# 3 Acoustic Phonetics

JONATHAN HARRINGTON

## 1 Introduction

In the production of speech, an acoustic signal is formed when the vocal organs move resulting in a pattern of disturbance to the air molecules in the airstream that is propagated outwards in all directions eventually reaching the ear of the listener. Acoustic phonetics is concerned with describing the different kinds of acoustic signal that the movement of the vocal organs gives rise to in the production of speech by male and female speakers across all age groups and in all languages, and under different speaking conditions and varieties of speaking style. Just about every field that is covered in this book needs to make use of some aspect of acoustic phonetics. With the ubiquity of PCs and the freely available software for making spectrograms, for processing speech signals, and for labeling speech data, it is also an area of experimental phonetics that is very readily accessible.

Our knowledge of acoustic phonetics is derived from various different kinds of enquiry that can be grouped loosely into three areas that derive primarily from the contact of phonetics with the disciplines of engineering/electronics, linguistics/phonology, and psychology/cognitive science respectively.

1 *The acoustic theory of speech production*. These studies assume an idealized model of the vocal tract in order to predict how different vocal tract shapes and actions contribute to the acoustic signal (Stevens & House, 1955; Fant, 1960). Acoustic theory proposes that the excitation signal of the source can be modeled as independent from the filter characteristics of the vocal tract, an idea that is fundamental to acoustic phonetics, to formant-based speech synthesis, and to linear predictive coding which allows formants to be tracked digitally. The discovery that vowel formants can be accurately predicted by reducing the complexities of the vocal tract to a three-parameter, four-tube model (Fant, 1960) was one of the most important scientific breakthroughs in phonetics of the last century. The idea that the relationship between speech production and

acoustics is nonlinear and that, as posited by the quantal theory of speech production (Stevens, 1972, 1989), such discontinuities are exploited by languages in building up their sound systems, is founded upon models that relate idealized vocal tracts to the acoustic signal.

2. *Linguistic phonetics* draws upon articulatory and acoustic phonetics in order to explain why the sounds of languages are shaped the way they are. The contact with acoustic phonetics is in various forms, one of which (quantal theory) has already been mentioned. Developing models of the distribution of the possible sounds in the world's languages based on acoustic principles, as in the ground-breaking theory of adapative dispersion in Liljencrants and Lindblom (1972), is another. Using the relationship between speech production and acoustics to explain sound change as misperception and misparsing of the speech signal (Ohala, 1993, this volume) could also be grouped in this area.
3. *Variability*. The acoustic speech signal carries not only the linguistic structure of the utterance, but also a wealth of information about the speaker (physiology, regional affiliation, attitude and emotional state). These are entwined in the acoustic signal in a complex way acoustically both with each other and with background noise that occurs in almost every natural dialogue. Moreover, speech is highly context-dependent. A time slice of an acoustic signal can contain information about context, both segmental (e.g., whether a vowel is surrounded by nasal or oral sounds) and prosodic (e.g., whether the vowel is in a stressed syllable, in an accented word at the beginning or near the end of a prosodic phrase). Obviously, listeners cope for the most part effortlessly with all these multiple strands of variability. Understanding how they do so (and how they fail to do so in situations of communication difficulty) is one of the main goals of speech perception and its relationship to speech production and the acoustic signal.

As in any science, the advances in acoustic phonetics can be linked to technological development. Present-day acoustic phonetics more or less began with the invention of the sound spectrograph in the 1940s (Koenig et al., 1946). In the 1950s, the advances in vocal tract modeling and speech synthesis (Dunn, 1950; Lawrence, 1953; Fant, 1960) and a range of innovative experiments at the Haskins Laboratories (Cooper et al., 1951) using synthesis from hand-painted spectrograms underpinned the technology for carrying out many types of investigation in speech perception. The advances in speech signal processing in the 1960s and 1970s resulted in techniques like cepstral analysis and the linear prediction of speech (Atal & Hanauer, 1971) for source-filter separation and formant tracking. As a result of the further development in computer technology in the last 20–30 years and above all with the need to provide extensive training and testing material for speech technology systems, there are now large-scale acoustic databases, many of them phonetically labeled, as well as tools for their analysis (Bird & Harrington, 2001).

A recording of the production of speech with a pressure-sensitive microphone shows that there are broadly a few basic kinds of acoustic speech signal that it will be convenient to consider in separate sections in this chapter.

- *Vowels and vowel-like sounds*. Included here are sounds that are produced with periodic vocal fold vibration and a raised velum so that the airstream exits only from the mouth cavity. In these sounds, the waveform is periodic, energy is concentrated in the lower half of the spectrum, and formants, due to the resonances of the vocal tract, are prominent.
- *Fricatives and fricated sounds*. These will include, for example, fricatives and the release of oral stops that are produced with a turbulent airstream. If there is no vocal fold vibration, then the waveform is aperiodic; otherwise there is combined aperiodicity and periodicity that stem respectively from two sources at or near the constriction and due to the vibrating vocal folds. I will also include the silence that is clearly visible in oral stop production in this section.
- *Nasals and nasalized vowels*. These are produced with a lowered velum and in most cases with periodic vocal fold vibration. The resulting waveform is, as for vowels, periodic but the lowered velum and excitation of a side-branching cavity causes a set of anti-resonances to be introduced into the signal. These are among the most complex sounds in acoustic phonetics.

