influences our ability to find systematic patterns in speech rate and pause durations is left here to empirical inquiry. The extent to which we can develop statistical models that can account for the data given a set of social and linguistic criteria will make much progress towards an answer.

The data for this and the remaining chapters come from transcribed recordings made from diverse sociolinguistic research projects on American English. These recordings, and their transcripts and other annotations, are housed in the Sociolinguistic Archive and Analysis Project (SLAAP), which was described at length in Chapter 3. I begin by providing an overview of the data and where they come from (§5.2). I then examine speech rate and pause duration variation in these data, first at a per-utterance level (§5.3) and then at a per-speaker level (§5.4). As I will explain, at the utterance level there is a large amount of data available (~30,000 tokens for each feature), and a valuable question to pursue from an interest in corpus-based or large-scale sociophonetics is whether all of these data actually yield insights that could not be gleaned from much smaller datasets (§5.5). Then, I end by assessing what these statistical analyses actually indicate in terms of our larger interest in sociolinguistic patterns in speech rate and silent pause (§5.6).

5.2 The data

The data for these analyses come from the transcribed portions of many interviews within the SLAAP archive. The transcripts within SLAAP are of widely varying lengths, ranging from short excerpts of a minute or so to complete interviews of over 70 minutes. For the main analysis here, I have selected English language, conversational interview speech from speakers in as many transcripts as possible coming from four separate areas of the US (North Carolina, Ohio, Texas, and Washington, DC). I have excluded speakers from the analyses who appear in the transcripts with very limited talk in English or who only appear in very short transcripts. In later chapters, I draw on some data from SLAAP not included in the main analyses here.

Further, it also must be mentioned that for some speakers used in these analyses we have only a single transcript of, say, five minutes from a single interview, while for others we have multiple transcripts spanning a number of interviews. The mean length for all transcripts used in these analyses is 14.2 minutes and the median is 10.0 minutes. The shortest is 1.8 minutes long (though this transcript is for a speaker for which there is also another, 15 minute long, transcript used). For some

speakers there are quite a lot of data available. The longest transcript is 74.8 minutes. For the interviewer in the Washington, DC recordings with adolescent African Americans, for example, we have over 12 hours of transcripts, with over 171 minutes of uttered talk – actual phonation – by the interviewer herself (not including silent pauses).[1] More important than the length of each transcript is the amount of talk available for each of the speakers within a transcript. It is of course the case that some speakers talk more than others, and this is especially the case in interviews with multiple interviewees. In this chapter, we examine the speakers independently from their interlocutors and the larger discourse and social context of the interviews. In a number of cases only some of the speakers contained in a transcript have been selected for analysis. The decision of which speakers to include in the analysis and which to exclude was most often determined based on the amount of talk by the speakers. The speaker with the least amount of analyzed talk had only 16 phonetic utterances, but the median number of utterances across all speakers is 102, the mean is 190. (Recall from Chapter 2 that numerous studies have drawn on very small samples of speech. I take up the question of how using different amounts of speech from the various informants affects the analysis in the next chapter.)

As is implied by my mention of the Washington, DC interviewer, in addition to examining the interviewees in the SLAAP recordings I am occasionally examining the interviewers as well. Nine of the 159 speakers examined in this chapter are interviewers rather than interviewees. There are benefits to treating the interviewers in a recording as "speakers" for analysis (cf. Hazen 2000b, Schilling-Estes 2004a, Kendall 2010b). There are also many instances in the SLAAP archive where the interviewers are in fact locals of the research site and seem appropriate for inclusion (these are the only interviewers included in the analyses of this chapter). Some of the time the interviewers are as talkative as the interviewees. Also, as will be explored in Chapter 6, looking at the interviewers' speech can tell us important things about the interviews as interactions.

Table 5.1 provides a summary of the demographic breakdown for the speakers examined here, with respect to ethnicity, sex, and region. There are four ethnicities represented in the data, which I label as African American, European American, Latino/a, and Lumbee. The Lumbee are a Native American tribe in southern North Carolina. The Lumbee lost their native language prehistorically but have been shown to speak a unique variety of Southern American English (Wolfram, Dannenberg, Knick, and Oxendine 2002). In regional terms, the speakers come from

four main areas of the US, the states of North Carolina, Ohio, and Texas, and the city of Washington, DC. North Carolina is divided in these data into four subregions, Western NC, Central NC, Eastern NC, and Southern NC. This subdivision was made for several reasons. First, the data from the other regions, such as Ohio and Texas, are not geographically dispersed within those regions, but rather come from fairly circumscribed locales. The Texas speakers come from a single research project conducted in a single fairly small town in southern Texas (Thomas and Ericson 2007, Wolford and Carter 2007, Kendall and Thomas 2010). The Ohio data come from two collections of interviews conducted by Erik Thomas (Moreton and Thomas 2007, Thomas 2010) throughout the state, but all of the interviews used here come from the part of Ohio within the Inland North dialect region (Labov et al. 2006). Thus, although Ohio can be problematic to characterize in terms of regional dialects, these data all come from northern Ohio and are taken here to represent the Inland North. Finally, the Washington, DC speakers all come from the same summer camp for teenagers of a specific background in the same city (see §7.3.2 and §8.6, and Mallinson and Kendall 2009, for a fuller discussion of the Washington, DC interviews).

Second, thanks to the extensive work of Walt Wolfram, Erik Thomas, and the past and present members of the NCLLP (cf. Wolfram and Schilling-Estes 1995, Wolfram et al. 2000, Hazen 2000a, Torbert 2001, Wolfram and Thomas 2002, Mallinson and Wolfram 2002, Carpenter 2005, Mallinson and Childs 2007), a huge amount of speech data is available from throughout North Carolina. Ninety-nine speakers are included in this analysis from communities throughout that state. With this many speakers available, we are able to more finely subdivide

Table 5.1 Speaker demographics

Region	African American		European American		Latino/a		Lumbee		*Totals*
	Female	Male	Female	Male	Female	Male	Female	Male	
Central NC	12	9	8	7	7	5	–	–	*48*
Eastern NC	4	5	3	3	–	–	–	–	*15*
Ohio	1	1	4	4	–	–	–	–	*10*
Southern NC	3	3	4	3	–	–	5	9	*27*
Texas	–	–	2	8	19	9	–	–	*38*
Washington, DC	10	2	–	–	–	–	–	–	*12*
Western NC	4	1	1	1	2	–	–	–	*9*
Totals	*34*	*21*	*22*	*26*	*28*	*14*	*5*	*9*	*159*

the region than we can for the other areas. Further, North Carolina's settlement history and current dialect variability would make lumping all 99 of these speakers a priori into a single dialect region problematic (Wolfram 1999). For instance, the coastal areas of eastern North Carolina have very different settlement histories and dialect backgrounds from, say, the Appalachian region of western North Carolina. Finally, and perhaps most importantly, the limited research on regional variation in speech rate (e.g. Ray and Zahn 1990, Jacewicz et al. 2009, 2010) has extrapolated from single communities larger regional patterns, such as "Western American English" versus "Southern American English." If it is the case that speech timing is regionally stable across individual communities and "subregions," we should see fewer differences within North Carolina than between North Carolina's subregions and, say, Ohio. For all of these reasons I have separated the North Carolina speakers into four separate regional groups.

Western NC includes speakers from the North Carolina mountain and foothill communities of Texana, Beech Bottom, and Hickory. Central NC includes speakers from central and northern North Carolina, including Raleigh, Durham, Princeville, and Warren County.[2] Southern NC is entirely comprised of speakers from the triethnic Robeson County, where the NCLLP conducted a large-scale survey in the 1990s (cf. Wolfram et al. 2002). Eastern NC includes speakers from Hyde County, Roanoke Island, and Wilmington, all communities on the Atlantic coast.

I am interested in evaluating the degree to which pause and speech rate may vary along social parameters, such as region, but I am not solely interested in the question of what regional differences exist. Better understanding, for example, the degrees to which pause and speech rate distributions may vary between geographically proximate locations (such as different areas within North Carolina) is an interesting question and, I think, potentially more important to answer than questions focused on whether Southerners and Northerners or men and women or Blacks and Whites talk faster or slower than one another. Readers may also wonder more broadly what motivated the use and definition of the "region" category at all (and the same could be said about the other social categories of ethnicity and sex). It would clearly be preferable, on the one hand, if there were enough data available from each specific community, to group speakers by specific community location, for example, instead of the broader (and admittedly somewhat arbitrary) "region" category or, on the other, to consider these data more robustly in terms of criteria such as "cultural orientation." As readers will see throughout these analyses, I am ultimately more interested in the

possibilities of explaining the data by social categories. The social categories used here, like region, are intended as useful heuristics. Whether these specific categories are somewhat arbitrary is less the point than discovering the extent to which sequential temporal features of talk pattern when examined from social vantage points.

Finally, it is clear from Table 5.1 that the data examined here are not well balanced across the social categories. The fact that all of the Washington, DC speakers are young African Americans, and all but two are female, and the fact that all of the Lumbee speakers examined come from Southern NC do present some confounds for analyzing and interpreting aspects of these data. These are noted when relevant, but for the most part do not cause problems for analysis. The statistical methods described in Chapter 4 and expanded later in this chapter are robust against data sparsity issues.

Figure 5.1 displays these speakers by age, organized by sex and ethnicity, with region indicated by the shape of the plotted symbols. Following common conventions, "F" stands for female and "M" stands for male. Ethnicity is labeled as "AA" for African American, "EA" for

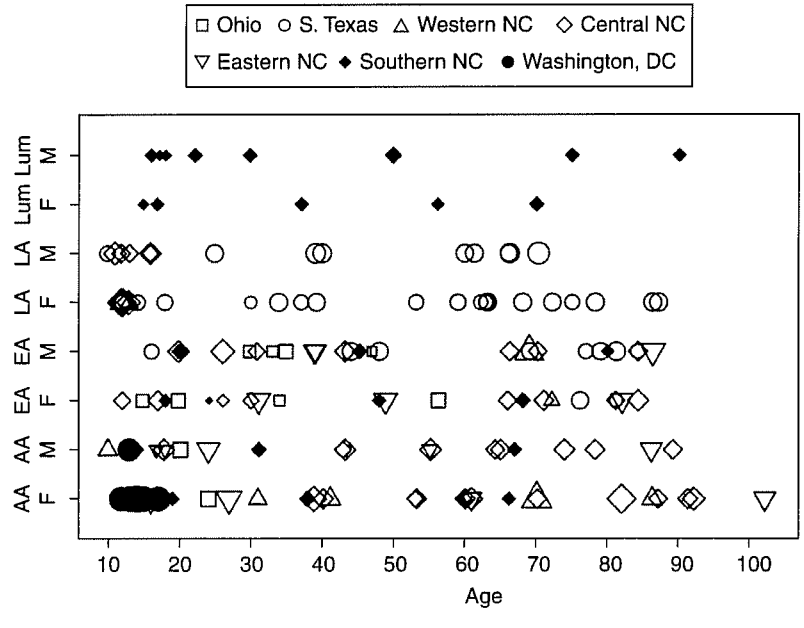

Figure 5.1 All speakers plotted by age

European American, "LA" for Latino/a, and "Lum" for Lumbee. The interviews examined here span 20 years of research, so it should be noted that a plot by actual year of birth would show a slightly different distribution than depicted in Figure 5.1.[3] The shapes of the plotting points are used throughout Part II to indicate speakers' regional affiliation. The size of the symbols reflects the (logged) average number of measurements (articulation rate N + pause duration N / 2) available for each speaker. The full information for all of these speakers – including central tendencies for pause and speech rate – can be found in Table 5a on the book's website.

Since these data come from diverse sociolinguistic projects, a number of different speaker-naming conventions have been used in the archive. Thus, I primarily refer to the speakers by an alphanumeric code. This is based on the first two or three characters of the speakers' identifier in SLAAP hyphenated with their age. The alphanumeric codes are used primarily to regularize the labeling of the speakers. Occasionally, I use the speakers' full identifications as they are given in SLAAP or a completely different identifier. Only in the case of some of the interviewers or fieldworkers are SLAAP's identifiers possibly a person's real name. All other "names" are pseudonyms. Many speakers are referred to by initials in SLAAP but even these may be pseudonymous or simply codes used in the original research.

5.3 Modeling speech rate and pause durations at the measurement level

In this section we examine the speech rate and pause duration measurements at the level of the individual measurements. Practically speaking, this approach follows closely the mixed-effect modeling approach that was described in the previous chapter. Models for articulation rate (not speaking rate) and pause (log-)duration are developed to determine just how well the available social factors can account for variability in the data, at the utterance (for articulation rate) and pause (for pause durations) level. This analysis makes use of a large dataset – about 30,000 measurements for each dependent variable – and a part of the interest here, in terms of this project's methodological contributions to corpus-based large-scale (socio)phonetic analysis, is asking just how useful it really is to have all these data at hand for analysis. In §5.4, we will look at these same data at a per-speaker level, with each speaker's median articulation rate and pause duration as the dependent variables. Which view tells us more about social variation in speech timing?

5.3.1 Speech rate at the utterance level

A total of 30,136 measured utterances are available for the 159 speakers examined here. While this yields an average of 190 tokens per speaker (each utterance contributes one measurement token to the data), the actual contributions by the speakers are much more variable. The median *N* for the talkers is 102, with a minimum *N* of 16 (a Mexican American from south Texas). The highest *N*s come primarily from the African Americans from Washington, DC (maximum *N* of 1939), whom we will examine at greater length in the following chapters, thanks to the large amount of data available for them.

Similar to the brief analysis of the last chapter, we focus our primary attention on the available social factors but also include the important nonsocial factor of the number of syllables in each utterance.[4] The following factors are considered:

- The number of syllables in the utterance (NUMSYLS; continuous, with a mean of 6.96 σ)
- The speaker's ETHNICITY (one of African American, European American, Latino/a, or Lumbee)
- The speaker's SEX (female or male)
- The speaker's AGE (continuous, with a mean value of 36.1 years old)
- And, the speaker's REGION (one of Central NC, Eastern NC, Ohio, Southern NC, Texas, Washington DC, or Western NC)

A summary of the mean articulation rates for these factors is shown in Figure 5.2. This view, similar to a cross-tabulation, but for a continuous dependent variable gives us a rough sense of the variability in the data.[5] It is important to keep in mind that each line in Figure 5.2 does not account for the influence of the other factors and that the plot is showing mean values for each factor without other indications of the distribution for the given factor. (In §4.5.1, I reviewed the raw data primarily using boxplots and gave a more detailed view of the data before modeling since that was our first look at speech rate and pause data. I do not do this here as there are more potentially relevant factors and I believe the summary plot in Figure 5.2 provides a thorough overview of the raw data.) Before proceeding to the statistical analysis, it is worth making some brief observations about the articulation rate data from this summary and I will now do this for each factor in turn.

First off, from Figure 5.2, we see indications of a strong effect of utterance length, NUMSYLS, on articulation rate. Utterances with between one and four (not including four) syllables have a mean articulation rate of

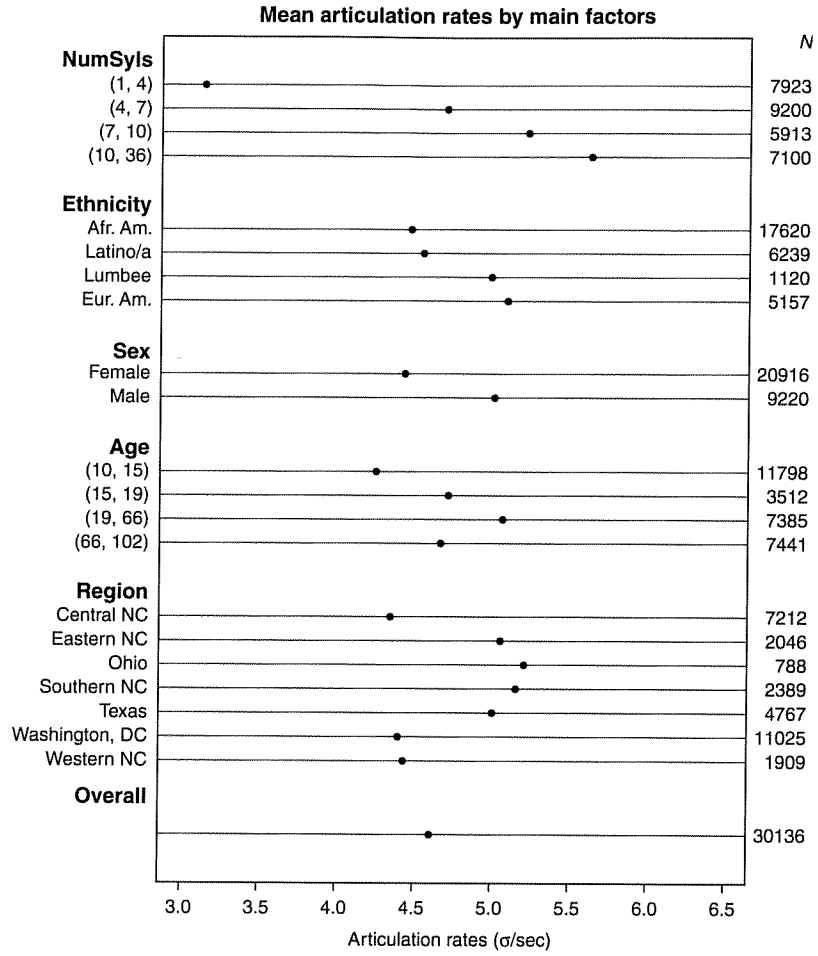

Figure 5.2 Mean utterance articulation rates by main factors

just above 3 σ/sec while longer syllable lengths and all other factors have rates above 4.25 σ/sec. In fact, utterances four syllables and longer have much faster rates, at 4.70 σ/sec, with articulation rates becoming even faster for longer utterances. This appears in line with expectations based on our knowledge of both phenomena of phrase-final lengthening and compensatory shortening. Short utterances should show a larger influence of the lengthened (i.e. slowed) final syllable or syllables, since those syllables likely comprise a larger proportion of the utterance. This is examined

to some length later, in §6.4. Compensatory shortening, whereby syllable durations are reduced as the number of syllables increase, would also be in line with the increasing rates of longer utterances indicated here (as in Quené 2008). Finally, in view of the fact that many one- and two-syllable utterances are comprised of single-word responses (e.g. "Yeah."), discourse markers (e.g. "Well"), and filled pauses (e.g. "Um"), we would further expect that these short utterances are elongated as a part of their discourse function. (Recall from Chapter 2, many previous studies prune these kinds of utterances from their analysis.)

For the social factors of ETHNICITY, SEX, AGE, and REGION we first note that the data are imbalanced; we have many more tokens, for example, by young speakers and by females (in particular by the young African American females from Washington, DC) than by other social categories. I will attempt to show that this is not a problem for the analysis (thanks to modern statistical techniques, the large overall amount of data available, and to the stability of the measures across the data; see §6.2). We also notice, for ETHNICITY, SEX, and REGION, that there appears to be somewhat of a clustering of the mean articulation rates – rates for African Americans and Latinos/as, females, and speakers from Central NC, Western NC, and Washington, DC are around 4.5 σ/sec while rates for Lumbees and European Americans, males, and the other regions cluster around 5 σ/sec. For AGE, we note indications of a curvilinear pattern. The youngest speakers exhibit the lowest mean articulation rates while the adults aged 19–66 show the fastest rates. For the oldest speakers, articulation rates slow again. Overall, the 30,136 tokens yield a mean articulation rate of 4.6 σ/sec.

We now assess the results of mixed-effect linear regressions on the data. (Mixed-effect modeling, and my general approach to building these models, was discussed in Chapter 4.) In the mixed-effect models presented here, intercepts (the baseline articulation rate values) and slopes for the NUMSYLS effect are allowed to vary for each individual. These are the random effects, while factors outlined above and in Figure 5.2 are the fixed effects. Results from the best model of the articulation rate data are presented in Table 5.2.[6] The displayed model was generated by trimming the data to remove outliers (removing 1.8 percent of the data, leaving 29,600 measurements) after the determination of the best model. As in Chapter 4, outliers were determined from the standardized residuals of an initial regression and were those more than 2.5 standard deviations from zero (cf. Baayen 2008: 256–8).

