16 New Directions in Speech Production

Jonathan Harrington, Phil Hoole and Marianne Pouplier

Chapter Overview

Introduction	242
Coarticulation	244
Assimilation	249
Consonant Clusters	254
Conclusions	258

1 Introduction

Studying speech production from a phonetic perspective can be considered on the one hand to form part of the task in cognitive science and psycholinguistics of explaining how a speaker's intention to produce an utterance is related to its physical instantiation, that is, to the movements of the vocal organs that give rise to an acoustic signal. But it also shares with phonology and linguistics the goal of explaining why the sounds of languages are shaped the way that they are. These two obviously related aims can nevertheless be differentiated by the kinds of questions that are asked in studying speech production. For example, a question that is more directly relevant to the first task of explaining how the sounds of speech are transmitted between a speaker and a hearer in conversation might be: what kinds of control structures are invoked by the speaker so that a listener perceives sounds to have both a certain serial order and a grouping into structures such as syllables and words? Questions that are more applicable to the second goal of finding the physiological bases to the way that sounds are distributed and pattern in the world's languages might be: how do aerodynamic and voicing constraints contribute to the relative paucity in languages of voiced velar stops? And how do articulatory and perceptual characteristics influence the relatively greater likelihood of a vowel and consonant being blended in a sound change when the consonant is syllable-final than syllable-initial (e.g. Ohala, 1990)?

Both of these approaches to speech production are evident in studies of *coarticulation* in which the task is to develop a model of the way in which sounds overlap, or are blended, with each other in time. Research on *assimilation* deals with broadly similar issues to those of coarticulation, but they have given somewhat greater emphasis methodologically to the types of variability that occur to the production of speech across the major types of juncture (in particular word boundaries) and theoretically to whether segments that are perceived to be deleted really are deleted in speech production. These issues are also central to analyses of *consonant clusters*: but here there has been somewhat greater emphasis on the way that prosodic (and especially syllabic) structure influences the relative overlap between consonants and vowels.

One of the reasons why coarticulation, assimilation and consonant clusters remain important for understanding speech production is because all such analyses require both an explicit modelling of the dynamics of speech as well as an understanding of some of the different ways that speech dynamics can be phonologized. Currently, we lack sufficient information about how dynamics are incorporated into a language's phonology, partly because there are too few empirical studies of the dynamics of speech in the different languages of the world, but also because speech production has been shaped by the idea inherent in much of generative phonology (Clements, 1985) and psycholinguistics (Levelt, 1989) that phonological information precedes and is mapped onto an independent component that deals with the biophysical aspects of speech production (see also Pierrehumbert, 1990, for a critique of this position). Considerable progress against this view has been made through the development of articulatory phonology (AP) in the last 20 years in which gestures function both as phonological primitives and as dynamic action units of the vocal tract: in this model, time is inherent in phonological representations and there is no division between the specification of a phonological plan and its execution in speech production (see also Fowler, 1981, 1984). One of the major challenges in the future will be to extend the AP model so that it can incorporate the subtle ways in which the temporal control and coordination of speech vary between languages and its varieties. Another challenge which is important in the context of modelling the relationships between synchronic variation and diachronic change will be to evaluate more closely than before how dynamic information is transmitted between a speaker and a hearer. As will be discussed below, these are some of the reasons why recent studies of speech dynamics are beginning to give greater emphasis to crosslinguistic physiological comparisons of typologically rare combinations of consonants and vowels (e.g. Pouplier & Beňuš, 2011), and why all three sections presented below are at various points concerned with the perceptual consequences of the production of coarticulation, assimilation and of consonant clusters.

2 Coarticulation

Modelling coarticulation can be considered to be part of the wider problem of how phonology and phonetics are connected. The phonetics-phonology dichotomy comes about because on the one hand it seems undeniable that words can be combined from a smaller set of abstract units – the features, phonemes and syllables of a language – which can be permuted in a rule-governed way to create new words. A central aspect of this combinatorial possibility and indeed of the flexibility to add new words to the lexicon is that the units are *context-independent*: thus, the same units – the phonemes – are presumed to form part of the constitution of the words *tip* and *pit*. However, almost any analysis of the speech signal shows that speech communication is highly *context-dependent*: that is, the ways in which the raising of the tongue dorsum for the vowel and the closure and release of /t/ are timed relatively to each other are very different even when these monosyllables are produced in isolation by the same speaker (Krakow, 1999). Modelling coarticulation is central to understanding how the context-independent units suggested by phonology and the dynamic, context-dependent characteristics of speech communication are related to each other.

For various reasons, it is not possible to explain the production of coarticulation without also considering its perception. Moreover, the relationship between them also forms a key part of relating synchronic variation in speech to diachronic sound change. Some of these issues are discussed in further detail below.

