CHAPTER 6

MESSAGE-RELATED VARIATION

SEGMENTAL WITHIN-SPEAKER VARIATION

MIRJAM ERNESTUS

TONAL VARIATION

YIYA CHEN

The authors in this chapter provide parallel discussions of segmental and tonal variations. In each case attention is given to the nature of categorical vs. gradient effects and the question of the degree to which effects are automatic consequences of the production/perception system vs. under speaker control. Ernestus discusses the case of segmental variation, focusing on the rich and complex literature on assimilation and reduction. Chen discusses tonal variation, focusing on coarticulatory effects in languages with lexical tones and global effects of the prosodic encoding of information.

6.1 SEGMENTAL WITHIN-SPEAKER VARIATION

Mirjam Ernestus

6.1.1 Introduction

It has long been known that a word's pronunciation may be different in connected speech than when carefully produced in isolation. The differences concern the segmental as well as the suprasegmental properties of words. This section focuses on segmental adaptations, while Chen (this chapter) discusses tonal variation.

Segmental differences largely result from the adapation of word-initial and word-final segments to adjacent segments (assimilation) and from reduction (segment lenition and deletion). Within the traditional phonological framework, these types of pronunciation variation were mostly investigated on the basis of data obtained from introspection and impressionistic observation. New technical developments of the last few decades, such as the enormous increase in computer memory and the ability to analyze speech files automatically, have made it easier to study pronunciation variation on the basis of large quantities of real speech (see Cole and Hasegawa-Johnson, this volume). Moreover, the ability to store and search large speech corpora helps researchers find stimuli for psycholinguistic experiments, which facilitates the study of the comprehension process as well.

A strong indication of the importance of these new studies is the finding that simple phonological processes described in the literature, such as assimilation, are not as pervasive as had been thought, while other processes are much more frequent. For instance, Dilley and Pitt (2007) studied place assimilation of alveolar segments to following bilabial or velar segments (e.g. the pronunciation of *green boat* as *gree*[m b]*oat*) in a corpus of American English, a process that has been described as highly productive (e.g. Harris 1994). Contra standard analyses, they found that place assimilation is relatively rare and that deletion, glottalization, or canonical pronunciations of the alveolar consonant are more frequent. Similarly, Ernestus et al. (2006) reported that in a corpus of Dutch, obstruents followed by voiced plosives show regressive voice assimilation (e.g. *we*/t + b/*oek* is pronounced as *we*[db]*oek* 'law book'), as described in many theoretical studies, but they also found that these clusters can show deletion of the first obstruent (*we*[b]*oek*), and most importantly, progressive voice assimilation (*we*[tp]*oek*), a process that has been claimed not to apply to such clusters in Dutch (e.g. Booij 1995).

These new types of empirical studies also show that reduction processes, which had received only limited attention in the phonological literature, are widely attested in informal speech styles. Traditional phonological descriptions mention reduction rules, such as full vowel to schwa reduction and /t/-deletion (see e.g. Booij 1995 for Dutch), but underestimate the frequency and variety of reduction.

Deletion of unstressed vowels in American English, for instance, affects 25 percent of the possible word tokens even in news interviews on television (Dalby 1984), and in casual conversations 25 percent of all word tokens show lenition or deletion of at least one segment (Johnson 2004). Moreover, speakers delete complete syllables, a phenomenon hardly described at all in the phonological literature, resulting in pronunciations like [pʰɛrĩ] for English *apparently* and [wɛs] for Dutch *wedstrijd* /wɛtstrɛit/ 'contest' (see Johnson 2004 for English; Kohler 1990 for German; and Ernestus 2000 for Dutch examples). These reduction processes lead to a vast number of pronunciation variants for one and the same word, as exemplified by the Dutch word *natuurlijk* [natyrlək] 'of course,' which may be pronounced as [natyləkˌ], [ntyləkˌ], [ntykˌ], [tyrləkˌ] (which also has an orthographic representation), [tyləkˌ], [tylkˌ], [tykˌ], [tygˌ], [dykˌ], and [dygˌ], among others (Ernestus 2000).