Figure 5.3 provides a graphical representation of the fixed effects of the model. Note that the *y*-axis is scaled differently for the NUMSYLS effect,

Table 5.2 Best mixed-effect model for (trimmed) utterance-level articulation rates

Factor	Estimate	Std. err.	*p*
(Intercept)	−0.336	0.314	–
NUMSYLS	1.043	0.012	0.0001
NUMSYLS'	−5.692	0.118	0.0001
NUMSYLS"	15.069	0.422	0.0001
NUMSYLS"'	−10.205	0.414	0.0001
AGE	0.065	0.018	0.0002
AGE'	−2.668	0.725	0.0002
AGE"	2.881	0.785	0.0002
REGION = Eastern NC[7]	0.228	0.111	0.0288
REGION = Ohio	0.262	0.140	0.0454
REGION = Southern NC	0.238	0.118	0.0372
REGION = Texas	0.284	0.104	0.0036
REGION = Washington, DC	0.118	0.146	[0.3650]
REGION = Western NC	0.047	0.132	[0.6790]
SEX = Male	0.333	0.104	0.0010
ETHNICITY = Latino/a	0.219	0.136	[0.0788]
ETHNICITY = Lumbee	−0.096	0.205	[0.6280]
ETHNICITY = European Am.	0.345	0.110	0.0008
ETHNICITY = Latino/a x SEX = Male	0.023	0.157	[0.8588]
ETHNICITY = Lumbee x SEX = Male	0.005	0.234	[0.9704]
ETHNICITY = Eur. Am. x SEX = Male	−0.398	0.151	0.0042

$R^2 = 0.647$.

which has a much larger effect range, than for the other predictors. We find significant main effects for all of the available factors – NUMSYLS (modeled with a five-knot spline[8]), AGE (modeled with a four-knot spline), REGION, SEX, and ETHNICITY – and a significant interaction for ETHNICITY x SEX. These are discussed in turn, beginning with the social factors.

The Central NC speakers, the baseline factor for REGION, have the slowest rates, followed by Western NC and Washington, DC, both of which, however, are not significantly different than Central NC. The speakers from Eastern NC, Southern NC, Ohio, and then Texas, are increasingly faster (in that order) and are significantly different from the Central NC speakers (*p* from 0.045 for Ohio to 0.004 for Texas). This confirms the visual clustering in the summary display of Figure 5.2 – speakers from Central NC, Western NC, and Washington, DC are slower than the speakers from the other regions. Males are significantly faster than females (estimated as 0.33 σ/sec faster, $p = 0.001$), an outcome that also confirms the visual summary from earlier.

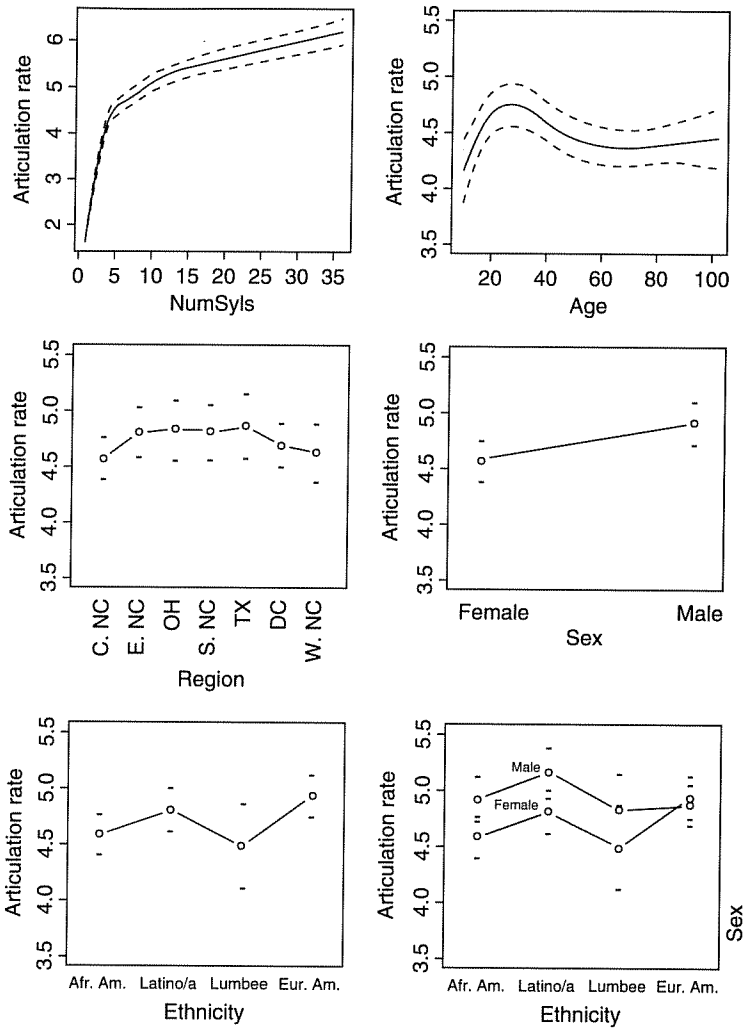

Figure 5.3 Effects in the mixed-effect model for articulation rates

For ETHNICITY, the baseline factor of African Americans is 0.10 σ/sec faster than the Lumbees, and 0.22 σ/sec slower than Latinos, although both of these differences are nonsignificant. European Americans are the fastest speakers, 0.35 σ/sec faster than African Americans (p = 0.0008). The significant interaction between ETHNICITY and SEX sheds further light

on the influence of speakers' ETHNICITY and SEX in these data. Namely, for African Americans, Latinos, and Lumbees, males have articulation rates about a 0.3 σ/sec faster than females of the same ETHNICITY. This pattern is not true for European Americans, where in fact, females have equivalent rates to males. This difference is easier to observe from the figure than the tabular results.

AGE is highly significant, and follows an interesting curvilinear pattern. This was somewhat indicated by the summary plot earlier, but here the pattern has much better resolution. We see that the youngest speakers have the slowest rates, speakers between about 20 and 40 years old have the highest rates, and then rates decrease and level off for older speakers, above around 60 years old. The confidence intervals in the age panel of Figure 5.3 widen for the oldest speakers, as relatively less data are available for speakers above 85 years old.

The number of syllables in an utterance, NUMSYLS, has a large effect and is similar to though less "lumpy" than was found for the reading passage data last chapter. The shortest utterances are quite slow (and note from Figure 5.3 that they have extremely narrow confidence intervals), yet quickly increase in rate until around 5 σ in length at which point rates continue to increase as utterances get longer, but at a much shallower slope. As I commented in Chapter 4, the dip in the effect of NUMSYLS seen at about 10 σ for the earlier reading passage data is not apparent in conversational, interview speech. The pattern here is smoother and as such seems more reliable. Finally, as noted earlier, we can interpret the slow rates for the shortest utterances as likely influenced by the relatively large proportion of phrase-final lengthened material and by discourse pragmatics (e.g. the fact that many short utterances are of types like "Well..." and "I see." which often have discourse-specific, and slower, prosody).

While mixed-effect modeling has allowed us to model these data in detail, it does have some downsides in comparison to traditional regression methods. One of these drawbacks is that its mathematics are much more complicated than those of analogous fixed-effect models (i.e. models with the same fixed effects but no random effects), and it is much less straightforward to assess how well a mixed-effect model fits the data – how much of the variance in the data is accounted for the model – in a meaningful way. These kinds of measures are easily generated (and often generated automatically by the modeling software) for fixed-effect only models, but less direct techniques are necessary for mixed-effect models. We can, however, approximate a measure of model fit for the mixed-effect model in Table 5.2 by assessing the

correlation between the fitted output of the model and the real-world, original data points. Determined this way, the model of Table 5.2 yields an R^2 of 0.647. A bootstrap validation method does not decrease the R^2 value, indicating that the fit is quite robust.[9]

We can also examine another simple way to approximate this measure that seems to me appropriate if the mixed-effect model's outcome correlates highly with that from an analogous fixed-effect model (i.e. a model with the same fixed effects but no random effects). If this is the case, we can use the fixed-effect model as a baseline to understand the overall fit. This relies on the fact that the mixed-effect model accounts for some higher amount of the overall variance than its fixed-effect counterpart. (If it did not, the model building and model criticism maneuvers would indicate that the inclusions of the random effects are spurious.) Random effects, by adjusting the model's intercept, and, here, slope of NumSyls, for each speaker, allow the model to more accurately fit the data, and thus create a more explanatory model for the fitted data. Provided the mixed-effect model's fixed-effect coefficients are highly correlated with the analogous fixed-effect model's coefficients, then that fixed-effect model's R^2 value should tell us something about the overall variance that we can account for with the mixed-effect model. Doing this for the model presented in Table 5.2, we obtain an extremely high correlation between the analogous fixed-effect model and the full, mixed-effect model (Pearson's $r = 0.998$; $p < 0.000001$). These coefficients are displayed in Table 5.3. This fixed-effect model has an adjusted R^2 of 0.578, confirming that these models are able to account for a large portion of the variability in the data.

Finally, we can ask about how much of the model fit is achieved by the obviously influential factor of utterance length in syllables (NumSyls) versus the social factors. We can examine the influences of the different factors to some degree by comparing the complete best model against a submodel, one which contains only the random intercept for speaker and the social factors. The model with only the social factors yields an R^2 of 0.197. This R^2 is quite a bit worse than the complete model; in fact, 69 percent smaller than that for the complete model. It is clear that when we examine the data at the utterance level, utterance syllable length is massively influential and dwarfs the significance of the social factors in determining articulation rate. Nonetheless, we still can interpret the social factors as important explanatory factors behind articulation rate realizations. As Baayen (2008: 258–9) notes, just because the effects we are most interested in are often smaller in size than the effects we are not, it does not decrease the importance of their identification and validation.

Table 5.3 Mixed-effect (M-E) and analogous fixed-effect (F-E) model fixed-effect coefficients

Factor	Estimate from M-E model	Estimate from F-E model
(Intercept)	–0.336	–0.862
NumSyls	1.043	1.070
NumSyls'	–5.692	–5.858
NumSyls''	15.069	15.517
NumSyls'''	–10.205	–10.499
Age	0.065	0.088
Age'	–2.668	–3.504
Age''	2.881	3.781
Region = Eastern NC	0.228	0.285
Region = Ohio	0.262	0.253
Region = Southern NC	0.238	0.417
Region = Texas	0.284	0.213
Region = Washington, DC	0.118	0.328
Region = Western NC	0.047	0.199
Sex = male	0.333	0.309
Ethnicity = Latino/a	0.219	0.262
Ethnicity = Lumbee	–0.096	–0.223
Ethnicity = European Am.	0.345	0.370
Ethnicity = Latino/a x Sex = male	0.023	0.100
Ethnicity = Lumbee x Sex = male	0.005	0.070
Ethnicity = Eur. Am. x Sex = male	–0.398	–0.174

We will consider the patterns in these data further, after examining the pause duration data.

5.3.2 Pause duration at the pause level

We now turn to look at the pause duration data on a per-measurement level. A total of 29,614 pauses have been extracted and measured for the 159 speakers. This yields an average of 186 measurements per speaker, but, as for articulation rate, the actual number of contributions by the speakers is quite variable. The median N for the talkers is 105, with a minimum N of 7 (a European American from Southern NC). As with the articulation rate data, the highest Ns come primarily from the African Americans from Washington, DC (maximum N of 1449). The pause data from Chapter 4 were limited to silent durations 200 ms or above. The pause data here include much shorter pauses, silences 60 ms and above (with a maximum pause value of 5000 ms).

As was discussed in Chapter 4, pause durations distribute in a log-normal fashion and, as such, are modeled in units converted to log-ms. (Recall, Appendix II provides a brief conversion table between log-ms and ms.) The mean over all of the pause measurements is 562 ms, or 5.98 log-ms (in all cases, figures for the log-transformed durations are based on the log-transformed values themselves – e.g. 5.98 is the mean of the individually log-transformed durations, it is not the log of 562). The median value for the untransformed durations is 410 ms, while the median over the log-transformed values is 6.02. In the next chapter, §6.3 provides a further discussion of the pause data distributions, and considers how different THRESHOLDS for which measurements are included impact the analysis.

In order to provide a large-scale view of the pause duration data, only the following factors are considered here:

- The speaker's ETHNICITY (one of African American, European American, Latino/a, or Lumbee)
- The speaker's SEX (female or male)
- The speaker's AGE (continuous, with a mean value of 37.6^{10})
- And, the speaker's REGION (one of Central NC, Eastern NC, Ohio, Southern NC, Texas, Washington, DC, or Western NC)

We have reasons to expect that discourse- and event-related factors also influence pausing (see Chapter 2 for a review of the literature). In §6.5, we will examine a subset of these pauses more closely from this perspective, but here we focus on the social factors and on fully leveraging the large amount of data available. As was discovered for the reading passage data in Chapter 4, we will see that pause duration data are much less patterned according to the available social factors than the speech rate data have proven to be. A summary of the mean pause durations for these factors is shown in Figure 5.4.

As we have noted several times, the data are not well balanced across social categories, with different factor levels contributing different amounts of data. For ETHNICITY, SEX, and AGE we see some variability across the speakers but this is contained within the range between about 5.90 and 6.05 log-ms. For REGION we see a much larger spread in the data, with Ohioans having by far the shortest pauses (5.66 log-ms) and speakers from Washington, DC having the longest (6.10 log-ms). The best mixed-effect regression model, treating speaker as a random intercept, is shown in Table 5.4. As was done for articulation rate earlier, the model shown was generated by trimming the outliers based

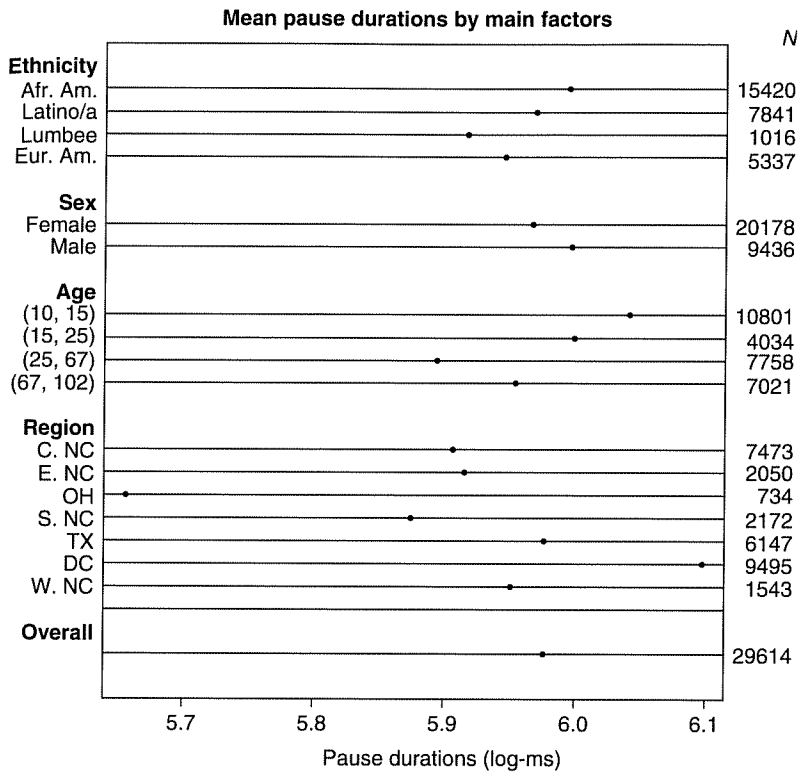

Figure 5.4 Mean pause durations by main factors

Table 5.4 Best mixed-effect model for (trimmed) pause-level pause durations

Factor	Estimate	Std. err.	*p*
(Intercept)	5.829	0.040	–
REGION = Eastern NC	0.004	0.060	[0.9394]
REGION = Ohio	−0.264	0.074	0.0004
REGION = Southern NC	−0.092	0.064	[0.1268]
REGION = Texas	0.001	0.049	[0.9818]
REGION = Washington, DC	0.264	0.064	0.0002
REGION = Western NC	−0.033	0.073	[0.6410]
SEX = male	0.099	0.033	0.0018
ETHNICITY = Latino/a	0.103	0.054	/0.0530/
ETHNICITY = Lumbee	0.125	0.082	[0.1156]
ETHNICITY = European Am.	0.060	0.043	[0.1616]

$R^2 = 0.053$.

on the residuals from an initial regression (removing 0.4 percent of the data, leaving 29,489 measurements) and then refitting the model to the trimmed dataset.

REGION and SEX emerge as strongly significant in the model. ETHNICITY is included here as well, although none of the comparisons are, strictly speaking, significant; the factor was just barely included in the model based on likelihood ratio tests. These effects are plotted in Figure 5.5. For ETHNICITY, the model p values indicate that Latinos are on the cusp of being significantly different from African Americans (the baseline factor for ETHNICITY; $p = 0.053$) but that the other comparisons are not significant. The Lumbees, who appear to have the longest pauses based on their coefficient, have too large an error term to reach significance. For SEX, we see that males have pauses that are slightly but significantly longer than females. For REGION, we see a confirmation of the patterns visible in the raw data. The Ohioans have the shortest pauses while the speakers from Washington, DC have the longest. It bears remembering

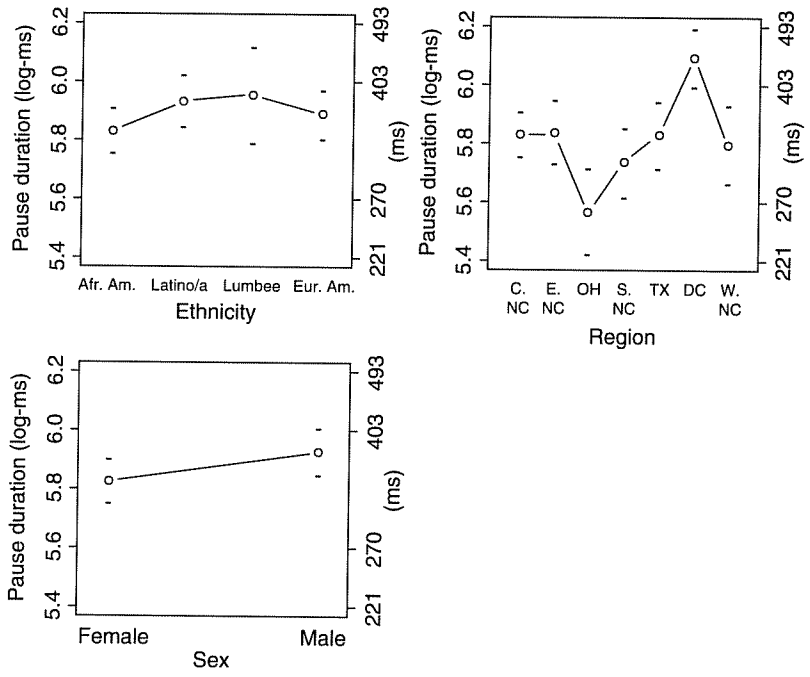

Figure 5.5 Effects in the mixed-effect regression model for pause durations

that the DC speakers are all young African Americans, so the DC region confounds several factors. The four North Carolina groups, and Texas, are quite similar to one another.

More important than the significant effects in this model is, perhaps, the acknowledgment that the model overall achieves a poor fit for the data. We are only able to account for a very small amount of variance in the pause duration data from the factors examined here. In our consideration of the articulation rate data we saw that although much of the model's good fit was based on the within-utterance factor of utterance length, many of the social factors arose as significant and we still were able to account for a sizable portion of the variance through the social factors alone. Here the social factors – the only factors we are examining – provide an R^2 of only 0.053. Remarkably, the same bootstrap validation method used in the previous section for articulation rate *increases* the R^2 to 0.058. This increase (especially coupled with such a low value) can be taken as evidence of just how poorly this "best" model fits the data.

While we were able to find some significant social effects on pause duration realization, we can end this section by reiterating that pause durations appear poorly predicted by social information.

5.4 Modeling speech rate and pause durations at the speaker level

We have thus far shown that articulation rates are fairly well predicted by social factors in a large-scale corpus-based analysis of individual utterances drawn from SLAAP's archive of sociolinguistic interview recordings. Pause durations, on the other hand, are not. §5.3 focused on examining the approximately 30,000 measurements available for each feature through mixed-effect regression models to leverage all of these data. In a sense, I have started with a "the more the merrier" approach to the data, assuming that all of those data points, coupled with the sophisticated techniques now available to analyze them, can shed deeper insight into the patterns of these features. Here we back up a bit to look at the speaker central tendencies for articulation rate and pause duration and to ask whether we gain as much, or even more, insight by modeling these features at the speaker level than we did at the utterance and pause level.

By examining the data one value per speaker, we actually generate from one perspective a much simpler dataset – one that is most appropriately modeled through traditional, fixed-effect regression – and, from

another, a more complex dataset – in that we can assess in a straightforward way some additional factors, such as whether a speaker's median articulation rate is a predictive factor for her or his median pause duration and vice versa. The speaker-level analysis is appropriately modeled used fixed effects because there are no nested factors; each speaker contributes only one data point and, therefore, does not contribute her or his own individual variance to the data.

5.4.1 Speech rate at the speaker level

Since there are 159 speakers in the dataset, we now examine a dataset with 159 values, with each representing the central tendencies for a speaker. For this model of the articulation rates, the following potentially predictive factors are considered:

- The speaker's median number of syllables per utterance (MEDSYLS; continuous, with a mean of 6.04 σ)
- The speaker's median pause duration (MEDPAUSEDUR; continuous, with a mean value of 411.6 ms)[11]
- The speaker's ETHNICITY (one of African American, European American, Latino/a, or Lumbee)
- The speaker's SEX (female or male)
- The speaker's AGE (continuous, with a mean value of 44.1)
- And, the speaker's REGION (one of Central NC, Eastern NC, Ohio, Southern NC, Texas, Washington, DC, or Western NC)

Figure 5.6 displays a summary of the main factors in this per-speaker assessment of articulation rate. Examining the summary plot, we note first that the mean of the articulation rates over all 159 speakers is 4.8 σ/sec. Once again, we will briefly review each factor in turn.

