown to affect vowel duration as well (Crystal and House 1988a): 1) back vowels should be longer before velars and shorter before non-velars, and 2) all vowels should be longer before bilabial stops. Three additional binary parameters with corresponding duration factors are added into this model at this point. The first is a parameter added to the base duration for whether the vowel is a diphthong. This is a first approximation to modeling intrinsic vowel duration completely, which would instead require a separate parameter for each vowel. The second parameter added is whether the following consonant is bilabial. Finally, because the two children may have different base durations, a factor is added to account for the difference.

```
> est <- mle(data, data$vowel_duration, function(data, D0monoph = 0.1,
+     D0diph = 0, D0lily = 0, prepausal = 1, bilabial = 1) {
+     m = 0.45
+     D0 = D0monoph + ifelse(!data$diphthong, 0, D0diph) + ifelse(data$child ==
+         "alex", 0, D0lily)
+     (D0 - m * D0) * prepausal^data$prepausal * bilabial^(data$placemanner ==
+         "p/b") + m * D0
+ }, nullhyp = list(D0monoph = NA))
```

```
          estimate stderr  null pval
D0monoph  0.138    0.00549 NA   NA
D0diph    0.0362   0.00687 0    1.36e-07
D0lily    0.0140   0.00464 0    0.00248
prepausal 1.60     0.0918  1    4.48e-11
bilabial  0.701    0.0526  1    1.37e-08
σ_ε       0.0837   0.00191 NA   NA
r^2  0.0988
```

The results show a strong effect in the expected direction for diphthongs (lengthened by 0.036 ms). It also shows a very strong effect of preceding a bilabial stop (shorter by a factor of 0.7), but in the direction opposite of what has been reported in the literature — thus it will be dropped from future models. (Additional model comparisons showed no effect for velars and back vowels.)

In order to interpret Figures 6.3–6.4 without being caught by the confounding variables mentioned so far, instead of the raw vowel durations the residuals against a model should be plotted. This has the effect of removing the factors that were included in the model. If all of the factors influencing duration that are correlated with [voice] have been included in the model, then the residuals that are left should reflect the effect of [voice]. The calculations were performed separately for each child, and plots of the residuals plus a 100 ms base duration are shown in Figures 6.6–6.7. Fortunately there are no important differences between the two pairs of figures in any case, validating that the sample was in fact relatively well balanced with respect to time and post-vocalic voicing. The PVD duration difference between voiced and voiceless tokens, computed every three months based on the vowel duration residuals, is shown in Figure 6.8.

```
> for (child in c("lily", "alex")) {
+     fltr <- (data$child == child)
```

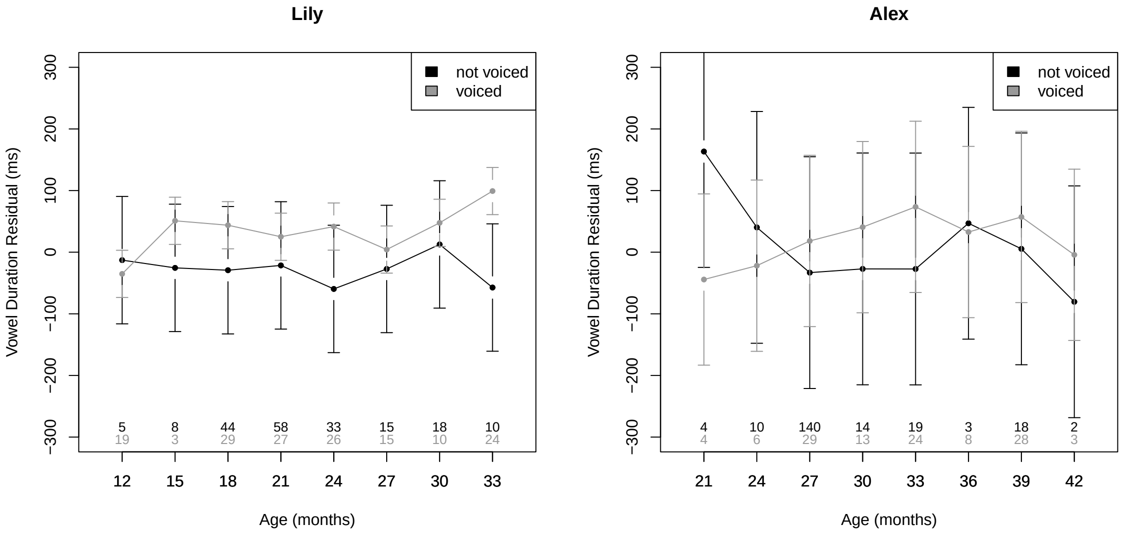


Figure 6.6: Vowel duration residuals by age, post-vocalic voicing, and child in pre-pausal position.

```
+     data$normalized_vowel_duration[fltr] = mle(data[fltr, ],
+         data$vowel_duration[fltr], function(data, D0monoph = 0.1,
+             D0diph = 0, prepausal = 1.5) {
+             m = 0.45
+             D0 = D0monoph + ifelse(!data$diphthong, 0, D0diph)
+             (D0 - m * D0) * prepausal^data$prepausal + m * D0
+         }, getresids = T)
+ }
```

Both sets of figures show a clear effect on vowel duration of post-vocalic voicing in both pre-pausal and non-pre-pausal contexts. If we incorporate post-vocalic voicing into the model we find a large and highly significant effect. (Note that the two children are pooled.) The model and results are shown below:

```
> est <- mle(data, data$vowel_duration, function(data, D0monoph = 0.1,
+     D0diph = 0, D0lily = 0, prepausal = 1, voiced = 1) {
+     m = 0.45
+     D0 = D0monoph + ifelse(!data$diphthong, 0, D0diph) + ifelse(data$child ==
+         "alex", 0, D0lily)
+     (D0 - m * D0) * prepausal^data$prepausal * voiced^data$voiced +
+         m * D0
+ }, nullhyp = list(D0monoph = NA))
```

```
          estimate stderr  null pval
D0monoph  0.118    0.00468 NA   NA
D0diph    0.0134   0.00485 0    0.00579
D0lily    0.00446  0.00356 0    0.21
prepausal 1.68     0.0824  1    1.85e-16
voiced    1.55     0.062   1    5.27e-19
σ_ε       0.0806   0.00184 NA   NA
r^2  0.164
```

The magnitude of the PVD effect in the Klatt model, now taking into account the two children's different speaking rates and that diphthongs have longer intrinsic duration, is estimated at 1.55. As

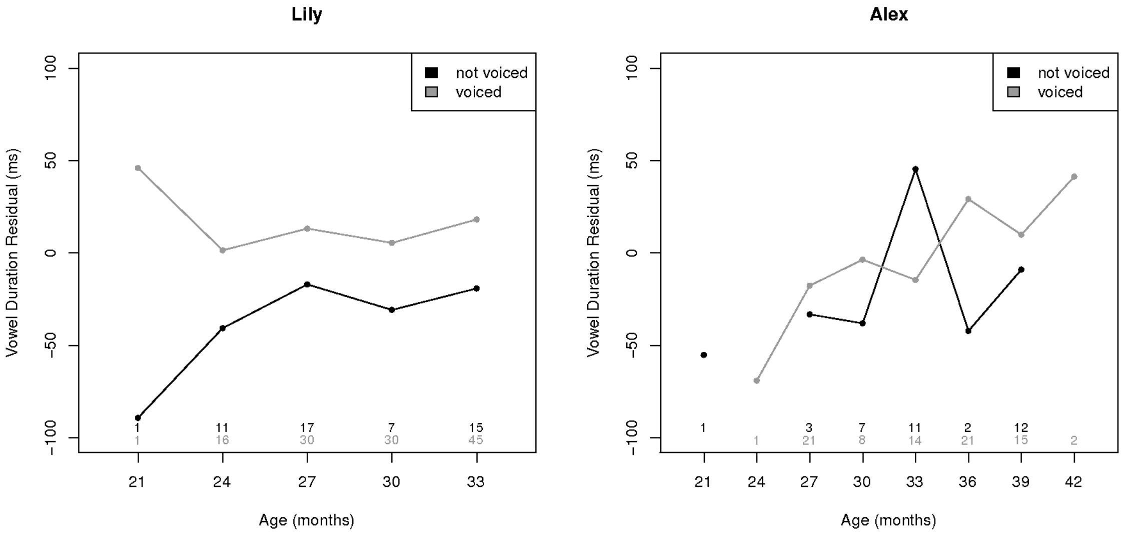


Figure 6.7: Vowel duration residuals by age, post-vocalic voicing, and child in non-pre-pausal position.

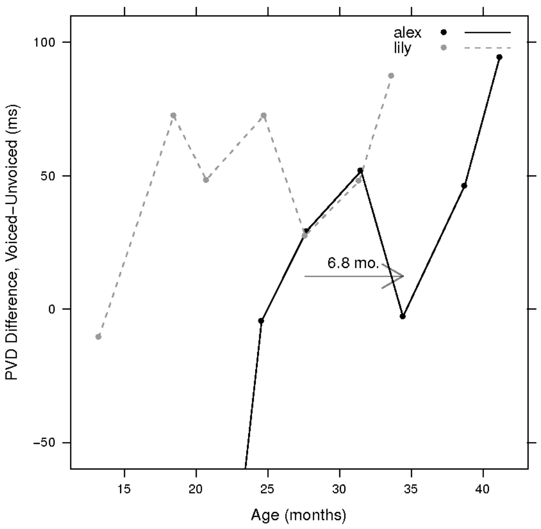


Figure 6.8: Voiced-minus-unvoiced PVD difference by age, with tokens grouped in three-month bins, based on vowel duration residuals.

discussed above, Klatt model estimates are not duration ratios. Computing a duration ratio in the absence of the effects of other factors, using the formula given earlier, we find the PVD ratio to be 1.3. This is a better estimate of the PVD effect in infants that we find by dividing the mean duration of all of the voiced tokens by the mean duration of all of the voiceless tokens, which yields a ratio of 1.25. (The corresponding multiplicative model computes a ratio of 1.36.)

To test for a developmental trend, age can be added into the model by revising the voiced term so that it varies linearly with age: $k_{\text{voiced}} = k_{\text{voiceint}} + k_{\text{voiceage}}(age - \max(\text{age}))$. max(age) is a constant. In this model, k_{voiceint} (i.e. voice intercept) gives the PVD effect factor at the child's maximum age and k_{voiceage} gives the increase in the factor per month leading up to that point. For both children we find a highly significant effect of age, with an estimated 4-percentage-point increase in the PVD effect per month. This validates the trend we see in Figure 6.6 for Alex, though no developmental trend is obvious from the figures for Lily. (A multiplicative model with a voice-by-log(age) interaction term does indicates a statistically significant effect of age for Lily but not Alex.) The results of the Klatt model estimations are given below:

```
function (data, D0monoph = 0.1, D0diph = 0, prepausal = 1, voiceint = 1,
    voiceage = 0)
{
    m = 0.45
    D0 = D0monoph + ifelse(!data$diphthong, 0, D0diph)
    voiced = voiceint + (data$age - max(data$age)) * voiceage
    (D0 - m * D0) * prepausal^data$prepausal * voiced^data$voiced +
        m * D0
}
```

```
alex
          estimate stderr  null pval
D0monoph  0.116    0.00665 NA   NA
D0diph    0.0176   0.00689 0    0.0107
prepausal 1.77     0.134   1    9.17e-09
voiceint  1.83     0.137   1    1.27e-09
voiceage  0.0435   0.0116  0    0.000186
σ_ε       0.0777   0.00261 NA   NA
r^2  0.161

lily
          estimate stderr  null pval
D0monoph  0.115    0.00571 NA   NA
D0diph    0.0422   0.0101  0    3.12e-05
prepausal 1.75     0.113   1    4.94e-11
voiceint  1.96     0.133   1    4.19e-13
voiceage  0.0381   0.00959 0    7.05e-05
σ_ε       0.0805   0.0025  NA   NA
r^2  0.218
```

6.4.3 First formant frequency

A mixed effects, linear model was used to analyze F_1 at the vowel midpoint. Random effects are vowel identity, to model the different vowel spectral targets, and consonant identity, to model the different spectral loci of the consonants at offset and any earlier effect. The fixed effects include age (as the vocal tract lengthens over time, we expect lower formant frequencies), the child (they may have different vocal tract lengths), post-vocalic voicing, and an interaction between child and age (to account for different rate of vocal tract development). The results of the model fit are given below:

```
Linear mixed model fit by REML
Formula: f1m ~ child * age + voiced + (1 | vowel) + (1 | placemanner)
   Data: data[!data$diphthong, ]
   AIC   BIC logLik deviance REMLdev
 10952 10989  -5468    10968   10936
Random effects:
 Groups      Name        Variance Std.Dev.
 vowel       (Intercept) 45804.5  214.020
 placemanner (Intercept)  1955.4   44.220
 Residual                42094.5  205.169
Number of obs: 811, groups: vowel, 7; placemanner, 5

Fixed effects:
              Estimate Std. Error t value
(Intercept)    674.456    115.835   5.823
childlily      422.142     97.261   4.340
age              8.258      2.522   3.274
voicedTRUE     -47.910     17.327  -2.765
childlily:age  -11.221      3.139  -3.575

Correlation of Fixed Effects:
            (Intr) chldll age    vcTRUE
childlily   -0.580
age         -0.676  0.815
voicedTRUE   0.030 -0.076 -0.140
childlily:g  0.530 -0.980 -0.761  0.066
```

All of the fixed effects are significant (assuming the number of degrees of freedom is high enough that a t value of 2.5 is significant). A main effect of age is found: a surprising increase rather than a decrease of 8.3 Hz per month. However the interaction term was also significant, meaning this estimate applied for Alex and a decrease of 3 Hz per month is estimated for Lily. Voicing also showed a significant effect, a reduction of F_1 by 48 Hz in the voiced context (shown in Figure 6.9 with raw F_1 and Figure 6.10 with normalized F_1 by plotting residuals with the same model but with voicing removed as a factor). This is roughly in line with the adult pattern. Summers (1987) reported that F_1 is lower in the voiced context by around 10–20 Hz at the onset of the vowel, 35–45 Hz during its steady-state, and 90–140 at the onset of the consonant, as discussed above. When a age-by-voicing interaction term is added to the model to test for a developmental trend, neither the voiced main effect nor the interaction term come out significant.