My emphasis will be on describing the acoustic phonetic characteristics of speech sounds, drawing upon studies that fall into the three categories described earlier. Since prosody is covered elsewhere in two chapters in this book, my focus will be predominantly on the segmental aspects of speech. I will also not cover vowel or speaker normalization in any detail, since these have been extensively covered by Johnson (2005).

## 2 Vowels, Vowel-Like Sounds, and Formants

### 2.1 *The F1 × F2 plane*

The acoustic theory of speech production has shown how vowels can be modeled as a straight-sided tube closed at one end (to model the closure phase of vocal fold vibration) and open at the lip end. Vowels also have a point of greatest narrowing known as a constriction location (Stevens & House, 1955; Ladefoged, 1985) that is analogous to place of articulation in consonants and that divides the tube into a back cavity and a front cavity. As Fant's (1960) nomograms show, varying the constriction location from the front to the back of the tube causes changes predominantly to the first two resonant frequencies. The changes are *nonlinear* which means that there are regions where large changes in the place of articulation, or constriction location, have a negligible effect on the formants (e.g., in the region of the soft palate) and other regions such as between the hard and soft palate where a small articulatory change can have dramatic acoustic consequences. Since there are no side-branching resonators – that is, since there is only one exit at the mouth for the air expelled from the lungs – the acoustic structure of a vowel is determined by resonances that, when combined (convolved) with the source signal, give rise to *formants*. The formants are clearly visible in a spectrographic display and they occur on average at intervals of $c/2L$, where $c$ is the speech of sound and $L$ the length of the vocal tract (Fant, 1973) – that is, at about 1,000 Hz intervals for an adult male vocal tract of length 17.5 cm (and with the speed of sound at 35,000 cm/s). As far as the relationship between vocal tract shape and formants are concerned, some of the main findings are:

- All parts of the vocal cavities have some influence on all formants and each formant is dependent on the entire shape of the complete system (see, e.g., Fant, 1973).
- A maximally high F1 (the first, or lowest, formant frequency) requires the main constriction to be located just above the larynx and the mouth cavity to be wide open. An increasing constriction in the mouth cavity results in a drop in F1 (see also Lindblom & Sundberg, 1971).
- A maximally high F2 is associated with a tongue constriction in the palatal region. More forward constrictions produce an increase in F3 and F4 that is due to the shortening of the front tube (Ladefoged, 1985) so that there is a progressive increase first in F2, then in F3, then in F4 as the constriction location shifts forward of the palatal zone. F2 is maximally low when the tongue constriction is in the upper part of the pharynx.
- Either a decrease of lip-opening area or an increase of the length of the lip passage produces formant lowering. Lip-protrusion has a marked effect on F3 in front vowels and on F2 in back vowels – see, e.g., Lindblom and Sundberg (1971) and Ladefoged and Bladon (1982).

The acoustic theory of speech production shows that there is a relationship between phonetic height and F1 and phonetic backness and F2, from which it follows that if vowels are plotted in the plane of the first two formant frequencies with decreasing F1 on the *x*-axis and decreasing F2 on the *y*-axis, a shape resembling the articulatory vowel quadrilateral emerges. This was first demonstrated by Essner (1947) and Joos (1948), and since then the F1 × F2 plane has become one of the standard ways of comparing vowel quality in a whole range of studies in linguistic phonetics (Ladefoged, 1971), sociophonetics (Labov, 2001), and in many other fields.

Experiments with hand-painted spectrograms using the Pattern Playback system at the Haskins Laboratories showed that vowels of different quality could be accurately identified from synthetic speech that included only the first two or only the first three formant frequencies (Delattre et al., 1955). In the 1970s and 1980s, experimental evidence of a different kind, involving an analysis of the pattern of listeners' confusions between vowels (e.g., Klein et al., 1970; Shepard, 1972) showed that perceived judgments of vowel quality depend in some way on the F1 × F2 space. The nature of these experiments varied: in some, listeners were presented with a sequence of three vowels and asked to judge whether the third is more similar to the first or to the second; or listeners might be asked to judge vowel quality in background noise. The pattern of resulting listener vowel confusions can be transformed into a spatial representation using a technique known as *multidimensional scaling* (Shepard, 1972). Studies have shown that up to six dimensions may be necessary to explain adequately the listeners' pattern of confusion between vowels (e.g., Terbeek, 1977), but also that the two most important dimensions for explaining these confusions are closely correlated with the first two formant frequencies (see also Johnson, 2004, for a discussion of Terbeek's data). These studies are important in showing that the F1 × F2 space, or some auditorily transformed version of it, represents the principal dimensions in which listeners judge vowel quality. Moreover, if listener judgments of vowel quality are primarily dependent on the F1 × F2 space, then languages should maximize the distribution between vowels in this space in order that they will be perceptually distinctive and just this has been shown in the computer simulation studies of vowel distributions in Liljencrants and Lindblom (1972).

Even in citation-form speech, the formants of a vowel are not horizontal or "steady-state" but change as a function of time. As discussed in section 2.5, much of this change comes about because preceding and following segments cause deviations away from a so-called *vowel target* (Lindblom, 1963; Stevens & House, 1963). The vowel target can be thought of as a single time point that in monophthongs typically occurs nearest near the temporal midpoint, or a section of the vowel (again near the temporal midpoint) that shows the smallest degree of spectral change and which is the part of the vowel least influenced by these contextual effects. In speech research, there is no standard method for identifying where the vowel target occurs, partly because many monophthongal vowels often have no clearly identifiable steady-state or else the steady-state, or interval that changes the least, may be different for different formants. Some researchers (e.g., Broad & Wakita, 1977; Schouten & Pols, 1979a, 1979b) apply a Euclidean-distance metric to the vowel formants to find the least-changing section of the vowel, while others estimate targets from the time at which the formants reach their maximum or minimum values (Figure 3.1). For example, since a greater mouth opening causes F1 to rise, then when a nonhigh vowel is surrounded by consonants, F1 generally rises to a maximum near the midpoint (since there is greater vocal tract constriction at the vowel margins) and so the F1-maximum can be taken to be the vowel target (see van Son & Pols, 1990 for a detailed comparison of some of the different ways of finding a vowel target).