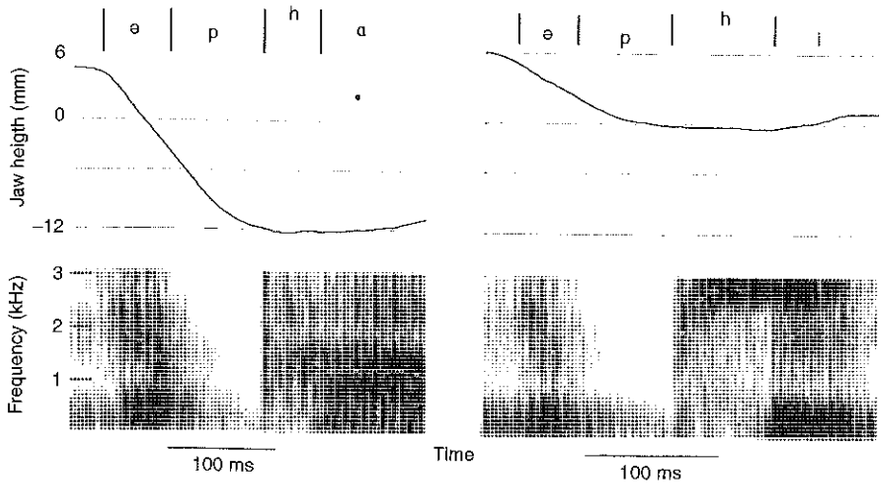

Figure 16.1 Jaw height trajectories and synchronized spectrogram showing the first two formant frequencies with superimposed acoustic phonetic segment boundaries of /əpɑ/ (left) and /əpi/ (right) produced by a female speaker of Standard German

2.1 The Production of Coarticulation

Figure 16.1 shows an instance of so-called transconsonantal vowel coarticulation, first studied systematically by Öhman (1966) and Perkell (1969), in which the vowels influence each other in a VCV sequence across an intervening consonant. The evidence for VCV coarticulation in Figure 16.1 is that the differences in jaw height in the final vowels of /əpɑ/ and /əpi/ are already anticipated across the medial /p/ and during the initial schwa: thus the jaw height is lower throughout the extent (and certainly by the offset) of the schwa of /əpɑ/ than that of /əpi/. Öhman's (1966) study showed that VCV coarticulation is possible (also across non-labial consonants) because vowel and consonant movements are generally controlled by different sets of muscles.

In so-called translation accounts of coarticulation of the 1960s and 1970s (e.g. Henke, 1966; Daniloff & Hammarberg, 1973), the nasalization of /a/ in a nasal context such as *man* was effected by a context-sensitive rule that changed /a/ into a nasal variant [ã]. By contrast, Fowler (1981) developed a context-invariant explanation of coarticulation as coproduction in which consonants and vowels are controlled by autonomous articulatory coordinative structures that overlap with each other in time. Vowel nasalization in this example comes about according to Fowler (1981) not because the vowel is modified by context, but instead because the nasal consonant is overlaid at some phase during the production of the vowel and without these articulatory strategies actually influencing or changing each other.

Many of these ideas have been incorporated into Browman and Goldstein's (1991, 1992) model of Articulatory Phonology in which coarticulation is explained as the temporal layering of gestures. Articulatory strength, which defines how resistant a segment is to coarticulatory influences (Fowler & Saltzman, 1993), is a key part of this model: in general, the more resistant a segment, the greater its influence on neighbouring segments (i.e. coarticulatory resistance and dominance are roughly inversely proportional). Articulatory resistance is quantified by Recasens (1999) in terms of a consonant's degree of articulatory constraint (DAC). Fowler and Brancazio (2000) provide support for one of the findings in Recasens (1984) that although consonants with a high DAC may affect the magnitude of coarticulation, they have little effect on its temporal extent. They also show that a consonant restricts vowel-to-vowel coarticulation only in the vicinity of the medial consonant (see also Fowler, 2005).

Data such as these form the basis for Fowler's argument that the coarticulatory influence of the consonant V_1CV_2 sequences is unlikely to be directly planned, since otherwise a speaker would have to plan first for a large degree of V_2 on V_1 coarticulatory influence (at the onset of V_1) but then plan to reduce it close to the consonant. It seems instead more likely that the extent of observable coarticulation is a function of the waxing and waning of the consonant (Fowler,

1984): that is, in the middle of the VCV, the vowels' gestures are largely covered up or hidden by the consonant during the phase at which it is maximally prominent; but they are then more in evidence at the consonant's margins during which its articulation is less prominent. It is this type of articulatory waxing and waning that is also used to explain many types of acoustic shortening. For example, polysyllabic shortening, in which /beɪ/ is shorter in *baby* than in *bay* comes about, not because the speaker plans a shorter first syllable, but instead because of the coproduction of the two syllables as a result of which the first syllable is progressively covered up by the second as it wanes towards its margin (Fowler & Thompson, 2010).

Coarticulation resistance also forms a central part of Keating's (1990) window model of coarticulation. A window in this model defines the range of allowable articulatory variation of different features that make up a segment: in the production of /u/, there is, then, a window specifying the extent to which the tongue body can vary along a front-back dimension, and a separate window that defines how much allowable variation there can be in lip-rounding and so on for all of the features that make up /u/. Essentially, the narrower the window for any feature, the greater the coarticulatory resistance. The width of windows is presumed to be influenced by phonology: thus the window width for the feature nasal is much narrower in French than for English (as a result of which the coarticulatory influence of flanking nasal consonants on oral vowels is small) because French, but not English, makes a phonological contrast between oral and nasal vowels.