We see indications from the summary plot of a relationship between a speaker's median syllables per utterance and her or his articulation rate. Speakers with shorter average utterance lengths also have slower average speech rates, while speakers with longer utterance lengths have faster speech rates. (The simple correlation is significant with Pearson's $r = 0.51$, $p < 0.0001$.) This is demonstrated more clearly in Figure 5.7, the left panel of which plots the 159 speakers' articulation rate medians against their utterance length medians (MEDSYLS).

The view obtained for MEDPAUSEDUR in the summary of Figure 5.6 indicates that articulation rate may in fact be correlated with pause durations, in that the speakers with longer median pauses, around a half-second or more, also have the slowest speech rates, around 4.6 σ/sec,

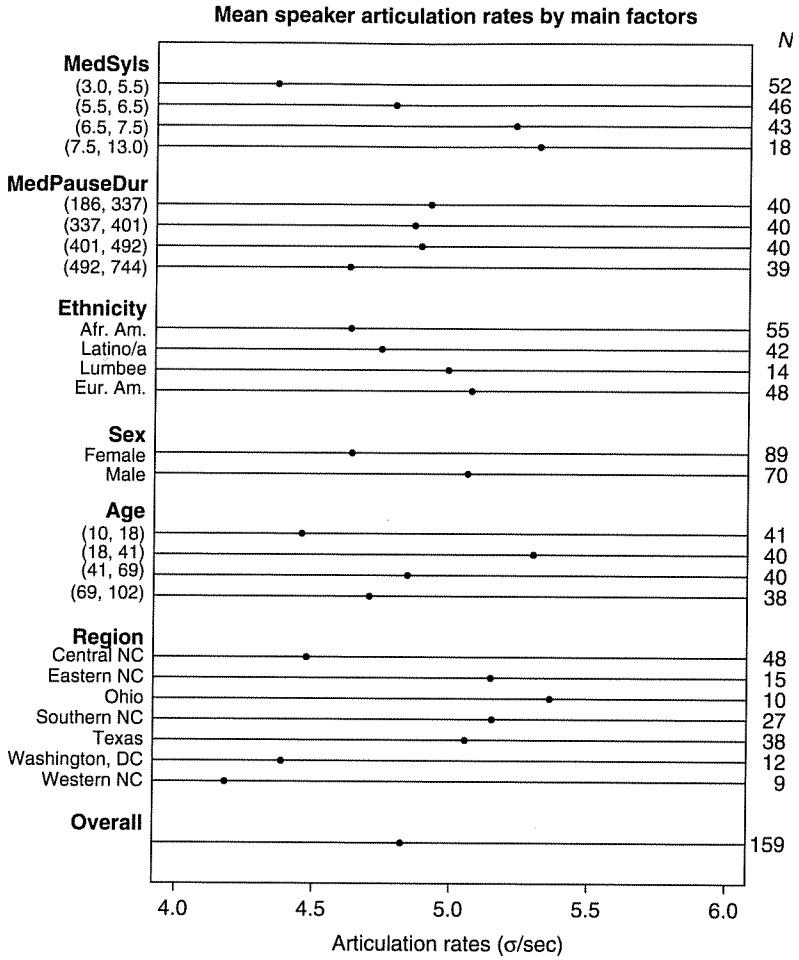

Figure 5.6 Mean speaker (median) articulation rates by main factors

while speakers with shorter pauses have rates around 4.9 σ/sec. The right panel of Figure 5.7 displays this relationship more closely. This correlation is also significant, although much less strongly than for MEDSYLS (Pearson's $r = -0.16$, $p = 0.045$).

Turning to the social factors, we see quite a bit of differentiation in the data. For ETHNICITY, we see that African American and Latino speakers have articulation rate medians in the 4.7 σ/sec range, while Lumbees and European Americans have rates around 5.0 σ/sec. For SEX, we see

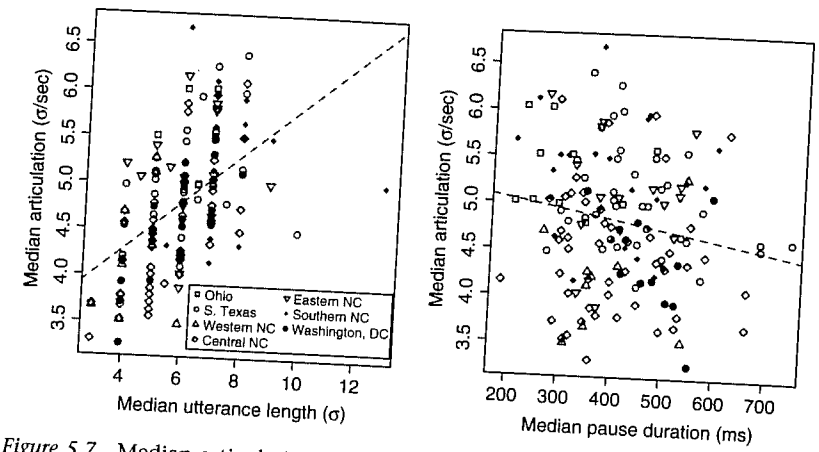

Figure 5.7 Median articulation rates by median utterance lengths (MEDSYLS, on left) and median pause durations (MEDPAUSEDUR, on right)

that the 89 female speakers have a mean rate of 4.6 σ/sec, while the 70 males have a mean rate close to 5.1 σ/sec. For AGE, we see the youngest speakers, age 10 to (but not including) 18, have the slowest rates, around 4.5 σ/sec, and that the next group, speakers age 18 to (but not including) 41, have the fastest rates at 5.3 σ/sec. The older speakers decrease again, with speakers between 41 and (not including) 69 having rates around 4.8 σ/sec and the oldest speakers, age 69 and above, have rates around 4.7 σ/sec. Finally, for REGION, we see that the seven regional categories appear to sit in roughly two clusters. Speakers from Central NC, Western NC, and Washington, DC have rates between 4.2 and 4.5 σ/sec, while speakers from Eastern NC, Southern NC, Texas, and Ohio have rates between 5.0 and 5.4 σ/sec. Overall, these patterns are similar to those for the utterance-level data seen earlier in Figure 5.2 and discussed in §5.3.1.

To model these data, we make use of ordinary least-squares linear regression modeling, a standard form of fixed-effect linear regression modeling. (The specific tools used are the `ols()` function and the supporting tools of the Design library in R, Harrell 2009.) Table 5.5 presents the best model from the factors discussed above.

For the most part, the speaker-level model is extremely similar to the model obtained from the utterance-level articulation rates earlier. Figure 5.8 displays the model main effects as well as the significant interaction between ETHNICITY and SEX. I have presented the panels in the figure in the same order as was done earlier, in Figure 5.3. This

Table 5.5 Best fixed-effect model for speaker-level articulation rate

Factor	Estimate	Std. err.	*p*
(Intercept)	0.380	0.758	–
MEDSYLS	0.514	0.155	0.0011
MEDSYLS'	−1.886	0.928	0.0441
MEDSYLS''	6.040	2.806	0.0331
MEDSYLS'''	−8.307	3.464	0.0178
AGE	0.094	0.026	0.0004
AGE'	−1.660	0.475	0.0006
AGE''	2.223	0.655	0.0009
AGE'''	−0.665	0.253	0.0096
REGION = Eastern NC	0.398	0.152	0.0097
REGION = Ohio	0.354	0.187	[0.0599]
REGION = Southern NC	0.416	0.164	0.0121
REGION = Texas	0.326	0.146	0.0270
REGION = Washington, DC	0.173	0.206	[0.4017]
REGION = Western NC	0.107	0.183	[0.5610]
SEX = male	0.517	0.145	0.0005
ETHNICITY = Latino/a	0.280	0.188	[0.1388]
ETHNICITY = Lumbee	−0.146	0.282	[0.6056]
ETHNICITY = European Am.	0.367	0.149	0.0154
ETHNICITY = Latino/a x SEX = male	−0.099	0.216	[0.6483]
ETHNICITY = Lumbee x SEX = male	−0.321	0.317	[0.3123]
ETHNICITY = Eur. Am. x SEX = male	−0.548	0.207	0.0092

$R^2 = 0.590$; adjusted $R^2 = 0.527$.

graphical view is a little easier to interpret than the tabular format, and comparing it with Figure 5.3, the plots of the mixed-effects model effects from earlier, gives an immediate sense of just how similar these two models are.[12] The predicted articulation rates are a bit lower in this fixed-effect model for speaker medians than they were in the mixed-effect model for individual utterances, but the social factor effects are all quite similar.

As before, it must be remembered that the categorical variables, like REGION and ETHNICITY, are assessed against a baseline factor, so each line in Table 5.5 for those factors is a comparison between the current predictor and that baseline. For ETHNICITY, the baseline is again African American, so we see in the table that the Latino speakers are estimated as having articulation rates 0.28 σ/sec faster than the African Americans and the Lumbees as having rates 0.15 σ/sec slower than the African Americans, though neither of these differences are significant. European Americans, however, do have significantly faster rates than African

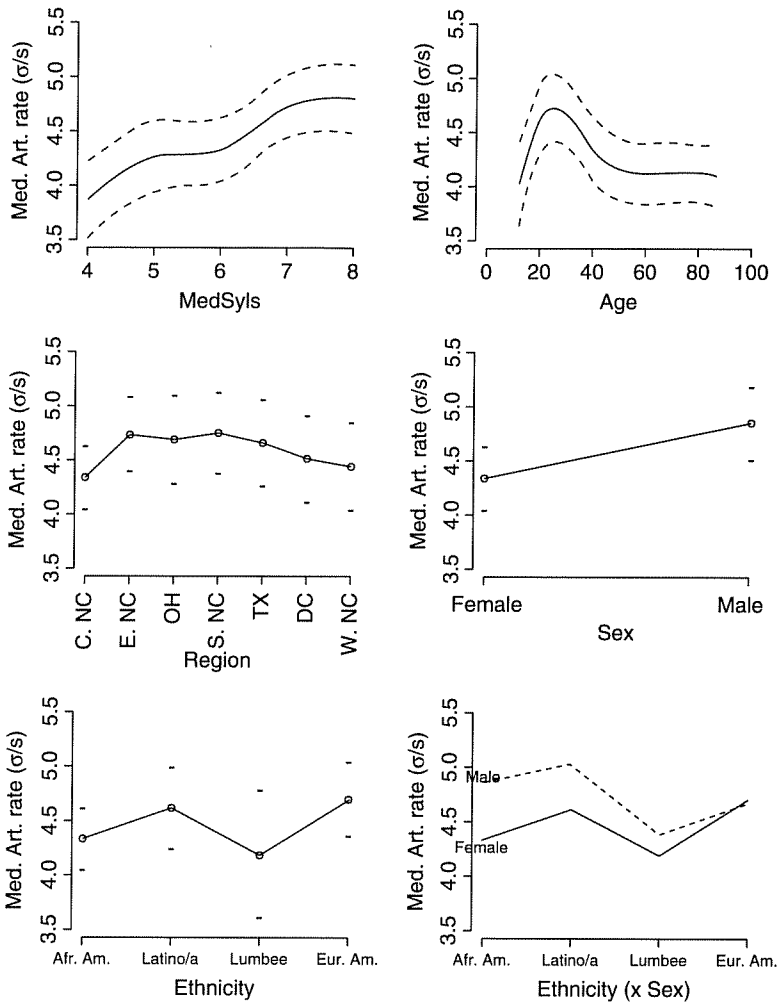

Figure 5.8 Effects in the fixed-effect regression model for articulation rates

Americans (estimated at 0.37 σ/sec faster, $p = 0.015$). As is clear from the plot of the REGION effects in Figure 5.8, most of the regions are not all that different from one another. The relative positions of the regions are slightly different between Figures 5.8 and 5.3 but these differences are so minute as to seem inconsequential. Central NC, the baseline factor for REGION, has the lowest predicted rates again. All of the other regions

have positive estimates (i.e. faster rates than the baseline), ranging from 0.11 σ/sec (Western NC) to 0.42 σ/sec (Southern NC). Only Texas ($p = 0.027$), Southern NC ($p = 0.012$), and Eastern NC ($p = 0.0097$), however, reach significance in their comparison to the slower rates of Central NC. Ohio, despite sizably faster rates (0.35 σ/sec), just fails to reach significance ($p = 0.0599$).

SEX is highly significant in these data with males predicted as having rates 0.52 σ/sec faster than females ($p = 0.0005$). We also see the same important interaction between SEX and ETHNICITY that was found for the per-utterance model. Male European Americans do not show the same increase in rates over female European Americans as males of other ethnicities do and, in fact, have slightly slower rates than the females (estimate: 0.52 – 0.55 = –0.03 σ/sec). This is nicely illustrated in the bottom right-hand panel of Figure 5.8.

The AGE factor has a sharper peak than it did in the per-utterance model and, actually, the age spline has an additional significant knot than it did in the per-utterance model. (This minor difference is the only difference in parameters between the two models – otherwise, both models Table 5.2 and Table 5.5 select the same parameters.) Overall, the substantive effect of AGE is the same despite the additional knot. Speakers in their late twenties have the fastest speech rates, at about 4.7 σ/sec, and children have the slowest rates, with rates about 4.0 σ/sec for the youngest speakers. Starting at about 50 years old, speakers' rates level out and stabilize at about 4.2 σ/sec.

The only truly noticeable difference between the two models lies in the effect of the number of syllables factor. Earlier we found the syllable length factor, NUMSYLS, to have a strong and large effect on articulation rate. Here the MEDSYLS factor has, for lack of a better word, a "squiggly" effect. In general terms, we see that median articulation rates increase as speakers' median utterance lengths increase, but this happens in a less curvilinear and less drastic way than it did for the direct utterance-level syllable length to rate relationship. MEDSYLS also has a narrower effect size than we saw for the utterance-level model, no larger than that for AGE.

From its appearance in the figure, it looks as if the MEDSYLS effect could have been modeled without the nonlinear terms (i.e. with a straight line modeling its general tendency), but various tests indicated that the model was better when the nonlinearity was included. A model considering MEDSYLS as a simple linear predictor (rather than a nonlinear predictor) also yields significance and a similar outcome to the model in Table 5.5, but obtains an overall worse fit (the simple linear term obtains an adjusted R^2 of 0.49 instead of the 0.53 achieved with

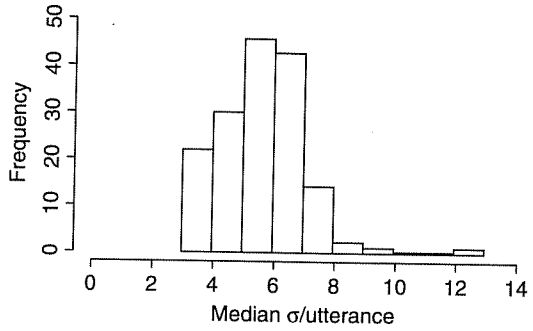

Figure 5.9 Median syllables per utterance for the speakers

this best model). While this syllable length effect difference between the two models stands out, it also makes sense given that the relationship between the number of syllables in an utterance (NUMSYLS) and that utterance's articulation rate is direct and specific and the relationship between a speaker's median syllables per utterance (MEDSYLS) and her or his median articulation rate is less direct. Further, the median utterance lengths for the speakers fall within a tight range, which I illustrate through a histogram in Figure 5.9, and, as central tendencies, limit the amount of possible variability.

The model does not include MEDPAUSEDUR, which in exploratory models was found to be just under significance ($p \approx 0.07$). In addition to not being significant, the inclusion of MEDPAUSEDUR knocked the ETHNICITY x SEX interaction out of significance in models that tested both together, and, overall, resulted in less well-fitting models. Thus, it appears that speakers' pause durations are not predictive of their articulation rates, despite the correlation seen earlier.

Most importantly, perhaps, for the general project here is the observation that the model provides a fairly good fit for the data, with an adjusted R^2 of 0.53. As a large part of our interest here is in the social factors behind articulation rate variability, I reran this best model on the data without the MEDSYL predictor to gather a sense of how well the social factors alone account for the data. This social-factors-only model yielded an adjusted R^2 of 0.40. The speakers' median utterance length is clearly an important factor behind their median articulation rate, but it only accounts for 25 percent ((0.53 – 0.40)/0.53) of the model's success. Most of the fit is yielded by the social factors. This is a major difference between the outcome of this model and the per-utterance

model of Table 5.2. That model achieved a better overall R^2, 0.647 for the mixed-effect model (and an adjusted R^2 of 0.578 for an analogous fixed-effect model on the utterance-level data without the random effects), but we also saw how that R^2 dropped off when we removed the NUMSYLS effect. The fixed-effect, speaker-level model indicates that, at the level of speaker central tendencies, articulation rates are quite socially influenced.

An important downside to the per-speaker analysis I have just presented is that we have been examining a fairly large regression (i.e. testing a high number of predictors) for only 159 data points. Baayen (2008: 195) recommends that a dataset should have at least 15 times the number of observations than predictors. Bootstrap validation indicates that the model is overly optimistic with the R^2 falling to 0.45 (the R^2 of the comparable social-factors-only model falls to 0.33).[13] This is a sizable reduction and likely related to the high parameter to token ratio tested here, but it still indicates that the model provides a meaningful fit for the data.

5.4.2 Pause duration at the speaker level

Following the strategy for the speech rate analysis in the previous section, we now examine the pause duration data at the per-speaker level, where the object of analysis is each speaker's median pause duration (log-transformed and modeled as log-ms). As before, since there are 159 speakers, we examine a dataset with 159 data points. We consider the following factors:

- The speaker's median number of syllables per utterance (MEDSYLS; continuous, with a mean value of 6.04 σ)
- The speaker's median articulation rate (MEDARTRATE; continuous, with a mean value of 4.83 σ/sec)[14]
- The speaker's number of pauses, normalized to a per-100-word measure (PP100WDS; continuous, with a mean value of 10.76)[15]
- The speaker's ETHNICITY (one of African American, European American, Latino/a, or Lumbee)
- The speaker's SEX (female or male)
- The speaker's AGE (continuous, with a mean value of 44.1)
- And, the speaker's REGION (one of Central NC, Eastern NC, Ohio, Southern NC, Texas, Washington, DC, or Western NC)

Recall that for the pause data the earlier statistical analyses for individual measurements were largely unsuccessful. For the reading passage data in Chapter 4, no factors were found to statistically influence

pause duration. For the pause-level version of these data in §5.3.2, we did obtain significant effects for REGION, SEX, and to a lesser degree ETHNICITY, although the overall model poorly fitted the data. Considering speaker central tendencies for pauses has some immediately obvious advantages – it makes available several additional potential predictors that we have thus far not considered and we can see if these additional predictors help account for the variation in the data. Figure 5.10 displays the summary of these main factors. As before, we

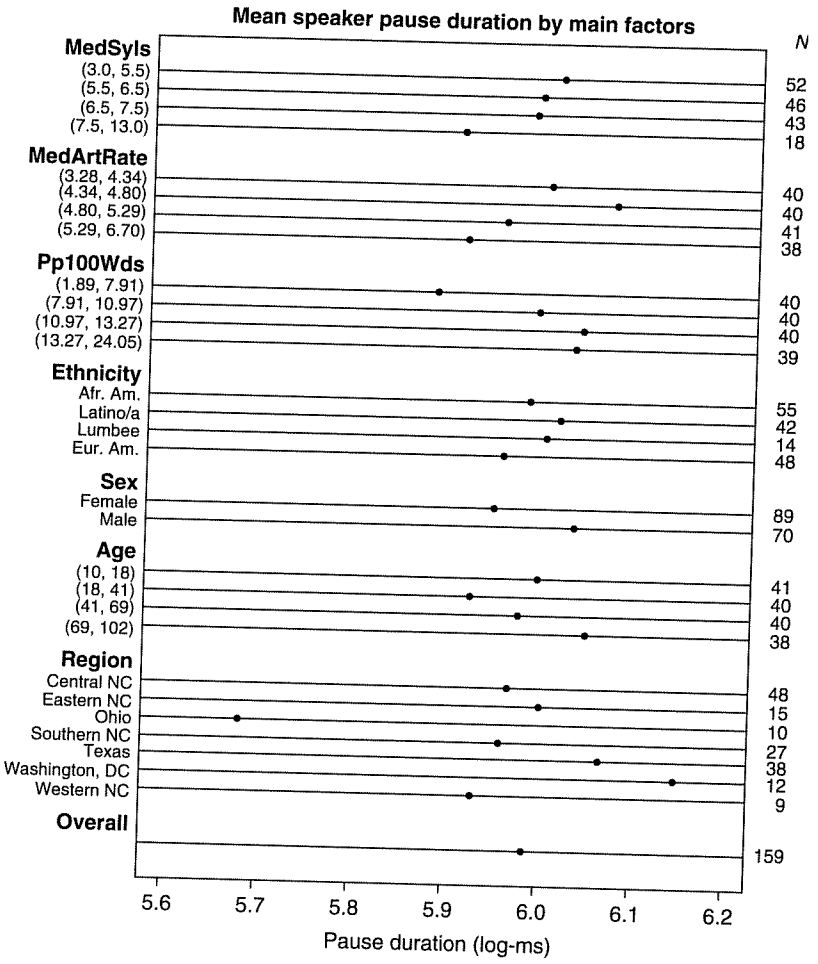

Figure 5.10 Mean speaker (median) pause durations by main factors

examine log-transformed pause durations and Appendix II provides a brief table of correspondences between ms and log-ms values.