All these studies provide evidence that pronunciation variation at the segmental level is much more pervasive and above all more complex than previously thought. The detailed characteristics of pronunciation variants and the factors conditioning these variants have important implications for linguistic and psycholinguistic theory. Below, I first discuss the articulatory and acoustic properties of pronunciation variation, focusing on the theoretically important difference between categorical and gradient processes (Section 6.1.2); and how recent findings can be accounted for within generative grammar (Section 6.1.2), Articulatory Phonology, and exemplar-based models (also Section 6.1.3). Then, I discuss different accounts of how listeners process segmental pronunciation variation, including underspecification theory, a perceptual account, and a learning model (Section 6.1.4). Finally, I give an overview of the most important variables conditioning segmental pronunciation variation, which informs phonological theory and psycholinguistic theories of speech production and comprehension (Section 6.1.5).

6.1.2 Categorical versus gradient variation

Traditionally, categorical variation is distinguished from gradient variation. Variation is categorical if it can be well described with the categorical values of phonological features (e.g. [+voice] and [−voice]). Variation is gradient if the acoustic characteristics of the variants reflect values between these categorical values (e.g. partly voiced). The distinction between categorical and gradient variation is theoretically important, since within the generative framework it distinguishes between phonological and phonetic processes: Phonological processes are considered to be categorical while phonetic processes are typically gradient (e.g. Keating 1990b; Cohn 1993). These definitions of phonological and phonetic processes replace earlier definitions that see phonological processes as language-specific and phonetic processes as language-universal and automatically resulting from articulatory

mechanisms (e.g. Kenstowicz and Kisseberth 1979). The distinction between categorical phonological processes and gradient phonetic processes plays a role in both linguistic and psycholinguistic (e.g. Levelt 1989) models.

Within the generative framework (starting with Chomsky and Halle 1968), processes such as assimilation, vowel reduction, and segment deletion are generally assumed to be categorical. Assimilation involves the "spreading" of a phonological feature from one segment to another segment, and this latter segment subsequently cannot be distinguished from segments that have the same feature value in their underlying specifications. For example, [m] has exactly the same surface phonological representation and phonetic characteristics in *a ru*[m p]*icks you up* when the speaker intends *run* or *rum*. Similarly, vowel reduction implies replacement (or deletion) of the phonological features specifying the vowel's quality, and this vowel consequently cannot be distinguished from underlying schwas. Finally, segment deletion implies the complete loss of a segment in the surface phonological representation. Thus, if words differ only in the presence versus absence of a segment in their underlying representations (e.g. English *sport* and *support*, or *miss* and *mist*), deletion of this segment leads to identical pronunciations (see Coetzee, this volume for further discussion of deletion).

Detailed articulatory and acoustic studies have cast serious doubt on the classification of many connected speech processes as categorical. The evidence for gradient rather than categorical variation is especially strong for place assimilation, since the exact location of the closures for plosives can relatively easily be measured by means of electropalatography (EPG, Hardcastle 1972). Several EPG studies have shown that palatalization (as in *hi*/t j/*ou*) may be gradient, with the obstruent becoming more palatal over time, which distinguishes a palatalized obstruent from underlying palatals (e.g. Barry 1992 for Russian; Zsiga 1995 for post-lexical palatalization in American English). The same type of gradience has been documented for place assimilation of /t/ and /d/ in American English (e.g. in *la*/t k/*alls*): These obstruents often start out alveolar and only then gradiently assimilate to the place of articulation of the following consonant (Nolan 1992). Similar results, but showing considerable inter- and intraspeaker variation, have been found for place assimilation of alveolar nasals in American English (e.g. in *gree*/n b/*oat*, Ellis and Hardcastle 2002). All these data suggest that place assimilation, especially assimilation crossing word boundaries, cannot be simply accounted for by the spreading of a phonological place feature from one segment to another.