2.2 The Perception of Coarticulation

A study by Alfonso and Baer (1982) showed that listeners could identify V_2 when presented with the initial /əC/ from /əCV_2/ sequences. Compatibly, listeners' reactions to identifying V_2 in V_1CV_2 were found to be slowed (Martin & Bunnell, 1982) when V_1 provided conflicting as opposed to valid coarticulatory information about V_2 (see also Fowler & Smith, 1986). These experiments provide evidence both that the source of coarticulation is perceptible, and that it contributes to phonetic identification (since identification is impaired when it is removed, as in these cross-spliced stimuli). As discussed in Fowler (2005), there are two different ways in which listeners could make use of this coarticulatory information about an upcoming segment. One of them is context-sensitive in which the different variants of schwa due to coarticulation are perceived to be different. The other is context-invariant in which listeners perceive directly the layered gestures that are produced by the speaker. Based on a series of discrimination tests, Fowler (2005) provides evidence for the second context-independent mode of perception: thus, listeners hear the different phonetic variants as the

same because they subtract, or factor out, the variation in attributing it to its source, the final vowel.

Experiments on the compensation for coarticulation provide further evidence for this context-independent mode of perception. For example, Mann and Repp (1980) showed that when a continuum between /s/ and /ʃ/ is synthesized by lowering the spectral centre of gravity of the fricative and prepending it to a vowel, then listeners' responses are biased towards /s/ when the vowel is rounded (in /su-ʃu/) compared with when it is not (/sa-ʃa/). This is because the spectral centre of gravity is not only a positive cue for /ʃ/ but is also a consequence of the anticipatory coarticulatory influence of a following rounded vowel. Thus listeners attribute some of the spectral centre of gravity lowering in the fricative to /u/ and factor it out: as a result, they hear more /s/ tokens from the same /s-ʃ/ continuum when it is prepended to rounded /u/ than to unrounded /a/. Similar effects of perceptual compensation for nasal coarticulation are demonstrated in Beddor et al. (1986).

2.3 Language-specific Coarticulatory Effects

The earlier discussion of Keating's window model suggests that coarticulation is influenced by a language's phonology. Öhman (1966) also showed that the extent of transconsonantal vowel coarticulation is less in Russian than in English or Swedish: his explanation is that a large displacement of the tongue dorsum in the production of the consonant due to vowel coarticulation might compromise the phonological opposition which exists in Russian (but not English or Swedish) between palatalized and non-palatalized consonants. There is also some evidence that V-on-V coarticulatory influences are related to the size of the vowel inventory: since phonetic variation in the production of a vowel needs to be contained if a language has a phonologically crowded vowel system (in order for vowels not to encroach upon each others' space), then languages with a large number of phonemic vowel oppositions are predicted to be affected less by coarticulation than those with fewer contrasts (Manuel, 1999 for a review); however other studies (Beddor et al., 2002; Mok, 2010) have found no evidence in support of this relationship between inventory size and the magnitude of coarticulation (see also Jones, this volume).

One of the difficulties in comparing the size of coarticulation crosslinguistically is that observed differences in coarticulation between two languages may come about because of phonetic differences in the segments that gives rise to coarticulation, rather than to differences in the magnitude and timing of coarticulation per se. Suppose it were found that the measured differences between schwas in /əCi/ and /əCu/ are greater in German than English. This may come about, not necessarily because of learned, language-specific coarticulatory differences,

but instead because the vowels in English are less peripheral than in German, as a result of which the anticipatory front-back influence on schwa is less. One way to begin to resolve this difficulty is with perception experiments in order to test whether listeners of different languages respond differently to coarticulation when presented with the same stimuli. This issue was addressed in a study by Beddor and Krakow (1999) who showed that native listeners of Thai both compensated less for the anticipatory nasalization in vowels than did English listeners to the same continua and showed less anticipatory nasal coarticulation in their production. More recent crosslinguistic studies of VCV coarticulation in Shona and English and other studies of nasal coarticulation in English and Thai have provided some further evidence that coarticulation is language-specific in production and that listeners are sensitive to these language-specific effects in perception (Beddor et al., 2002). This different sensitivity to perceived coarticulation depending on production differences has also been shown for two age groups of the same variety of Standard British English in Harrington et al. (2008): in this study, older subjects both compensated more for the coarticulatory influences on /u/ and (compatibly) exhibited a greater influence of context in their production of /u/ than younger subjects of the same variety.

2.4 Coarticulatory Variation and Change

Coarticulation has been shown to be speaker-specific (van den Heuvel et al., 1996; Magen, 1997; Grosvald, 2009). Part of the reason for this type of variability may well be because coarticulation is affected by speaking style (Krull, 1989). Another is that if phonological generalizations over phonetic detail depend statistically on learning experience (Pierrehumbert, 2003, 2006), then, given that no two speakers are ever exposed exactly to the same speaking situations, phonetic detail and therefore coarticulation also are likely to vary slightly from speaker to speaker.

Variation in speaking style and coarticulation are linked in Lindblom's (1990) H&H theory because words that are unpredictable for the listener tend to be hyperarticulated with the result that they are produced with greater clarity (Wright, 2003). Hyperarticulation is a form of segment strengthening which, as discussed earlier, can be linked to coarticulation resistance. Thus since in Germanic languages prosodically accented words often signal new and unpredictable information in an utterance, they tend to be hyperarticulated (de Jong, 1995) and their segments are less prone to coarticulatory influences than those in unaccented words (Harrington et al., 1995; Cho, 2004; Lindblom et al., 2007). Another prosodic variable that influences coarticulation is speaking rate (Bakran & Mildner, 1995) and sometimes in a way that can also be related to H&H theory (Agwuele et al., 2008).