We see indications from this view of the raw data that speakers with the longest utterances (in terms of median syllables per utterance, MEDSYLS) may have shorter pauses than other speakers. There is also an indication of a relationship between speakers' median articulation rates (MEDARTRATE) and pause durations, although we see that speakers binned into the second slowest articulation rate group (4.34–4.80 σ/sec) have longer pauses than the group with the slowest articulation rates and disrupt what might otherwise look like a linear trend. Figure 5.11 examines these potential correlations further.

The left panel of Figure 5.11 displays a plot of the median pause durations (in log-ms) against MEDSYLS for the 159 speakers. Although there may appear to be some tendency for pause durations to decrease as utterance lengths increase, the correlation is not significant (Pearson's $r = -0.12$, $p = 0.142$). The relationship between median pause duration and median articulation rate, shown on the right panel of Figure 5.11, is significant however, with pause durations decreasing with increasing articulation rates. The correlation is nonetheless a weak one (Pearson's $r = -0.17$, $p = 0.037$).

Finally, for the number of pauses realized per 100 words by each speaker, PP100WDS, we see some indication of a relationship. Speakers with the fewest pauses, between 1.89 and 7.91 pauses per 100 words, have much shorter pauses than speakers with more pauses. Figure 5.12

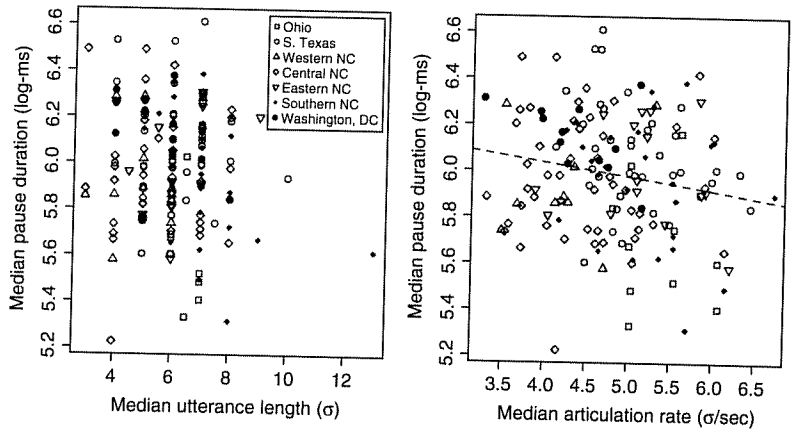

Figure 5.11 Median pause durations by median utterance lengths (MEDSYLS, on left) and median articulation rates (MEDARTRATE, on right)

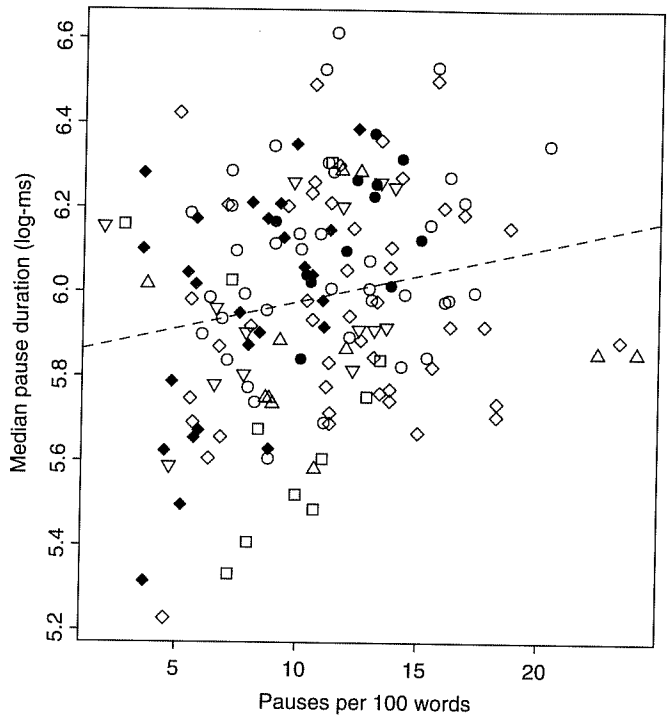

Figure 5.12 Median pause durations by number of pauses per 100 words (PP100WDS)

displays this relationship more closely. Although far from strong, this correlation is significant (Pearson's $r = 0.20$, $p = 0.013$). It appears that people who have more pauses tend to have longer pauses as well. The multivariate analysis will momentarily indicate whether this trend holds up when all of the factors are taken into account.

In terms of the social factors, we see some indications of differences among the speakers. For ETHNICITY, African Americans and European Americans have slightly shorter pause durations than Latinos and Lumbees. For SEX, females have shorter pauses than males. We see a U-shaped pattern for AGE, with speakers between 18 and 41 having the shortest pauses, shorter than speakers younger and old than them. The oldest speakers appear to exhibit the longest pauses.

Finally, for REGION, we see quite a large range. The four North Carolina groups hover somewhat close together in the center of the distribution.

Table 5.6 Best fixed-effect model for speaker-level pause durations

Factor	Estimate	Std. err.	p
(Intercept)	6.147	0.165	–
MEDARTRATE	–0.081	0.032	0.0130
PP100WDS	0.010	0.005	[0.0504]
REGION = Eastern NC	0.098	0.072	[0.1778]
REGION = Ohio	–0.195	0.086	0.0246
REGION = Southern NC	0.074	0.062	[0.2368]
REGION = Texas	0.153	0.054	0.0050
REGION = Washington, DC	0.208	0.076	0.0067
REGION = Western NC	–0.037	0.085	[0.6583]
SEX = male	0.139	0.039	0.0005

$R^2 = 0.248$; adjusted $R^2 = 0.202$.

The Ohioans have a much shorter group mean, while the speakers from Washington, DC (though, recall, all young African Americans) have the longest mean pause durations. Altogether, the mean across all 159 speakers' medians is 5.9 log-ms, which corresponds to 398 ms.

As with the articulation rate data in the previous section, we now make use of ordinary least-squares linear regression modeling to better understand the extent and significance of each of these factors when all these factors are considered together. The best statistical model for the speaker-level pause data is presented in Table 5.6.

Recall that the model for the full pause duration data developed in §5.3.2 was quite poor in its fit of the data but that it did find significant effects for REGION, SEX, and, although marginally, ETHNICITY. The model here finds SEX and REGION to be significant, but not ETHNICITY. Two of the three additional factors are found to be important as well. Speakers' median articulation rates, MEDARTRATE, arise as significant, and the number of pauses per 100 words, PP100WDS, is bordering on significant with $p = 0.05$. The median number of syllables per utterance, MEDSYLS, which was a significant predictor in the speech rate data, is not significant here. Figure 5.13 displays the significant effects graphically, and I now discuss each in turn.

Speaker SEX is a significant predictor, with males having pauses estimated at 0.14 log-ms longer than females ($p = 0.0005$). REGION is also significant, with Ohioans having significantly shorter pauses (estimated at 0.20 log-ms shorter) than speakers from Central NC, who are the baseline for the statistical comparison ($p = 0.025$). Texans and the Washington, DC speakers have pauses significantly longer than Central

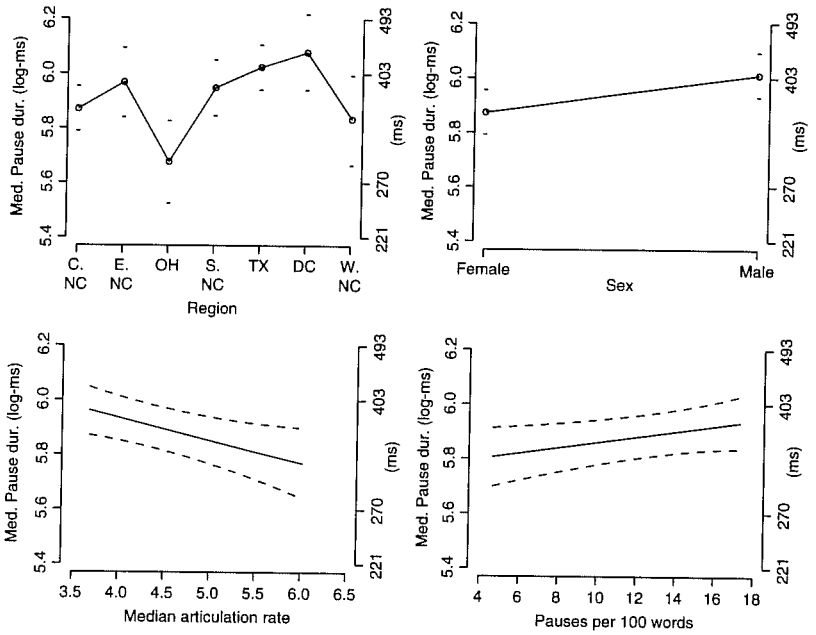

Figure 5.13 Effects in the fixed-effect regression model for pause durations

NC (+0.15 log-ms, $p = 0.005$, and +0.20 log-ms, $p = 0.007$, respectively). Similar to the per-pause analysis, but especially interesting in light of the findings on articulation rate where we did see regional differentiation within North Carolina, none of the North Carolina regional groups are found to be significantly different from the Central NC baseline. For speakers' median articulation rates, MEDARTRATE, we see an inverse relationship – as articulation rates increase pause durations decrease. This indicates that – at an overall, per-speaker level – talkers do not compensate for faster articulation rates by having longer pauses. Instead, it appears that faster talkers have both faster articulation rates and shorter pauses.[16] The indication of a potential nonlinear relationship between MEDARTRATE and pause duration, which was visible in Figure 5.10, does not arise in the statistical analysis; nonlinear components for MEDARTRATE were tested but found to be nonsignificant.

Finally, the best model also includes the number of pauses (per 100 words, PP100WDS) factor, although it falls on the cusp of significance ($p = 0.05$). While it must be taken with a grain of salt, given its borderline status, its pattern supports that of MEDARTRATE. Faster talkers appear

to be faster throughout all of the relevant metrics; speakers with fewer pauses also have shorter pauses, not the reverse, which might perhaps have been expected. No interactions were found to be significant for any of these main effects, and, as indicated by their absence in Table 5.6 and Figure 5.13, AGE, ETHNICITY, and each speaker's median utterance length (MEDSYLS) were not significant despite some indications in the raw data seen in Figure 5.10.

Comparing the pause model to the articulation rate model, we again see that the pause duration data are much less accounted for by the available factors. The pause model here yields an adjusted R^2 of 0.202. This is much better than the fit for the pause-level model which had an R^2 of only ~0.05, even with the random effect for speaker. The better fit is not a result of just the additional factors. In fact, a model fit to the pause data with just speaker SEX and REGION – the two social factors significant in both models – obtains an adjusted R^2 of 0.15, still much better than the R^2 from the pause-level mixed-effect model. The pause duration central tendencies are quite a bit more patterned by the available factors than individual pauses are, but are still much less so than articulation rates.

5.5 Which approach is better?

Having looked at the results of both per-speaker and per-measurement analyses of the data, we can make some conclusions about the value of each kind of analysis, and, relatedly, the scope of variation in these features. I do this here before moving on to consider the meaningfulness of the findings from a substantive, sociolinguistic perspective in the next section.

From a methodological perspective, it is worth asking: How much better is the mixed-effect model than the per-speaker fixed-effect model? Or, even, are the per-measurement data more useful than the speaker central tendencies? That is, have we learned more about speech rate and pause by examining ~30,000 individual measurements rather than the central tendencies for the 159 speakers? Obviously, the utterance-level and pause-level analyses provide us with a view of the data on a per-measurement level and this is important if we are interested in factors that influence the individual utterances. However, our principal interest here has been in the social factors and these are factors that relate to the speakers overall, not the individual utterances. The NUMSYLS predictor has shed insight into the influences on articulation rate beyond what was visible in the per-speaker model, but, beyond this, we find roughly

similar results for both modeling approaches. For articulation rate, the two approaches tell very similar stories – the significant factors are roughly the same and have similar effects. In fact, the only discernible difference between the substantive results of the two articulation rate models lies in very subtle differences between the REGION results. For both models, Washington, DC and Western NC are found to be not significantly different than the baseline factor, Central NC. In the speaker-level model, the Ohio region just fails to reach significance, despite that group having faster rates than Central NC. The other, significantly different, REGION groups are ordered slightly differently in the two models. The utterance-level model of Table 5.2 indicates that the Texas group has the fastest rates, while the speaker-level model of Table 5.5 points to higher rates for the Southern NC group. Ultimately, though, these differences are minor and I believe attempting to order the regional groups by their rates based on the only slight differences between many of the regions would be a wasteful endeavor.

For the pause data, we obtain less similar results across the two models. The mixed-effect model for the pause-level data found significant differences for REGION and SEX, and a marginal effect for ETHNICITY. The significance of REGION was driven by the shorter pauses of the Ohio group and the longer pauses of the Washington, DC group. In the speaker-level model, ETHNICITY is not significant although, for REGION, Texas, in addition to Washington, DC, is found to have significantly longer pauses than the baseline of Central NC. Nonetheless, as for speech rate, these differences seem to be the results of small shifts in the effects rather than qualitatively different results. We also note that for pauses, the inclusion in the speaker-level model of each speaker's median articulation rate (MEDARTRATE) and pause frequency (PP100WDS) resulted in significant effects and, overall, the speaker-level model was able to achieve a much better (though still ultimately poor) fit to the data.

To a large degree, we can conclude that the two approaches to the data resulted in similar outcomes and neither is markedly better than the other. In general, the mixed-effect utterance-level and pause-level models have some advantages over the per-speaker models. Through them, we see a more precise window into the influence of, for instance, utterance syllable length on articulation rate. From a general perspective, through the mixed-effect models, we are able to better leverage the huge amount of data (~30,000 tokens) in the statistical analysis. Using all those measurements to come up with speaker central tendencies and then modeling those fewer data points (here, 159) obtains strong results but loses detail. The measurement-level mixed-effect models,

by making use of so much data, are also more robust against overfitting, which can occur when models fit too many parameters to too few observations.[17]

While I did not address this earlier in the chapter, the mixed-effect modeling approach to the individual utterances has an additional benefit over the speaker-level analysis. It is rather robust against differences in the amount of data available for each speaker. Recall that some speakers in the analysis contributed a huge amount of data (many hundreds of measurements) but many speakers contributed much less (just 20 or 30). I will return to the question of how much this matters and how stable the results are in the next chapter (§6.2).

To a large extent, however, the greater appropriateness of one approach or the other will be based on the ultimate goal of the study. I have been interested in assessing how well the variability in articulation rate and silent pause duration can be modeled, but I have been doing so with a close eye to the influence of social factors on the modeling. As Chapter 2 surveyed, other factors have been found to influence these features – especially cognitive factors like word predictability, information load, task difficulty, and so forth. These factors are properties of the individual utterances, or even subutterance units like words, and as such should be examined at a per-utterance level. Thus, with more data available about each utterance, such as utterance type (e.g., declarative, interrogative), topic, narrative position, information load (e.g. Shannon's entropy; Shannon 1948), etc., or about each pause (syntactic location, pause type, etc.; cf. §6.5), we would want to pursue an utterance- or pause-level analysis to understand the impact of those factors. Without these kinds of factors, though, and with primary interest in the influence of macrosocial factors, the speaker-level, fixed-effect models seem quite adequate.

5.6 The sociolinguistic patterns of speech rate and pause duration

Through this analysis, we have seen that social factors like speakers' regional affiliations, ethnicities, and sexes do indeed influence the realization of silent pauses and articulation rates. We turn now to consider the substance of these results and do so in terms of both the popular stereotype that "Southerners talk slow" and the findings from the previous investigations of social patterns in speech rate and pause.

We begin by considering the regional patterns, where we see, for pauses, some indication of the existence of larger regional trends.

Speakers from the four subregions of North Carolina exhibit more similar pause durations to one another than they do to speakers from other regions. Neither of the pause models showed significant differences among any of the North Carolina regional groups, but Ohio was found to have shorter pauses and Washington, DC to have longer pauses than Central NC. For the speaker-level model, Texas was also found to have longer pauses than NC. This regional finding is interesting and indicates that there may be substantial (i.e. real) differences in pause patterns across larger regions. For articulation rate, on the other hand, differences have emerged for the REGION factor, but ones which are not as clearly separate across regions. Ohio has among the fastest rates according to both modeling approaches, but these rates are in actuality not very dissimilar to Eastern NC, Southern NC, and Texas. In fact, the model of speech rate at the speaker level indicates that Southern NC and Central NC are the most different from one another, despite their proximate location. This is especially notable in terms of the literature that has attempted to assess larger regional patterns of speech rate variability in the US (e.g. Ray and Zahn 1990; Jacewicz et al. 2009, 2010), as it seems to indicate that there may be more variability within larger regional areas (i.e. "the US South") than between areas (i.e. "Inland North" vs "South"). This also contradicts (or at least complicates) the common stereotype that Southerners have slower rates of speech than Northerners.

One possible interpretation of the regional differences can be made in terms of Goldman-Eisler's claim that listeners attend more to pause patterns than actual articulation rates when judging rates of speech. The pause data do indicate that Southerners have longer pauses than the one Northern group, Ohio, and these longer pauses are likely interpreted by listeners as "slower talk." What makes less sense according to this logic is the great extent to which the articulation rate was modelable and the poorness of the pause models. While the pause models show this regional pattern, it must be remembered that they perform badly and, actually, fail to account for much of the data. The models indicate, on the other hand, that articulation rates are quite systematic. Further, according to the JND of 5 percent proposed by Quené (2007), fairly small changes in articulation rate should be noticeable by listeners. For example, assuming a base articulation rate of 4.5 σ/sec, a change of ±0.225 σ/sec should be noticeable by a listener and this difference is well within the models' predictions across regions.

In terms of speaker sex, we found highly significant effects for both models and both features. Interestingly, the effects run in opposing

directions. Males are found to have faster articulation rates than females (+0.5 σ/sec at the speaker level, +0.33 σ/sec at the utterance level) but also longer pauses than females (~ +0.1 log-ms). This possibly indicates that males and females distribute their speaking time differently – with males speaking faster but with longer pauses than females – but beyond making this suggestion I resist further speculation.

For ethnicity we find only marginal patterns. There are some differences for both speech rate and pause but the differences are minor. For speech rate, we find that European Americans speak significantly faster than the other ethnic groups and that there is a significant interaction between sex and ethnicity, so that the sex effect for speech rate described above does not exist for European Americans. For pauses, ethnic-based differences are even less apparent with the only putatively significant effect the comparison between African Americans and Latinos in the pause-level data (which obtains $p = 0.05$).

For age, we do not find effects for pause duration. This null result runs counter to some of our potential arguments, such as that older speakers experience a slowdown in their cognitive processes or that children develop adult-like pausing patterns as they age. In their study of "pausological" development over a number of age and educational levels (kindergarten through college), Sabin, Clemmer, O'Connell, and Kowal, for instance, found that "with increasing age, less time (shorter total length of unfilled pauses) is involved in the production of more fluent speech, while unfilled pauses also occur less frequently" (Sabin et al. 1979: 44). We find no evidence of this in the conversational data examined here. (Redford's (in press) recent study also does not find serious differences between children's and adults' pause patterns.) For articulation rate, we do, however, find a robust nonlinear effect for age. The youngest speakers have the slowest rates, but rates increase quickly and peak for speakers around age 30. From there, rates decrease again and then level off for speakers older than about 50. This pattern is quite interesting and could be interpreted in several ways. First, it possibly indicates that articulation rate variation is an AGE-GRADED phenomenon (a feature that changes over the course of a person's life but that ultimately shows stability in the community over time). Like many age-graded sociolinguistic variables, articulation rate might show a sensitivity to the pressures of the LINGUISTIC MARKETPLACE (Bourdieu and Boltanski 1975, Sankoff and Laberge 1978). As the studies reviewed in Chapter 2 indicate, rates of speech have been implicated in listeners' judgments of competence and so forth and speakers, especially in the first decade or two of adulthood, might speak faster for this reason. An alternative explanation

might come from the nature of these data, as measurements coming from sociolinguistic interviews. Young children might be more shy and reticent in these interviews than the adults. We also might expect that people in their twenties, thirties, and forties would be more comfortable while being recorded, while older and younger speakers might have less experience with interview-like situations and recording gadgetry. Further, most of the NCLLP interviewers (and, I would add, most sociolinguistic interviewers in general) are graduate students in their twenties or early thirties. The fastest rates for interviewees of this age range might be an accommodation effect. These proposals are, of course, necessarily speculative and cry out for further and future investigation, but I return to the idea that there are accommodation effects in these data in §7.2. Regardless of their cause, it is interesting that the age effects only surface for articulation rate and not for pause.