6.4.4 Consonant duration

Consonant duration in intervocalic context was also measured, except for the alveolar stops as they were often (and, according to the adult grammar, supposed to be) realized as flaps in this context. Durations were measured as the total of closure, burst, and aspiration. Unfortunately, there was a very uneven distribution of consonants over both [voice] and time (see Figure 6.11) making it very difficult for any conclusions to be drawn about the relation between consonant duration and either of these factors. In other words, assessing a correlation between duration and time is clearly confounded by a) velars occurring generally later in time than bilabials and b) velars known to have generally a longer duration (Crystal and House 1988a). Likewise, the fricatives in this context were overwhelmingly voiceless, making for a bad comparison between voiced and voiceless duration.

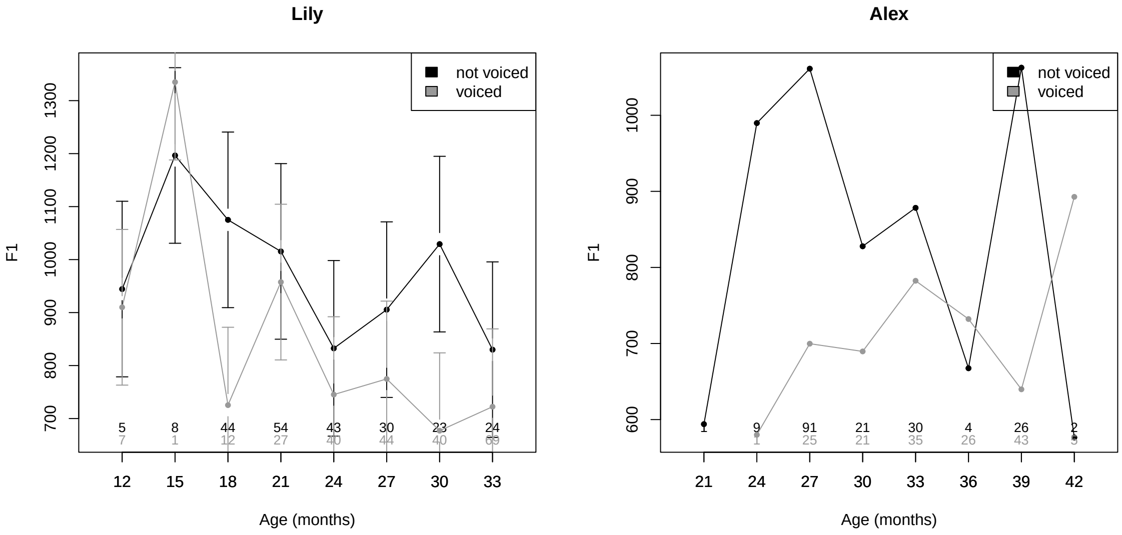


Figure 6.9: F_1 by age, post-vocalic voicing, and child, for monophthongs.

Normalization R code:

```
> for (child in c("lily", "alex")) {
+     f <- (data$child == child) & !data$diphthong & !is.na(data$f1m)
+     model <- lmer(f1m ~ age + (1 | vowel) + (1 | placemanner),
+         data[f, ])
+     data$normalized_f1m[f] <- resid(model)
+ }
```

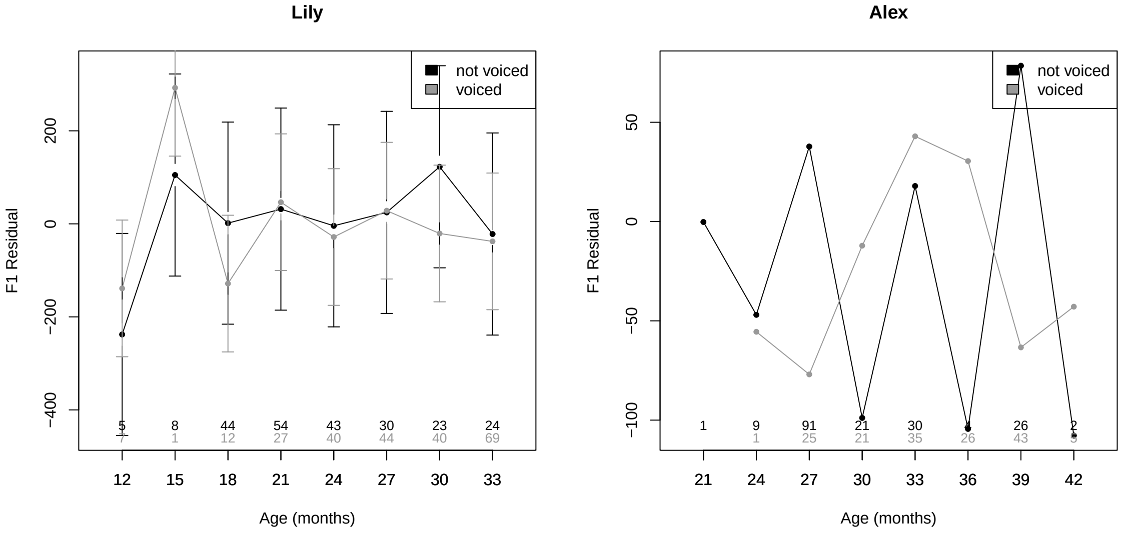


Figure 6.10: F_1 residual by age, post-vocalic voicing, and child, for monophthongs.

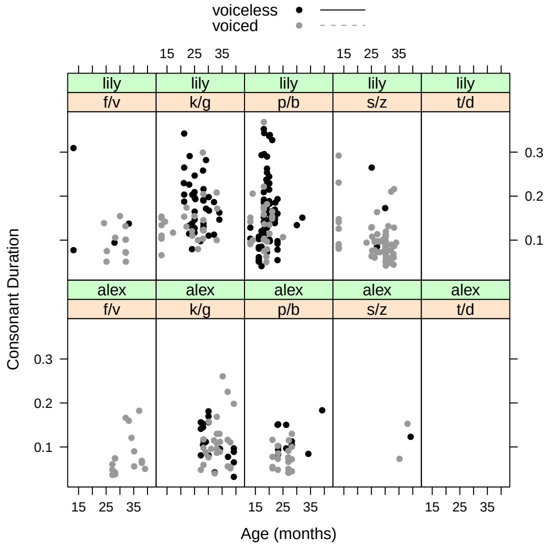


Figure 6.11: Consonant duration by age, voicing, place of articulation, and child.

The dearth of /s,f,v/ tokens from Lily cannot be attributed to her linguistic environment — Lily had far fewer of these phones in intervocalic context than her mother. See Figure 6.2 above. Likewise for /z/ for Alex. But although Alex had few intervocalic /f,s/ tokens in the sample, these phones do not stand out in Figure 6.2, meaning that the scarcity of these tokens in the intervocalic same probably did reflect his linguistic environment.

Alex's /p,b/ distinction and Lily's /k,g/ distinction through consonant duration were each statistically significant ($p < 0.003$), and in the correct direction, according to two one-tailed t-tests.

6.4.5 Closure voicing intensity

The closure voicing intensity in intervocalic stops was measured qualitatively and quantitatively. These two measurements were reliably correlated, as shown by the positive slopes and relatively small error bars in Figure 6.12, as well as the fact that the two lines for the voiced and voiceless consonants are very close.

As with consonant duration, the data for voicing intensity was not amenable to an analysis involving time. Voicing intensity is plotted against age, with [voice] separated by shading, in Figure 6.13. The quantitative voicing intensity measure was different between /p/ and /b/ ($p < 0.0001$), and in the correct direction, according to a one-tailed t-test. The other differences were not significant.

6.4.6 Interactions between acoustic correlates

Interactions between the five correlates of [voice] measured in this experiment were assessed by first dividing the tokens into the voiced and voiceless groups. Without dividing them first, we expect all of the measures to be correlated because they are all correlated with [voice]. By splitting up the tokens first by their [voice] feature, we can see whether there are any connections between these acoustic dimensions that go beyond the phonological featural specification. The correlations are also performed on normalized measures in a same manner as discussed throughout, with the R code

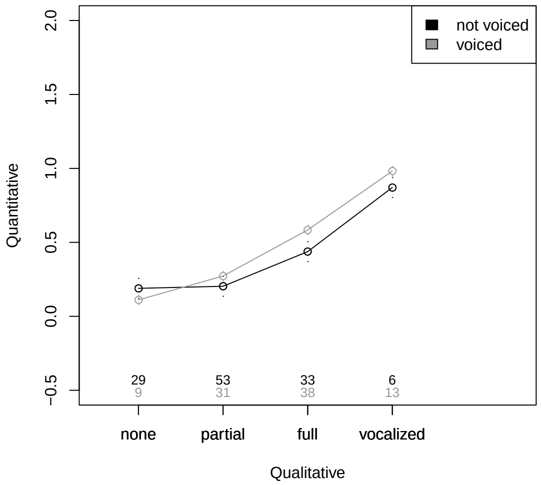


Figure 6.12: Comparison of the two measurements of closure voicing intensity, the impressionistic qualitative measure and the quantitative and automatic measure of signal intensity.

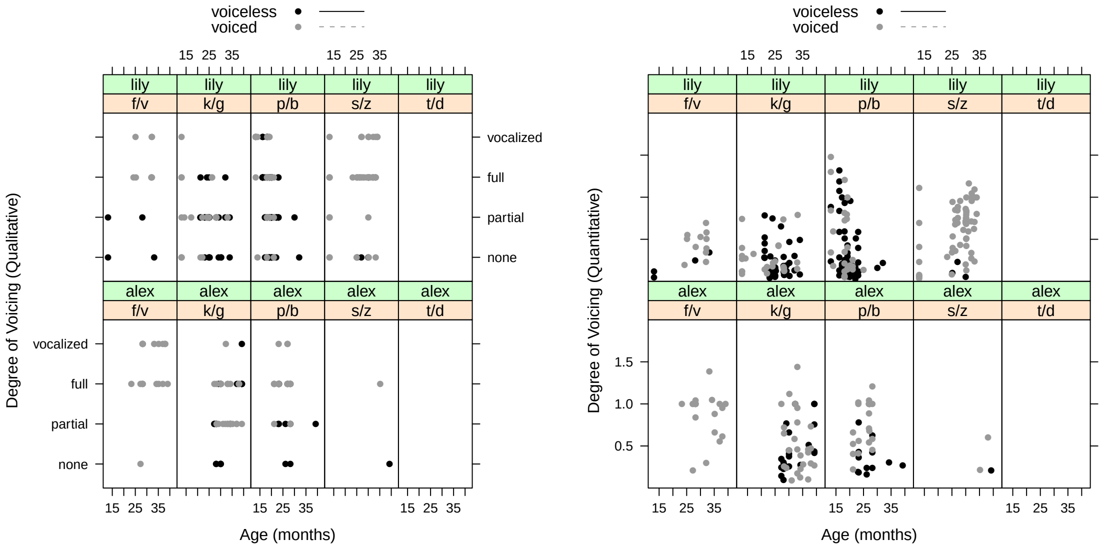


Figure 6.13: Closure voicing by age, according to the two measures (left: impressionistic, right: signal intensity). The same measurements for the fricatives are also shown but are not used.

below. The correlation coefficients are corresponding p-values are reported in Table 6.2 for each pair of measures.

```
> data$normalized_vowel_duration = mle(data, data$vowel_duration,
+     function(data, D0alex = 0.1, D0lily = 0, D0diph = 0, prepausal = 1) {
+         m = 0.45
+         D0 = D0alex + ifelse(data$child == "alex", 0, D0lily) +
+             ifelse(!data$diphthong, 0, D0diph)
+         (D0 - m * D0) * prepausal^data$prepausal + m * D0
+     }, getresids = T)
> f <- !is.na(data$consonant_duration)
> model <- lmer(log(consonant_duration) ~ child + (1 | vowel) +
+     (1 | placemanner), data[f, ])
> data$normalized_consonant_duration[f] <- resid(model)
> f <- (!is.na(data$f1m) & !is.na(data$f1e))
> model <- lmer(f1m ~ child * age + (1 | vowel) + (1 | placemanner),
+     data[f, ])
> data$normalized_f1m[f] <- resid(model)
```

Where formant frequencies are used, only tokens with monophthong vowels are included. Where consonant duration or voicing intensity are used, only intervocalic tokens are included. Voicing intensity refers to stops only.

The only interesting highly significant correlation is a negative correlation between consonant duration and voicing intensity. This is not surprising: It is more difficult to maintain voicing during longer closure, and as a result we expect less voicing intensity over-all in longer consonants. (Also highly significant is the correlation between F_1 at vowel midpoint and offset, which is to be expected.) Also significant was a positive correlation between vowel duration and F_1 frequency at vowel midpoint in the voiceless tokens only, depicted in Figure 6.14. Note that while [-voice] is correlated with higher F_1 and shorter vowels, we see here within the [-voice] category that higher F_1 is correlated with longer vowels. Approaching significance was a positive correlation between F_1 offset and consonant duration among the [+voice] tokens, also shown in Figure 6.14.