Listeners have been shown to vary in their perception of coarticulation even when they respond to the same stimuli: for example, some experiments have shown that listeners are not consistent in the extent to which they compensate for nasal coarticulation in VN (Fowler & Brown, 2000) nor for vowel-to-vowel coarticulation in VCV stimuli (Beddor et al., 2001, 2002). Thus different listeners perceive the same coarticulated segment differently, if they vary in the extent to which they factor out coarticulation perceptually (Beddor, 2007; Kleber et al., 2012).

According to Ohala (1981, 1993), it is just this type of perceptual ambiguity that is the origin of sound-change: for example, the diachronic development of contrastive oral-nasal vowel phonemes in French and vowel harmony are two types of sound change that originate from under-parsing in which insufficient coarticulatory information is attributed to the source. A sound change such as the loss of the first /w/ from Latin /kwɪnkwe/ in its evolution into Italian /tʃɪnkwe/ (*five*) is presumed to come about because listeners overcompensate for coarticulation: thus, the presence of the initial /w/ is erroneously attributed by the listener to the anticipatory, coarticulatory lip-rounding and backing influence of the second /w/ and then (erroneously) factored out (ultimately resulting in its deletion if these perceptual effects are carried over to production).

3 Assimilation

Assimilation is a phenomenon related to coarticulation, in that it captures how the underlying specification of a sound may change under the influence of neighbouring sounds, either through lexical derivation (e.g. *in – probable → improbable*) or conditioned by fluent speech processes (e.g. *Paris show → Pari*[ʃʃ]*ow*). Here we will mainly be concerned with the latter type of assimilation. There have been (at least) three broad approaches to understanding assimilation. For one, assimilation has been modelled as a symbolic phonological restructuring process independent of phonetics. Another approach, primarily represented by Articulatory Phonology sees the origins of assimilation in the overlap of articulatory gestures, while a third view emphasizes the perceptual roots of assimilation. Although these three views have their origins in different phonological and phonetic models, they are by no means mutually exclusive, and no single approach has as of yet been able to account for the whole range of assimilation phenomena that is observed empirically.

In non-linear approaches to phonology such as Autosegmental Phonology or Feature Geometry, assimilation occurs when a distinctive feature (or subset of features) within a segment changes to agree with the feature(s) of an adjacent segment. This is achieved through linking and delinking of features (Goldsmith, 1976; Clements, 1985; McCarthy, 1988). By way of an example, in fluent speech, the word boundary cluster /d#b/ as in the phrase *road boy* may be (audibly)

pronounced with the final coronal being assimilated to the following labial. Schematically this can be represented as in Figure 16.2: The place feature [labial] spreads to the preceding Place node and the feature [coronal] is delinked, with the result of the assimilated sequence being specified as [labial] only.

In this type of model, assimilation is by definition categorical (all-or-none) and happens prior to the computation of the physical properties of the utterance. Therefore in articulation, the coronal is predicted to be categorically absent, that is, it has been replaced completely by the labial feature specification and is not produced. Articulatory recordings of assimilated sequences have shown that this prediction is not necessarily borne out: for example, articulatory records in the phrase *perfect memory* revealed that although *perceptually* the final /t/ was completely assimilated to the following labial, the coronal constriction was in fact still produced, but came to be hidden by the temporally overlapping labial articulation (Browman & Goldstein, 1990a; Tiede et al., 2001). This is schematically illustrated in Figure 16.3, where each box represents the time during which

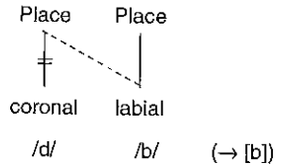

Figure 16.2 A representation of assimilation as spreading and delinking in Feature Geometry

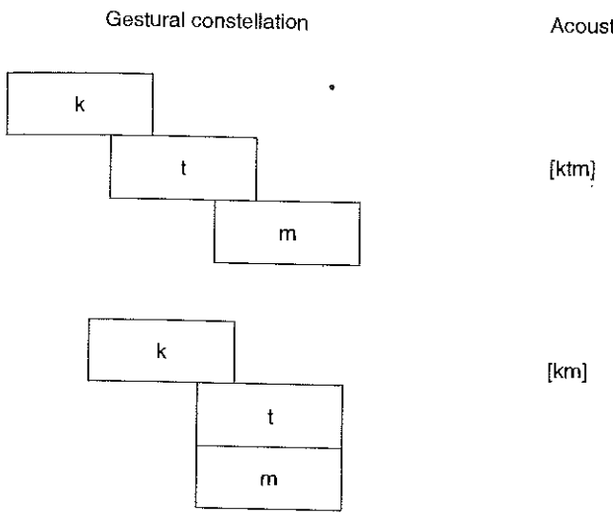

Figure 16.3 Overlap of gestures in Articulatory Phonology

a given constriction gesture (lips, tongue tip, etc.) is active. The tongue tip and the lips can perform their constrictions independently of each other and compatibly they occupy different 'tiers' in the gestural score. This means that even though the gestures may overlap in time in fluent speech, they can both be fully articulated, but it is the acoustic and perceptual consequences of the gestures that will change, resulting in *perfe[km]emory*.