More importantly than the individual substantive results, for my present interests we must note the lack of success of our attempts to statistically model the pause data. For articulation rate, we were able to build quite explanatory models. Granted, a large part of the utterance-level model's success came from the important influence of utterance syllable length on articulation rate realizations, but even without this factor we still find reasonably good predictive power for models based on only the social factors. For pause durations, while we do find significant effects for several social factors and we do not have a powerful linguistic factor like utterance length as a potential source for improved model fit, we are only able to build models that poorly predict the data (i.e. with extremely low R^2 values). We can, I think, interpret this outcome as roughly in line with the views of pause and articulation rate in the literature reviewed earlier, such as the work by Goldman-Eisler, Kowal, O'Connell, Feldstein, and others. The finding that articulation rate is quite socially mediated leaves less room for its variability to relate to cognitive aspects of speech production. Pause, on the other hand, does not appear to be (very) socially influenced and, as such, there is more space for its variability to relate potentially to cognitive and speech production processes. I will return to this consideration in Chapter 8, but first, in Chapter 6, we look more closely at these data and then, in Chapter 7, we visit more evidence of social patterns in speakers' speech rates and pause durations.

6

Closer Looks at Speech Rate and Pause Variation: Methods and Findings

6.1 Introduction

The previous chapter examined a large amount of data from sociolinguistic interviews to determine whether and how speech rate and pause duration variability pattern with social factors in conversational speech and to assess, from a more methodological perspective, what we can gain from a large-scale sociophonetic corpus analysis. Numerous other questions can be asked of speech rate and pause variability, of the methods I have employed, and of these particular data. I turn now to pursue several questions building on the large-scale analyses of the last chapter.

I begin these closer looks, in §6.2, by continuing a line of questioning that was central to the previous chapter – is having all of these data actually all that beneficial? I consider this by looking closely at the speech rate data available in the main dataset for the 15 speakers from that analysis for whom the most data are available and ask what happens to their central tendencies, and to the patterns in the data, as we massively reduce the number of tokens examined. In §6.3, I continue to think about how much data matter, but from a different perspective, with a view to the pause data. Here I consider what has been called the THRESHOLD PROBLEM, the question of what length of silence to include in an analysis of pause, and I attempt to shed new light on this decision by conducting a simulation over the available data to see how different thresholds impact the analysis. Can these corpus-based data shed light on the question of what length of silence is actually a pause? Then, in §6.4, I return to speech rates to consider the implication of the use of phonetic utterances as the unit of speech for analysis rather than some other "chunk" of speech. Does an analysis by Intonational Phrase yield

different results? I also use this as an opportunity to look more closely at the nonlinear effect of utterance length (in syllables) on articulation rate, asking how this finding relates to the widely known phenomenon of phrase-final lengthening. Finally, in §6.5, I turn the attention back to pause to investigate a subset of the main dataset that has been coded for several additional factors relating to hesitancy and pause types. This more comprehensive analysis allows us to ask whether we can build better models of pause duration if we include relevant cognitive and linguistic factors. It also lets us ask whether new social findings (different effects or interactions) emerge when we include better nonsocial information in the models.

6.2 How many speech rate measurements yield stable patterns?

We saw in the last chapter that most of the substantive, social questions about pause and speech rate were actually more adequately studied through the simpler fixed-effect linear regressions and measures of speaker central tendencies than by examining the individual measurements themselves. The 30,000 measurements shed some additional light, for instance on the specific influence of utterance length on articulation rate, but we perhaps would have learned as much about the social influences behind these features from a speaker-level analysis alone. If large-scale, corpus-based sociophonetic research is going to grow as a linguistic pursuit, it is worth knowing what the advantages really are, and, further, whether there really are advantages over more traditional, smaller dataset-based approaches. One important question here then is how many measurements yield stable patterns?

Jaffe and Breskin (1970) considered how much talk is required for accurately measuring a speaker's temporal patterns and indicated that five minutes can yield enough data to obtain stable measurements. While at first glance this seems like a sensible figure, at closer inspection and in the context of sociolinguistic interviews it appears somewhat vague. For example, is this five minutes of actual, continuous speech by a single talker or simply five minutes of recorded interview? If the former, the figure seems to me to be more than necessary as five minutes of actual talk is quite a lot of data. But, if the latter, how much actual talk do we need for each speaker? Loquacious talkers and/or those who are the target of an interview may generate a great deal of speech in a five-minute excerpt but more reticent talkers, or those who are not the focus of a given interview, might only make a small number of contributions in five minutes. To

the best of my knowledge the question of just how many measurements provide stable results has not been examined to further depth, and the 30,000+ speech rate measurements available here seem like a perfect dataset to revisit this question. It is meanwhile notable that some of the large studies of rate and/or pause – such as Dankovičova (2001) – have based their studies on actually fairly small amounts of data.

To consider this we look more closely at two subsets of the main dataset. First, we examine the median speech rate values for the 15 speakers in the dataset with the largest number of measurements and examine how stable their median values are as we decrease the number of measurements used to calculate their central tendencies. Since central tendencies were used for the speaker-level analysis – and yielded quite robust results – the stability of these central tendencies is important. Then, we will examine a larger subset of the data to ask how the utterance-level mixed-effect model changes as we decrease the data. We then consider how correlated the results from the reduced datasets are with the original model results.

6.2.1 The stability of central tendencies

The 15 speakers with the most data range from having 508 tokens (i.e. measured phonetic utterances) to 1939 measured utterances. The median number of utterances for these speakers is 922 and the mean is 902.9. To examine how stable articulation rate central tendencies are, I have iterated over each of these 15 speakers' data, starting with each speaker's full data and then decreasing the amount of data examined by 10 samples each iteration, stopping once there are less than 20 remaining data points. To do this, I randomly resampled smaller subsets of measurements from the original data.[1] The fluctuations in the speakers' median speech rate values are shown in Figure 6.1, which provides a simple plot of the number of tokens against the median rate value for each speaker. The speakers are labeled in the top left of each plot, so they can be compared with their full information in Table 5a on the website. Each plot displays the speaker's original median value as the point on the left and shows decreasing *N*s to the right. Since different speakers have different amounts of initial data, the *x*- and *y*-axes for the plots are different, based on the individual speakers' data.

For the most part, we see a picture of general stability for the central tendencies across most of the speakers' data. There is a quite a bit of individual variability in terms of how the speakers' values change as their data decrease, but most of the fluctuations occur below *N*s of about 200. I have placed dashed lines at $N = 200$ on each plot to help

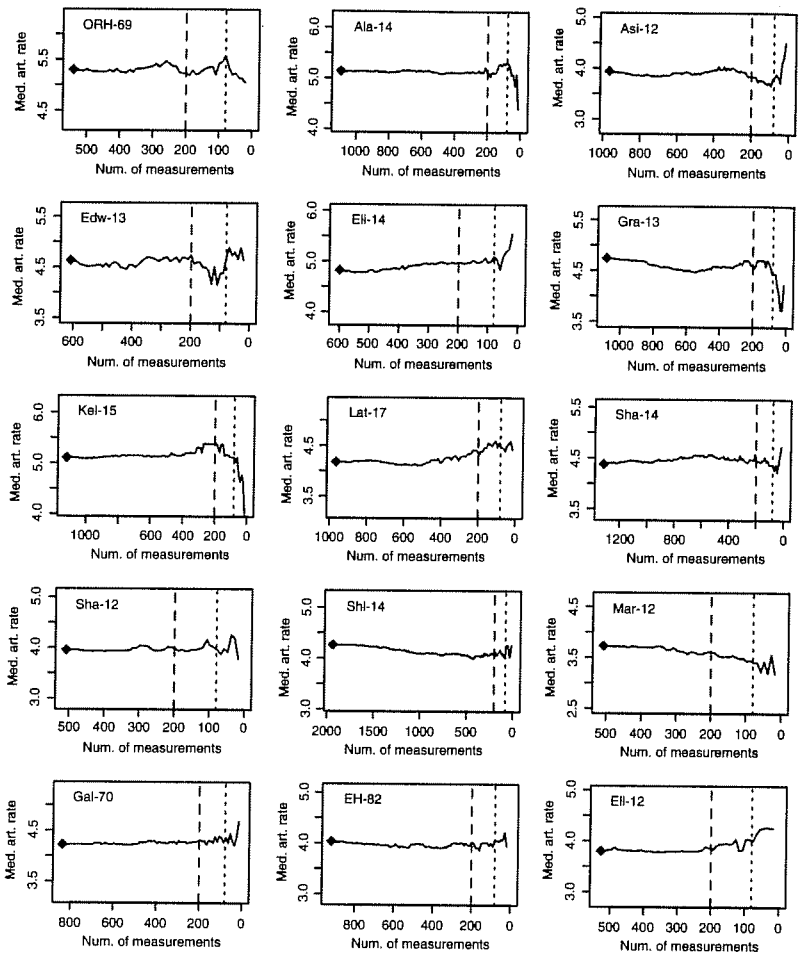

Figure 6.1 Changes in median articulation rates as sample size is decreased

indicate this pattern. Despite greater fluctuations after about $N = 200$, the values are still reasonably comparable. At about 80 measurements, a few speakers (e.g. Kei-15 and Gra-13) change much more drastically and I have indicated this location in each plot with a dotted line.

Following Quené's (2007) figure of a 5 percent value for the JUST NOTICEABLE DIFFERENCE (JND) for rates, we will use a difference of greater

difference of more than 5 percent their overall median at 200 tokens (Kei-15, with 5.5 percent difference). At 80 tokens, five of the 15 speakers have differences greater than 5 percent (Edw-13: 5.5 percent, Gra-13: 6.5 percent, Lat-17: 7.1 percent, Mar-12: 8.6 percent, and Eli-12: 5.3 percent; Kei-15, with a difference of 1.8 percent, is no longer >5 percent different at 80 tokens). At the final computation, when the data for each speaker go below 20 tokens, 10 of the 15 are more than 5 percent different than their original median, and 8 of these have differences greater than 10 percent. Kei-15, the only speaker to exceed a 15 percent difference at this final computation, has a median articulation rate of 3.91 σ/sec, 23.5 percent lower than the rate from her full data, 5.11 σ/sec. However, it must also be noted that Kei-15's final measurement is based on only 11 tokens (recall, the number of tokens is reduced by 10 per iteration; the final measurement in the iteration is taken when the number of tokens goes below 20, for Kei-15 this is 11; her second to last measurement taken with 21 tokens yields a median value of 4.59 σ/sec, not quite as bad at *only* 10.2 percent lower than her actual median). Ultimately, we can say that the number of tokens does matter: 200 utterances appear to be enough data to yield stable values for analysis. Starting at about 80 tokens, the differences may be more consequential.

6.2.2 Measurement size and the stability of the statistical models

We now turn to the second of these investigations, where we examine how decreasing the number of measurements impacts the stability of the statistical analysis. The data used here are the 80 speakers from the original 159 who have more than 100 measured utterances in the dataset each. Since these are the speakers who contribute the most data to the dataset, this subset contains 24,682 articulation rate measurements, 81.9 percent of the total rate data examined in the last chapter. The number of speakers from each region, ethnicity, and sex included in this subset are displayed in Table 6.1. Table 5a on the website, which provides the summary information for each speaker, is ordered by the number of utterances in the dataset and can be used to look more closely at these 80 speakers.

A mixed-effect model fit with the same factors as earlier (NUMSYLS, with five knots, AGE, with four knots, REGION, ETHNICITY, SEX, and an ETHNICITY x SEX interaction) was tested on the subset data but it did not fit the data well. REGION and the ETHNICITY x SEX interaction are found not

Table 6.1 Speaker demographics for the speakers who contribute more than 100 utterances

Region	African Am.		European Am.		Latino/a		Lumbee	
	Female	Male	Female	Male	Female	Male	Female	Male
C. NC	6	4	2	3	5	2	–	–
E. NC	3	2	3	1	–	–	–	–
OH	1	1	–	–	–	–	–	–
S. NC	1	2	1	2	–	–	1	3
TX	–	–	1	5	10	6	–	–
DC	10	2	–	–	–	–	–	–
W. NC	1	–	–	1	1	–	–	–

half of the speakers and, therefore, contain much less social variability. This, of course, matters in terms of modeling the data, as we see here by the failure of the earlier model to fit the new dataset adequately. The best model for these 80 speakers and 24,862 measurements is shown in Table 6.2. It obtains an R^2 of 0.611, not quite as good as the full model fit on the full data, but still quite good. I have not plotted these model results, but comparing the fixed-effect results from the model here with the full model of Table 5.2 shows that these effects are highly correlated ($r = 0.996$, $p < 0.000001$).

We now compare the model in Table 6.2, as a baseline, to models generated on the same speakers, but where I have randomly selected smaller numbers of tokens from each speaker. This is done three times – first with 80 tokens per speaker (as this was identified above to be a point at which central tendencies become less stable), and then with 40 tokens per speaker, and, finally, only 20 tokens per speaker. Thus, we build three additional models with the same parameters as Table 6.2 on a subset of these data, the same 80 speakers, but with only 80, 40, or 20 tokens from each. For the 80-token subsample, this involves a dataset of 6400 measurements, 25.9 percent of the tokens used for the model of Table 6.2. By the time we extract only 20 tokens per speaker, we have a dataset of 1600 measurements, only 6.5 percent the size of the data modeled in Table 6.2. How good are the models for these massively reduced datasets? The fixed-effect estimates and p values for each of these models, along with the original model from Table 6.2, are shown side by side in Table 6.3.

The model based on 80 tokens is almost as good as the model fitted

Table 6.2 Mixed-effect model for the 80 speakers with the most data

Factor	Estimate	Std err.	p
(Intercept)	–0.623	0.348	–
NumSyls	1.053	0.014	0.0001
NumSyls'	–5.634	0.140	0.0001
NumSyls''	14.677	0.501	0.0001
NumSyls'''	–9.699	0.493	0.0001
Age	0.091	0.023	0.0001
Age'	–3.617	1.017	0.0001
Age''	3.941	1.112	0.0001
Sex = male	0.181	0.084	0.0216
Ethnicity = Latino/a	0.290	0.948	0.0016
Ethnicity = Lumbee	–0.089	0.190	[0.6274]
Ethnicity = European Am.	0.104	0.105	[0.2746]

parameters reach significance. In terms of the models' significances, the only noticeable difference is that the comparison between African American and European American ethnicity almost reaches significance, with $p = 0.087$, in the smaller model, while it was far from significant, at $p = 0.275$, in the larger model. For the model generated from 40 tokens per speaker, we again get roughly equivalent results. The p values have increased some, but the same factors remain significant and the R^2 value is equivalent and still quite good, at 0.587. The model generated from 20 tokens per speaker still yields significant results for the NumSyls factor, the Age factor, and marginally for the Latino/a comparison against the African American Ethnicity baseline, but it no longer finds a significant Sex effect and, in general, we see that the p values are quite a bit higher than the other models. Somewhat remarkably, this 20-token model still results in an R^2 value of 0.581, but at this point we are likely overfitting the data (see §5.5).

We get a better view of the differences between the different models when we compare the plots of the fixed effects. Figure 6.2 displays these four factors (along rows) for each of the four models (along columns). Here we see that while the relative positions and significances for the different factors are similar across the four models, in fact the randomly sampled smaller datasets predict higher speech rates for the social fac-

Mixed-effect models for the full data, 80, 40, and 20 tokens sampled from each of the 80 speakers

	Full model		80 Tokens		40 Tokens		20 Tokens	
	Estimate	*p*	Estimate	*p*	Estimate	*p*	Estimate	*p*
	−0.623	–	0.168	–	0.038	–	0.352	–
	1.053	0.0001	0.948	0.0001	0.972	0.0001	0.894	0.0001
	−5.634	0.0001	−8.589	0.0001	−8.960	0.0001	−7.477	0.0001
	14.677	0.0001	18.099	0.0001	18.901	0.0001	15.147	0.0001
	−9.699	0.0001	−10.429	0.0001	−10.871	0.0001	−7.975	0.0001
	0.091	0.0001	0.041	0.0012	0.048	0.0006	0.044	0.0060
	−3.617	0.0001	−0.306	0.0022	−0.344	0.0012	−0.309	0.0120
	3.941	0.0001	0.413	0.0024	0.463	0.0018	0.413	0.0150
	0.181	0.0216	0.240	0.0052	0.224	0.0284	0.119	[0.3020]
= Latino/a	0.290	0.0016	0.295	0.0020	0.273	0.0172	0.232	[0.0538]
= Lumbee	−0.089	[0.6274]	−0.005	[0.9562]	0.057	[0.8018]	0.024	[0.9444]
= Euro. Am.	0.104	[0.2746]	0.184	[0.0870]	0.101	[0.4526]	0.210	[0.1290]

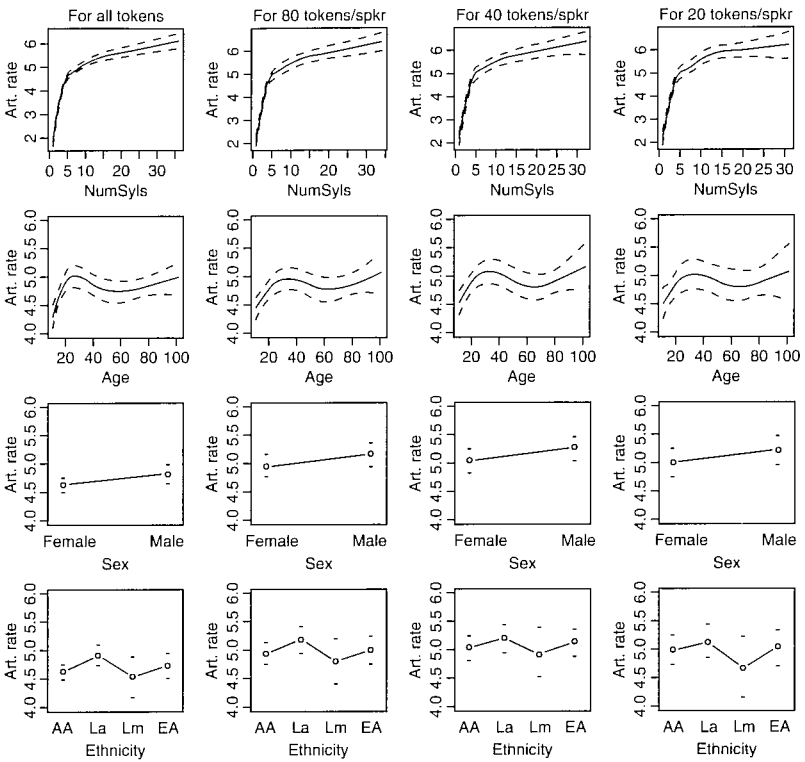

Figure 6.2 Comparison of model results for four sample sizes

models based on more data points, with larger confidence intervals and noisier shapes (for the nonlinear continuous terms).

6.2.3 Making sense of conflicting results

This section has left us with some apparently conflicting findings. The differences discovered here for the (in)stability of the median values could be interpreted as problematic for the main analysis of Chapter 5. Only half of the speakers I analyzed there have over 100 tokens and only 111, or 69.8 percent, of the 159 speakers have more than 80 tokens. Ultimately, this seems to point towards the importance of having large datasets for studying speech timing features, even if the large number of

some evidence that, despite possible fluctuations in speaker medians, the mixed-effect modeling based analysis of the individual measurements is robust against sampling limitations. Ultimately, the combination of methods, as done in Chapter 5, provides some assurances that the findings from the speaker-level analysis (§5.4), despite using many speakers for whom only small numbers of measurements were available, are meaningful and reasonable. Thus, the two approaches can be seen to buttress one another.