As part of testing whether the PVD effect is determined by phonetic or phonological factors, we can see which is a better predictor of vowel duration. A comparison of two models — one of which uses the phonological feature as a predictor and the other the quantitative measure of voicing intensity — shows through the value of r^2 that the phonological feature model ($r^2 = 0.174$) is nominally a better fit than the phonetic model ($r^2 = 0.144$). The models have the same number of parameters, although even the slightly extra power in the phonetic model's continuous-valued voice parameter did not help it. (Using four dummy variables for the qualitative measure of voicing intensity increases the r^2 value but not enough to make it a better fit, and not more than would be expected from merely adding additional parameters.)

```
> d <- data[!is.na(data$rmsvoice) & !data$fricative, ]
> est <- mle(d, d$vowel_duration, function(data, D0monoph = 0.1,
+     D0diph = 0, D0lily = 0, prepausal = 1, voice_feature = 1) {
+     m = 0.45
+     D0 = D0monoph + ifelse(!data$diphthong, 0, D0diph) + ifelse(data$child ==
+         "alex", 0, D0lily)
+     (D0 - m * D0) * prepausal^data$prepausal * voice_feature^data$voiced +
+         m * D0
+ }, nullhyp = list(D0monoph = NA))


              estimate stderr  null pval
D0monoph      0.100    0.00837 NA   NA
```

Voiced

	F_1 Midpoint	F_1 Offset	Consonant Duration	Voicing Intensity
Vowel Duration	0.06 (0.25)	0.05 (0.35)	0.15 (0.036)	-0.05 (0.6)
F_1 at Midpoint		0.45 (0)	0.12 (0.17)	-0.03 (0.8)
F_1 Offset			0.09 (0.34)	-0.10 (0.46)
Consonant Duration				-0.26 (0.011)

Voiceless

	F_1 Midpoint	F_1 Offset	Consonant Duration	Voicing Intensity
Vowel Duration	0.12 (0.017)	0.02 (0.69)	0.01 (0.89)	0.16 (0.058)
F_1 at Midpoint		0.54 (0)	0.02 (0.84)	-0.01 (0.92)
F_1 Offset			0.24 (0.0049)	-0.03 (0.69)
Consonant Duration				-0.42 (3e-07)

Table 6.2: Correlation coefficients (and p-values in parentheses) between vowel duration residuals, F_1 residuals at vowel midpoint and offset, consonant duration residuals, and (the quantitative measure of) voicing intensity. The voiced and voiceless tokens are separated first.

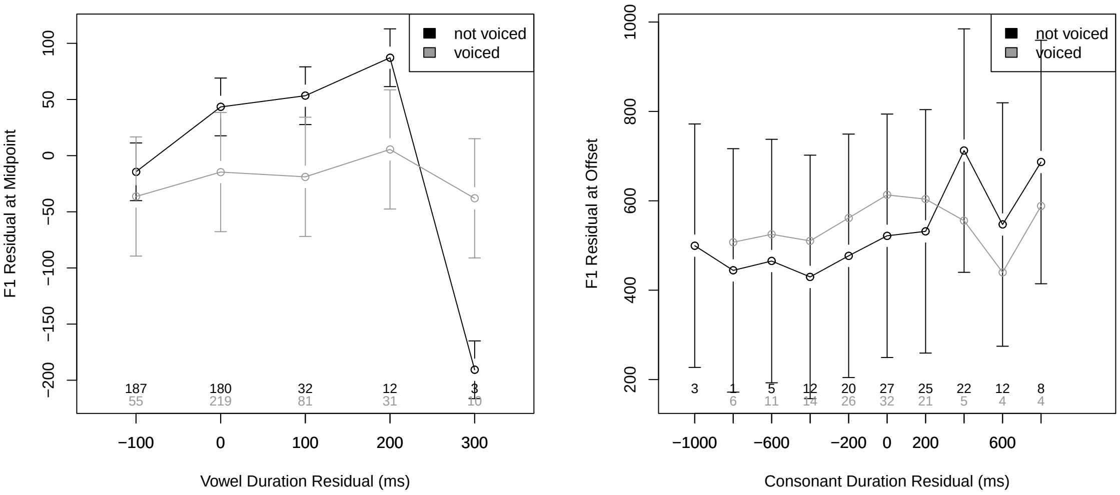


Figure 6.14: Left: First formant frequency at vowel midpoint by vowel duration. Right: First formant frequency at vowel offset by consonant duration. (All measurements normalized as described in the text.)

```
D0diph        0.0198   0.0116  0    0.0875
D0lily        0.0269   0.00918 0    0.00333
prepausal     1.34     0.158   1    0.0328
voice_feature 1.50     0.156   1    0.00142
σ_ε           0.0723   0.00329 NA   NA
r^2  0.174
```

```
> est <- mle(d, d$vowel_duration, function(data, D0monoph = 0.1,
+     D0diph = 0, D0lily = 0, prepausal = 1, voice_intensity = 0) {
+     m = 0.45
+     D0 = D0monoph + ifelse(!data$diphthong, 0, D0diph) + ifelse(data$child ==
+         "alex", 0, D0lily)
+     (D0 - m * D0) * prepausal^data$prepausal + m * D0 + voice_intensity *
+         data$rmsvoice
+ }, nullhyp = list(D0monoph = NA))
```

```
                estimate stderr  null pval
D0monoph        0.0963   0.0123  NA   NA
D0diph          0.0416   0.0127  0    0.00108
D0lily          0.0328   0.0109  0    0.00255
prepausal       1.33     0.19    1    0.078
voice_intensity 0.031    0.0166  0    0.0621
σ_ε             0.0736   0.00335 NA   NA
r^2  0.144
```

6.5 Discussion

In this chapter an experiment was discussed whose results bear on the questions of a) whether infants' phonetic implementation of post-vocalic [voice] uses the same acoustic dimensions as in adults, b) how the use of those acoustic dimensions changes over time during the infant's linguistic development, and c) whether those acoustic dimensions are specified in the grammar as part of the phonetic implementation of [voice] or are physiological consequences of other aspects of voice.

In the introduction to this chapter I raised a number of methodological concerns about using a corpus to ask these questions. The first concern was that extralinguistic factors such as prominence or an inherent child-excitability of animate words would confound the results. This is a particularly difficult problem to face since there is no a priori way to judge prominence without making use of some of the same acoustic dimensions that are also correlated with [voice], such as duration. Other factors such as intensity and pitch could help determine prominence, but they are difficult to normalize in a sample such as this one where utterances are short and vary widely along these dimensions. It is not possible to rule out prominence as influencing the results of this experiment. But at least Table 6.1, which listed the 15 most frequent words in the sample, did not indicate an obvious trend for the most excitable (i.e. animate) words to have longer durations. Other concerns included incidental correlations between [voice] and vowel quality, age, and position in the utterance.

A way in which these concerns were addressed was by using appropriate models for each variable, taking into account factors such as vowel quality, whether the vowel is a diphthong, and place in the utterance. For the two duration variables (vowel and consonant duration), a model based on Klatt's incompressibility was used. This model struck a balance between linguistic plausibility and complexity or over-fitting. The use of maximum likelihood estimation to fit a Klatt model appears to be a new, practical technique in linguistic phonetics. We had reason from adult speech data to believe it would be more appropriate than a multiplicative model fit by linear regression (in log space), and the MLE-estimated Klatt model found a statistically significant effect of age on the PVD effect for both children, while the multiplicative model seemed to be powerful enough only to find an effect for one child. (In other cases the two models appeared to account for the variability in the data about the same, based on the r^2 values.) The residuals from these models were used in so-called "normalized" plots of these factors by age, so that confounding effects could be removed. This was especially important in the case of plotting first formant frequency.

The analysis revealed that infants aged roughly 1;3–3;5 use vowel duration in making the [voice] distinction. This is in line with past research on the PVD effect. Here it was shown that the voiced-to-voiceless PVD ratio was roughly 1.3 overall — but if there is indeed a learning curve this is merely an average of the PVD effect during the child's development during this time. The ratio is in line with that observed in adults. The PVD ratio found in adult lab speech in Section 5.2 varied between 1.2 and 1.5 depending on speaking rate and word length. In Section 5.1, dialectal variation in American English revealed a range from 1.15–1.26. But task and measurement methods have a large effect on the PVD difference. The lab speech elicited in Section 5.2 differed from the mixed-task recordings in Section 5.1, and these differed from observations of spontaneous speech in infants at varying stages of linguistic development.

The analysis similarly revealed a statistically significant effect of [voice] on monophthong F_1, the direction and magnitude of which were in line with what has been observed in adults. Unfortunately, the plots of these differences were much less convincing than the p-values.

Little could be said about consonant duration and voicing intensity during closure/frication because of the sparsity of the data. The /p,b/ distinction in one child was differentiated by consonant duration and voicing intensity, but other distinctions were not found. However, the missing acoustic contrasts for some of the other phonemes which were not so sparse (the /k,g/ distinction for both children) may indicate that not all phonemes are alike when it comes to implementing [voice]. This could be for several reasons. At the level of articulation, it may be harder to voice some phonemes than others (and this may explain why the cross-linguistic preference for voiceless stops in a language's stop inventory varies from place-of-articulation to place-of-articulation, Maddieson 1984, p35). At the level of phonology, it may indicate that there is no inventory-wide phonetic implementation of [voice] in the grammar after all. (It would have been interesting as well to see if the magnitude of the PVD effect varies by place of articulation, but the same problem here prevented an analysis of other aspects of the consonants, the the places of articulation are not well distributed by age. Thus the p/b contrast showed a much smaller PVD ratio than the p/b contrast, likely on account of the p/b tokens tending to occur at the earliest of ages.)

On a token-by-token basis, there were few correlations between the acoustic dimensions that suggested any were dependent on the others. The exception was a correlation between consonant voicing and duration, which is to be expected since voicing is difficult to maintain and so more likely to cease the longer closure or frication goes on. It is also compatible with a model in which closure and frication duration differences between [±voice] are due not to differences in supra-laryngeal articulator timing but to differences in the glottal state (something suggested, for fricatives, by Stevens et al. 1992). Vowel duration and F_1 were weakly correlated, but in the direction opposite to what would explain the F_1 component of [voice]. The direction of the correlation is instead probably attributable to differences in intrinsic vowel duration: low vowels with high F_1 tend to have longer durations. Beyond this, vowel duration, formant frequency, and consonant duration and voicing intensity appeared to be independent at least as far as [voice] is concerned. Comparing models of vowel duration based on either phonological voicing or phonetic voicing intensity, the phonological model fit the data better. Preceding vowel duration was not dependent on whether the consonant was phonetically voiced, and so vowel duration could not be a simple physiological consequence of the glottal state. Likewise for formant frequency, and vowel duration could not be the result of a rule of syllable duration constancy. These factors are each under separate linguistic and articulatory control by the infant and their use in [voice] must be specified in the grammar.

Evidence for a developmental trend (i.e. an interaction between [voice] and age) was found for preceding vowel duration based on maximum likelihood estimation, though it was less obvious from the figures, and no trend was found for first formant frequency. (The data was too sparse to perform an analysis on consonant duration and voicing intensity.) The magnitude of the vowel duration difference increases over time, indicating that the PVD effect is a learned phenomenon. There had been some question of the direction of the development of the PVD effect. Buder and Stoel-Gammon (2002), for instance, proposed that the unmarked state was the presence of a vowel duration difference, and that infants learning languages with small or no PVD must unlearn the difference. The results here (if indicative of linguistic and not mere motor learning) suggest otherwise, that the unmarked state is no PVD. As for whether it is the voiced tokens that increase in duration or the voiceless tokens that decrease, or a mix, it is hard to say. Figures 6.6–6.7, at least for Alex, suggest that it is a mix.

An interesting but tentative observation is that the developmental trend of the PVD difference is not only similar for the two children but also what differs between the two children may be explainable in terms of the gross linguistic development of the children. For both children, the PVD developmental curve in Figure 6.8 starts at or below zero, peaks around a 50–70 ms difference 3–9 months later, then drops substantially for 3–6 months, and finally recovers and ends on a large difference of roughly 75–90 ms duration difference. But, Alex's curve appears delayed relative to Lily's. It might be mere chance that the curves are similar in this way, given the high variance in the data. But what makes this interesting is that the delay between Alex and Lily's curves here is similar to a delay observed only incidentally in Figure 6.5, which showed the proportion of tokens which were pre-pausal over time — an indication of the complexity of the child's utterances at each recording session. (When the child begins complex, multi-word utterances, the number of pre-pausal tokens declines.) In both cases, Alex was behind Lily by close to six months. The delays are indicated by arrows in the figures. And for both children, the drop in the PVD difference (at 27 and 32 months) was just around the completion of the child's transition to complex utterances (at 28 and 36 months). If the timing is not a coincidence, perhaps the changes indicate a reanalysis of the PVD effect. As the child learns the proper prosodic structures for complex utterances, the PVD effect may be put aside until it can be integrated in the grammar with suprasegmental effects. It is too early to make much of this either way, but it is perhaps an area of high potential for future research.

The analysis of F_1's correlation with the glottal state and the developmental trend of the F_1 component of [voice] provide conflicting perspectives on the nature of F_1's role in [voice]. The most likely explanation for F_1 in adults was that it is a physiological consequence of the glottal state (see Chapter 2). The lack of a developmental trend for F_1 that was found is expected if it is a physiological phenomenon and not a linguistic phenomenon. On the other hand, a correlation between F_1 and the glottal state was hoped for to substantiate this point, but no correlation was found.

Chapter 7

Properties of Fluent Speech and Modeling the Learning Process

Chapter 2 would have the reader believe that with so many acoustic differences between voiced and voiceless consonants that the problem of acoustically discriminating them ought to be trivial. At best this might be true of laboratory recorded speech, in which speech is slow and careful. But in continuous, fluent speech one finds that voiced and voiceless tokens are not easily separable on any given acoustic dimension, and some acoustic dimensions that separate tokens in laboratory speech may show no reliable value to discrimination in continuous speech.