## 2.2 *F3 and* $\mathbf{f_0}$

When listeners labeled front vowels from two-formant stimuli in the Pattern Playback experiments at the Haskins Laboratories, Delattre et al. (1952) found that they preferred F2 to be higher than the F2 typically found in the corresponding natural vowels and they reasoned that this was due to the effects of F3. This preferred upwards shift in F2 in synthesizing vowels with only two formants was subsequently quantified in a further set of synthesis and labeling experiments (e.g., Carlson et al., 1975) in which listeners heard the same vowel (a) synthesized with two formants and (b) synthesized with four formants, and were asked to

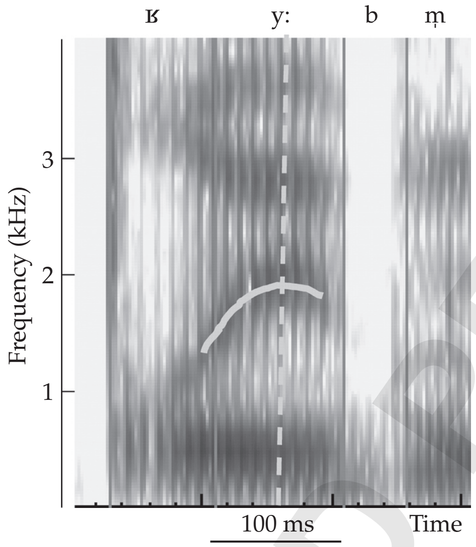

**Figure 3.1** Spectrogram of the German word *drüben*, [dʁy:bm̩], produced by an adult male speaker of German. The intersection of the vertical dotted line with the hand-drawn F2 is the estimated acoustic vowel target of [y:] based on the time at which F2 reaches a maximum.

adjust F2 until (a) was perceptually as close to (b) as possible. The adjusted F2 is sometimes referred to as an *effective upper formant* or *F2-prime*.

As discussed in Strange (1999), the influence of F3 on the perception of vowels can be related to studies by Chistovich (1985) and Chistovich and Lublinskaya (1979) showing that listeners integrate auditorily two spectral peaks if their frequencies are within 3.0–3.5 Bark. Thus in front vowels, listeners tend to integrate F2 and F3 because they are within 3.5 Bark of each other, and this is why in two-formant synthesis an effective upper formant is preferred which is close to the F2 and F3 average.

Based on the experiments by Chistovich referred to above, Syrdal (1985) and Syrdal and Gopal (1986) proposed F3 – F2 in Bark as an alternative to F2 as the principal correlate of vowel backness. In their studies, a distinction between front and back vowels was based on the 3.5 Bark threshold (less for front vowels, greater for back vowels). When applied to the vowel data collected by Peterson and Barney (1952), this parameter also resulted in a good deal of speaker normalization. On the other hand, although Syrdal and Gopal (1986) show that the extent of separation between vowel categories was greater in a Bark than in a Hertz space, it has not, as far as I know, been demonstrated that F3 – F2 Bark provides a more effective distinction between vowels than F2 Bark on its own.

In the post-alveolar approximant [ɹ] and the "r-colored" vowels in American English (e.g., *bird*), F3 is very low. F3 also contributes to the unrounded/rounded distinction in front vowels in languages in which this contrast is phonemic (e.g., Vaissière, 2007). In such languages, [i] is often prepalatal, i.e., the tongue dorsum constriction is slightly forward of the hard palate and it is this difference that is responsible for the higher F3 in prepalatal French [i] compared with palatal English [i] (Wood, 1986). Moreover, this higher F3 sharpens the contrast to [y] in which F3 is low and close to F2 because of lip-rounding.

It has been known since studies by Taylor (1933) and House and Fairbanks (1953) that there is an intrinsic fundamental frequency association with vowel height: all things being equal, phonetically higher vowels tend to have higher $f_0$. Traunmüller (1981, 1984) has shown in a set of perception experiments that perceived vowel openness stays more or less constant if Bark-scaled $f_0$ and F1 increase or decrease together: his general conclusion is that perceived vowel openness depends on the difference between F1 and $f_0$ in Bark. In their reanalysis of the Peterson and Barney (1952) data, Syrdal and Gopal (1986) show that vowel height differences can be quite well represented on this parameter and they show that high vowels have an F1 – $f_0$ difference that is less than the critical distance of 3 Bark.