6.3 How long is a pause? (An experiment in modeling)

The previous section examined the stability of the speech rate measurements, but not the pause duration measurements, even though this is also an area worth further consideration. In fact, the investigation of pause duration raises some more complicated questions about the stability of the measurements, and, in particular, about the criteria, or THRESHOLDS, used to determine which silences in speech are counted as pauses. Some silences during conversational speech are not "pauses" in language production, but simply silence – periods of time when no one is speaking or when nontalk action is taking place. At the same time, extremely short silences might simply be acoustic phenomena, such as the short silence occurring during voiceless stop consonants or between adjacent stops. Deciding what length of silences will be counted as pauses in a given analysis is a long-standing problem in research on pause (and articulation rate, for that matter, since the determination of what counts as a pause has ramifications on what is excluded from "speech"[2]). Several of the published examinations of pause have discussed the difficulty of, or relative arbitrariness typically used in, determining the length of pauses included in analyses of these speech timing phenomena (e.g. Robb et al. 2004). For instance, in his textbook on sociophonetics, Thomas (2011a), following Robb et al. (2004), explains,

> Silent intervals under 50 ms are reliably due to stop closure, those over 250 ms are reliably pauses, and those in the 50–250 ms range could be either one. These figures are reasonable, so the minimum value for what you count as a pause should be no less than 50 ms and no more than 250 ms. It's up to [the individual researcher] to decide where to set the threshold between those values, though. (Thomas 2011a: 185)

This can be thought of as the THRESHOLD PROBLEM (Campione and Véronis 2002): Does one include pauses that are extremely short (or long) as

relevant data for an analysis of pause? What is an "extremely" short (or long) pause? Typically, as indicated in the Thomas quote above, pause durations have been limited to pauses *above* a certain duration in length. Kowal and O'Connell, for example, adopted "as a convention the minimal cut-off point of 270 milliseconds for silent pauses" (Kowal and O'Connell 1980: 62), while Goldman-Eisler and colleagues adopted various low threshold values from 100 ms (e.g. Henderson et al. 1966: 208) to 250 ms (e.g. Goldman-Eisler 1968: 12), depending on the experiment. Redford (in press) provides a detailed, articulatorily motivated variable threshold based on the phonetic context of each silent interval. Such a detailed and rigorous approach is quite rare in the literature, but has the advantage of rooting out differences in silence that are likely primarily articulatory or perceptual from those that are cognitively real. Very few papers (beyond Campione and Véronis 2002) discuss their maximum pause thresholds.

It is likely immediately clear to many readers how this is important. In sociolinguistic terms this is the classic problem of DELIMITING THE ENVELOPE OF VARIATION OR CIRCUMSCRIBING THE VARIABLE CONTEXT (Tagliamonte 2006). If we are interested in searching for systematicity in pause duration variation, the ordered heterogeneity of Weinreich et al. (1968), the outcome of this search will surely be a function of what potential pause data we include in our analysis and what we exclude!

In Chapter 4, for the reading passage analysis, I somewhat arbitrarily chose a low threshold at the high end of the conventionally used values, 200 ms, and, in Chapter 5, I used a low threshold of 60 ms aiming to include as much data in the analysis as possible. The transcripts in SLAAP are time-aligned to the phonetic utterance at an extremely fine level. Pauses are delimited in the act of transcription based on a criterion of silence longer than 60 ms. As I pointed out at the start of §5.3.2, the main examination of pause there included all of the within-talker-turn silences in these transcribed data, including those pauses that are only 60 ms in duration; 60 ms is extremely short – and those shortest pauses are likely extremely "noisy" data, in the sense that transcribers and analysts have the hardest time accurately identifying the shortest pauses, especially those not cued by intonation or syntactic information.

The question of what is the most meaningful threshold for pause duration analyses is likely one best examined through experimental work. Such research – like that by Duez (1982, 1985) – has shed light on the perceptibility of pauses. From a speech production perspective, however, the question of disentangling what we might think of as "cognitive" pauses from other kinds of short silences (nonplanning-related breaks

for breath, stop-gaps, and so forth) is a trickier issue, although one that, as we saw in Chapter 2, is also an area that has had a long history of research within psychology and psycholinguistics. Sticking to the corpus-based focus of this book, we consider this problem here by assessing just what thresholds provide the most systematic view of the pause data. This proceeds from the hypothesis that pauses outside of some "core" durational range may distribute in different, or noisier, ways than pauses that are more purely, well, "pauses."

Let us start by examining the distribution of the pause duration data analyzed earlier. As was explained, pause duration data distribute in a roughly log-normal fashion, as illustrated in Figure 6.3. The distribution has a long tail consisting of the pauses longer than about 2 s (7.6 log-ms). Since the pause durations are bounded by 60 ms at the low end there can be no tail on the left. The bulk of the pause data, 86 percent, fall below 1 s in length (6.9 log-ms).

So, what happens if we adjust the edges of the distribution?

In order to determine whether the "extremely" short or long pauses negatively (or positively, for that matter) impact the orderliness of the data, I conducted a stepwise comparison of models using the same factors that arose as significant in the model of §5.3.2 (i.e. ETHNICITY, SEX, and REGION) over a systematically manipulated subset of the data. I iterated over subsamples which increased the low threshold from the minimal measurement length of 60 ms in 20 ms increments until a maximum value, excluding those pause durations below the low threshold each time. At the same time, I ran this stepwise comparison over numerous maximums, ranging from 5000 ms, the measurement maximum, down to 1000 ms, in 400 ms increments. In order to assess some measure of model fit – that is, to compare how well we are able to model each of these subsets – I calculated two values for each model. I determined the model's simple R^2 value – how well the model's fitted values correlated with the actual observations – and a sum of the model's t values (the quotient of each estimate over its standard error). Larger t values indicate more confident estimates (regardless of the estimate size), and thus a sum of each model's t values seems to me a useful way to compare the models' success.[3] While I adjust the data examined for each model based on thresholds on nontransformed pause durations (in ms), the actual models, like that in §5.3.2, are run against log-duration values (log-ms).

Surprisingly perhaps, removing the (few) large values by decreasing the maximum threshold reduced the quality of the modeling, so the data I present here include all of the long durations, that is, pauses all the way up to 5000 ms. This is empirically grounded, in that models

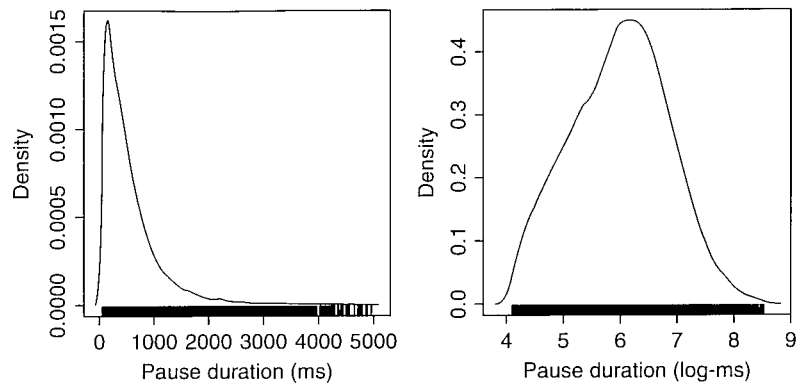

Figure 6.3 Pause distributions

without a lowered maximum threshold performed better than models with one, but, in some ways, the maximum threshold value seems not as important as the minimum threshold since there are relatively few long pauses in the dataset, as seen in Figure 6.3. There are only 720 pauses longer than 2000 ms (2.4 percent of the total data), only 176 pauses longer than 3000 ms (0.6 percent of the total data), and 52 pauses longer than 4000 ms (0.18 percent of the total data). For the very short pauses, on the other hand, since there are so many of them, this is a more important question and likely has a larger impact on the analysis. Figure 6.4 shows the results of this stepwise movement through the data, increasing the minimal values of examined pauses. We see that both the sum of t values and the model R^2 increase over the first few hundred milliseconds. At about 400 ms, where the t values sum to 22.41 and the R^2 equals 0.087, they begin to peak and at 560 ms we have the largest t value sum, of 22.71, and a relatively high R^2 (0.090). Of course, this R^2 is quite low, and it does increase later in the duration increments, but at the expense of most of the tokens and most of the models' relative surety (i.e. summed t values). This t value maximum and R^2 peak are indicated in Figure 6.4 with a vertical dot–dash line. We note that the threshold suggested here by this experiment is a quite a bit higher than the standard cutoff used in the literature, 150–250 ms, and that it also decreases our total data from 29,614 to 10,619, a reduction of almost two-thirds of the total data. In fact, this high peak value is quite surprising as it goes against the majority of research on pause, which has, implicitly when not explicitly, assumed and often found that much shorter stretches of silence can be usefully studied as pauses.

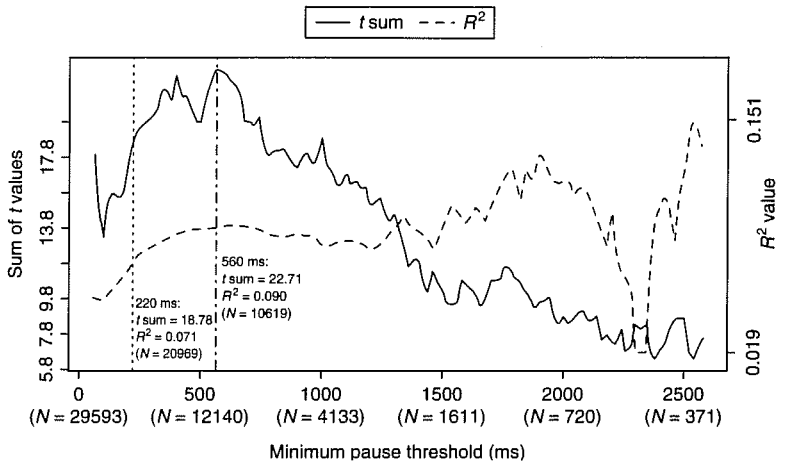

Figure 6.4 Stepwise comparison of minimum threshold increases on pause duration modeling

Especially for the *t* values, Figure 6.4 paints a picture of high volatility in the pause modeling. This makes sense, as all of the R^2 values are quite low and none of these models fit the pause duration data very well. It is perhaps unjustified – or at least unnecessary – to take the peak here as an absolute determination of the best threshold. If we relax the view of what constitutes the "peak" in the models' outcomes, we note that at about 220 ms the models *begin* to peak, with a *t* value sum of 18.78 and an R^2 of 0.071. This point is indicated by the vertical dotted line in the figure. It is quite close to the kinds of thresholds used in most research on pause and is, perhaps, a better place to "draw the line" for further investigation.

In order to assess how well this longer pause minimum improves the modeling of the data, let us now consider new statistical models of the pause duration data using these two new threshold points, greater than or equal to 220 ms and greater than or equal to 560 ms. For further exploration, I also report the results of the same model parameters fit to just the very short pauses, those under 560 ms. The fixed-effect components of these three models, along with the fixed effects from the main pause duration model of Table 5.4, are shown in Table 6.4. These models have been fit to the trimmed version of the pause dataset from earlier in order to make their results most comparable with the original model's predictions. As such, the models for pauses above 560 and 220 ms

Table 6.4 Mixed-effect models for full data and three different threshold levels

Factor	Full model		≥ 560 ms model		≥ 220 ms model		< 560 ms model	
	Estimate	*p*	Estimate	*p*	Estimate	*p*	Estimate	*p*
(Intercept)	5.829	–	6.747	–	6.245	–	5.437	–
Region = Eastern NC	0.004	[0.9394]	0.023	[0.5390]	0.056	[0.2482]	–0.047	[0.2024]
Region = Ohio	–0.264	0.0004	0.034	[0.5136]	–0.090	[0.1274]	–0.163	0.0001
Region = Southern NC	–0.092	[0.1268]	–0.022	[0.5842]	–0.049	[0.3198]	–0.041	[0.3084]
Region = Texas	0.001	[0.9818]	–0.033	[0.2544]	0.017	[0.6560]	–0.029	[0.3390]
Region = Washington, DC	0.264	0.0002	0.174	0.0001	0.192	0.0001	0.079	0.0294
Region = Western NC	–0.033	[0.6410]	0.115	0.0114	0.121	0.0408	–0.110	0.0136
Sex = male	0.099	0.0018	0.052	0.0098	0.071	0.0056	0.023	[0.2658]
Ethnicity = Latino/a	0.103	/0.0530/	0.081	0.0132	0.072	[0.0878]	0.057	[0.0788]
Ethnicity = Lumbee	0.125	[0.1156]	0.151	0.0048	0.162	0.0170	–0.045	[0.3796]
Ethnicity = European Am.	0.060	[0.1616]	0.036	[0.1934]	0.054	[0.1300]	0.025	[0.3592]

here yield slightly different R^2 values than they did in the thresholding experiment above.

When we look at the specific predictions of the different models, we note that there are both similarities and differences among the subsampled models and the main model (shown in the leftmost columns of Table 6.4). The comparison between Latinos and African Americans is on the cusp of significance in the full model ($p = 0.053$). In the ≥560 ms model it is much more significant ($p = 0.013$), but it is not in the ≥220 ms model and the model based on only the shortest pauses. We can infer from this that the difference between pause durations by African American (the baseline) and Latino/a speakers is a result of differences in the longer pauses only. The shorter pauses obscure this pattern. Although the pattern is different, we can also make a similar inference for the difference between the Lumbees and the African Americans. This comparison does not yield significance in the main model or in the model for the shortest pauses, but the two models that exclude the shortest pauses do show that the Lumbees have significantly longer pauses than the African Americans. It seems that the very shortest pauses disguise this effect as well.

When we turn to the REGION differences, we see the opposite pattern for the Ohioans in their comparison with the speakers from Central NC (the baseline) from what we see for the Lumbees in ETHNICITY. The finding of significantly shorter pauses for Ohioans in the main model is driven by the significant difference for the shortest pauses. Their short pauses are significantly shorter than those from Central NC, but excluding pauses shorter than 220 ms removes this effect. The Western North Carolinians have at first glance the pattern that is hardest to explain. Their comparison to Central NC is not (even close to) significant in the main model, but does obtain significance in all three of the subsample models. This striking difference makes sense when examined closely: The Western NC speakers have significantly *longer* pauses in both the ≥560 ms ($p = 0.011$) and the ≥220 ms ($p = 0.041$) models and significantly *shorter* pauses in the <560 ms model ($p = 0.014$). These opposing tendencies cancel one another out in the main model! Finally, we observe that the Washington, DC speakers produce significantly longer pauses than the baseline of Central NC in all four models, indicating that they have longer pauses throughout the durational range, although we note that the comparison is much less significant for the shortest pauses alone than for any of the other models ($p = 0.029$ compared to $p \leq 0.0002$ for the others). The last factor, SEX, shows a significant effect for all of the models except for the model of the shortest pauses.

Figure 6.5 displays these model results graphically, with the three rows of the figure displaying the different factors and the four columns displaying the results from the four different datasets. The difference in the absolute values of the model predictions are, of course, a result of the different thresholds applied. This necessarily adjusts the overall range and central tendencies of the values, as we will explore shortly. This graphical view helps to reinforce the observations just made. For instance, comparing the ETHNICITY plots, we can see the pattern discussed above for the Lumbees (identified as "Lm"), where the model for the shortest pauses (rightmost column) contrasts with the models which exclude the shortest pauses.

All in all, the three new models are quite instructive. As we saw above, the model based on pauses above 560 ms obtains the best fit with the actual data ($R^2 = 0.094$), while the model based on pauses above 220 ms does slightly worse ($R^2 = 0.073$) but still outperforms the full model's fit ($R^2 = 0.053$). The model generated on just the shortest pauses does the poorest, obtaining an R^2 of only 0.038. Yet, examining this poor model has still shed some important light on the pause duration data. While we have seen that we get the best model fit (relatively speaking) by

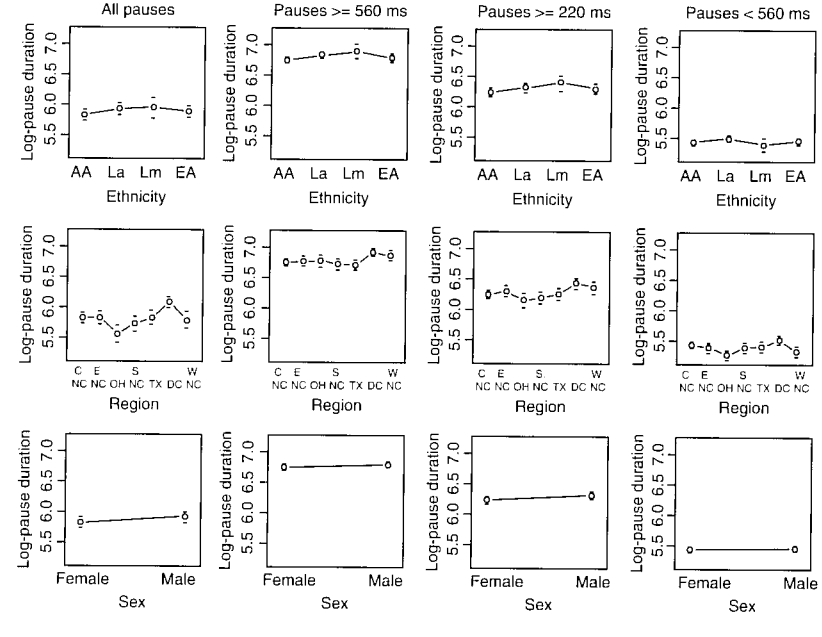

Figure 6.5 Comparison of pause model results for different threshold values

excluding the shortest pauses, we have also seen here that the shorter pauses do pattern with social factors. Most importantly, we have seen that pause durations are not influenced by social factors in a unidimensional manner. Speakers, and groups of speakers, can have productions that differ according to the overall length of the pause. This seems to indicate that there is not a single threshold for what makes a pause a "pause," but rather several cutoff points or durational "preference regions" and (groups of) speakers may be differentially sensitive to these different regions.

This points to the importance of work on the perceptibility of pauses, such as that by Duez (1982, 1985, 1993). Further research may also be able to tease apart production differences depending on the function or functions of the pauses. Some pauses, especially the shortest ones, likely play a primarily or solely cognitive role (e.g. planning or word-searching) – and be primarily a speaker-oriented behavior – while others, especially presumably longer pauses, play a primarily or solely discourse-related role (e.g. pausing for emphasis) – and are more listener-oriented – and some pauses, obviously, have multiple functions and purposes (e.g. planning and breathing and emphasis). I will return to considering pause modeling in §6.5, but many of these ideas will have to wait for further research and especially for experimental study.

6.4 Articulation rates in Intonational Phrases and the effect of phrase-final lengthening

We saw from the analyses of Chapter 5 that articulation rates are variable across social categories and, more specifically, that ethnicity, sex, age, and region do significantly impact articulation rates. However, we also saw, especially at the per-utterance level, that the largest effect by far is from the one nonsocial factor, the number of syllables in the utterance. This is not entirely surprising; several researchers have noted the fact that there are within-utterance linguistic factors that influence variability in articulation rates. Namely, it is quite clear that syllables at the end of utterances/Intonational Phrases are lengthened (a phenomenon called PHRASE-FINAL LENGTHENING) and this, for obvious reasons, may play a large role in an utterance's overall rate of articulation (Dankovičova 2001, Yuan et al. 2006, Quené 2008). We turn now to look more closely at this effect by examining the data in terms of INTONATIONAL PHRASES (IPs)[4] rather than phonetic utterances (stretches of uninterrupted phonation surrounded by silence; see §3.3), and by paying close attention to the distribution of syllables in the IPs.

This extended look also lets us examine how much the specific quantitative results of the main analysis are a result of the unit of analysis – the phonetic utterance. Does an analysis at the IP level yield different findings, either in relative terms (comparisons across speakers) or absolute terms (the actual values obtained)? I here examine much of the same source data from one of the regions from earlier – the South Texas data – but chunked and coded at the IP level rather than at the phonetic utterance level. This closer examination also has the benefit of allowing us to assess the degree to which the heavily automated methods of the rest of this chapter and the previous chapter are accurate. Those data came from automatically extracted syllable counts from orthographic transcripts in SLAAP. The transcripts were finely time-aligned with attention paid, in particular, to delimiting pauses, but most of the transcripts were not developed specifically with the view of extracting syllable counts for an analysis of speech rate. The smaller dataset we turn to now, while still syllable-counted using SLAAP's automatic syllable-counting algorithm, was designed specifically for this analysis, and extreme attention was paid to ensuring that the transcripts would be accurate for the syllable counter and for the durational measures.

The present discussion arises from work with Erik Thomas (Kendall and Thomas 2010), in which we examined these questions in detail in order to understand whether variation in utterance-level articulation rate measures is connected, possibly even spuriously, to variation in phrase-final lengthening.[5] To do this, we examined data from 36 of the speakers from the South Texas site used in the primary analyses of Chapter 5. For this closer look we delimited IP and intermediate phrase boundaries for all of the transcribed talk, rather than delimiting the talk by phonetic utterance, or inter-pause stretch. We still segmented the speech from the silence at the same finely accurate level as before, with a 60 ms pause threshold, but here we also subsegmented the utterances into final feet and pre-final feet. Figure 6.6 shows a Praat Editor window demonstrating how the utterances were chunked relative to the phonetic utterances of the SLAAP transcripts (and the earlier analyses). The top tier shows the phonetic utterances as they are delimited in the original SLAAP transcripts. The middle tier shows the same text (and boundaries against silent intervals) but has the individual IPs delimited. Thus the utterance "I moved into the hospital as a unit secretary" is a single "chunk" in the phonetic utterance tier and was analyzed in Chapter 5 as a single utterance, but here is split into two IPs, "I moved into the hospital as a" and "unit secretary," with the "a" ending the first IP showing clear phrase-final lengthening. The bottom tier delimits

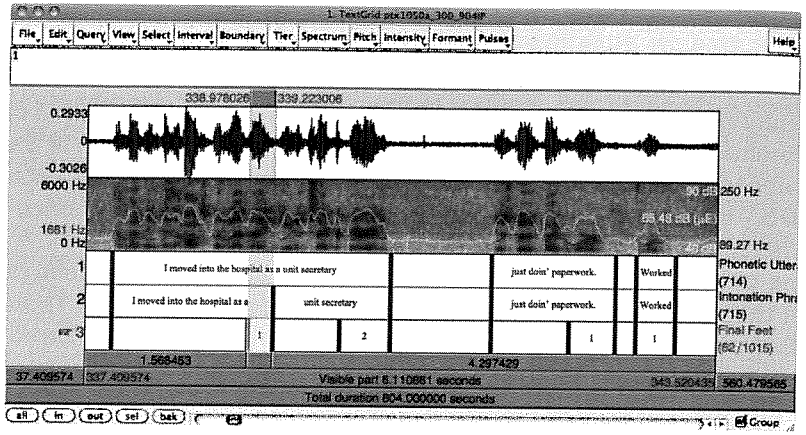

Figure 6.6 Praat Editor window showing an IP-coded transcript for data analysis

the temporal span of the final foot, and the number in the final foot interval is a manually assessed count of the number of syllables in this span. The analysis scripts, which compute the syllable counts for the entire IP, then use this final foot syllable count to determine the distribution of syllables over the full IP.