In other cases of assimilation it could be shown that the spatial magnitude of the overlapped tongue tip gesture is reduced along a continuum of values (Surprenant & Goldstein, 1998). These findings seem to refute purely symbolic approaches to assimilation in which one feature specification is in a wholesale fashion replaced by another one. Note that a specification of a Place node as both coronal and labial, that is, linking without delinking, would specify a contour segment rather than gradient assimilation (Hayes, 1992; Nolan, 1992).

The *perfect memory* example illustrates how an articulatorily unassimilated sequence may be *perceived* as assimilated. Specifically *articulatory* assimilation is seen when the temporally coproduced gestures control the same articulator, that is when the gestures overlap spatially. For example, in the phrase *ten themes* both the final and initial consonant are produced with a tongue tip gesture. If the gestural activation intervals overlap in time, conflicting demands will govern the tongue tip with the result of a blended output, in this case a dental /n/ (more examples and a detailed discussion can be found in Browman & Goldstein, 1990a). Another case in point are English [ʃ#j] sequences, such as *miss you*, which can be pronounced as [mɪʃju] in connected speech. This [ʃ]-like percept is caused by the temporal overlap of the tongue tip gesture for /s/ and a tongue body raising gesture for /j/. Articulatorily and acoustically, the assimilated [ʃ] differs from the production of an underlying /ʃ/, since the assimilated fricative is a blend of the two simultaneously realized targets /s, j/ (Zsiga, 1995). For lexically derived forms, however, such as *impression*, Zsiga finds no evidence of [ʃ] arising from a coproduction of /s#j/ gestures. She therefore proposes that assimilation in lexically derived forms arises through symbolic (de)linking of features, while postlexical assimilation arises from gestural overlap.

According to the gestural overlap account of assimilation, (apparent) deletions, assimilation and weakening are all traced back to a single underlying principle: different degrees of gestural overlap, which may or may not be accompanied by a gradient reduction of the overlapped gesture. Depending on whether the overlapping gestures involve the same or different articulators, different consequences are observed. The hypothesis that fluent speech phenomena never involve a symbolic restructuring of phonological units (most explicitly stated in Browman & Goldstein, 1992), but can be exclusively understood as variation in timing and spatial magnitude of overlapping gestures has been quite controversial.

For one, several studies have shown that assimilation may indeed be categorical in that the assimilated gestures may be consistently and categorically absent (not produced), as predicted by the symbolic (de)linking account. For example, the tongue tip gesture for word-final alveolar /n/ in Castilian Spanish (e.g. *diga*[*n*] → *diga*[*m*] *paja*) is categorically reduced and the lip aperture gesture is temporally extended (Honorof, 1999). However, the categorical assimilation only occurs when there is a following non-coronal, whereas for following coronals, a blended output is observed in line with the gestural overlap approach. For Korean, Son et al. (2007) showed that in word-medial /pk/ clusters, the /p/ is either fully present or categorically not produced (but see Jun, 1996). While these cases of consistent and categorical assimilation might be viewed from a diachronic perspective in that a formerly gradient, fluent-speech assimilation process has become lexicalized and is independent of postlexical factors such as speech rate which usually condition gradient assimilation, there are several studies demonstrating that categorical and gradient assimilations truly coexist for connected speech processes.

Ellis and Hardcastle (2002) investigated /n#k/ sequences in English and found that some speakers produced an assimilatory continuum between [nk] and [ŋk], yet others showed a binary opposition between unassimilated [nk] or fully assimilated [ŋk], with no evidence for a non-velar target contributing to the output articulation. Experiments on English [s#ʃ] sibilant assimilation by Nolan et al. (Holst & Nolan, 1995; Nolan et al., 1996) confirmed that for some tokens a blended articulation between /s/ and /ʃ/ could be observed, as predicted by the gestural overlap account. Yet for other tokens, /s/ assimilated to /ʃ/ such that there was neither acoustically nor articulatorily any trace of a partially articulated [s] (as judged by tongue-palate contact data, EPG). The key argument against the gestural view comes from the durational properties of the assimilated sequence: importantly, the duration of the assimilated sibilant was longer compared to an underlying singleton [ʃ]: therefore, so the argument goes, the assimilated fricative cannot be the result of complete gestural overlap but must instead arise through a symbolic restructuring of features. Figure 16.4 illustrates schematically why increasing gestural overlap predicts, all else being equal, a reduced duration in assimilated sequences (the question of what constitutes a reference duration is not straightforward however – see Kühnert & Hoole, 2004).

Interestingly Nolan and colleagues also observe intermediate assimilation patterns for many tokens as predicted by gestural overlap, but they do not offer

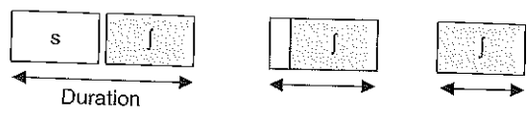

Figure 16.4 Schematic illustration of the durational effects of gestural overlap

an account of these assimilation patterns or how the occurrence of one or the other type of assimilation might be conditioned (for a recent discussion and extended study on s-ʃ assimilation, see Pouplier et al., 2011). Other studies describe a similar range of speaker behaviour, and assimilation data that are consistent with a symbolic linking-delinking view have emerged side-by-side with gradient assimilation and gestural hiding and blending phenomena (e.g. Barry, 1991; Nolan, 1999; Kühnert & Hoole, 2004; Kochetov & Pouplier, 2008).