## 2.3 *Dynamic cues to vowels*

Many languages make a contrast between vowels that are spectrally quite similar but that differ in duration. On the other hand, there is both a length and a spectral difference in most English accents between the vowels of *heed* versus *hid* or *who'd* versus *hood*. These vowel pairs are often referred to as "tense" as opposed to "lax." Tense vowels generally occupy positions in the F1 × F2 space that are more peripheral, i.e., further away from the center than lax vowels. There is some evidence that tense–lax vowel pairs may be further distinguished based on the proportional time in the vowel at which the vowel target occurs (Lehiste & Peterson, 1961). Huang (1986, 1992) has shown in a perception experiment that the cross-over point from perception of lax [ɪ] to tense [i] was influenced by the relative position of the target (relative length of initial and final transitions) – see also Strange and Bohn (1998) for a study of the tense/lax distinction in North German. Differences in the proportional timing of vowel targets are not confined to the tense/lax distinction. For example, Australian English [i:] has a late target, i.e., long onglide (Cox, 1998) – compare for example the relative time at which the F2 peak occurs in the Australian English and Standard German [i:] in Figure 3.2.

Another more common way for targets to differ is in the contrast between monophthongs and diphthongs, i.e., between vowels with a single as opposed to two targets. Some of the earliest acoustic studies of (American English) diphthongs were by Holbrook and Fairbanks (1962) and Lehiste and Peterson (1961). Gay (1968, 1970) showed that the second diphthong target is much more likely to be undershot and reduced than the first. From this it follows that the first target and the direction of spectral change may be critical in identifying and distinguishing between diphthongs, rather than whether the second target is actually attained.

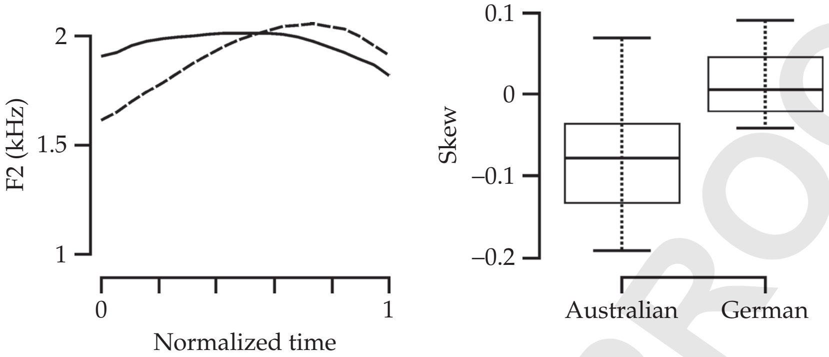

**Figure 3.2** Left: Linearly time-normalized plots of F2 averaged across 57 [i:] vowels produced by a male speaker of Australian English (dotted) and across 38 [i:] vowels produced by a male speaker of Standard German (solid). All vowels were extracted from lexically stressed syllables in read sentences. Right: The distribution of these [i:] vowels on a parameter of the F2-skew for the Australian and German speakers separately, calculated with the third statistical moment (see (8) and section 3.1).

Gottfried et al. (1993) analyzed acoustically in an F1 × F2 logarithmic space three of the different hypotheses for diphthong identification discussed in Nearey and Assmann (1986). These were that (a) both targets, (b) the onset plus the rate of change of the spectrum, and (c) the onset plus the direction, are critical for diphthong identification. The results of an analysis of 768 diphthongs provided support for all three hypotheses, with the highest classification scores obtained from (a), the dual target hypothesis.

Many studies in the *Journal of the Acoustical Society of America* in the last 30 years have been devoted to the issue of whether vowels are sufficiently distinguished by information confined to the vowel target. It seems evident that the answer must be no (Harrington & Cassidy, 1994; Watson & Harrington, 1999), given that, as discussed above, vowels can vary in length, in the relative timing of the target, and in whether vowels are specified by one target or two. Nevertheless, the case for vowels being "dynamic" in general was made by Strange and colleagues based on two sets of data. In the first, Strange et al. (1976) found that listeners identified vowels more accurately from CVC than from isolated V syllables; and in the second, vowels were as well identified from so-called silent center syllables, in which the middle section of CVC syllable had been spliced out leaving only transitions, as from the original CVC syllables (Strange et al., 1983). Both sets of experiments led to the conclusion that there is at least as much information for vowel identification in the (dynamically changing) transitions as at the target. Compatibly, human listeners make more errors in identifying vowels from static (steady-state) synthetic vowels compared with synthetic vowels that include formant change (e.g., Hillenbrand & Nearey, 1999) and a number of acoustic experiments have shown that vowel classification is improved using information other than just at the vowel target (e.g., Hillenbrand et al., 2001; Huang, 1992; Zahorian & Jagharghi, 1993).

## 2.4 Whole-spectrum approaches to vowel identification

Although no one would dispute that the acoustic and perceptual identification of vowels is dependent on formant frequencies, many have also argued that there is much information in the spectrum for vowel identity apart from formant center frequencies. Bladon (1982) and Bladon and Lindblom (1981) have advocated a whole-spectrum approach and have argued that vowel identity is based on gross spectral properties such as auditory spectral density. More recently, Ito et al. (2001) showed that the tilt of the spectrum can cue vowel identity as effectively as F2. On the other hand, manipulation of formant amplitudes was shown to have little effect on listener identification of vowels in both Assmann (1991) and Klatt (1982); and Kiefte and Kluender's (2005) experiments show that, while spectral tilt may be important for identifying steady-state vowels, its contribution is less important in more natural speaking contexts. Most recently, in Hillenbrand et al. (2006), listeners identified vowels from two kinds of synthesized stimuli. In one, all the details of the spectrum were included while in the other, the fine spectral structure was removed preserving information only about the spectral peaks. They found that identification rates were higher from the first kind, but only marginally so (see also Molis, 2005). The general point that emerges from these studies is that formants undoubtedly provide the most salient information about vowel identity in both acoustic classification and perception experiments and that the rest of the shape of the spectrum may enhance these distinctions (and may provide additional information about the speaker which could, in turn, indirectly aid vowel identification).