Several acoustic studies suggest that voice assimilation may be gradient as well. For instance, Ernestus et al. (2006) have shown that Dutch obstruent clusters expected to be subject to regressive voice assimilation (which voices the initial segment) may be produced without any glottal vibration, with glottal vibration throughout the whole cluster, or during only part of the cluster. Even though many acoustic characteristics co-determine the perceptual voiced-voiceless distinction, this result is telling since glottal vibration is considered the most important cue to voicing in Dutch obstruent clusters (van den Berg 1986). In addition, voice

assimilation appears gradient since, instead of affecting all acoustic characteristics cueing the voiced-voiceless distinction, it may affect only some of them. This results in segments that are neither completely voiceless nor completely voiced. For instance, Kuzla et al. (2007) showed that progressive devoicing assimilation in German (e.g. in *ha*/t v/*älder* 'had woods') results in shorter intervals of glottal vibration, while it hardly affects the duration of the fricative, which is the most important cue to the voice (or fortis/lenis) specification of fricatives in German. Similar results were obtained by Ernestus and colleagues (2006), who showed that Dutch obstruent clusters expected to be subject to regressive voice assimilation tend to be shorter (cueing more perceptual voicing), to be produced with longer periods of glottal vibration (also cueing more perceptual voicing), but also with longer release noises (cueing less perceptual voicing) in words of high compared to low lexical frequencies (probably because speakers produce high-frequency words with less articulatory effort). Consequently, in more frequent words, some acoustic characteristics signal more and others less voicing. These data provide additional evidence that assimilation is more complex than the spreading of a phonological feature.

In addition to assimilation, many reduction processes appear gradient. For instance, vowels may show any realization between unreduced pronunciation variants (with formants distinguishing the vowels maximally from the other vowels in the language) and schwas (e.g. Mooshammer and Geng 2008). Simultaneously, vowels may vary in their duration from values typical for accented full vowels to zero, showing all durations in between (e.g. gradient deletion of the first vowel of a sequence of two in Plains Cree, Russell 2008). They may thus have clear, some, or no cues in the acoustic signal. Also, obstruents may show several types of reduced realizations in addition to being fully present or absent (e.g. Mitterer and Ernestus 2006 for /t/ in Dutch). In many cases, consonant deletion thus appears to be the natural endpoint of gradient reduction processes, rather than to result from categorical phonological processes.

Taken together, these studies suggest that most connected speech processes are gradient and thus, according to the definition of the phonological component as containing only categorical processes, they belong to the phonetic component. In other words, the new findings move most of post-lexical phonology from the phonological to the phonetic component in the generative framework. These findings therefore raise the questions of whether the division within generative grammar between phonology and phonetics should be revisited again and how mechanisms responsible for gradient variation should be formalized.

6.1.3 Processing models naturally incorporating gradient variation

The gap between the phonological (i.e. categorical) and physical (i.e. gradient) structure of speech in generative models has stimulated the development of new

models. These alternative models make fundamentally different assumptions and naturally incorporate gradience in pronunciation variation.

One of these models is Articulatory Phonology (see Gafos and Goldstein, this volume), developed by Browman and Goldstein (e.g. 1986, 1992). It assumes that phonological representations consist of abstract articulatory gestures, rather than segments or features. Articulatory Phonology can account for many phenomena that are well explained by non-linear phonology, since the temporal alignment of gestures may be changed, which may result in the overlap of different gestures in time. For instance, regressive place assimilation in *gree*[m b]*oat* may result from the early onset of the bilabial closure, before the realization of the preceding nasal, which then hides the alveolar closure. Importantly, this retiming of gestures may likewise account for gradient assimilation. In addition, it may explain the complete absence of segments in the acoustic signal. For instance, a word-final /t/ may appear absent before bilabial stops (as in *perfec*/t m/*emory*), because speakers close their lips before the /t/ is released, which makes the release of the /t/ inaudible (Browman and Goldstein 1990a). The hypothesis that speakers may produce the articulatory gestures for inaudible segments, as assumed in these accounts, is supported by several X-ray studies (e.g. Browman and Goldstein 1992). Finally, gestures may be reduced in size, which results in the lenited realizations often encountered in casual speech, also a form of gradience. This notion of size reduction, however, has only recently begun to be robustly modelled in Articulatory Phonology (see Gafos and Goldstein, this volume).