While the gestural approach has sought to explain how different assimilation patterns may follow from the spatio-temporal overlap of gestures, neither Articulatory Phonology nor non-linear phonologies inherently make predictions about why certain types of assimilations are more frequently observed than other types.[1] For example, coronal stops are most likely to assimilate, while fricatives are less likely to assimilate. Several researchers (among others, Kohler, 1990; Ohala, 1990; Jun, 1995; Steriade, 2001, 2009) have pointed out that there is a dimension of assimilatory behaviour which cannot be captured on the basis of articulatory considerations only. Rather, it seems that the perceptibility of the consonant undergoing assimilation is inversely correlated with the propensity for assimilation to occur. A common view therefore holds that assimilation will be most frequently observed if the assimilated consonant is perceptually 'weak': fricatives do not assimilate where stops do because the former are perceptually more salient. Moreover, the regressive nature of assimilation is seen as falling out from perceptual factors, since word-final sounds are less perceptually salient compared to word-initial sounds. That perceptibility may be a predictor of place assimilation patterns is shown by Hura et al. (1992) (see also Ohala, 1990). In a perception experiment they show that the consonants that are generally most likely to be assimilated are the ones that were most frequently misperceived in their study, even though the misperceptions revealed in their study are mostly non-assimilatory in nature.

Opinions differ in the interpretation of these types of results. Some ascribe these perceptually determined assimilation patterns to functional considerations of the speaker. The speakers 'knows' about the contextually conditioned differences in the perceptibility of sounds and will choose to overlap and reduce articulations only in circumstances in which the change is likely to go unperceived or does not endanger lexical contrast. Therefore, word-final sounds are more likely to assimilate than word-initial sounds – word-initial sounds are crucial for lexical access. Conservation of articulatory effort is seen as the driving force of the speaker's behaviour (Kohler, 1990; Lindblom, 1990; Jun, 2004). To predict the circumstances under which assimilation may not be perceptually salient is of course not entirely straightforward, but concepts like Steriade's P-Map (Steriade, 2009) represent ways of turning this general concept into testable hypotheses. Others take a non-teleological view on the perceptual origins of assimilation, but relate assimilation to inadvertent perceptual errors.

Assimilation then becomes an inevitable by-product of (psycho)acoustic ambiguities resulting from nonlinearities in the articulatory-acoustic relationship. If a consonant is perceptually ambiguous in a certain context, it may be perceived as assimilated (even though potentially still produced by the speaker). Since speakers imitate each other, listeners will perpetuate the perceived assimilation in their own productions (Ohala, 1981; Chen, 2003).

In sum, assimilation as a pervasive phenomenon in spoken language remains a touchstone issue for different approaches to speech production in phonology as much as in phonetics. Overall, it has become clear that numerous factors influence the occurrence of assimilation for any given utterance, such as the phonetic context (Farnetani & Busà, 1994; Recasens & Pallarès, 2001; Kochetov & Pouplier, 2008), assumed casualness of speech (Lindblom, 1990; Jun, 1995) and also factors not discussed here such as lexical and co-occurrence frequency (Bybee, 2001; Pierrehumbert, 2001; Stephenson, 2003; Jaeger & Hoole, 2011).

4 Consonant Clusters

Consonant clusters represent some of the motorically most complex behaviour in speech, and their analysis has contributed to our understanding of the relationship between linguistic structure and speech as coordinated behaviour. Much evidence has shown that the coordination of onset clusters with the following vowel is different from that of coda clusters with the preceding vowel (Byrd, 1995; Honorof & Browman, 1995). An influential model of these effects was proposed by Browman and Goldstein (2000) and further developed in a computational framework in Goldstein et al. (2010). In their model, phasing relations between gestures make heaviest use of patterns that are motorically intrinsically stable, namely in-phase and anti-phase. For coda consonants, it is assumed that only the left-most consonant is directly coordinated with the preceding vowel (in an anti-phase relation) and that additional coda consonants are coordinated (also anti-phase) with the preceding one. By contrast, all onset consonants are coordinated in-phase with the vowel. To prevent the onset consonants being synchronous (and thus unrecoverable by the listener) they are assumed to be coupled anti-phase with each other. This results in a competitive coupling topology that expresses itself in a compromise timing pattern at the articulatory surface often referred to as the C-centre pattern (Browman & Goldstein, 2000). The arithmetic mean of the temporal location of the midpoints of all consonants stays in a stable timing relation to the vowel gesture: put another way, the right edge of the right-most consonant in the onset overlaps the vowel more and more as consonants are added to the onset (see Figure 16.5).