Once again, the evidence that the primary information for vowel identification is contained in the formant frequencies emerges when data reduction techniques are applied to vowel spectra. In this kind of approach (e.g., Klein et al., 1970; Pols et al., 1973), energy values are summed in auditorily scaled bands. For example, the spectrum up to 10 kHz includes roughly 22 bands at intervals of 1 Bark, so if energy values are summed in each of these Bark bands, then each vowel's spectrum is reduced to 22 values, i.e., to a point in 22-dimensional space. The technique of *principal components analysis* (PCA) finds new axes through this space such that the first axis explains most of the variance in the original data, the second axis is orthogonal to the first, the third is orthogonal to the second, and so on. Vowels can be distinguished just as accurately from considerably fewer dimensions in a PCA-rotated space of these Bark-scaled filter bands as from the original high-dimensional space. But also, one of the important findings to emerge from this research is that the first two dimensions are often strongly correlated with the first two formant frequencies (Klein et al., 1970). (This technique has also been used in child speech in which formant tracking is difficult – see Palethorpe et al., 1996.)

This relationship between a PCA-transformed Bark space and the formant frequencies is evident in Figure 3.3 in which PCA was applied to Bark bands

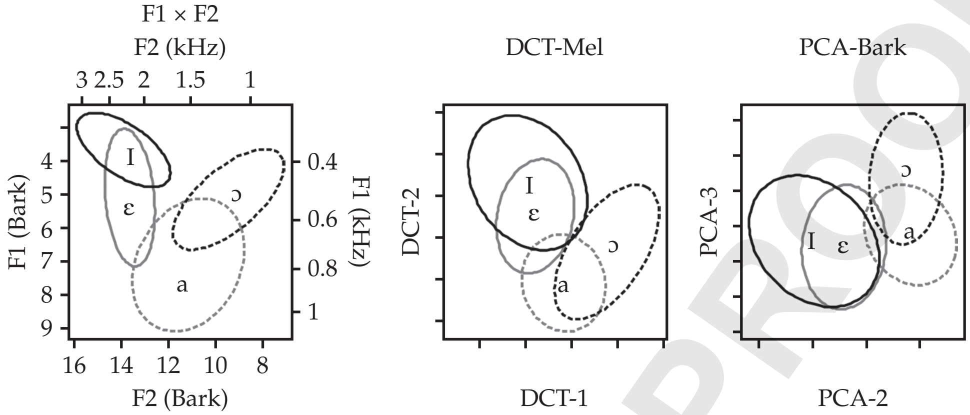

**Figure 3.3** 95% confidence ellipses for four lax vowels extracted from lexically stressed syllables in read sentences and produced by an adult female speaker of Standard German in the planes of F2 x F1 in Bark (left), the first two DCT coefficients (center), and two dimensions derived after applying PCA to Bark bands calculated in the 200–4,000 Hz range (right). The numbers of tokens in the categories [ɪ, ɛ, a, ɔ] were 85, 41, 63, and 16 respectively.

spanning the 200–4,000 Hz range in some German lax vowels [ɪ, ɛ, a, ɔ]. Spectra were calculated for these vowels with a 16-ms window at a sampling frequency of 16 kHz and energy values were calculated at one Bark intervals over the frequency range 200–4,000 Hz, thereby reducing each spectrum to a point in a 15-dimensional space. The data were then rotated using PCA. As Figure 3.3 shows, PCA-2 is similar to F1 in separating vowels in terms of phonetic height while [a] and [ɔ] are separated almost as well on PCA-3 as on F2. Indeed, if this PCA space were further rotated by about 45 degrees clockwise, then there would be quite a close correspondence to the distribution of vowels in the F1 × F2 plane, as Klein et al. (1970) had shown.

We arrive at a similar result in modeling vowel spectra with the discrete cosine transformation (DCT; Zahorian & Jagharghi, 1993; Watson & Harrington, 1999; Palethorpe et al., 2003). As discussed in more detail in section 3.1 below, the result of applying a DCT to a spectrum is a set of DCT coefficients that encode properties of the spectrum's shape. When a DCT analysis is applied to vowel spectra, then the first few DCT coefficients are often sufficient for distinguishing between vowels, or the distinction is about as accurate as from formant frequencies (Zahorian & Jagharghi, 1993). In Figure 3.3, a DCT analysis was applied to the same spectra in the 200–4,000 Hz range that were subjected to PCA analysis. Before applying the DCT analysis, the frequency axis of the spectra was converted to the auditory mel scale. Again, a shape that resembles the F1 × F2 space emerges when these vowels are plotted in the plane of DCT-1 × DCT-2. (It should be mentioned here that DCT coefficients derived from mel spectra are more or less the same as mel-frequency cepstral coefficients that are often used in automatic speech recognition – see, e.g., Nossair & Zahorian, 1991; and Milner & Shao, 2006.)

## 2.5 *Vowel reduction*

It is important from the outset to make a clear distinction between phonological and phonetic vowel reduction: the first is an obligatory process in which vowels become weak due to phonological and morphological factors, as shown by the alternation between /eɪ/ and /ə/ in *Canadian* and *Canada* in most varieties of English. In the second, vowels are phonetically modified because of the effects of segmental and prosodic context. Only the second is of concern here.