In this chapter machine learning techniques are applied to the problem of classifying tokens as either voiced or voiceless. Three techniques are compared: semi-supervised automated clustering, clustering plus cluster merging, and supervised support vector machines. The features used in each case were a selection of automated acoustic measurements on the consonant and preceding and following vowel. Machine learning tells us something about the difficulty of the problem and what information the child has available is sufficient, necessary, or perhaps even unhelpful for determining whether any given consonant is voiced or voiceless.

There are two main differences between the methodologies of the studies discussed in Chapter 2 and the methodology of this chapter. First, in Chapter 2 most of the acoustic data was of laboratory speech, and of isolated tokens in the most extreme case. In this chapter, continuous fluent speech is considered and it is found that, not surprisingly, the differences between voiced and voiceless tokens on any given acoustic dimension is much less than what has been found in the past.

Second, in this chapter the task of discrimination between voiced and voicelessness (at the phonological level) is considered with multidimensional input. That is, while studies in Chapter 2 mostly sought to describe the differences on individual dimensions here the goal is to give the hypothetical learner as best a chance as he can have by giving him all of the available acoustic data. For instance, in Edwards (1981), the acoustic distributions of pre-stress, intervocalic stops in a word-list reading task were examined on a dimension-by-dimension basis. Following-vowel F_0, voicing duration during closure, and segment duration (that is, including aspiration) each independently allowed for the placement of a decision boundary that correctly classified around 90% of tokens. Edwards also found that VOT alone yielded a classification accuracy of as much as 98% — with knowledge of place of articulation, the error rate dropped to 1 in the 240 observations. But the overlap of the distributions of preceding vowel duration was “so extensive” that it had little value in discrimination. Closure duration also had a very small voiced–voiceless difference. In this chapter, we will see whether VOT proves to be as useful in continuous speech and whether the same acoustic correlates have the same relative importance.

Although the focus of this dissertation has been on language learning, the speech data used in this chapter is adult-directed speech rather than infant-directed speech (IDS). The two modes of speech are well known to have significant differences. For instance, in English IDS has higher overall pitch, larger pitch excursions, long vowel durations, and vowel peripheralization, and these exaggerations may support learning. One might hypothesize that exaggerations would extend to the voiced–voiceless difference as well. But that does not seem to be the case. Narayan, Gorman, and Swingley (2008) reported that in IDS, voiced and voiceless tokens has *less* of a separation on the VOT dimension than in adult-directed speech, with IDS having less pre-voicing, and that the magnitude of VOT was not significantly different between IDS and adult-directed speech. Likewise, there was no difference in the following-F_0 dimension of voicing between IDS and adult-directed speech. Thus, it seems appropriate to work on machine learning of [voice] beginning with the more well-understood register: adult-directed speech.

7.1 Methods

7.1.1 Corpora and Alignment

Two corpora were used in this chapter. Both were portions of audio-book recordings downloaded from LibriVox.org, a project to create a library of public domain recordings of public domain books. The first corpus was *Around the World in 80 Days* ("AW") by Jules Verne, read by Mark Smith who identified himself as from South Carolina, which had a total recording time of 429 minutes. The second corpus was the first 17 chapters of *Pride and Prejudice* ("PP") by Jane Austen, read by Annie Cole who identified herself as from St. Louis, Missouri. This corpus had a total recording time of 175 minutes. Both readers spoke naturally. The recordings were down-sampled to 11,025 Hz for the purposes of forced alignment.

The Penn Phonetics Lab Forced Alignment Toolkit (P2FA, Yuan and Liberman 2008) was used to determine the phone boundaries based on the book texts, which were adapted into a transcription of the audio, and the Carnegie Mellon University Pronouncing Dictionary (cmudict) plus additional pronunciation entries for words in the corpora not found in cmudict. As described several times in this dissertation, P2FA is a Hidden Markov Model-based tool based on HTK (Young et al. 2006), using multiple gaussian mixture monophone models on 39 PLP coefficients. No manual correction was performed on the alignments.

7.1.2 Features

The set of consonant tokens analyzed were stops that either immediately preceded or followed a stressed vowel. Alveolar stops following a stressed vowel and preceding an unstressed vowel were excluded as they would be flaps. Stops following an /s/ were also excluded to put aside the cases of unaspirated stops in onset clusters. Praat (Boersma 2001) was used for acoustic analysis, with no manual correction. The acoustic measurements made were:

1. Closure Intensity: The RMS intensity during the consonant's middle third. Consonant intervals include closure, burst, and aspiration. The rationale behind this measurement was to

approximate a degree of voicing during closure, and to exclude acoustic energy toward the end of the segment due to the burst and aspiration noise.

2. Aspiration Duration: The duration of aspiration (or Voice Onset Time), using a novel method to locate the burst. The burst location is defined as that time which maximizes the ratio of the RMS intensity from the burst to the end of the consonant interval to the RMS intensity in the 1/10th of the consonant interval preceding the burst. A minimum ratio of 2.5 was chosen to prevent small changes in intensity in intervals lacking a burst (e.g. because of aligner error) from being recorded as a burst. For sample results of this procedure, see Figure 7.1.

3. Closure Pitch: The maximum pitch during the middle 50% of the interval of the consonant. When there is no periodic noise (i.e. when Praat indicates there is no pitch), zero is recorded.

4. Vowel Duration and Intensity: The duration and RMS intensity of the preceding and following vowel.

5. Vowel F_1: The median F_1 value of the preceding and following vowel.

6. Vowel F_0: The maximum pitch in the first quarter of the following vowel.

Each of these acoustic dimensions are correlated with the voicing contrast, as discussed in Chapter 2.

Three subsets of the tokens were used separately, and different feature sets where used for each:

1. CV Tokens: Closure duration (total interval duration minus aspiration duration), aspiration duration, the log of the closure RMS intensity, closure pitch, following vowel F_1 and following vowel F_0 were used as features. Following vowel F_0 was normalized by subtracting the mean F_0 for each vowel (i.e. its intrinsic pitch). F_1 was normalized by replacing the raw F_1 values with their residuals after they are fit to a linear model using vowel identity (including stress level) and consonant place of articulation as factors (i.e. its expected target and offset). Additionally, a feature for the consonant locus was used, which was computed as the mean of the F_1 values by the place of articulation of the consonant.

2. VC Tokens: The ratio of closure duration to preceding vowel duration ("relative closure duration"), aspiration duration, the log of closure intensity, preceding vowel duration, F_1, and locus were used as features. In the *Pride and Prejudice* corpus, preceding vowel duration was normalized by replacing it with its residuals when log durations are fit to a linear model using vowel identity and utterance position as factors. Utterance position was defined as the log of one more than the number of stressed vowels between the consonant and the subsequent punctuation mark in the book text. Formant frequencies were normalized as described above. A locus feature was used here as well. Because the effect of [voice] on preceding vowel F_1 differs between monophthongs and diphthongs, diphthongs were excluded.

3. Intervocalic Tokens (VCV): All of the features above were used, except for the relative closure duration feature. (The raw closure duration was used instead.)

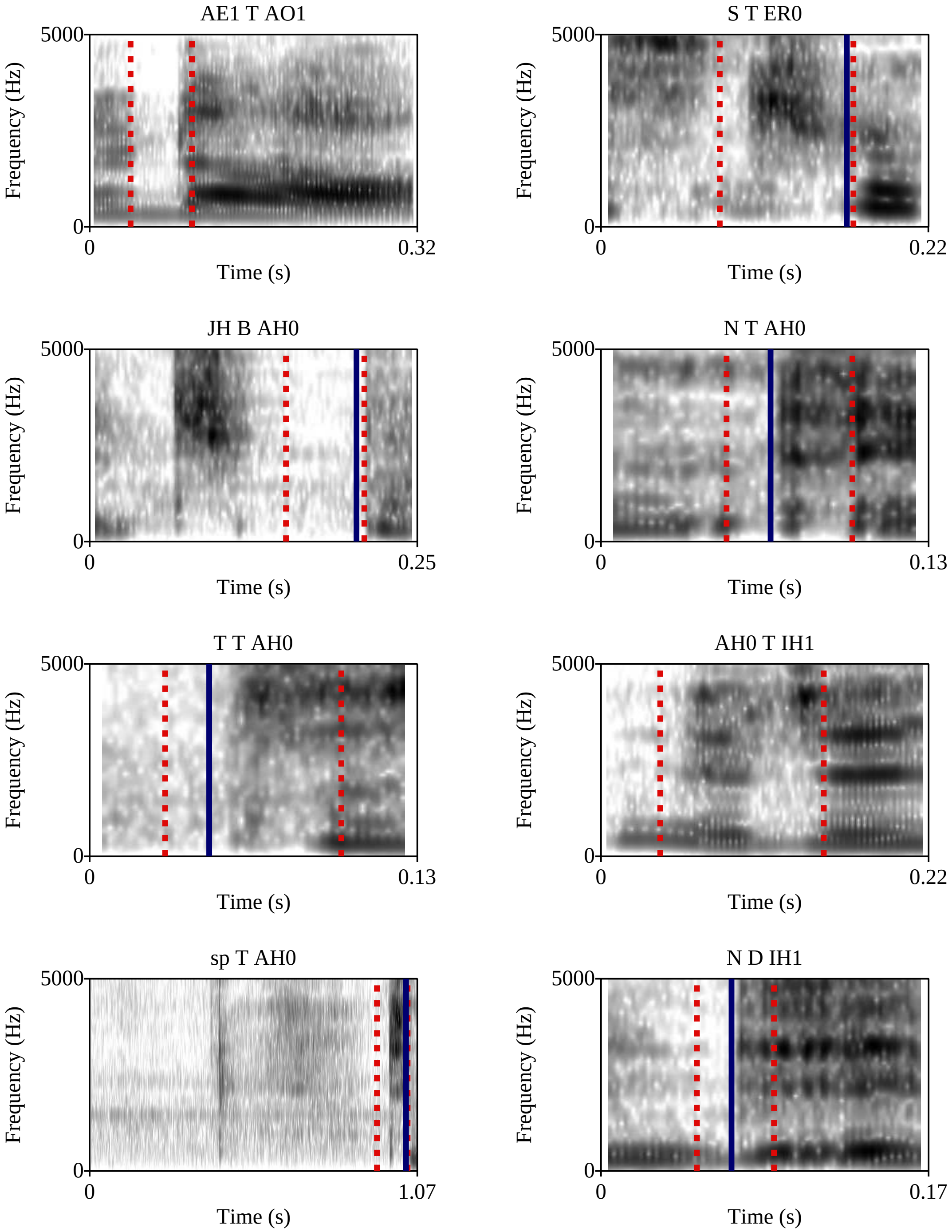


Figure 7.1: Spectrograms showing the automated detection of the start of aspiration (solid blue line) within the interval for the consonant determined by forced alignment (bounded by dashed red lines), and the surrounding phonemes. A missing aspiration line indicates the procedure failed to find a suitable location and such tokens were dropped from analysis. This is a random sample of pre-vocalic stops from the Around the World corpus. The identity of the phonemes are indicated above each spectrogram ('sp' stands for a pause).

7.1.3 Acoustic Distribution

The distribution of voiced and voiceless tokens overlaps on all of the acoustic dimensions in continuous speech, and in some cases acoustic dimensions seen as useful for discrimination in lab speech appear with a voiced/voiceless difference opposite to what is expected from past research.

Figure 7.1 shows the differences in mean acoustic measurements (in some cases normalized as described above) between the voiced and voiceless tokens, separately for pre-stress CV tokens, post-stress VC tokens, and post-stress VCV tokens[1], and separately for the two audio-book corpora. The table also includes the number of standard deviations that separate the means (using the average of the s.d. of the voiced tokens and the s.d. of the voiceless tokens), as an indication of how separable the two distributions are. Many of the differences reach statistical significance, but as can be seen from the histograms in Figure 7.2 the distributions of voiced and voiceless tokens overlap so much that for any particular value on the acoustic dimension (say, aspiration duration of .05 ms or zero closure pitch) a token may be equally likely to be either voiced or voiceless. The larger the separation of the means, in terms of standard deviations, the more accurately a token can be labeled voiced or voiceless based on the known distributions.

No acoustic dimension has a consistently high separation across contexts and corpora (again in Figure 7.1). While aspiration duration is relatively highly separable in CVs (1.2–1.3 standard deviations), it is much less separable in the other contexts. In the VCs (and VCVs), closure pitch and (in AW) closure RMS become more useful to the listener. This again points to the lack of acoustic constancy across contexts, a premise that underlies some frameworks of feature theory (see Chapter 4). While many of the other acoustic dimensions show differences that are in the right direction as would be expected from Chapter 2, their separability is often negligible. In the case of the CV tokens from Pride and Prejudice, two dimensions show promising separability: aspiration duration and closure pitch. Figure 7.3 shows the corresponding scatter plot, which seems to indicate at first that the combination of the two features may be more useful at discrimination than any one alone. For those tokens which show closure voicing, aspiration duration clearly separates the few phonologically voiceless tokens (with high aspiration duration) from the majority of phonologically voiced tokens (with low aspiration duration). On the other hand, for those tokens with no closure voicing the voiced and voiceless tokens show essentially identical distributions of aspiration duration (the right half of the figure). In other words, in an area of one acoustic dimension where the two categories overlap, the distributions in the next most useful dimension also overlap.

7.1.4 Simple Clustering

Methods

The first machine-learning algorithm used was a variant of k-means clustering (in particular, the Partitioning Around Medoids (pam) function of R in the cluster package). Clustering is one solution to a labeling problem. In a binary classification task such as discriminating voiced from voiceless stops, the algorithm groups the data points into two clusters. In standard k-means, each data point is identified with the closest of k clusters, with distance measured straightforwardly in Euclidean space. An iterative process is used to locate the best locations for the centers of the clusters to minimize the overall distance between a cluster's center and its associated data points.

[1] Note that the post-stress VCV tokens are a subset of the VC tokens.