The analysis here examines more data than did the main analysis for the Texas fieldsite. Thirty-four of the 36 speakers examined here were speakers included in the main analysis, but here we include two additional speakers, Mexican American women born in 1922 (JV-85) and in 1937 (CP-70), for whom we have hand coded IP-level data. Further, since this closer analysis involved coding the data at the IP level, rather than the phonetic utterance, this closer analysis also examines a larger number of smaller units. On average, there are 133 more measurements per speaker for the closer analysis than for the main analysis presented earlier, and, again on average, the IPs examined here are 1.6 σ shorter than the phonetic utterances examined earlier.

Figure 6.7 shows the correlations between the speaker-level median articulation rates for the 34 speakers who are analyzed in both the main analysis of Chapter 5 and here. The solid line shows the best-fit line through the data. The dashed line shows a best-fit line when three speakers who have substantially slower rates in this IP-based analysis are removed from the correlation test. The dotted line shows the line with a slope of 1, the $X = Y$ line, or what would be the perfect relation between the two datasets. The legend in the figure indicates the intercepts for the dashed and solid

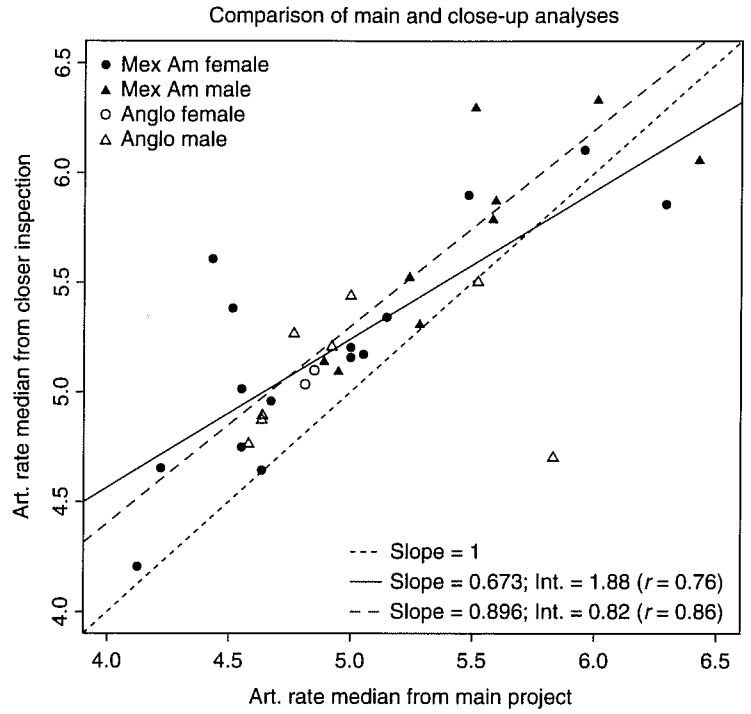

Figure 6.7 Correlation between rates from the main analysis of Chapter 5 and the IP-based analysis

lines, as well as their r values. The two datasets are generally correlated ($r = 0.76$, $p < 0.000001$), although we also note that a few speakers stand out as having quite different rates between the two analyses and that only a few speakers' median values fall perfectly on the $X = Y$ line. While the best-fit line is not parallel to the perfect relation line, we still can note that the majority of the speakers actually do cluster in a dimension that parallels $X = Y$ (i.e. the dashed line in the figure almost runs parallel to the $X = Y$ line). For most of the speakers, this closer chunking of the data results in slightly faster speech rates, especially for speakers in the slower range (~4.0–5.5 σ/sec). There are a few speakers who stray further from this and there is some larger scatter for the speakers with the fastest rates. The difference arising for most of the speakers makes sense, however, in terms of the different units used in computing the articulation rates. Many of the phonetic utterances used in the main analysis span multiple IPs, and, as such, have multiple phrase-final lengthened segments. This would, we

might imagine, contribute to a generally slower rate compared to the rates for single IPs, which have only one lengthened foot.

So, while there are some definite differences between the two sets of measures, the fact that most of the speakers when analyzed by IP fall on a line roughly parallel to the "perfect relation" line indicates that, especially in relative terms, the two ways of chunking the data for speech rate analysis are somewhat equivalent. The median difference between each speaker's rates as calculated in these two ways is 0.235 σ/sec, with rates faster for the IP-based measure.

Figure 6.7 displayed the articulation rate medians for the complete IPs for each speaker. But let us turn now from the relationship between the hand-delimited IPs and larger, phonetic utterance chunks created by the transcribers for SLAAP to the relationship between each IP's phrase-final lengthened material, what we will refer to hereafter as its final foot (FF), and the spoken material preceding the final foot, or the pre-final foot (PFF). Every complete IP by a speaker is composed of a PFF followed by an FF in the dataset. (We do not include here one-syllable utterances or utterances that only contain a stressed FF.) Thus, we can assess the impact of the FF (and PFF) on the overall rate by including their lengths as factors in the statistical analyses. We can also include a factor based on the proportion of the IP that is comprised of the FF in an IP-level statistical model (i.e. an utterance-level model, but here an individual "utterance" is an IP). There are two relevant proportions – one based on the proportion of an IP's *syllables* that are in the FF and one based on the proportion of an IP's *duration* that is in the FF. We will examine both to see which performs better in the statistical analysis.

First, as we saw in the Praat screenshot of Figure 6.6, it should be noted that final feet can be different lengths. They are primarily one (N = 6047) or two (N = 2764) syllables long, but longer final feet are also possible (3 σ, N = 505; 4 σ, N = 27).[6] Altogether, final feet contain 34 percent of the total syllables spoken in these data and make up 46 percent of the total temporal duration. This difference between syllable length and duration makes sense, given that FF are spoken at a slower rate than PFF. Figure 6.8 displays the distribution of the syllables in final feet by total IP length and the *N*s in these data for the various IP lengths, reinforcing the view that a huge amount of the data come from very short IPs. It also confirms that the proportion of syllables in the FF of the IPs decreases systematically as IP length increases.

The statistical modeling for these data was closely based on the utterance-level model of speech rate from §5.3.1, with some adjustments to account for the fact that this subset of data had fewer relevant social

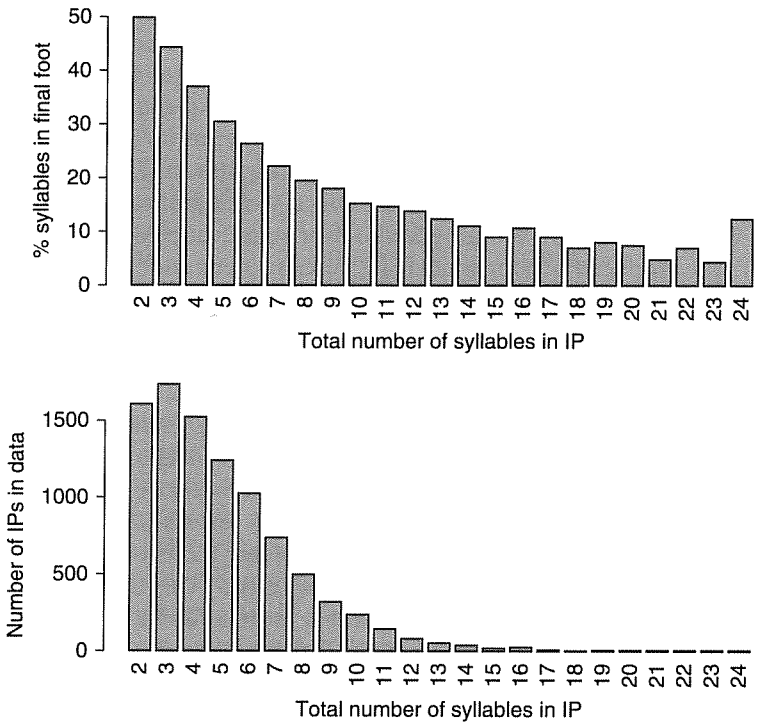

Figure 6.8 Syllable distribution in all IPs

factors, since it comes from one region with only two ethnicities. Specifically, we examine the following factors:

- The number of syllables in the IP (NumSyls; continuous, with a mean of 4.98 σ)
- The number of syllables in the IP's final foot (FFSyls; integers from 1 to 4)
- The proportion of the IP's total syllables which fall in the final foot (FFSylProp; continuous)
- The proportion of the IP's total duration which falls in the final foot (FFDurProp; continuous)
- The speaker's Ethnicity (either Anglo or Hispanic[7])
- The speaker's Sex (female or male)
- The speaker's Age (continuous, with a mean value of 57.1)

Modeling began with an equivalent model to that in Table 5.2, the best utterance-level model for the entire dataset, but without the REGION factor since that factor is not relevant for this subset of the data. That model did quite poorly for these data, with none of the social factors arising as significant and most of the nonlinear terms for NUMSYLS coming out as nonsignificant. Removing the nonsignificant ETHNICITY x SEX interaction (since neither of the main effects were significant) improved the model somewhat, bringing SEX to near-significance with $p = 0.058$, but still produced a poor model. In order to determine whether this failure of the original model to fit these data was a result of the new set of measurements (again, IP-level measurements rather than phonetic utterance-level measurements, plus two additional speakers) or whether this was simply a function of this subset of speakers, I also reran the original model (again without REGION) on just the Texas data from the original, phonetic utterance, data. This model also performed poorly. But, after removing the (nonsignificant) ETHNICITY x SEX interaction from that model, it did yield significance for SEX ($p = 0.006$) and ETHNICITY ($p = 0.022$) and for all of the nonlinear terms of the NUMSYLS factor (although AGE remained nonsignificant). This indicates that the poor model fit for the IP-level data is more a function of the new measurement technique (and possibly slightly different speakers) than a property of these particular speakers more generally.

To determine what patterns were in these IP-level data, modeling began anew, starting as we did earlier with a base, null model and adding factors that improved the model (according to likelihood ratio tests) one by one, in the order in which they had the largest improvement in the model. The best model for these IP-level Texas articulation rate data, after trimming, is presented in Table 6.5 and Figure 6.9 (this model has a random intercept and random slope for NUMSYLS for speaker).

We obtain significant main effects for NUMSYLS, SEX, and ETHNICITY. Comparing Figure 6.9 with Figure 5.3, the model for the full data, we see that SEX has roughly the same effect, with females having slower rates than males. The effect of ETHNICITY is somewhat similar as before, although the full model found a significant interaction between ETHNICITY and SEX and showed that only male Latinos had faster rates than European Americans, though here all Hispanics are predicted to have faster rates than the Anglos. AGE was not found to be significant in the final model, though preliminary models indicated a nearly significant linear effect, with rates decreasing slightly as AGE increased. The proportion factors, PROPFFSYLS and PROPFFDUR, were quickly outperformed by the combination of NUMSYLS and FFSYLS. The factors NUMSYLS

Table 6.5 IP-level mixed-effect model for Texas articulation rates

Factor	Estimate	Std err.	*p*
(Intercept)	3.081	0.231	–
NUMSYLS	0.282	0.040	0.0001
NUMSYLS'	–0.159	0.061	0.0070
FFSYLS	–0.066	0.127	[0.6033]
SEX = male	0.586	0.148	0.0001
ETHNICITY = Hispanic	0.518	0.165	0.0017
NUMSYLS x FFSYLS	0.069	0.031	0.0270
NUMSYLS' x FFSYLS	–0.104	0.046	0.0229

$R^2 = 0.366$.

and FFSYLS together capture related information to PROPFFSYLS and would not be expected to occur in the same model. The model makes it apparent that the actual values of these syllable counts matter more than their proportional distribution in the IPs. Overall, the social factors in this best model end up rather similar to those in the model described a moment ago for the utterance-level Texas data.

The model obtains an R^2 value of 0.366 in its fit with the actual data. This is quite a bit worse than the earlier articulation rate models were able to achieve. In fact, the model described above (though not shown), in which I fitted a model to the original, utterance-level Texas data alone and obtained significant results for NUMSYLS, ETHNICITY, and SEX, yielded an R^2 of 0.526. This seems to indicate that the phonetic utterance-level data, rather than the IP-based data, are most amenable to an analysis by social factors. This is, admittedly, lucky for large-scale analysis – the IP-level coding and the delimitation of final feet must be undertaken by hand and is much more time intensive than the chunking at pause boundaries.

The most important outcome of the model – and where, even with the lower R^2, it is most useful – is in what it tells us about the role of phase-final lengthening and final feet. There is an important interaction between the total number of syllables in the IPs and the number of syllables in the final feet, and this helps explain the nonlinear influence of the length of the utterance in syllables found earlier in §5.3.1. NUMSYLS in this model has a nonlinear component, but it is much smaller, much less curved, than was found earlier. The interaction shows that it is short utterances with long final feet that contribute to the curve in the syllable length effect. Also, we note that, while included in the model, FFSYLS, the number of syllables in the final foot, was not significant as

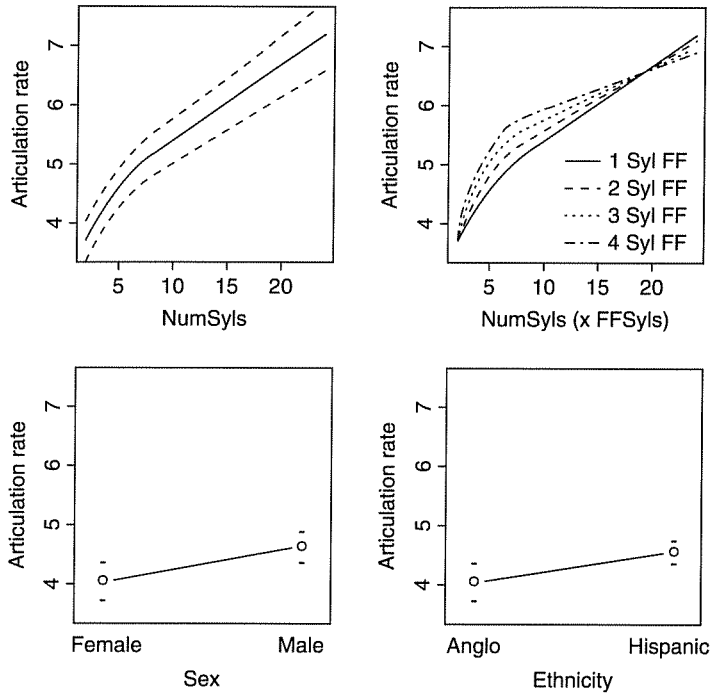

Figure 6.9 Effects in the mixed-effect regression model for IP-level articulation rates

a main effect when its significant interaction with NumSyls was added to the model. (Earlier models, without the NumSyls x FFSyls interaction, do have significant main effects for FFSyls, but these models are outperformed by the inclusion of the interaction.) Ultimately, the length of the final feet is important only in how it interacts with the total utterance length.

Another way to look at the influence of phrase-final lengthening and final feet in these data is to ask: How correlated are the FF and PFF rates with the overall rates? Testing the correlation, separately for each IP length, provides perhaps a clearer view of how the FF interact with the overall rates. Figure 6.10 displays the R^2 values for these correlations from IP lengths of 2 σ up to 16 σ (above this *N*s get too small for meaningful correlation testing). What we see here is that for the shortest IPs, of two or three syllables, the overall articulation rate is highly correlated with the FF rate. As IPs get longer, the influence of the FF decreases and

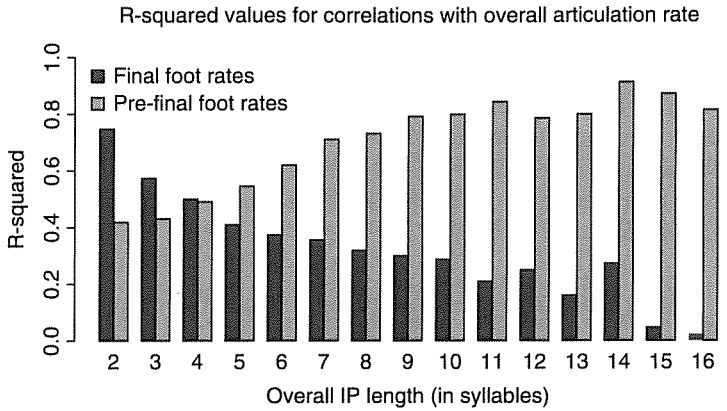

Figure 6.10 Correlation coefficients for the relationship between FF and PFF articulation rates and overall utterance rates

the overall rates correlate more highly with the rates of the PFF. This makes sense, since final feet are necessarily a high proportion of the total IP for short utterances, but the model above and the view in Figure 6.9 confirm that the phrase final foot plays an important role in determining the rate of short utterances. This is important to keep in mind, as much conversational speech is comprised of very short utterances (as demonstrated in Figure 6.8).

To review, this section has sought to better understand the role of utterance and IP length on articulation rate, and to assess whether similar results are obtained from a manual analysis at the IP-level as from the mostly automated analysis at the phonetic utterance-level. We have seen that phrase-final lengthening does indeed play an important role in influencing articulation rates, primarily for short IPs, which importantly constitute a large amount of conversational data. By including the length (in syllables) of each IP's final foot we can better account for the nonlinear effect of the number of syllables on articulation rate. As Figure 6.7 indicated, rates are fairly similar between IP-based and utterance-based articulation rate measures, though there are some individuals who stray from the overall correlation. We also saw from Figure 6.7 that articulation rates are on average slightly faster when computed at the IP level than at the utterance level. Phonetic utterances, stretches of speech separated by silence, can span multiple IPs and thus can contain multiple phrase-final lengthened final feet. This in turn decreases the articulation rate measures. At the same time, in terms of sociolinguistic

patterns and the influence of social factors, we have not seen here a very different picture from the IP-based analysis than we do from an analysis of the utterance-level data from the Texas speakers alone. The overall findings demonstrating the extent to which social factors explain differences in articulation rates and the relative strength of the statistical models (by that I mean the goodness-of-fit of the models) indicate that these speech rate patterns are robust against methodological differences.

6.5 Pause duration variability as a function of pause type

For the final "closer look" of this chapter, we revisit some of the pause duration data to ask whether including additional factors in the analysis can add clarity to the poor models we have thus far been able to develop. Up until now we have focused on the pause duration data solely in terms of the potential social factors available in the data. At this point, we will briefly consider the extent to which including several cognitive and linguistic factors improves our ability to model the pause data, for which we have thus far found only marginal effects of the social factors. Do we find additional, or stronger, social factors when we account for more cognitive and linguistic factors in the pause models?

It is clear from the previous research on pauses (discussed in Chapter 2) that pause realization is often a function of cognitive activity and, thus, considering factors that relate to the cognitive and linguistic status of the individual utterances should find significant effects and, thus, should improve our ability to model the data. To investigate many cognitive aspects directly (like task difficulty and sentence processing) requires experimental work and is outside the scope of the kind of corpus-based studies I pursue here. However, several additional factors beyond what I examined earlier can be coded from the corpus data. Pauses can be coded for where they occur in syntactic constituents and whether they occur with other hesitation markers, like filled pauses. The distinction between pauses that occur at grammatical junctions versus nongrammatical junctions, for instance, may help to improve the overall analysis.