The assumption that gestures may overlap in time and be reduced in size, even to zero, makes Articulatory Phonology a very powerful theory. It can account for the absence of any acoustic cue under any condition. Obviously, research is necessary to properly constrain the theory such that it accounts only for those pronunciation variants that really occur. Furthermore, if lexical representations consist of abstract gestures, listeners should extract these gestures from the acoustic signal. Some data suggest that this is indeed what listeners do (e.g. Fowler et al. 2003), but other experiments cast doubt on these results. For instance, Mitterer and Ernestus (2008) found that the speed with which participants shadow words containing the phoneme /r/ is independent of whether participants produce the /r/ with different or with the same articulatory gestures as those used in the words. Furthermore, participants imitate phonetic detail more closely if it is phonologically relevant. These results are unexpected if the basic units of lexical representations are gestures rather than more abstract phonological symbols. More research is necessary also to settle this issue.

Another type of model naturally incorporating the gradience of pronunciation variation is the exemplar-based model (see Chapter 8 this volume). Exemplar models assume that the mental lexicon contains a representation for every pronunciation variant of a word (possibly even one for every token ever heard or

uttered by the language user), with detailed information about all phonetic properties of the variant. Johnson (2004), for instance, following Klatt (1979), proposed that lexical representations can be considered as sequences of spectra with no categorical information at all. The assumption of different lexical representations for pronunciation variants is supported by both production and comprehension data showing that speakers' and listeners' response latencies are affected by the frequency of the given pronunciation variant compared to the frequencies of the other variants for the same word (Ranbom and Connine 2007; Bűrki et al. 2010). These results demonstrate that language users store frequency information about pronunciation variants, which suggests that they store the variants themselves as well. Further specificity within exemplar-based models is necessary to clarify to what extent an actual realization needs to be in line with a corresponding stored exemplar and to what extent its phonetic detail may result from the phonetic implementation of an exemplar. Moreover, future studies have to implement exemplar-based models computationally to test which additional assumptions may be necessary to account for the full range of available data (Ernestus forthcoming).

6.1.4 Comprehension of pronunciation variation

The comprehension of pronunciation variation may be accounted for within the processing models mentioned above (see Nguyen, this volume and Holt, this volume for discussion of the perception of canonical pronunciation variants). Psycholinguistic models based on generative models may assume that the acoustic input is reconstructed to the canonical pronunciation stored in the mental lexicon by means of rules or phonological constraints (e.g. Boersma 1998). This reconstruction may be based, for instance, on the grouping of feature cues distributed over time (Gow 2003). Articulatory Phonology assumes that listeners retrieve the underlying gestures from the gradient acoustic input, while exemplar-based models assume that the mental lexicon contains representations for many pronunciation variants and that an acoustic input is recognized if it is sufficiently similar to one of the stored exemplars. In addition to these models, several other mechanisms have been proposed to account for the comprehension of pronunciation variation.

Underspecification theory (Lahiri and Reetz 2002) assumes, like most models in the tradition of generative grammar, that the mental lexicon contains only one lexical representation for every word (see Lahiri, this volume). In order to explain the recognition of words with assimilated segments (such as *green* in *gree*[m b]*oat*), the theory assumes that phonological features subject to assimilation (e.g. the place feature of alveolar nasals in English) are lexically unspecified and do not contribute

to word recognition. Thus, assimilation does not hinder word recognition, as it does not lead to mismatches with stored phonological representations. Underspecification theory is supported by language acquisition data, which show that young children confuse some words and not others, which is taken as evidence that some phonological segments are lexically underspecified (Fikkert 2005). In contrast, underspecification theory is challenged by perception studies showing that listeners only recognize a pronunciation variant that may result from assimilation if it occurs in the appropriate segmental context (Gaskell and Marslen-Wilson 1996, 1998; Gow 2002). Thus, listeners recognize *gree*[m] as *green* only before bilabial plosives, as in *gree*[m b]*oat*.