Vowel reduction is generally of two kinds: *centralization* and *coarticulation*, which together are sometimes also referred to as *vowel undershoot*. The first of these is a form of paradigmatic vowel reduction in which vowels become more schwa-like and the entire vowel space shrinks as vowels shift towards the center. Coarticulation is syntagmatic: here there are shifts in vowels that can be more directly attributed to the effects of preceding and following context.

The most complete account of segmental reduction is Lindblom's (1990, 1996) model of hyper- and hypoarticulation (H&H) in which the speaker plans to produce utterances that are sufficiently intelligible to the listener, i.e., a speaker economizes on articulatory effort but without sacrificing intelligibility. Moreover, the speaker makes a moment-by-moment estimate of the listener's need for signal information and adapts the utterance accordingly. When the listener's needs for information are high, then the talker tends to increase articulatory effort (hyperarticulate) in order to produce speech more clearly. Thus when words are excised from a context in which they are difficult to predict from context, listeners find them easier to identify than when words are spliced out of predictable contexts (Lieberman, 1963; Hunnicutt, 1985, 1987). Similarly, repeated words are shorter in duration and less intelligible when spliced out of context than the same words produced on the first occasion (Fowler & Housum, 1987).

As far as vowels are concerned, hyperarticulated speech is generally associated with less centralization and less coarticulation, i.e., an expansion of the vowel space and/or a decrease in coarticulatory overlap. There is evidence for both of these in speech that is produced with increased clarity (e.g., Picheny et al., 1986; Moon & Lindblom, 1994; Smiljanić & Bradlow, 2005). Additionally, Wright (2003) has demonstrated an H&H effect even when words are produced in isolation. He showed that the vowels of words that are "hard" have an expanded vowel space relative to "easy" words. The distinction between hard and easy takes account both of the statistical frequency with which words are used in the language and the lexical *neighborhood density*: if a word has a high value on neighborhood density, then there are very many other words which are phonemically identical to it based on substituting any one of the word's phonemes. Easy words are those which are high in frequency and low in neighborhood density. By contrast, hard words occur infrequently in the language and are confusable with other words, i.e., have high neighborhood density.

There have been several recent studies exploring the relationship between redundancy and hypoarticulation (van Son & Pols, 1999, 2003; Bybee, 2000; Bell et al., 2003; Jurafsky et al., 2003; Munson and Soloman, 2004; Aylett & Turk, 2006). The study by Aylett and Turk (2006) made use of a large corpus of citation-form speech including 50,000 words from each of three male and five female speakers. Their analysis of F1 and F2 at the vowel midpoint showed that vowels with high predictability were significantly centralized relative to vowels in less redundant words.

Many studies have shown an association between vowel reduction and various levels of the stress hierarchy (Fry, 1965; Edwards, Beckman, & Fletcher, 1991; Fourakis, 1991; Sluijter and van Heuven, 1996; Sluijter et al., 1997; Harrington et al., 2000; Hay et al., 2006) and with rate (e.g., Turner et al., 1995; Weismer et al., 2000). The rate effects on the vowel space are not all consistent (van Son & Pols, 1990, 1992; Stack et al., 2006; Tsao et al., 2006) not only because speakers do not all increase rate by the same factor, but also because there can be articulatory reorganization with rate changes.

As far as syntagmatic coarticulatory effects are concerned, Stevens and House (1963) found that consonantal context shifted vowel formants towards more central values, with the most dramatic influence being on F2 due to place of articulation. More recently, large shifts due to phonetic context have been reported in Hillenbrand et al. (2001) for an analysis of six men and six women producing eight vowels in CVC syllables. At the same time, studies by Pols (e.g., Schouten & Pols, 1979a, 1979b) show that the size of the influence of the consonant on vowel targets is considerably less than the displacement to vowel targets caused by speaker variation and in the study by Hillenbrand et al. (2001), consonant environment had a significant, although small, effect on vowel intelligibility. Although consonantal context can cause vowel centralization, Lindblom (1963), Moon and Lindblom (1994), and van Bergem (1993) emphasize that coarticulated vowels do not necessarily *centralize* but that the formants shift in the direction of the loci of the flanking segments.

Lindblom and Studdert-Kennedy (1967) showed that listeners compensate for the coarticulatory effects of consonants on vowels. In their study, listeners identified more tokens from an /ɪ–ʊ/ continuum as /ɪ/ in a /w_w/ context than in a /j_j/ context. This comes about because F2 lowering is a cue not only for /ʊ/ as opposed to /ɪ/, but also because F2 lowering is brought about by the coarticulatory effects of the low F2 of /w/. Thus, because of this dual association of F2 lowering, there is a greater probability of hearing the same token as /ɪ/ in a /w_w/ than in a /j_j/ context if listeners factor out the proportion of F2 lowering that they assume to be attributable to /w/-induced coarticulation.

Based on an analysis of the shift in the first three formants of vowels in /bVb, dVd, gVg/ contexts, Lindblom (1963) developed a mathematical model of vowel reduction in which the extent of vowel undershoot was exponentially related to vowel duration. The model was founded on the idea that the power, or articulatory effort, delivered to the articulators remained more or less constant, even if other factors – such as consonantal context, speech tempo, or a reduction of stress – caused vowel duration to decrease. The necessary outcome of the combination of a constant articulatory power with a decrease in vowel duration is, according to this model, vowel undershoot (since if the power to the articulators remains the same, there will be insufficient time for the vowel target to be produced).