The main analysis of pause made use of almost 30,000 pause duration measurements to assess the influence of the social factors on pause duration. Here, we examine a subset of the total pause data including the following factors in *addition* to the factors outlined in §5.3.2:

- Pause Type (PType), whether the pause is at a grammatical juncture (gr); whether it is "nongrammatical" (ng) and interrupts a syntactic constituent or a "normal" intonational contour; whether it occurs with a syntactic reset (rs); or, whether it occurs with a lexical restart or repetition (rl).
- Verb Phrase constituency (VP), whether the pause occurs directly before, after, or within a verb phrase, or n/a.
- Noun Phrase constituency (NP.PP), whether the pause occurs directly before, after, or within a noun phrase (NP) or prepositional phrase (PP; NPs and PPs are coded separately, but grouped together within this factor group), or n/a.
- Filled Pause Adjacency (FPAdj), whether the silent pause falls directly before a filled pause, after a filled pause, between two filled pauses, or n/a. As is the case throughout this study, filled pauses are not examined separately as dependent variables. They are of interest here particularly as they are thought to influence (or at least relate to) the duration of adjacent silent pauses (cf. Clark and Fox Tree 2002).
- Filled Pause (FP), for the actual type of adjacent filled pause. This was coded as one of: none, complex (for the cases where the silent pause was between two filled pauses), other (a catchall for discourse particles used as fillers, like "like," "you know," "well," etc.), uh, and um.

Since the coding of these factors takes extensive manual work and this consideration was conceived as an addendum to the main analysis, only 10 percent of the total data were targeted for coding in this manner. In actually, 3282 pauses (11.1 percent of the total 29,614 pause measurements) were coded.[8] To obtain this subsample, 6000 pauses were randomly selected from the full dataset using the random sampling function in R (`sample()`). Of these 3500 were coded, in the order they were retrieved by the randomization function; 218 tokens were discarded due to unsure contexts or other ambiguities. The 3282 tokens examined here come from 155 of the 159 speakers, with a mean *N* of 21.2 and a median *N* of 12 per speaker. Based on the discussion of threshold values in §6.3, we could imagine examining these data in several different ways, such as examining different ranges of pause durations separately and looking for more nuanced patterns, as in §6.3. For sake of time and space, I have decided only to follow the practice of the main pause duration analysis of Chapter 5, however, and proceed with this closer analysis using the same thresholds as were used there (60 ms low threshold and 5000 ms high threshold).

Figure 6.11 displays a summary of these factors and their effect on the pause duration data. The *N*s from Figure 6.11 (in comparison, for example, to those in Figure 5.4) indicate that the random sampling

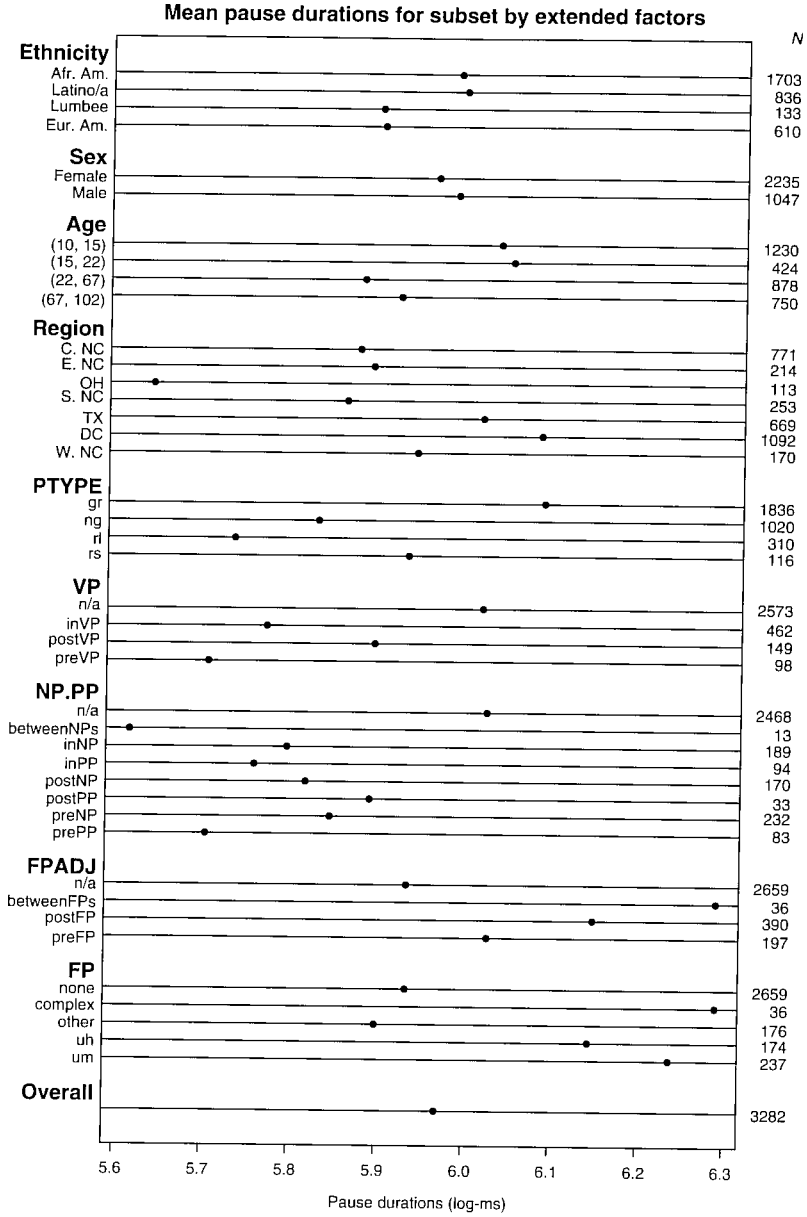

Figure 6.11 Mean pause durations for subset data by extended factors

achieved a highly proportional subset to the main data. For instance, both data sets contain 68.1 percent data from females and 31.9 percent data from males. The data for ethnicity and region are not quite as balanced but the *N*s still yield correlations above $r = 0.99$ ($p < 0.001$) with the full dataset. The comparison of *N*s for region is shown in Table 6.6, for further illustration.

Looking first at the influence of the social factors in comparison to the full dataset (i.e. comparing Figures 6.11 and 5.4), we see that there are some visible differences. (Note that the scale of Figure 6.11 is different than that for Figure 5.4; the inclusion of the additional factors extends the range of possible mean pause durations in the summary.) While we also saw lower pause durations for European Americans and Lumbees than African Americans and Latinos in the raw summary earlier, here we see a more bimodal distribution with African Americans and Latinos clustered closely together at 5.99 log-ms and European Americans and Lumbees clustered closely together at 5.90 log-ms. The AGE factor also appears to be a bit a more bimodally distributed for this subset, but this is also possibly an artifact of the summary plot, a result of the redistribution of some of the speakers in their early twenties by the algorithm that bins the continuous predictors for the summary plot. The earlier summary (Figure 5.4) based the second youngest age group on speakers between 15 and 25 (not including 25), while Figure 6.11 forms the second group from speakers between 15 and 22 (not including 22). Since younger speakers appear to have longer overall pauses, this reshuffling could account for the second group's higher mean pause duration. The REGION and SEX factors appear quite similar to their raw patterns in the full dataset. Overall, we can be comfortable that the subset data reflect the range of social variability in the full dataset reasonably well.

Turning to the new, additional factors, we see indications of some important patterns. For the PAUSE TYPE (PTYPE) factor, we see that grammatical pauses (gr) are substantially longer than the other pause types. Nongrammatical pauses (ng) are about 0.26 log-ms shorter than grammatical pauses. Lexical restarts and repetitions (rl) are the shortest

Table 6.6 Proportion of data and *N*s for region for main data and subset

	C. NC	E. NC	OH	S. NC	TX	DC	W. NC	Total
Full data *N*	7,473	2,050	734	2,172	6,147	9,495	1,543	29,614
Full data %	25.23	6.92	2.48	7.33	20.76	32.06	5.21	100
Subset *N*	771	214	113	253	669	1,092	170	3,282
Subset %	23.49	6.52	3.44	7.71	20.38	33.27	5.18	100

pause category, while syntactic resets (rs) are about halfway between grammatical and nongrammatical pauses. For the VP category, we see that pauses that are not adjacent to or within VPs (coded as "n/a") have the longest durations. Pauses within (inVP) and just before (preVP) have some of the shortest means among the data. The NP.PP category has eight different factor levels. The n/a category comprises 81 percent of the data, while the seven other levels code where in relation to an NP or PP the pause falls. Figure 6.11 gives some indication that pauses adjacent to and within NPs are shorter than pauses not near NPs. Pauses between NPs are shortest (though note the low *N*), followed by pauses within NPs and PPs. Pauses following NPs and PPs and preceding NPs are relatively longer. Pauses right before PPs have a shorter raw mean than most of the other categories. These two syntactic constituency factor groups are complex categories and could be analyzed and reanalyzed in several different ways (such as, for example, within NP.PP pauses versus other pauses). However, none of my explorations of these factor groups yielded significant results through statistical analysis, so NP.PP and VP factor groups and their derivatives are not included in the final statistical models and I do not discuss them further.

Finally, Figure 6.11 also presents the influences of adjacent filled pauses on (silent) pause duration. The FILLED PAUSE ADJACENCY (FPADJ) factor group indicates that silent pauses not collocated with a filled pause are shortest at about 5.93 log-ms, while silent pauses falling between two filled pauses (like "uh" or "um") or fillers (such as "like" or "I mean") are longest (e.g. "...uh [silent pause] uh...," "...uh [silent pause] um...," "...like [silent pause] um..."; although we note that only 36 pauses in the randomly sampled subset fall into this category). Silent pauses following a filled pause appear to be longer than pauses preceding a filled pause. We also see evidence that the type of filled pause or filler matters as well. The FILLED PAUSE (FP) factor group shows a relationship between "uh" and "um" in line with Clark and Fox Tree's (2002) finding, whereby "uh" indicates a shorter period of hesitation/processing than "um." We also see that the other fillers ("hmm," "I mean," "like," "well," and "you know") actually occur with shorter pauses than the none-FP adjacent pauses. It does appear from this that the filled pauses "uh" and "um" are different than the other, more discourse particle-like fillers.

Having explored these raw data, let us now turn to the results of regression modeling. Modeling here began with a comparable mixed-effect linear regression model to that used for the full pause dataset, described above and shown in Table 5.4. This social factor model is worse than the model was for the entire dataset, with significant effects

Table 6.7 Initial mixed-effect model for (trimmed) subset pause duration data

Factor	Estimate	Std err.	*p*
(Intercept)	5.810	0.065	–
REGION = Eastern NC	0.035	0.099	[0.7260]
REGION = Ohio	–0.217	0.119	0.0500
REGION = Southern NC	–0.094	0.109	[0.4060]
REGION = Texas	0.047	0.078	[0.5380]
REGION = Washington, DC	0.293	0.091	0.0020
REGION = Western NC	0.015	0.119	[0.7980]
SEX = male	0.087	0.054	[0.0700]
ETHNICITY = Latino/a	0.143	0.086	[0.0900]
ETHNICITY = Lumbee	0.147	0.140	[0.3400]
ETHNICITY = European Am.	0.029	0.074	[0.6700]

$R^2 = 0.091$.

for only two of the REGION comparisons (the shorter Ohio pauses and the longer Texas pauses) and none of the other social factors. This baseline model is provided in Table 6.7 for completeness, although we will now turn to build on this model using the additional linguistic and cognitive factors, without spending further time considering its details.

This model was extended by testing various possible combinations of the additional factors, building in a step-up fashion from the initial model. This investigation quickly added the PAUSE TYPE factor group, which a likelihood ratio test shows to greatly improve the model fit over the social factors alone. From there, both FILLED PAUSE ADJACENCY and the type of FILLED PAUSE further improve the model, but the type of FILLED PAUSE outperforms the FPADJ factor and likelihood ratio tests indicate that FILLED PAUSE type alone provides a better model than both FILLED PAUSE and FPADJ. Numerous interactions were tested for the main effects, but only one, the interaction between PAUSE TYPE and ETHNICITY, was found to improve the model. The best model includes this significant interaction. Despite showing the same tendency as earlier, with males having longer pauses than females, SEX never obtained significance as a main effect or interaction with the new factors for any of these models and was dropped from the best model. As with the main models of §5.3, outliers, defined as data points with standardized residuals greater than or equal to 2.5 standard deviations from zero (17 tokens), were removed and the model was refit to the remaining data. This final model is presented in Table 6.8 and Figure 6.12. The model obtains an R^2 of 0.15, still not a good fit, but better than the social-factor-only model of Table 6.7 and of the main model of Table 5.4.

Table 6.8 Best mixed-effect model for (trimmed) subset pause duration data

Factor	Estimate	Std err.	*p*
(Intercept)	5.976	0.068	–
PTYPE = ng (nongram.)	–0.319	0.045	0.0001
PTYPE = rl (lexical restart)	–0.481	0.075	0.0001
PTYPE = rs (syntactic reset)	–0.307	0.114	0.0070
FILLED PAUSE = Complex	0.414	0.134	0.0030
FILLED PAUSE = Other	–0.076	0.063	[0.2258]
FILLED PAUSE = Uh	0.266	0.064	0.0001
FILLED PAUSE = Um	0.355	0.056	0.0001
REGION = Eastern NC	0.023	0.104	[0.7776]
REGION = Ohio	–0.271	0.124	0.0192
REGION = Southern NC	–0.130	0.112	[0.2448]
REGION = Texas	0.106	0.081	[0.1526]
REGION = Washington, DC	0.233	0.097	0.0060
REGION = Western NC	–0.047	0.125	[0.7910]
ETHNICITY = Latino/a	0.010	0.097	[0.9156]
ETHNICITY = Lumbee	0.144	0.159	[0.3790]
ETHNICITY = European Am.	–0.002	0.085	[0.9532]
PTYPE = ng x ETHNICITY = Latino/a	0.133	0.076	[0.0914]
PTYPE = ng x ETHNICITY = Lumbee	–0.007	0.167	[0.9986]
PTYPE = ng x ETHNICITY = European Am.	0.072	0.085	[0.4254]
PTYPE = rl x ETHNICITY = Latino/a	0.266	0.113	0.0282
PTYPE = rl x ETHNICITY = Lumbee	0.574	0.259	0.0220
PTYPE = rl x ETHNICITY = European Am.	–0.001	0.141	[0.9928]
PTYPE = rs x ETHNICITY = Latino/a	0.498	0.177	0.0060
PTYPE = rs x ETHNICITY = Lumbee	–0.254	0.306	[0.4088]
PTYPE = rs x ETHNICITY = European Am.	0.322	0.225	[0.1584]

$R^2 = 0.151$.

While the inclusion of the nonsocial factors of PAUSE TYPE and FILLED PAUSE and their high significance leads to a better fitting model overall, they do not help the social factors fit to the data. The REGION main effect is roughly in line with the effects from the model of the full dataset, with only the comparisons between Ohio and Central NC and Texas and Central NC obtaining significance. ETHNICITY does not arise as a significant main effect. And, again, SEX was dropped from the model entirely. However, the interaction between ETHNICITY and PAUSE TYPE yields significance and is interesting in that it indicates that there may be, in fact, more striking social differences behind pause duration realization than was suggested by the social factor models of Tables 6.7 and 5.4. This is best seen in the bottom-left panel of Figure 6.12.

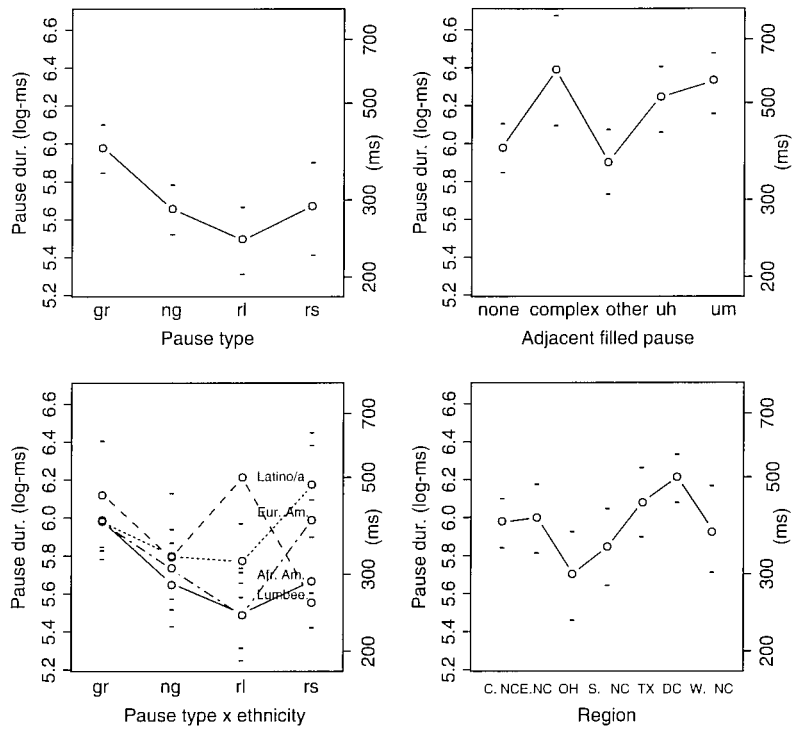

Figure 6.12 Effects in the mixed-effect model for the pause duration subset data

All four ethnic groups have roughly similar pause durations for grammatical and nongrammatical pauses. Yet, for the two restart categories, lexical repetition/restart and syntactic reset, we see some striking differences. African Americans and European Americans have similar pause durations for lexical restarts (rl) but Latinos and Lumbees do not. African Americans realize similar durations to these rl pauses for their syntactic reset (rs) pauses, although European Americans realize longer pauses for rs, roughly as long as their grammatical pauses (gr). Latinos have longer pauses across the board than the African American and European American speakers, although the shape of this pattern is fairly similar to the European Americans. Latinos also show the longest pauses for the rs category of all the groups. Lumbees exhibit a different pattern altogether, with long pauses occurring with lexical restarts and short pauses with syntactic resets. (The main effect for PAUSE TYPE, seen in the top-left panel of Figure 6.12, parallels the effects for the European

American and African American speakers, since the majority of the data come from those speakers; see the summary in Figure 6.11 for *N*s.) The difference for Lumbees, especially in terms of rl pauses, perhaps sheds some light on why their longer pauses in the earlier analysis of §5.3.2 (cf. Figure 5.5) were too variable to yield significance.

This closer analysis provides some evidence that there are, in fact, further social differences in pause realization beyond simply the relationship between social factors and pause duration. The finding that different ethnic groups' pause durations are influenced differently by different types of pauses – as investigated here in terms of the collocated linguistic material (lexical, syntactic, and prosodic) – is an exciting one and one which future research will hopefully shed more light on. As in §6.3, where we found some differences in the effects on longer pauses from shorter pauses, we see here that there is somewhat more systematicity underlying the massive noisiness of the pause data than seen in the last chapter. Nonetheless, despite our close looks, the pause durations in these data remain extremely variable and the bulk of that variability continues to escape us. It remains the case that only a relatively small amount of the variability in the pause data can be accounted for through statistical models of the available predictive factors, even when we expand the set of factors beyond the social factors alone.

6.6 Summing up

The sections in this chapter have refined the views of speech rate variation and pause duration variation developed in Chapter 5.

First, §6.2 explored the stability in the speech rate measurements by examining how speaker central tendencies and the utterance-level statistical models change when the number of measurements examined is drastically reduced. This examination provided a sense of both stability and instability in the results obtained. On the one hand, at token numbers below about 200, speaker medians begin to vary to large enough degrees to potentially impact the results of analyses based on those medians. On the other hand, utterance-level statistical models, especially through mixed-effect modeling, provide stable results even when we sample many fewer tokens per speaker.

Then, §6.3 asked a similar question about the pause duration measurements by examining the effect of different threshold criteria on the determination of the pauses to be included in the data. By slowly altering the minimum threshold value during iterations over the entire dataset, we saw that thresholds between 220 and 560 ms obtained the

best models based on the social factors. This analysis also allowed us to look more closely and separately at the short pauses (those < 560 ms) and the longer pauses and indicated that differences between short and long pauses should be studied more in the future. Further, this lent some empirical support for the common but rarely justified practice of using a low threshold of somewhere around 0.25 s in studies of pause.

§6.4 examined articulation rate data from just one regional group (Texas) to examine simultaneously whether an articulation rate analysis of IPs yields different results from the analysis of phonetic utterances conducted throughout the rest of this book and how phrase-final lengthening might affect articulation rates. This section shed some light on the nonlinear effect of utterance length (in syllables) by showing that the nonlinearity of this effect is mostly driven by short utterances with long final feet. Primarily, however, it demonstrated that the social patterns behind articulation rate are robust against these differences in methodology.

The last section, §6.5, examined the pause data in light of additional internal factors that were coded for approximately a tenth of the main dataset. Not surprisingly, we found there evidence that the addition of typological information about the pauses and information about adjacent filled pauses significantly improved the models. More interestingly for our purposes, the inclusion of pause type in the analysis yielded a significant interaction between speakers' ethnicity and pause type, which provided some additional evidence that social factors pattern pause variability in ways more substantial and complex than the simple (and generally poorly fitting) pause duration models have so far indicated. Nonetheless, that section ended by acknowledging that even with these additional factors pause variability is poorly predicted by statistical models of the available factors.

Overall, the findings of these closer looks have reinforced one of the most important findings from Chapter 5: speech rates are highly patterned and highly modelable while pause durations are not. In the next chapter we continue our closer looks at speech rate and pause variability by examining accommodation and interlocutor-related topics.