Another account for the comprehension of assimilated segments assumes that the human auditory system is not highly sensitive to the differences between assimilated and non-assimilated segments in assimilation contexts (e.g. between *gree*[n] and *gree*[m] if followed by *boat*). As a consequence, assimilation does not pose problems for comprehension. This account explains the role of segmental context in comprehension and receives experimental support from several studies using simple discrimination tasks (e.g. Mitterer, Csépe, and Blomert 2006) and event-related potentials in the brain (e.g. Mitterer and Blomert 2003). Especially convincing is the finding that listeners with native languages that differ in whether they contain the assimilation process under investigation are equally bad in discriminating between assimilated and unassimilated segments in appropriate assimilation contexts (Gow and Im 2004; Mitterer, Csépe, Honbolygo, and Blomert 2006). This account of the comprehension of assimilated segments can be extended to the comprehension of acoustically weak segments. Mitterer et al. (2008) showed that insensitivity of the auditory system can partly explain listeners' ease in the processing of /st/-final words produced with acoustically weak /t/s.

Yet another mechanism that may contribute to the comprehension of pronunciation variation is listeners' learning of language-specific patterns, as advocated by Gaskell and Marslen-Wilson (1998; Gaskell 2003). These researchers showed that listeners are better at inferring the citation form of an assimilated pronunciation variant (i.e. *gree*[n] from *gree*[m b]*oat*) if the word is an existing word in the language (such as *green*) rather than a pseudo word (such as *breen*). Familiarity with language-specific patterns may also explain why Dutch listeners are slightly better than Japanese listeners in discriminating between some pronunciation variants of word-final /st/-clusters: Such clusters are frequent in Dutch, whereas they are phonotactically illegal in Japanese (Mitterer et al. 2008).

In conclusion, the literature contains several proposals, most of them supported by experimental data, to account for the comprehension of pronunciation variation. Probably, the comprehension process results from the interaction of several mechanisms (as also concluded in Mitterer et al. 2008) and further research

should show how these mechanisms interact. Interestingly, some of the relevant mechanisms, including the sensitivity of the human auditory system, are not part of the grammar and are therefore traditionally assumed not to be interesting for linguistic theory. However, in order to define the contributions of linguistic mechanisms to speech behavior, we have to know the contributions of the extragrammatical mechanisms, which together with the grammar will provide us with a complete picture of human speech processing.

6.1.5 Factors conditioning pronunciation variation

All complete models of human language processing should account for the conditions under which pronunciation variation is likely to occur and is best understood by listeners. Data on these conditions are therefore crucial for linguistic and psycholinguistic theories. Whereas data on the comprehension of segmental variation are still relatively scarce, much more is known about conditions favoring the production of non-canonical forms.

Among these conditions, speech style is probably the most important: Non-canonical pronunciation variants tend to be more frequent in less formal speech. For instance, place assimilation of alveolar plosives to velar plosives in English is more common in less formal speech styles (e.g. Kerswill 1985; Barry 1992), and highly reduced pronunciation variants, such as [p^hɛrĩ] for *apparently*, are attested only in truly casual speech. Less formal speech is mostly uttered at relatively high speech rates, which may put speakers under time pressure. Speakers may attenuate this time pressure by deleting segments or by reducing the sizes of articulatory gestures and overlapping them more in time. Speech rate may therefore explain some of the phonetic characteristics of non-canonical forms attested in less formal speech. Importantly, however, a high speech rate does not necessarily lead to non-canonical forms, as documented, among others, by van Son and Pols (1990, 1992). It is therefore speech style rather than speech rate that conditions pronunciation variation, but speech rate may co-determine the type of non-canonical forms occurring in less formal speech.

Another important factor conditioning pronunciation variation is the prosodic structure of the utterance (see Frota, this volume and Turk, this volume). Consonants in the initial position of prosodic domains, such as the intonational phrase or the intermediate prosodic phrase, tend to be longer and to be produced with greater linguopalatal contact (initial strengthening, e.g. Fougeron and Keating 1997; Keating et al.) than consonants in domain-medial or domain-final positions. Domain-initial segments also appear less sensitive to connected speech processes, since initial consonants show less voice assimilation (Kuzla et al. 2007) and vowels in initial syllables show less coarticulation with neighboring vowels (Cho 2004). Among

the non-initial segments, domain-final ones tend to be longer than medial ones, a phenomenon called final lengthening (e.g. Wightman et al. 1992). Several studies suggest that listeners may use these prosodic strengthening cues in comprehension (e.g. Cho et al. 2007).