The superposition model of Broad and Clermont (1987) is quite closely related to Lindblom's (1963) model, at least as far as the exponential relationship between undershoot and duration is concerned (see also van Son, 1993: ch. 1 for a very helpful discussion of the relationship between these two models). Their model is based on the findings of Broad and Fertig (1970), who showed that formant contours in a CVC syllable can be modeled as the sum of $f(t) + g(t) + V_T$ where $f(t)$ and $g(t)$ define as a function of time the CV and VC formant transitions respectively and $V_T$ is the formant frequency at the vowel target. This superposition model is also related to Öhman's (1967) numerical model of coarticulation based on VCV sequences in which the shape of the tongue at a particular point in time was modeled as a linear combination of a vowel shape, a consonant shape, and a coarticulatory weighting factor.

In one version of Broad and Clermont (1987), the initial and final transition functions, $f(t)$ and $g(t)$, are defined as:

$$f(t) = K_i(T_v - L_i)e^{-\beta_i t} \tag{1}$$

$$g(t) = K_f(T_v - L_f)e^{\beta_f(t-D)} \tag{2}$$

where $K$ (i.e., $K_i$ and $K_f$ for initial and final transitions respectively) is a consonant-specific scale-factor, $T_v - L_i$ and $T_v - L_f$ are the target–locus distances in CV and VC transitions respectively, $\beta$ is a time-constant that defines the rate of transition, and $D$ is the total duration of the CVC transition. Just as in Lindblom (1963), the essence of (1) and (2) is that the greater the duration, the more the transitions approach the vowel target.

Figure 3.4 shows an example of how an F2 transition in a syllable /dɪd/ could be put together with (1) and (2) (and using the parameters in Table VI of Broad & Clermont, 1987). The functions $f(t)$ and $g(t)$ define F2 of /dɪ/ and /ɪd/ as a function of time. To get the output for /dɪd/, $f(t)$ and $g(t)$ are summed at equal points in time and then these are added to the vowel target, which in this example is set to 2,276 Hz. Notice firstly that the initial and final transitions are negative and asymptote to zero, so that when they are added to the vowel target, their combined effect on the formant contour is least at the vowel target and progressively greater towards the syllable margins. Moreover, the model incorporates the idea from Broad and Fertig (1970) that initial and final transitions can influence each other at *all* time points, but that importantly the mutual influence of the initial on the final transitions progressively wanes for time points further away from the target.

In the first row of Figure 3.4, the duration of the CVC syllable is sufficient for the target to be almost attained. In row 2, the CVC has a duration that is 100 ms

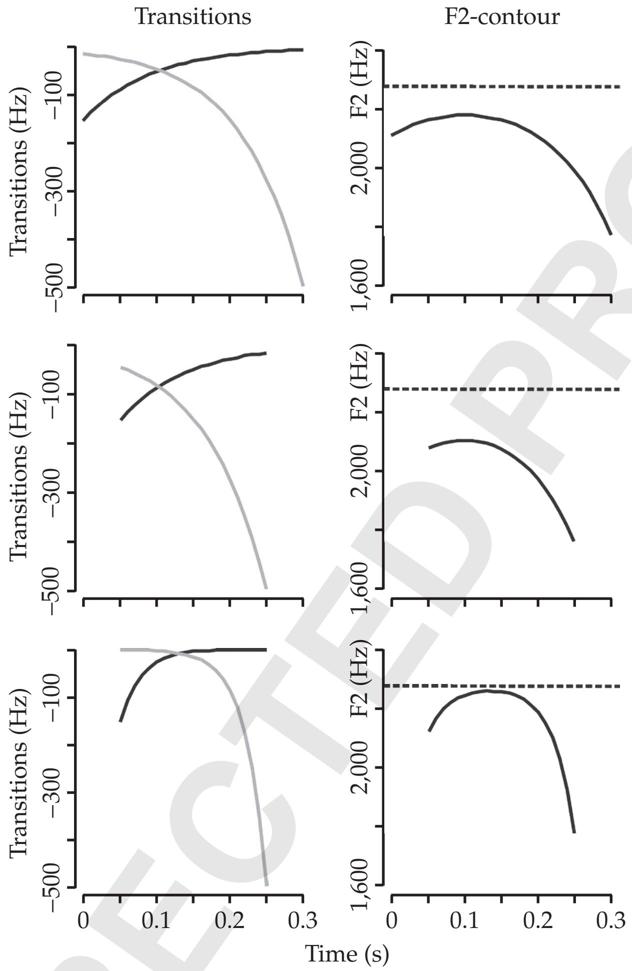

**Figure 3.4** An implementation of the equations (1) and (2) for constructing an F2-contour appropriate for the context [dɪd] using the parameters given in Table VI of Broad and Clermont (1987). Left: the values of the initial [dɪ] (black) and final [ɪd] (grey) transitions. Right: the corresponding F2 contour that results when the transitions on the left are summed and added to the vowel target shown as horizontal dotted line. Row 1: vowel duration = 300 ms. Row 2: the same parameters are used as in row 1, but the duration is 100 ms less resulting in greater undershoot (shown as the extent by which the contour on the right falls short in frequency of the horizontal dotted line). Row 3: the same parameters as in row 2, except that the transition rates, defined by *b* in equations in (1) and (2), are faster.

less than in row 1. The transition functions are *exactly the same*, but now there is less time for the target to be attained and as a result there is greater undershoot – specifically, the vowel target is undershot by about another around 100 Hz. This is the sense of undershoot in Lindblom (1963): the parameters controlling the transitions do not change (because the force to the articulators is unchanged) and the extent of undershoot is predictable from the durational decrease.