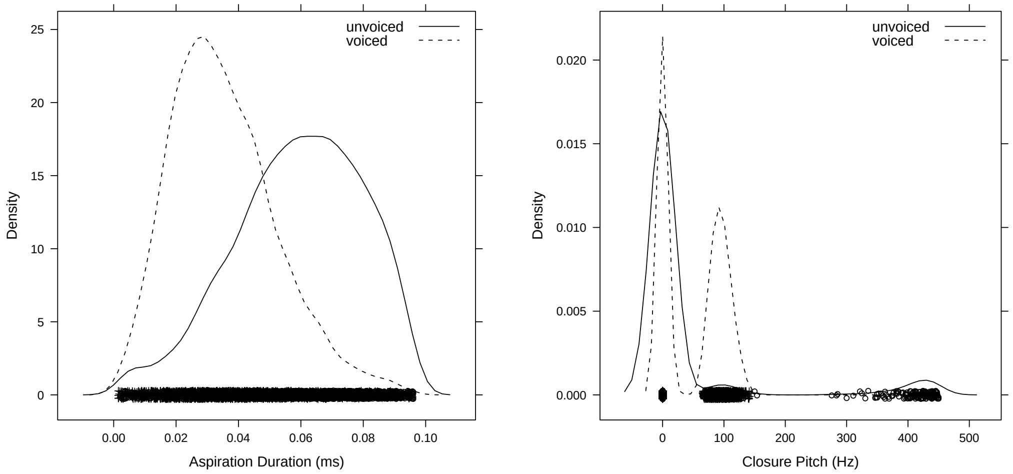


Figure 7.2: Histograms by aspiration duration (left) and closure pitch (right). Around the World CV tokens.

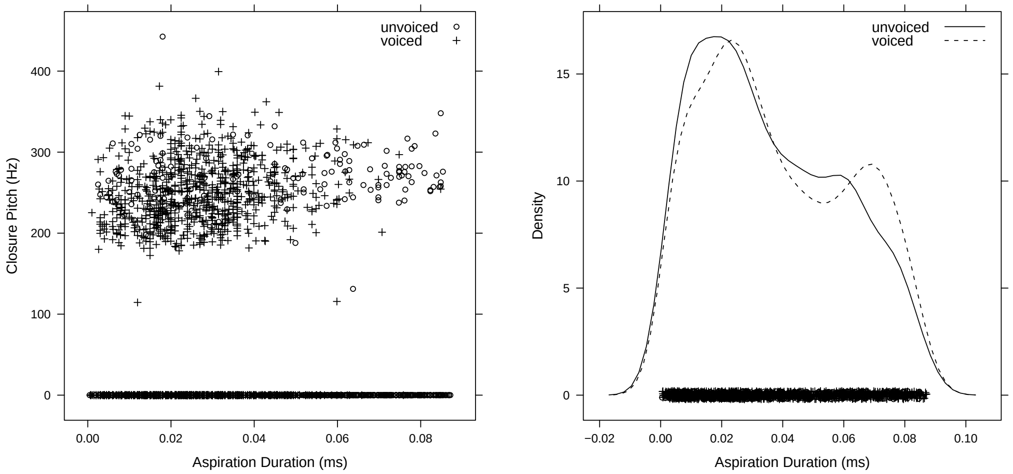


Figure 7.3: Left: Aspiration duration by closure pitch. Right: Histogram of aspiration duration for those tokens with zero closure pitch. CV tokens from Pride and Prejudice.

	Around the World in 80 Days					
	CV		VC		VCV	
N (voiced/voiceless)	2,704/2,277		161/275		140/208	
	$\Delta\bar{x}$	σ's	$\Delta\bar{x}$	σ's	$\Delta\bar{x}$	σ's
Closure Duration (ms)	.008	.37	—	—	-.012	.61
Relative Closure Duration	—	—	-.140	.32	—	—
Aspiration Duration (ms)	-.024	1.3	-.013	.77	-.013	.76
Closure RMS	-.12	.29	.54	1.4	.57	1.6
Closure Pitch (Hz)	13	.19	73	1.8	73	1.8
V_1 Duration (ms)	—	—	.016	.054	.055	.18
V_1 F_1 (Hz)	—	—	-38	.47	-32	.42
V_2 F_0 (Hz)	-4.9	.33	—	—	-1.6	.093
V_2 F_1 (Hz)	-.38	.08	—	—	-2.2	.047
	Pride and Prejudice					
	CV		VC		VCV	
N (voiced/voiceless)	1,620/992		113/344		86/263	
	$\Delta\bar{x}$	σ's	$\Delta\bar{x}$	σ's	$\Delta\bar{x}$	σ's
Closure Duration (ms)	-.0036	.16	—	—	-..0096	.47
Relative Closure Duration	—	—	-.200	.38	—	—
Aspiration Duration (ms)	-.023	1.2	-.0086	.51	-.0094	.56
Closure RMS	-.21	.51	.17	.39	.16	.38
Closure Pitch (Hz)	80	.71	110	.90	100	.86
V_1 Duration (ms)	—	—	.032	.092	.083	.24
V_1 F_1 (Hz)	—	—	-16	.21	-20	.27
V_2 F_0 (Hz)	-17	.63	—	—	-4.9	.15
V_2 F_1 (Hz)	-17	.23	—	—	3.6	.050

Table 7.1: Acoustic differences of voiced (VD) and voiceless (VL) stops, after transformation and normalization as described in the text. Each cell gives the mean value for voiced stops minus the mean value for voiceless stops followed by the number of standard deviations between the means.

This is an unsupervised learning algorithm: the algorithm does not make use of the "correct" label of a token (voiced or voiceless) in order to determine the clustering. It only uses the acoustic measurements to group nearby tokens together. Only after the clustering has been computed do we look back at the correct labeling to determine which clusters line up with which labels, if any, and how accurate the clustering matches the labels.

The only problem with applying clustering outright to the acoustic data is that a straightforward Euclidean distance on acoustic data will cause the clustering algorithm to weigh dimensions with large absolute measures more heavily than dimensions with small absolute measures. For instance, frequency values are in the range of 0–3,000 Hz but RMS values here are in the range of 0–1. As a result, clustering will pay far more attention to the frequency dimensions and essentially no attention to the RMS dimensions. The measurements must all be rescaled before starting clustering. A simple method is to transform all measurements to z-scores. What appears to work slightly better in practice is to scale the dimensions according to each's estimate in a linear regression with the feature dimensions predicting the classification label (Mark Liberman, p.c., who attributed the idea to Lyle Ungar). In other words:

```
model <- lm(cbind(labels, features))
for (c in names(features)) features[c] <- features[c] * model$coef[[c]]
```

Of course, this comes at the expense of the algorithm no longer being unsupervised since it is provided with the correct answer. But where the bulk of the work is being done in classifying the tokens — in the actual clustering — the method is still unsupervised.

The result of the clustering with $k = 2$ is a contingency table such as the hypothetical one below:

		k-means	
		A	B
True	Voiced	50	100
	Voiceless	90	60

In this example, cluster A lines up with voiceless tokens and cluster B with voiced tokens. There is no reason that the clusters should have come out this way — it is a matter of the algorithm's random start. In order to judge the accuracy of the clustering, each cluster is assigned to its majority *true* label (A → Voiceless, B → Voiced):

		Cluster	
		Voiced	Voiceless
True	Voiced	100	50
	Voiceless	60	90

and then accuracy is computed in the normal way ($(100 + 90)/(100 + 50 + 60 + 90) = 0.63$).

Multiple Clusters

The voiced/voiceless discrimination task is most easily thought of as a binary discrimination task. A token is either voiced or voiceless. Assuming this is true in the adult language, it might not be true in the language of the infant who is susceptible to both over- and under-generalization. Over-generalization is the more well-known phenomenon where a language learner applies a rule in more contexts where it is appropriate — because conditioning environments have not been separated enough on the relevant dimensions. Under-generalization is the inverse case, where a language learner applies a rule in too few cases because *too many* distinctions among conditioning environments have been made.

One way in which the voicing contrast could be undergeneralized is if not all of the pairs in the paradigm were distinguished by the same underlying feature. For instance, /p,b/ might hypothetically be distinguished by feature A and /k,g/ by feature B, each with separate acoustic correlates. A could be [voice] whereas B could be [tense/lax]. Or A could include a preceding vowel duration difference while B does not. Recall that in Chapter 4 I argued against a redundant feature model of voice precisely because it undergeneralizes in this way, allowing for different segments to make use of different correlates of the voicing contrast without there being a phonological representation of the entirety of the voice contrast. Though I argued against the redundant feature model, under-generalization is certainly a plausible state of being, especially for the language learner, and may be difficult to detect. In Chapter 6 it was not possible to examine this particular case in the developing children. Although the different places and manners of articulation of the contrastively voiced segments showed differences in acoustic correlates, these differences were also confounded by the point during language development where the tokens occurred (which independently is affecting factors such as PVD).

Assume the language user has undergeneralized the voice feature — be it a language learner or an adult speaker. In this case, two clusters are not enough. Voice perception is not a binary classification task but a multi-way classification task, i.e. $k > 2$. There is no difficulty in running the machine classifier with larger values for k; however, judging accuracy requires an extra logical step. When $k = 2$, judging accuracy requires us to map our two *a priori* categories to the two categories yielded by the classifier, as described above. This is done in the same spirit as when a phonologist applies a model to a set of natural observations: observations do not come pre-labeled with linguistic terminology. But in the case of $k > 2$, we have no *a priori* k-way labeling to compare the results of the classifier to. Instead, we can give the learner the benefit of the doubt that the clusters will still map sensibly onto the binary distinction and measure accuracy in the same manner as when $k = 2$, by comparing each label to the majority vote of its cluster.

Results

The clustering was run with 8-fold cross-validation, meaning it was run eight times, each time using 7/8ths of the data set for clustering (training) and then testing the accuracy on the left-out eighth. During testing, each token was assigned to the cluster with the nearest center (medoid) using Euclidian distance. The mean accuracy across the eight runs is reported. This process was repeated separately for each of the two audio-book corpora, and for each separately on the pre-stress CV, post-stress VC, and post-stress VCV tokens.

Some of the results are reported in Figure 7.2. In all but the case of the Pride and Prejudice corpus, binary clustering improved on baseline accuracy by at least 12 percentage points. The maximum accuracy was 83% for the Around the World VC tokens, where the baseline was 63%. (Baseline accuracy was the accuracy when all tokens were labeled with the most frequent label.) Accuracy continued to improve with additional clusters. From 2 to 8 clusters, accuracy improved by as much as 15 percentage points (for the PP CVs).

Corpus		N Tokens	Baseline	$k = 2$	$k = 3$	$k = 4$	$k = 8$
AW	CV	4,981	54%	66%	70%	73%	74%
AW	VC	436	63%	83%	86%	86%	86%
AW	VCV	348	60%	78%	81%	79%	83%
PP	CV	2,612	62%	57%	71%	70%	72%

Table 7.2: Results of clustering for $k = 2, 3, 4, 8$.

In the case of CV tokens in the AW corpus, it is interesting to look at what the clusters are when $k = 3$. This tells us something about the distribution of the acoustic properties. In a typical application of the clustering algorithm to one of the cross-validation subsets, we find two clusters with primarily voiced tokens and one cluster with primarily voiceless tokens. This seems to indicate that voicing has a less spherical distribution in acoustic space than non-voicing. The two voicing clusters fall into a less-voiced/more-voiced division. One cluster has the lowest aspiration, the highest pitch, and the lowest following vowel F_0, while the other has greater aspiration (but not as great as in the voiceless cluster), low pitch (in fact, the cluster medoid has no pitch like the voiceless cluster medoid), and higher following vowel F_0 (but again not as high as in the voiceless cluster). This second voiced cluster is, in a sense, the devoiced voiced tokens: those lacking phonetic voicing and reduced voice characteristics in other dimensions, but still somewhat separable from the phonologically voiceless tokens.

7.1.5 Cluster Merging

The improved accuracy when $k > 2$ indicates that voiced and voiceless tokens do not separate well into essentially two spherical regions in (scaled) acoustic space. Instead, the tokens form a more complex and irregular distribution which is better captured by additional clusters. That said, the learner has other tools at his disposal for avoiding unbounded undergeneralization without losing the benefit of classifying according to multiple clusters. If we provide the learner with additional information that it can use to divide the clusters into two groups, it may be able to have the best of both worlds.

Methods

The learner presumably knows that some tokens come from the same *type* (i.e. phonological/lexical class). In other words, multiple tokens of the /t/ in 'cat' must have the same label. The /t/ in 'cat' is a *type*. There could be any number of ways the learner might make use of this, but in this section a novel cluster merging technique has been employed based on the idea that if a type overlaps two clusters those clusters probably both represent voiced tokens or both represent voiceless tokens. Here we repeatedly merge pairs of clusters until only two remain. In each step, the pair of clusters merged is that pair with the highest overlap in *types*. Overlap is computed in terms of pointwise mutual information between the two clusters, where the "probability of occurrence" of a cluster is the probability that a type drawn at random is assigned to the cluster by drawing one of its tokens at random and finding the cluster with the nearest center. In terms of PMI, this is formalized as follows:

$PMI(x, y) = log\frac{p(x,y)}{p(x)p(y)}$
$p(x) = \sum_{\tau} p^{\tau}(x)$
$p(x, y) = \sum_{\tau} p^{\tau}(x)p^{\tau}(y)$
$p^{\tau}(x) = \frac{|\tau \cap x|}{|\tau|}$

where x and y are cluster and τ is a type. $|\tau|$ denotes the number of tokens observed of the type; $\tau \cap x$ denotes the set of tokens of type τ in the cluster x. To reduce k clusters to $k - 1$, on the order of k^2 pointwise mutual information computations must be performed, and as the number of tokens and types increases this becomes computationally prohibitive quickly.