One of the most extensive recent investigations that explicitly compares onset and coda clusters is that of Marin and Pouplier (2010). Although they

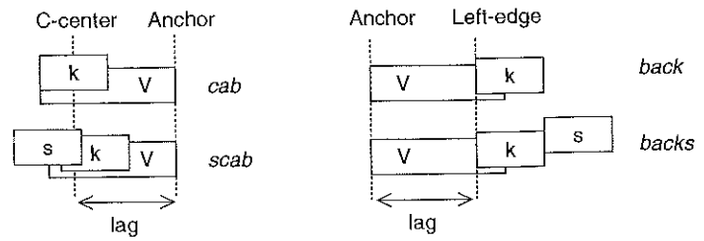

Figure 16.5 Illustration of hypothetical alignment of onset and coda consonants with the vowel and with each other (from Marin & Pouplier, 2010)

provided very clear evidence for the hypothesized C-centre timing pattern, they also found more variability in the coda timing patterns, that is, it was less clear that all possible codas prefer a purely sequential principle of organization. This greater stability of the onset is consistent with Nam's (2007) coupled oscillator model: a topology in which consonants are linked directly both to the vowel as well as to each other constitutes a more constrained topology than one in which the vowel and the consonants are simply linked serially to their neighbour. This may in turn be related to findings showing that codas are more sensitive to the influence of prosodic variation than are onsets (Bombien et al., 2010; Byrd & Choi, 2010; Hoole et al., 2010).

The idea of a C-centre has been used to advance arguments for or against syllable constituency. For example, Goldstein et al. (2007) found differences in timing between Berber and Georgian for superficially similar sequences of pre-vocalic consonants: since Georgian showed the C-centre pattern but Berber did not, they reasoned that the consonant sequences formed complex syllable onsets in Georgian, whereas Berber marked the presence of syllable divisions within the consonant sequences. Similarly, Shaw et al. (2009) found support for the assumption that in Arabic only simple syllable onsets are allowed, since forms such as /b/, /sb/, /ksb/ did not show the C-centre pattern (again a syllable boundary would be assumed before the /b/ in the clusters). Hermes et al. (2012) also argued that the so-called impure /s/ in Italian is not integrated into the C-centre pattern otherwise found for Italian syllable onsets, confirming its special status within Italian syllable structure.

The previous examples already show how relevant aspects of linguistic structure may be reflected in coordination patterns. Other recent work has indicated the need for further refinement in models of coordination relations. For example, based again on Georgian, Chitoran et al. (2002) documented an influencing factor on overlap patterns in consonants that has become known as the place-order effect (see also, for example, Gafos et al., 2010, for Arabic): in clusters of two plosive consonants, there is less overlap when C_1's place of articulation is posterior to that of C_2 (e.g. /tp/), than in the reverse case (e.g. /pt/). This

has been attributed to the necessity of ensuring the recoverability of information on C_1 by the listener: that is, there is a greater likelihood of the release of C_1 being acoustically obscured when the C_2 constriction is anterior to that of C_1.

In recent versions of the coupled oscillator model (Nam, 2007), the differences in overlap have been captured by expanding the possible range of coupling topologies, specifically by dividing consonantal gestures into a closure and a release gesture. For the high overlap (front-back) clusters, a default pattern can be used in which the closure gestures of C_1 and C_2 are coupled with the vowel. For the back-front clusters, the lower overlap can be achieved by coupling the *release* gesture of C_1 with the vowel (Goldstein et al., 2010). The differences in the coordination relations between the two cluster types can be associated with differences in phonological behaviour: specifically, the low-overlap clusters permit greater complexity in the laryngeal adjustments associated with the cluster. This may not be entirely surprising: a wider spacing of the consonants simply gives the speaker more time to change the laryngeal configuration and adjust to possibly conflicting aerodynamic requirements. The crucial point here is that an understanding of what are preferred patterns in one part of the speech production system (here the laryngeal specification) can depend on coordination relations in an apparently quite different part of the system (see Gafos et al., 2010 and Pouplier & Beňuš, 2011, for the related issue of cross-language differences in preferred coordination relations). In general, studies such as these provide a better understanding of why certain sound sequences are preferred in languages. Thus the sonority sequencing generalization in which consonants of increasing sonority are preferred nearer the vowel nucleus can be reinterpreted in terms of two phonetic principles: as discussed in Chitoran et al. (2002), preferred sound structures are those that allow a good compromise between parallel transmission of segmental information (efficient for the speaker) and clear modulation of the resulting acoustic signal (efficient for the listener).

The articulatory analysis of German consonant clusters in Hoole et al. (2010) provides further evidence that physiological and acoustic principles can explain both synchronic phonological patterns and diachronic change. Their results show that there is less overlap in /C_1n/ (e.g. German *Kneipe*) than in /C_1l/ (e.g. German *Claudia*) clusters: these differences may come about because premature lowering of the soft palate might destroy important acoustic properties of the plosive burst. The differences between these clusters could be modelled in a similar way to those for Georgian: for /C_1l/, there is a default coupling in which the closure of the initial stop is synchronized with the vowel whereas for /C_1n/ it is the initial consonant release that is synchronized with the nasal consonant. Thus, /C_1n/ clusters may be physiologically costly because they require a departure from the default coordination patterns (they may be additionally costly because of the need to increase the stiffness of the velar opening

gesture to ensure a sufficiently abrupt transition from closed position for the plosive to open position for the nasal). Compatibly, /C_1n/ initial clusters are not only rarer in the world's languages than /C_1l/, but they are also more prone to diachronic changes (Vennemann, 2000) such as the loss of the initial velar in /C_1n/ clusters (*knock, gnome*) around the seventeenth Century in English.