Less well documented, but probably an equally relevant factor is the word's function in the discourse. Local (2003) reported that the word combination *I think* is more reduced when it occurs in sentence-final position (even though final lengthening would be expected in these positions) and conveys a pragmatic meaning (e.g. in the sentence *they should be here by the time you come out next weekend I think*) than when it is followed by the complementizer *that* and has above all a lexical meaning (e.g. *I think that people have not yet woken up*). Similarly, Plug (2005) reported that the degree of reduction of the Dutch word *eigenlijk* 'actually' depends on whether the word signals contrast with what has been suggested before by the speaker or by the listener. So far, only a few studies have investigated the role of pragmatic function; future studies are needed to determine the exact mechanisms driving the effects and whether their relevance may be restricted to only a few word types.

A final important predictor of a word's pronunciation is its predictability within the context. Words tend to be more reduced when they are more predictable given the preceding or following words. For instance, Scheibman and Bybee (1999) reported that the English word sequence *don't* tends to be produced with a smaller number of segments if preceded by *I*, the word that most frequently precedes *don't*. Likewise, the sequence is more reduced before the words that most often follow *don't* (*know*, *think*, *mean*). In general, function words, like *don't*, are more reduced (in duration and in number of segments) the more predictable they are given the preceding words (e.g. Pluymaekers et al. 2005; Bell et al. 2009). Content words, in contrast, tend to be more reduced if they are more predictable given the following words (e.g. Bell et al. 2009). Thus, the English word *previous* has a higher probability of being reduced when followed by *year* than by *beer*. In addition, words tend to be shorter if they have been mentioned in the conversation before (e.g. Fowler and Housum 1987; Aylett and Turk 2004).

These predictability effects may automatically result from the production process: More predictable words are easier to plan and therefore do not require speakers to slow down their speech rate, which may result in reduction (see e.g. Pluymaekers et al. 2005; Bell et al. 2009). A planning account of the predictability effects has the advantage that it easily explains why reduction degree for content words is correlated especially with the predictability of the following word: Speech rate is determined by the planning of the next words rather than that of the preceding or current words. In addition, this account predicts correctly that words are less reduced if they are followed by hesitations, which indicate planning problems (e.g. Jurafsky et al. 2001).

In contrast to this speaker-driven account are two listener-driven accounts. The first one assumes that speakers would like to reduce their articulation effort as much as possible but adapt their reduction degree to the listeners' needs in order to guarantee smooth communication (in line with the Hyper- and Hypo-articulation theory by Lindblom 1990). More predictable units are easier to understand and speakers would therefore reduce especially highly predictable units. The second listener-driven account states that reduction degree facilitates comprehension as it indicates which information is given or predictable (e.g. Boersma 1998). The hypothesis that speakers adapt their degree of reduction to the listeners' needs is supported by the finding that in Dutch, English, German, and Italian, segmental sequences of medium durations are attested more frequently in corpora of spontaneous speech than sequences of relatively short duration. Probably the short combinations are more difficult to identify and are therefore used less often by the speaker (Kuperman et al. 2008). In contrast, listener-driven accounts are challenged by the observations by Bard and colleagues (Bard et al. 2000) that the second mention of a word tends to be more reduced independently of whether the listener has heard that word before in the conversation. Probably the documented predictability effects are both speaker- and listener-driven and future research has to investigate the relative relevance of the different mechanisms.

In conclusion, factors of very different natures appear to condition pronunciation variation. Speech-processing models can only account for the full range of data if they take the many different aspects of speech (including grammatical form, semantics, pragmatics, and planning) into account.

6.1.6 Conclusions

The last few decades have produced many linguistic studies based on corpus data and psycholinguistic experimentation. These studies have above all shown that speech is much more variable and gradient than has traditionally been assumed: natural speech shows more pronunciation variants than previously assumed, some well-known variants occur less often than expected in favor of others, and many variants show gradient properties. Moreover, speech processing involves mechanisms of very different natures (involving, among others, pragmatic function and speech planning) that appear to interact. So far, no existing model of speech processing can account for all findings. Further studies on the processing of pronunciation variation are necessary to formulate and evaluate comprehensive models of both speech production and comprehension. Since the mechanisms involved appear to be of very different natures, these studies will benefit from the multidisciplinary effort of the laboratory phonology approach.