Results

Again the clustering method was run with 8-fold cross-validation on each of the two corpora, and for each separately on the pre-stress CV, post-stress VC, and post-stress VCV tokens. The cluster merging approach does show a useful benefit to the learner, at least in some cases. In the corpus set that had the largest number of tokens and types, the accuracy of the classifier increased four percentage points from two clusters (the normal binary classification) to three clusters (one step of cluster merging). Unfortunately, adding additional clusters did not appreciably help and in many cases reduced accuracy. All of the four tests reported in Figure 7.3 show an improvement from two to three clusters (for the AW corpus) or two to four clusters (for the PP corpus). For the Pride and Prejudice CVs, while binary classification actually had an accuracy below the 62% baseline, four clusters with cluster merging had an accuracy of 70%. In no case was the accuracy of cluster merging better than that of undergeneralized clustering with $k = 8$, but cluster merging was fairly competitive.

Corpus		N Tokens	N Types	Baseline	$k = 2$	$k = 8$	Cluster Merging
AW	CV	4,981	959	54%	66%	74%	70%
AW	VC	436	230	63%	83%	86%	86%
AW	VCV	348	202	60%	78%	83%	81%
PP	CV	2,612	607	62%	57%	72%	70%

Table 7.3: Results of simple clustering with $k = 2, 8$ and cluster merging.

7.1.6 Support Vector Machine

To establish a rough upper bound on the accuracy of the classification task, a supervised learning algorithm — the support vector machine — was applied to the same acoustic data. (In a supervised machine learning method, the algorithm is given the correct labeling to determine the best slice through acoustic space that divides the two categories. Language learning is thought to be primarily unsupervised because of children's apparent lack of reliance on being explicitly told the correct way to speak.) The svm function in the e1071 R package was used. The accuracy of SVM classification is given in Figure 7.4. The SVM classifier consistently performed much better than any of the unsupervised clustering methods, as expected, with an accuracy reaching 91% for the AW VCVs.

Corpus		N Tokens	Baseline	$k = 2$	$k = 8$	SVM
AW	CV	4,981	54%	66%	74%	86%
AW	VC	436	63%	83%	83%	88%
AW	VCV	348	60%	78%	86%	91%
PP	CV	2,612	62%	57%	72%	87%

Table 7.4: Results of simple clustering with $k = 2, 8$ and SVM-based classification.

7.2 Summary

Today — when even non-linguists probably have some experience with the limitations of speech recognition — it should not really be any surprise that automated classification of voicing in fluent speech should be so difficult. However, we can learn several things from the attempt.

Compared to Edwards's (1981) use of a single segmental context, we see here that the value to discrimination of the nine acoustic dimensions varies from context to context. While aspiration duration was by far the most reliable dimension in pre-stress CVs based on the number of standard deviations that separates the means of the voiced and voiceless tokens, it trades its position with closure RMS intensity and closure pitch in post-stress VCs. (And as opposed to Edwards (1981), following-vowel F_0 had little value here.) The variation of the acoustic properties across contexts and especially their limitation in discrimination presents a problem for a theory of phonology which reduces paradigms to acoustics through narrow conceptions of features such as [tense/lax] or [asp]. In order to maintain a paradigm such as voicing with differences in acoustic realization, a more flexible (or "algebraic") model of the phonology-phonetics interface must be introduced, as was discussed in Chapter 4.

On the other hand, while this chapter attempted to have classification benefit from using multiple acoustic dimensions simultaneously, it was not readily apparent that this had any benefit. One would think that a production equivalent of cue trading might ensure that when one cue is absent another cue is present, if not exaggerated, to help the listener make the distinction. As discussed earlier, while aspiration duration separates voiced and voiceless tokens overall, when considering just those tokens with zero closure pitch (essentially voiceless and devoiced tokens) aspiration duration no longer has any discriminative value. (It would seem to be the case that the voiced tokens that have zero closure pitch are also aspirated. Perhaps in the end these tokens were phonologically devoiced.) Likewise, out of the six sample groups only in two cases were there at least two acoustic dimensions that seemed to be simultaneously useful for discrimination. These were the AW VCs and VCVs, in which both closure RMS and closure pitch were relatively highly separable according to the differences in the means. In other cases only one dimension seemed to stand out from the rest.

The fact that aspiration duration reached such a high level of discriminative ability validates the method introduced in this chapter for determining the start of aspiration within an otherwise unlabeled segment. The method defined the burst location roughly as that location which maximized the change in RMS intensity across that point. Although no comparison to hand annotations was made, the method clearly worked well enough to be more useful for discrimination of CV tokens' voicing than any of the long-standing acoustic measurements included in the feature set. (If the burst were simply placed randomly within each segment, we would have instead seen both aspiration duration and its complement — closure duration — as having high discriminating value. But instead closure duration had weak discriminating value.)

As for the distributions themselves, we saw that they are not best represented by two spherical regions in acoustic space. Accuracy with simple clustering increased as the number of clusters increased, indicating that a more refined model of the mapping between acoustics and phonology is useful — and that perhaps learners or even adult speakers use an undergeneralized model of the voice paradigm which involves more than two categories. On the other hand, additional information beyond acoustics — in particular the type membership of tokens — may allow for language users to learn a refined, multi-cluster map of acoustic space that still maps onto a binary distinction.

Chapter 8

Summary and Concluding Remarks

On the surface this dissertation could be seen as coming from the field of acoustic phonetics or, perhaps, experimental developmental psychology. But all along what has been of interest has been the phonetic implementation of features in the phonetics-phonology interface. Despite the centrality of feature theory to phonetics and phonology, discussion on the nature of features at the interface has been mostly confined to the extremes: either that phonological features are in a direct one-to-one mapping to acoustic or articulatory dimensions (e.g. Jakobson and Halle 1956; Stevens et al. 1986) or that phonological features bear only an indirect relation to the next lower level in the linguistic system (e.g. Keating 1980, 1984; Vaux 1998). Related infant language development research has fallen into a similar division, with studies mostly focusing either on particular acoustic dimensions or on phonemic contrasts while taking their acoustics for granted. For instance, in the mispronunciation studies discussed in Chapter 3 the mispronunciation was either along a single dimension (vowel duration) or was a change from one naturally produced segment to another without regard to the acoustic differences.

When it comes to stop and fricative [voice] in English, there are a number of acoustic dimensions that vary reliably across the paradigm, as reviewed in Chapter 2. Some of these dimensions encode targets that are a part of linguistic competence: certainly the state of the glottis (e.g. VOT) and perhaps closure duration in stops and preceding vowel duration, and possibly F_0 perturbation. Other dimensions seem to be a matter of linguistic performance only: the duration difference in fricatives and perhaps the effect on the F_1 of preceding monophthongs. I argued in Chapter 4 that the rich acoustic side to the [voice] paradigm must be encoded as a part of the specification of the [voice] feature itself, i.e. as its phonetic implementation or as the way it is passed from the phonological to the phonetic components of the grammar.

But while these dimensions are reliable in a laboratory setting, their usefulness for discriminating tokens is greatly reduced in fluent speech (Chapter 7). Aspiration duration was by far the most reliable dimension in pre-stress CVs, using standard deviations to guage the separability of the voiced and voiceless distributions. But in post-stress VCs, aspiration duration trades its position with signal intensity during closure and closure pitch. On the one hand, this supports the position that phonological features cannot be tied to individual acoustic dimensions. On the other hand, even if the phonetics-phonology interface allows for phonological features to map onto complex phonetic outputs, the fact that the ensemble of outputs depends on context remains just as unexplained as in a narrow model of features a la Jakobson and Halle (1956).

In Chapter 6, a corpus study was presented that investigated the phonetic realization of consonant [voice] in two English-learning infants aged 1;1–3;5. While preceding vowel duration has been studied before, the other correlates of post-vocalic voicing investigated here — preceding F_1, consonant duration, and closure voicing intensity — had not been measured before in infant speech. The study also made two methodological contributions. First, it was a large-scale corpus study. Although the final number of tokens used was only moderately large (around 1,000), they were drawn from a much larger pool using an automated process, and the full corpus was used to train a new forced alignment acoustic model specific to infant speech.

The second methodological contribution was the use of maximum likelihood estimation to fit Klatt models of segment duration based on his notion of "incompressibility". Klatt models strike a balance between the simpler and less empirically accurage multiplicative models and the more complex but linguistically implausible sums-of-products models. While a multiplicative model can be easily fit using linear regression, a Klatt model requires a more general solution technique. Maximum likelihood estimation proved to be effective to fit Klatt models, and it is especially useful for the experimental linguist because it can be used to compute standard errors for estimates (as with linear regression). The Klatt model for preceding vowel duration indicated a statistically significant result that the corresponding multiplicative model was not powerful enough to find.

Several results came out of the study. F_1 at the midpoint of a vowel preceding a voiced consonant was lower by roughly 50 Hz, which is in line with the effect found in adults. But while the effect has been considered most likely to be a physiological and nonlinguistic phenomenon, it actually appeared to be correlated in the wrong direction with other aspects of [voice], casting doubt on a physiological explanation. Some of the consonant pairs had statistically significant differences in duration and closure voicing. Additionally, a preceding vowel duration difference was found and additionally a preliminary indication of a developmental trend that suggests the preceding vowel duration difference is being learned.

Finally, two important aspects of [voice] were found through experimentation with unsupervised learning of [voice] categorization from multidimensional acoustic input. First, the voiced and voiceless tokens are best represented not as two spherical clusters in acoustic space. Rather, multiple clusters — especially for the voiced tokens — better captures the distribution of the tokens. This may reflect that the phonological representation of [voice] is undergeneralized (with respect to current phonological theory) or is likely to be undergeneralized by language learners (with respect to the adult language). On the other hand, information outside of the domain of acoustics can be used to correct for undergeneralization. Lexical/semantic information that groups tokens into types can be used to collapse a multi-cluster map of acoustic space into a binary classification that is almost competitive with a fully undergeneralized model in terms of classification accuracy.

Appendix A

Experiment 2 Word List

Below are the words used in the reading lists of Experiment 2, adult (adult-directed) speech production.

Monosyllabic real words:

spite	spied
chat	chad
thought	thawed
bout	bowed
leak	league
sack	sag
pick	pig
peck	peg
buck	bug
tap	tab
ape	Abe
cup	cub
rope	robe

Tautosyllabic real words:

neatness	needless
seatbelt	seedling
fraction	fragment
doctor	dogma
pectin	pegboard
crapshoot	crabmeat
optics	object

Heterosyllabic real words:

seater	cedar
catty	caddy
petal	pedal
coating	coding
vicar	vigor
backing	bagging
chucking	chugging
flocking	flogging
hokey	hoagie
sopping	sobbing
seaport	seabed
staples	stables
soapy	sober

Monosyllabic nonsense words:

geet	geed
jite	jide
zat	zad
fot	fod
spote	spode
jeek	jeeg
chack	chag
skik	skig
nuck	nug
pauk	paug
spap	spab
skop	skob
peip	peib
gup	gub
foup	foub

Tautosyllabic nonsense words:

geetmonk	geedmonk
jitehood	jidehood
zatback	zadback
fotful	fodful
spotestick	spodestick
jeekson	jeegson
chackpack	chagpack
skikmount	skigmount
nuckbon	nugbon
pauktill	paugtill
spapton	spabton
skoptrie	skobtrie
peipcat	peibcat
gupsnow	gubsnow
foupdram	foubdram

Heterosyllabic nonsense words:

geety	geedy
jiteing	jideing
zatting	zadding
fotins	fodins
spowtuck	spowduck
jeeker	jeeger
chackal	chagal
ziken	zigen
nuckist	nugist
paukam	paugam
spapale	spabale
skoping	skobing
peiper	peiber
gupomt	gubomt
foupest	foubest

Appendix B

Maximum Likelihood Estimation Function

The following is a new function for the R statistics program for performing maximum likelihood estimation, as described in Section 6.3.1. An example use of this function is given on page 70. The function reports the estimate, standard error, null hypothesis (as given by the user), and p-value for each of the model parameters.