At the same time, cross-linguistic comparisons suggest that it may be premature to conclude that speech production is constructed around a small set of basic coordination patterns. For example, although Hoole et al. (2010) found a similar trend for less overlap in /C_1n/ clusters in French, the effect was much less clear-cut than in German and showed greater between-speaker variation (Hoole et al., 2010). Perhaps these differences can be related to the different organization of voicing in French and German, which might produce differences in the acoustic transition between consonants of an initial cluster. Thus this could be another case in which details of coordination patterns emerge from the interaction of different parts of the speech production system. Articulatory synthesis should be able to play an important role in future in elucidating the perceptual consequences of different patterns of overlap.

We would like to conclude this section with a brief summary of four topics that do not directly involve clusters in the narrow sense, but which are likely to contribute to the overall theme of the relationship between linguistic structure and coordinated behaviour in the immediate future.

4.1 Tonal Aspects of Syllable Structure

Gao (2009) has recently proposed that tones in Mandarin Chinese can be considered as tone gestures (T) that enter into a C-centre topology with the consonant and vowel of CV syllables: C and T are both coupled in-phase to the vowel but anti-phase to each other; thus C and T together act like an onset cluster. This is intriguing given the relevance of consonants for the emergence of tones in tonogenesis (see Sun, 2003, for links between consonant cluster simplification and tonal development), but it will be an enormous task to determine how well this concept generalizes to other tone languages, given their number and variety.

4.2 Syllabic Consonants

An unresolved issue is whether the gestural patterning of syllabic consonants is similar to that of vowels. Research by Pouplier and Beňuš (2011) on Slovakian suggests that consonant sequences containing a syllabic consonant are kinematically quite like consonant clusters. But much remains to be learnt here (see also Fougeron & Ridouane, 2008 and Ridouane, 2008 on Berber).

4.3 The Production of Schwa

Studies of pre-tonic schwa deletion within a gestural framework (i.e. in English words such as *police/please, support/sport* – see especially Davidson, 2006) suggest that schwa deletion is a gradient process that may be better modelled as the outcome of changes in gestural overlap rather than segmental deletion. Under an even more radical proposal, schwa may emerge as a result of differences in gestural coupling instead of being specified as part of the underlying representation: thus *police* and *please* would share the same cluster under this approach whose hyperarticulation produces the greater asynchrony (and perceived schwa) in the former (see Geng et al., 2010, for a recent discussion and Browman & Goldstein, 1990b, on targetless schwa).

4.4 Coordination and Morphology

Gafos et al. (2010) found that the extent of overlap of homoganic clusters of Arabic was low (i.e. they were produced with two distinct closures) when produced within the same morphological template, but high (i.e. produced as a single long closure) across an affixal boundary. The influence of morphology on articulatory timing was also found for Korean by Cho (2001). We suspect that there are many more phenomena of this kind still to be investigated.

5 Conclusions

One of the main conclusions to emerge from the above brief review is that linguistic structure needs to incorporate time not just in what Gafos (2002) has called its trivial sense, that is, serial order, but in the much richer sense of coordinated behaviour. Indeed, studying coordinated behaviour has the potential not only to provide a common basis for both the production and perception of speech as research on coarticulation shows, but it can also – as the studies reviewed under assimilation suggest – challenge commonly held views about continuous speech processes that are based on hearing (and transcribing) speech as a sequence of serially ordered consonants and vowels. Finally, empirically based research on coordination and speech production can advance our understanding of phonological patterning and diachronic change as many of the studies summarized in the final section on clusters have shown.

A central theme in all of the preceding three sections is that modelling transmission of dynamical information between speakers and listeners forms an essential part of understanding the relationship between the infinite variation in speech signals and phonological categories. In the future, the listener's

parsing of coarticulatory patterns in speech production will need to be tested on speaker groups such as children (Zharkova et al., 2011) and in contexts such as weak syllables in which coarticulation is likely to be more extensive or more variable. Moreover, modelling production-perception relationships will continue to be important for understanding how consonants are subjected to assimilation across word and other prosodic boundaries as well as the ways in which consonant clusters are influenced by different prosodic constituents such as onsets and coda (Marin & Pouplier, 2010, Pouplier, in press). All of these issues can shed further light on why syllable types and phonotactic combinations vary in their frequency of occurrence in the world's languages.

Another main theme of the preceding three sections is that sound change may be more likely if the production and perception of speech provide different and divergent solutions about how dynamical information is associated with phonological categories. Thus the first two sections emphasized how sound change is perhaps more likely to come about when coarticulatory or assimilatory information is ambiguously transmitted between a speaker and a hearer. A further currently unresolved issue will be to explain how dynamic activity that is unstable in either or both production and perception can nevertheless be phonologized and acquired by children, albeit in a minority of languages. To do so will require a closer analysis of physiological-perceptual relationships and of a wider range of syllable types drawn from many more languages than have been studied in the past.

Note

1. Resistance to coarticulation (Recasens & Pallarès, 2001; Recasens, 2006) has offered some insights into asymmetries in assimilatory patterns based on articulatory considerations.