In addition, an extra parameter called σ_ϵ is always estimated. This is an estimate of the standard deviation of the normal distribution that the error terms (i.e. residuals) of the model are assumed to be drawn from. The estimated value is probably not as important as verifying that its standard error is sufficiently small to ensure the model fit is good enough.

```
mle <- function(x, y, f, nullhyp=list(), getresids=F) {
  # Parameters
  # ---------------
  #
  # x are the predictors. y are the observed data values.
  #
  # f should be a function that implements a model, e.g.:
  #   f <- function(x, intercept=.1, slope=.5) {
  #      intercept + x*slope
  #   }
  # where x, the first argument, is copied from the main call
  # to maxlikest.
  #
  # The default values for the parameters are required in
  # the function definition and are used as starting values
  # for the nonlinear optimization procedure and as the
  # null hypothesis when computing a p-value (unless
  # it is overridden with e.g. nullhyp=list(intercept = 0),
  # or nullhyp=list(intercept = NA) to turn off the test).
  #
  # if getresids==T, then this returns the residuals against
  # the best fit model, otherwise the result of the model
  # fit is printed out immediately and a list of the estimates
  # of the parameters is returned, e.g.:
  #     list(intercept = .203, slope = .395)
  #
  # An additional parameter called "sigma_epsilon" is returned
  # which is the estimate of the standard deviation of the
  # normal distribution from which the error terms (i.e. residuals)
  # of the model are drawn from.
  #
```

```
# Here is an example to fit a linear model:
# mle(x=c(0, 1, 2, 3), y=c(1, 3.1, 5.1, 7.1), function(x, m=1, b=1) { m*x + b });

# Ok then, here's the code:

# get the parameters we are estimating as a list, chopping
# off the first parameter which is the 'x' argument.
xparamname = names(formals(f))[1]
paramslist <- formals(f)[2:length(formals(f))]

build_param_list <- function(p) {
  # names of list arguments to f
  pp = list()

  # add the x argument
  pp[[xparamname]] = x

  # convert vector of args into named list for calling f,
  # except for p[length(p)] which is the implicit parameter
  # epsilon_sigma, which we strip off before calling f because
  # f doesn't know about it.
  for (pn in 1:length(paramslist))
    pp[names(paramslist)[pn]] = p[pn]

  return(pp);
}

# negloglik defines a function that returns the negative
# of the log likelihood of the parameters p given the
# residuals computed from the predicted values returned
# by f.
negloglik <- function(p) {
  residuals <- y - do.call(f, build_param_list(p)); # get residuals

  # we added an implicit parameter for the standard deviation
  # of the residuals
  residual_standard_deviation = p[length(p)];

  # return -log likelihood
  -sum(dnorm(residuals,
    sd=residual_standard_deviation, log=T))
};

# Use default value of f's arguments as initial values
for (p in paramslist)
  if (missing(p))
    stop("Provide a default value for each parameter in the model.\n");
initial_values <- sapply(paramslist, eval.parent)

# Add an implicit residual standard deviation variable that
# we will also estimate and report to the user, even though
# the user's function won't know about it. Here we add the
# variable and its initial guess.
initial_values = c(initial_values, list("σ_ε"=1));

# Maximize -log likelihood using the nlm function of R. While
# it is at it, get the Hessian matrix, which is used to compute
# standard errors later.

# nlm doesn't seem to work as well as optim. The hessian
# sometimes comes back singular when optim seems to
# be okay with it.
#solution <- nlm(negloglik, initial_values, hessian=T);
#if (solution$code > 2) { cat("...Solution Not So Good (code > 2)...\n") ; }
#estimate <- solution$estimate;
```

```
  solution <- optim(initial_values, negloglik, hessian=T, method="BFGS");
  #if (solution$code > 0) { cat("...Solution Not So Good (code > 0)...\n") ; }
  estimate = solution$par;

  # Get the residuals for these parameter estimates. First strip
  # off the residual standard deviation estimate because that's
  # not an argument to f. Then call f (see above) with the estimates
  # to get the predicted values, and subtract from y to get residuals.
  residuals = y - do.call(f, build_param_list(estimate));
  if (getresids) return (residuals); # get residuals and return, no output

  # Standard errors are the square roots of the diagonals of the
  # inverse of the Hessian matrix.
  stderr <- diag(solve(solution$hessian))^.5;

  # Get the null hypothesis for each parameter. If nullhyp is unspecified,
  # use the initial value as the null hypothesis. Otherwise, grab it from nullhyp.
  null = vector(mode="numeric", length=length(estimate))
  for (x in 1:length(names(paramslist))) {
    if (names(paramslist)[x] %in% names(nullhyp))
      null[x] = nullhyp[[ names(paramslist)[x] ]]
    else
      null[x] = initial_values[x]
  }

  # Clear out the null hypothesis for the standard deviation of the
  # residuals since we're not doing hypothesis testing on it.
  null[length(estimate)] = NA

  # Compute p value based on a two-tailed test.
  pval = 2*pnorm(-abs(unlist(estimate)-unlist(null)), sd=stderr);

  # These are the computed estimates of the parameters. Add
  # names to the estimates for row labels of the final output.
  names(estimate) <- names(initial_values);

  # Output estimates, standard errors, null hypotheses, and p values.
  print(cbind(estimate, stderr, null, pval), digits=3);

  # Output r^2.
  SStot = sum((y - mean(y)) ^ 2)
  SSerr = sum(residuals*residuals)
  rsquared = 1 - (SSerr / SStot)
  cat(paste("r^2 ", format(rsquared, digits=3), "\n"));

  # Output blank line.
  cat("\n");

  return(as.list(estimate));
}
```

Appendix C

Suggested Further Experiments

This dissertation raised many questions that could not be immediately answered but which could be addressed by further experimental work. I list here a summary of the questions raised in the hopes that this helps other linguists find interesting questions to answer. The questions are ordered according to the order of the dissertation.

1. Does the PVD effect occur across word boundaries — i.e. in vowel-final words followed by a consonant-initial word? This is interesting from the perspective of the question of whether the PVD is a phonologically productive process, although this would not answer that question.

2. It is important to revisit the claims for and against the compensatory adjustment explanations for the PVD effect. The reason, somewhat backwards, is that the best theoretical model for the effect — at least in my opinion — seems to be a gestural timing specification encoded in the [voice] feature. Is there any scale in which the vowel-consonant durations appear to maintain a constant sum?

3. A reevaluation of cross-linguistic comparisons of the PVD is necessary. Only in Laeufer (1992) was a fair comparison made between two languages (French and English). I have not read the original research on the PVD in Swedish (see Buder and Stoel-Gammon 2002) so it would be worthwhile to give that a second look and to replicate it. I also suspect that the PVD difference observed in Spanish is due to factors besides voice. Are there two types of PVD, or can all of the languages be classified either into the Polish/Czech category or the English/French category?

4. Danish and Hindi were claimed to have a reversed closure duration correlate of [voice] (i.e. longer for voiced consonants) and also a PVD effect. It would be useful to revist that literature (how large is the PVD effect?) and potentially replicate. These would make for especially interesting cases since syllables with a voiced coda would be lengthened throughout.

5. Huff (1980) did the novel experiment of having a word-final flap, which is a case that give language users good evidence of the underlying voice value of the flap. He found a vowel duration difference before flaps in this case but not in the case of word-internal flaps. It would be interesting to replicate the study with a larger sample and with speakers of a dialect that does not have short-a raising. It's especially interesting to look at the other correlates of [voice].

6. Studies are leaning toward universal following-vowel F_0 perturbation showing the pattern: voiced $\ll$ voiceless, aspirated $\ll$ unaspirated. But I noted that not all studies agreed.
7. Moreton's (2004) "Pre-Voiceless Hyperarticulation" has not been tested comprehensively for the whole of the vowel space.
8. Stevens and Klatt (1974) suggested that aspiration might be a spectral target, linked to F_1. Is there evidence for this?
9. Infant phone inventories are not a good measure of phone fluency, given the disparate rates of occurence of phones in infant-directed speech. What would we get if we looked at infant phone production rates relative to their occurrence in infant-directed speech?
10. There were three possible interpretations of Ko et al. (2009). One could be ruled out or strenghened if we know whether infants stared longer at longer vowels.
11. I raised the idea of a compositional phonetics, where the acoustics of a segment are predicted based on what we know about the phonetic correlates of its individual features.
12. Are there differences in the PVD effect in Boston (observed to be a large PVD) and Maine (no effect observed, but the difference was not significant)?
13. Do the acoustic correlates of [voice] develop at the same time in all phonemes in the [voice] paradigm (i.e. the acoustics of [voice] are fully generalized) or is [voice] undergeneralized at first?
14. What does the acoustic space of [voice] look like in adult, fluent speech? So far it seems as if in any given context there is only one or perhaps two acoustic dimensions that are useful for discriminating [voice] and that the use of multidimensional input was not beneficial.

Bibliography

BARAN, JANE, AND HARRY N. SEYMOUR. 1976. The influence of three phonological rules of Black English in the discrimination of minimal word pairs. *Journal of Speech and Hearing Research* 19:467–474.

BAUM, SHARI R., AND SHEILA E. BLUMSTEIN. 1987. Preliminary observations on the use of duration as a cue to syllable-initial fricative consonant voicing in English. *J. Acoust. Soc. Am.* 82:1073–1077.

BELL-BERTI, FREDERICKA. 1975. Control of pharyngeal cavity size for english voiced and voiceless stops. *J. Acoust. Soc. Am.* 57:456–461.

VAN BERGEM, DICK R. 1993. Acoustic vowel reduction as a fucntion of sentence accent, word stress, and word class. *Speech Communication* 12:1–23.

BOBERG, CHARLES, AND STEPHANIE M. STRASSEL. 2000. Short-a in Cincinatti: A change in progress. *Journal of English Linguistics* 28:108–126.

BOERSMA, PAUL. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5:341–345.

BRAUNSCHWEILER, NORBERT. 1997. Integrated cues of voicing and vowel length in German: A production study. *Language and Speech* 40:353–376.

BROWMAN, CATHERINE P., AND LOUIS M. GOLDSTEIN. 1986. Towards an articulatory phonology. *Phonology Yearbook* 3:219–252.

BUDER, EUGENE, AND CAROL STOEL-GAMMON. 2002. American and Swedish children's acquisition of vowel duration: Effects of vowel identity and final stop voicing. *J. Acoust. Soc. Am.* 111:1854–1864.

CARNEGIE MELLON UNIVERSITY. 1998. Carnegie Mellon University pronouncing dictionary (cmudict).

CASTLEMAN, WENDY A., AND RANDY L. DIEHL. 1996. Effects of fundamental frequency on medial and final [voice] judgments. *J. Phonetics* 24:383–398.

CHAMBERS, J. K. 1973. Canadian Raising. *Canadian Journal of Linguistics* 18:113–35.

CHARLES-LUCE, JAN. 1992. The effects of semantic context on voicing neutralization. *Phonetica* 50:28–43.

CHEN, MATTHEW. 1970. Vowel length variation as a function of the voicing of the consonant environment. *Phonetica* 22:129–159.

CHOMSKY, N., AND M. HALLE. 1968. *The Sound Pattern of English*. New York, NY: Harper & Row.

CIERI, CHRISTOPHER, DAVID MILLER, AND KEVIN WALKER. 2004. The Fisher corpus: a resource for the next generations of speech-to-text.

CLEMENTS, G.N., AND P.A. HALLÉ. 2010. Phonetic bases of distinctive features: Introduction. *J. Phonetics* 38:3–9. Phonetic Bases of Distinctive Features.

CRYSTAL, THOMAS H., AND ARTHUR S. HOUSE. 1988a. Segmental durations in connected-speech: Current results. *J. Acoust. Soc. Am.* 83:1553–1573.

CRYSTAL, THOMAS H., AND ARTHUR S. HOUSE. 1988b. Segmental durations in connected-speech: Syllabic stress. *J. Acoust. Soc. Am.* 83:1574–1585.

DAVIS, STUART, AND W. VAN SUMMERS. 1989. Vowel length and closure duration in word-medial VC sequences. *J. Acoust. Soc. Am.* 85:S28.

DEMUTH, K., J. CULBERTSON, AND J. ALTER. 2006. Word-minimality, epenthesis, and coda licensing in the acquisition of English. *Language & Speech* 49:137–174.

DENES, P. 1955. Effect of duration on the perception of voicing. *J. Acoust. Soc. Am.* 27:761–764.

DIETRICH, CHRISTIANE. 2006. The acquisition of phonological structure: Distinguishing contrastive from non-contrastive variation. Doctoral Dissertation, Max Planck Institute for Psycholinguistics.

DIETRICH, CHRISTIANE, DANIEL SWINGLEY, AND JANET F. WERKER. 2007. Native language governs interpretation of salient speech sound differences at 18 months. *Proceedings of the National Academy of Sciences* 104:16027–16031.

DISIMONI, FRANK G. 1974. Influence of consonant environment on duration of vowels in the speech of three-, six-, and nine-year-old children. *J. Acoust. Soc. Am.* 55:362–363.

DOWNING, LAURA, AND BRYAN GICK. 2001. Voiceless tone depressors in Nambya and Botswana Kalang'a. *BLS 27, February 16–18, 2001* .

EDWARDS, THOMAS J. 1981. Multiple features analysis of intervocalic English plosives. *J. Acoust. Soc. Am.* 69:535–547.

EILERS, REBECCA. 1977. Context-sensitive perception of naturally produced stop and fricative consonstants by infants. *J. Acoust. Soc. Am.* 61:1321–1336.

EIMAS, P. D., E. R. SIQUELAND, P. JUSCZYK, AND J. VIGORITO. 1971. Speech perception in infants. *Science* 171:303–306.

ESPOSITO, ANNA, AND MARIA GABRIELLA DI BENEDETTO. 1999. Acoustical and perceptual study of gemination in Italian stops. *J. Acoust. Soc. Am.* 106:2051–2062.

van der Feest, Suzanne V.H. 2007. Building a phonological lexicon: The acquisition of the Dutch voicing contrast in perception and production. Doctoral Dissertation, Radboud Universiteit Nijmegen.

Fergusun, Thomas S. 1996. *A course in large sample theory*. Roca Raton: CRC Press.

Fischer, Rebecca M., and Ralph N. Ohde. 1990. Spectral and duration properties of front vowels as cues to final stop-consonant voicing. *J. Acoust. Soc. Am.* 88:1250–1259.

Fitch, W. Tecumseh, and Jay Giedd. 1999. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *J. Acoust. Soc. Am.* 106:1511–1522.

Flege, James Emil, and W.S. Brown, Jr. 1982. The voicing contrast between English /p/ and /b/ as a function of stress and position-in-utterance. *J. Phonetics* 10:335–345.

Flege, James Emil, and James Hillenbrand. 1986. Differential use of temporal cues to the /s/–/z/ contrast by native and non-native speakers of English. *J. Acoust. Soc. Am.* 79:508–517.

Flege, J.E., and R. Port. 1981. Cross-language phonetic interference: Arabic to English. *Language and Speech* 24:125–146.

Fox, Robert, and Dale Terbeek. 1977. Dental flaps, vowel duration, and rule ordering in American English. *J. Phonetics* 5:27–34.

Fuchs, Susanne. 2005. Articulatory correlates of the voicing contrast in alveolar obstruent production in German. Doctoral Dissertation, Centre for General Linguistics (ZAS), Berlin and Queen Margaret University College (QMUC), Edinburgh.

Gandour, J., B. Weinberg, and D. Rutkowsky. 1980. Influence of postvocalic consonants on vowel duration in esophageal speech. *Language Speech* 23:149–158.

Gay, Thomas. 1968. Effect of speaking rate on diphthong formant movements. *J. Acoust. Soc. Am.* 44:1570–1573.

Hawkins, Sarah, and N.Noël Nguyen. 2004. Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English. *J. Phonetics* 32:199 – 231.

Hazan, Valerie, and Sarah Barrett. 2000. The development of phonemic categorization in children aged 6–12. *J. Phonetics* 28:377–396.

Hillenbrand, James, Dennis R. Ingrisano, Bruce K. Smith, and James E. Flege. 1984. Perception of the voiced-voiceless contrast in syllable-final stops. *J. Acoust. Soc. Am.* 76:18–26.

House, Arthur S., and Grant Fairbanks. 1953. The influence of consonant environment upon the secondary acoustical characteristics of vowels. *J. Acoust. Soc. Am.* 25:105–113.

Huff, C. 1980. Voicing and flap neutralization in New York City. *Research in Phonetics* 1:233–256.

Idsardi, William J. 2006. Canadian Raising, opacity, and rephonemicization. *Canadian Journal of Linguistics* 51:119–126.

ISHIHARA, SHUNICHI. 1998. Independence of consonantal voicing and vocoid F0 perturbation in English and Japanese. *Proceedings of the 5th International Conference on Spoken Language Processing* .

JACEWICZ, EVA, ROBERT A. FOX, AND JOSEPH SALMMONS. 2007. Vowel duration in three American English dialects. *American Speech* 82:367–385.

JAKOBSON, ROMAN, AND MORRIS HALLE. 1956. *Fundamentals of language*. The Netherlands: Mouton.

JANSEN, WOUTER. 2004. Laryngeal contrast and phonetic voicing: a laboratory phonology approach to English, Hungarian, and Dutch. Doctoral Dissertation, Rijksuniversiteit Groningen.

JANSEN, WOUTER. 2007. Phonological 'voicing', phonetic voicing, and assimilation in English. *Language Sciences* 29:270–293.

JESSEN, MICHAEL. 1998. *Phonetics and phonology of tense and lax obstruents in German*. John Benjamins.

JESSEN, MICHAEL. 2001. Phonetic implementation of the distinctive auditory features [voice] and [tense] in stop consonants. In *Distinctive feature theory*, ed. T. Alan Hall, volume 2 of *Phonology and Phonetics*, 237–294. Mouton de Gruyter.

JOHNSON, KEITH. 2003. *Acoustic and auditory phonetics*. Blackwell.

DE JONG, KENNETH. 2004. Stress, lexical focus, and segmental focus in English: Patterns of variation in vowel duration. *J. Phonetics* 32:493–516.

DE JONG, KENNETH, AND BUSHRA ZAWAYDEH. 2002. Comparing stress, lexical focus, and segmental focus: patterns of variation in Arabic vowel duration. *J. Phonetics* 30:53–75.

JOOS, MARTIN. 1942. A phonological dilemma in Canadian English. *Language* 18:141–144.

KAGAYA, RYOHEI, AND HAJIME HIROSE. 1975. Fiberoptic electromyographic and acoustic analyses of Hindi stop consonants. *Annual Bulletin. Research Institute of Logopedics and Phoniatrics. University of Tokyo.* 9:27–46.

KEATING, PATRICIA ANN. 1980. A phonetic study of a voicing contrast in polish. Doctoral Dissertation, Brown University.

KEATING, PATRICIA ANN. 1984. Phonetic and phonological representation of stop consonant voicing. *Language* 60:286–319.

KINGSTON, JOHN, AND RANDY L. DIEHL. 1994. Phonetic knowledge. *Language* 70:419–454.

KLATT, DENNIS H. 1973. Interaction between two factors that influence vowel duration. *J. Acoust. Soc. Am.* 54:1102–1104.

KLUENDER, KEITH R., RANDY L. DIEHL, AND BEVERLY A. WRIGHT. 1988. Vowel-length differences before voiced and voiceless consonants: an auditory explanation. *J. Phonetics* 16:153–169.

KO, EON-SUK. 2007. Acquisition of vowel duration in children speaking American English. In *Proceedings of Interspeech 2007*. Antwerp, Belgium.

KO, EON-SUK, MELANIE SODERSTROM, AND JAMES MORGAN. 2009. Development of perceptual sensitivity to extrinsic vowel duration in infants learning American English. *J. Acoust. Soc. Am.* 126:134–139.

KRAUSE, SUE ELLEN. 1982. Developmental use of vowel duration as a cue to postvocalic stop consonant voicing. *Journal of Speech and Hearing Research* 25:388–393.

LABOV, WILLIAM, SHARON ASH, AND CHARLES BOBERG. 2006. *The Atlas of North American English*. Mouton de Gruyter.

LAEUFER, CHRISTIANE. 1992. Patterns of voicing-conditioned vowel duration in French and English. *J. Phonetics* 20:411–440.

LEE, S., A. POTAMIANOS, AND S. NARAYANAN. 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *J. Acoust. Soc. Am.* 105:1455–1468.

LINDAU, MONA, AND PETER LADEFOGED. 1986. Variability of feature specifications. In *Invariance and variability in speech processes*. Lawrence Erlbaum Associates.

LISKER, LEIGH. 1957. Closure duration and the intervocalic voiced-voiceless distinction in English. *Language* 33:42–49.

LISKER, LEIGH. 1986. "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech* 29:3–11.

LISKER, LEIGH, AND ARTHUR S. ABRAMSON. 1964a. A cross-language study of voicing in initial stops: acoustical measurements. *Word* 20:384–422.

LISKER, LEIGH, AND ARTHUR S. ABRAMSON. 1964b. Stop categorization and voice onset time. In *Proceedings of the Fifth International Congress of Phonetic Sciences*, ed. E. Zwirner and W. Bethge. Karger.

LUCE, PAUL A., AND JAN CHARLES-LUCE. 1985. Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *J. Acoust. Soc. Am.* 78:1949–1957.

MADDIESON, IAN. 1984. *Patterns of sounds*. Cambridge: Cambridge University Press.

MAYO, C., AND A. TURK. 2005. The influence of spectral distinctiveness on acoustic cue weighting in children's and adults' speech perception. *J. Acoust. Soc. Am.* 118:1730–1741.

MAYOL, LAIA. 2009. Pronouns in catalan: information, discourse and strategy. Doctoral Dissertation, University of Pennsylvania.

MESSUM, PIERS RUSTON. 2007. The role of imitation in learning to pronounce. Doctoral Dissertation, University College London.

MORASSE, HÉLÈNE. 1995. Variations intrinsèques et co-intrinsèques de durée vocalique en français québécois. Doctoral Dissertation, Université du Québec à Chicoutimi.

MORETON, ELLIOTT. 2004. Realization of the English postvocalic [voice] contrast in F1 and F2. *J. Phonetics* 32:1 – 33.

MUGITANI, RYOKE, FERRAN PONS, LAUREL FAIS, CHRISTIANE DIETRICH, JANET F. WERKER, AND SHIGEAKI AMANO. 2009. Perception of vowel length by Japanese- and English-learning infants. *Developmental Psychology* 45:236–247.

NARAYAN, CHANDAN, KYLE GORMAN, AND DANIEL SWINGLEY. 2008. The microprosody of [voice] in infant- and adult-directed speech. Presented at LSA 2008–Chicago.

OHALA, JOHN J. 1997. Aerodynamics of phonology. *Proceedings of the 4th Seoul International Conference on Linguistics (SICOL)* 92–97.

OHDE, RALPH N. 1984. Fundamental frequency as an acoustic correlate of stop consonant voicing. *J. Acoust. Soc. Am.* 75:224–230.

PARKER, FRANK. 1977. Distinctive features and acoustic cues. *J. Acoust. Soc. Am.* 62:1051–1054.

PARNELL, MARTHA, JAMES D. AMERMAN, AND G. BEVERLY WELLS. 1977. Closure and constriction duration for alveolar consonants during voiced and whispered speaking conditions. *J. Acoust. Soc. Am.* 61:612–613.

PORT, ROBERT F. 1981. Linguistic timing factors in combination. *J. Acoust. Soc. Am.* 69:262–274.

PORT, ROBERT F., AND MICHAEL L. O'DELL. 1986. Neutralization of syllable-final voicing in German. *J. Phonetics* 13:455–471.

VAN ROOY, BERTUS, AND DAAN WISSING. 2001. Distinctive [voice] implies regressive voicing assimilation. In *Distinctive feature theory*, ed. T. Alan Hall, volume 2 of *Phonology and Phonetics*, 237–294. Mouton de Gruyter.

ROSEN, KRISTIN M. 2005. Analysis of speech segment duration with the lognormal distribution: A basis for unification and comparison. *J. Phonetics* 33:411–426.

VAN SANTEN, JAN P. H. 1994. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language* 8:95–128.

VAN SANTEN, JAN P. H., AND CHILIN SHIH. 2000. Suprasegmental and segmental timing models in Mandarin Chinese and American English. *J. Acoust. Soc. Am.* 107:1012–1026.

SCHISSEL, JAMIE. 2008. A preliminary investigation of word final voicing of /s/ and /z/ by l1 and l2 English speakers. Final project, LING 520 fall 2008, University of Pennsylvania, December 2008.

SCOBBIE, JAMES M., ALICE E. TURK, AND NIGEL HEWLETT. 1999. Morphemes, phonetics and lexical items: The case of the Scottish Vowel Length Rule. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, 1617–1220. San Francisco.

SHARF, D.J. 1964. Vowel duration in whispered and normal speech. *Language Speech* 7:89–97.

SIMON, CLAUDE, AND ADRIAN J. FOURCIN. 1978. Cross-language study of speech-pattern learning. *J. Acoust. Soc. Am.* 63:925–935.

SMITH, CAROLINE L. 1997. The devoicing of /z/ in American English: Effects of local and prosodic context. *J. Phonetics* 25:471–500.

SNOW, DAVID. 1997. Children's acquisition of speech timing in English: a comparative study of voice onset time and final syllable vowel lengthening. *J. Child Lang.* 24:35–56.

STEVENS, KENNETH N. 1998. *Acoustic phonetics*. MIT Press.

STEVENS, KENNETH N., SHEILA E. BLUMSTEIN, LAURA GLICKSMAN, MARTHA BURTON, AND KATHLEEN KUROWSKI. 1992. Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *J. Acoust. Soc. Am.* 91:2979–3000.

STEVENS, KENNETH N., SAMUEL JAY KEYSER, AND HARUKO KAWASAKI. 1986. Toward a phonetic and phonological theory of redundant features. In *Invariance and variability in speech processes*, ed. Joseph S. Perkell and Dennis H. Klatt, 426–463. Lawrence Erlbaum Associates.

STEVENS, KENNETH N., AND DENNIS H. KLATT. 1974. Role of formant transitions in the voiced-voiceless distinction for stops. *J. Acoust. Soc. Am.* 55:653–659.

STOEL-GAMMON, CAROL. 1985. Phonetic inventories, 15-24 months: A longitudinal study. *Journal of Speech and Hearing Research* 28:505–512.

STOEL-GAMMON, CAROL. 2002. Intervocalic consonants in the speech of typically developing children: emergence and early use. *Clinical Linguistics & Phonetics* 16:155–168.

SUH, CHANG-KOOK. 2001. Aspects of phonetics and phonology of Icelandic preaspiration. *Studies in Phonetics, Phonology, and Morphology* 7:63–83.

SUMMERS, W. VAN. 1987. Effects of stress and final-consonant voicing on vowel production: articulatory and acoustic analyses. *J. Acoust. Soc. Am.* 83:847–863.

SWINGLEY, DANIEL. 2009. Onsets and codas in 1.5-year-olds' word recognition. *Journal of Memory and Language* 60:252–269.

TAUBERER, JOSHUA, AND KEELAN EVANINI. 2009. Intrinsic vowel duration and the post-vocalic voicing effect: Some evidence from dialects of North American English. *Proceedings of Interspeech 2009* .

UMEDA, NORIKO. 1975. Vowel duration in American English. *J. Acoust. Soc. Am.* 58:434–445.

VANCE, TIMOTHY J. 1987. "Canadian Raising" in some dialects of the Northern United States. *American Speech* 62:195–210.

VAUX, BERT. 1998. The laryngeal specification of fricatives. *Linguistic Inquiry* 29:497–511.

VEATCH, THOMAS CLARK. 1991. English vowels: Their surface phonology and phonetic implementation in vernacular dialects. Doctoral Dissertation, University of Pennsylvania.

WARNER, NATASHA, ERIN GOOD, ALLARD JONGMAN, AND JOAN SERENO. 2006. Orthographic vs. morphological incomplete neutralization effects. *J. Phonetics* 34:285–293.

WEISMER, GARY, DANIEL DINNSEN, AND MARY ELBERT. 1981. A study of the voicing distinction associated with omitted, word-final stops. *Journal of Speech and Hearing Disorders* 46:320–328.

WIGHTMAN, COLIN W., STEFANIE SHATTUCK-HUNAGEL, MARI OSTENDORF, AND PATTI J. PRICE. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *J. Acoust. Soc. Am.* 91:1707–1717.

YOUNG, S. J., G. EVERMANN, M. J. F. GALES, T. HAIN, D. KERSHAW, G. MOORE, J. ODELL, D. OLLASON, D. POVEY, V. VALTCHEV, AND P. C. WOODLAND. 2006. *The HTK book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department.

YUAN, JIAHONG, AND MARK LIBERMAN. 2008. Speaker identification on the SCOTUS corpus. In *Proceedings of Acoustics '08*.

ZIMMERMAN, SAMUEL A., AND STANLEY M. SAPON. 1958. Note on vowel duration seen cross-linguistically. *J. Acoust. Soc. Am.* 30:152–153.