However, studies of speech production have shown that speakers can and do increase articulatory velocity when vowel duration decreases (Kuehn & Moll, 1976; Kelso et al., 1985; Beckman et al., 1992). As far as formulae (1) and (2) are concerned, this implies that the time constants can change to speed up the transition (see also Moon & Lindblom, 1994). An example of changing the time constants and hence the rate of transition is shown in the third row of Figure 3.4: in this case, the increase in transition speed (decrease in the time constants) easily offsets the 100 ms shorter duration compared with row 1 and the target is very nearly attained.

## 2.6 F2 locus and consonant place of articulation

The idea that formant transitions provide cues to place of articulation can be traced back to Potter, Kopp, and Green (1947) and to the perception experiments carried out in the 1950s with hand-painted spectrograms using two-formant synthesis at the Haskins Laboratories (Liberman et al., 1954; Delattre et al., 1955). These perception experiments showed that place of articulation could be distinguished by making F2 point to a "locus" on the frequency axis close to the time of the stop release. The Haskins Laboratories experiments showed that /b/ and /d/ were optimally perceived with loci at 720 Hz and 1,800 Hz respectively. An acceptable /g/ could be synthesized with the F2 locus as high as 3,000 Hz before nonback vowels, but no acceptable locus could be found for /g/ before back vowels.

In the 1960s–1980s various acoustic studies (Lehiste & Peterson, 1961; Öhman, 1966; Fant, 1973; Kewley-Port, 1982) explored whether there was evidence for an F2 locus in natural speech data. In general, these studies did not support the idea of an invariant locus; they also showed the greatest convergence towards a locus frequency for /d/.

F3 transitions can also provide information about stop place and in particular for separating alveolars from velars (Öhman, 1966; Fant, 1973; Cassidy & Harrington, 1995). As the spectrographic study by Potter et al. (1947) had shown, F2 and F3 at the vowel onset seem to originate from a mid-frequency peak that is typical of velar bursts: for example, F2 and F3 are much closer together in frequency at vowel onset following a velar than an alveolar stop, as the spectrograms in Figure 3.5 show.

In the last 15 years or so, a number of studies in particular by Sussman and Colleagues (e.g., Sussman, 1994; Sussman et al., 1993, 1995; Modarresi et al., 2005) have used so-called *locus equations* as a metric for investigating the relationship between place of articulation and formant transitions. The basic form of the locus equation is given in (3) and it is derived from another observation in Lindblom (1963) that the formant values at the vowel onset ($F_{ON}$) and at the vowel target ($F_T$) are linearly related:

$$F_{ON} = \alpha F_T + c \tag{3}$$

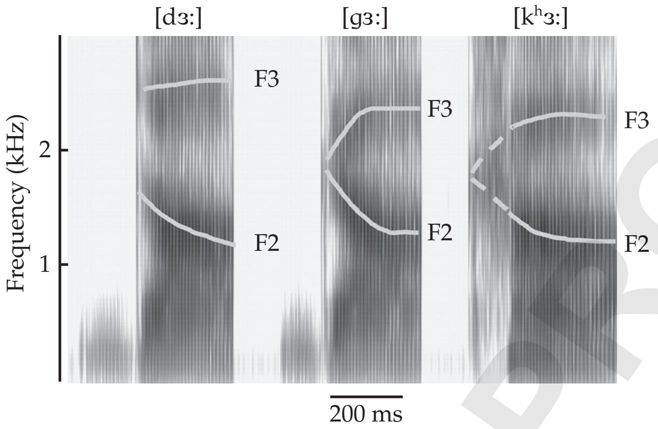

**Figure 3.5** Spectrograms, male speaker of Australian English, extracted from isolated productions of the non-word *dird* and the words *gird* and *curd* (Australian English is non-rhotic). The F2 and F3 transitions were traced by hand from the onset of periodicity in the first two words, and from the burst release in *curd*.

Krull (1989) showed that the slope, $\alpha$, could be used to measure the extent of V-on-C coarticulation. The theory behind this is as follows. The more that a consonant is influenced by a vowel, the less the formant transitions converge to a common locus and the greater the slope in the plane of vowel onset frequency by vowel target frequency. This is illustrated for two hypothetical cases of F2 transitions in the syllables [bɛ] and [bo] in Figure 3.6. On the left, the F2 transitions converge to a common locus: in this case, F2 onset is completely unaffected by the following vowel (the anticipatory V-on-C coarticulation at the vowel onset is zero). From another point of view, the vowel target could not be predicted from a knowledge of the vowel onset (since the vowel onsets are the same for [bɛ] and [bo]). On the right is the case of *maximum* coarticulation: in this case, the V-on-C coarticulation is so strong that there is no convergence to a common locus and the formant onset is the same as the formant target (i.e., the formant target is completely predictable for any known value of formant onset). In the panels on the right, these hypothetical data were plotted in the formant target by formant onset plane. The line that connects these points is the *locus equation*, and it is evident that the two cases of zero and maximal coarticulation differ in the lines' slopes which are 0 and 1 respectively.

It is possible to re-write (3) in terms of the locus frequency, $L$ (Harrington & Cassidy, 1999):

$$F_{ON} = \alpha F_T + L(1 - \alpha) \tag{4}$$

From (4), it becomes clear that when $\alpha$ is zero, $F_{ON} = L$ (i.e., the vowel onset equals the locus frequency as in Figure 3.6 left) and when $\alpha$ is 1, $F_O = F_T$ (i.e., the vowel

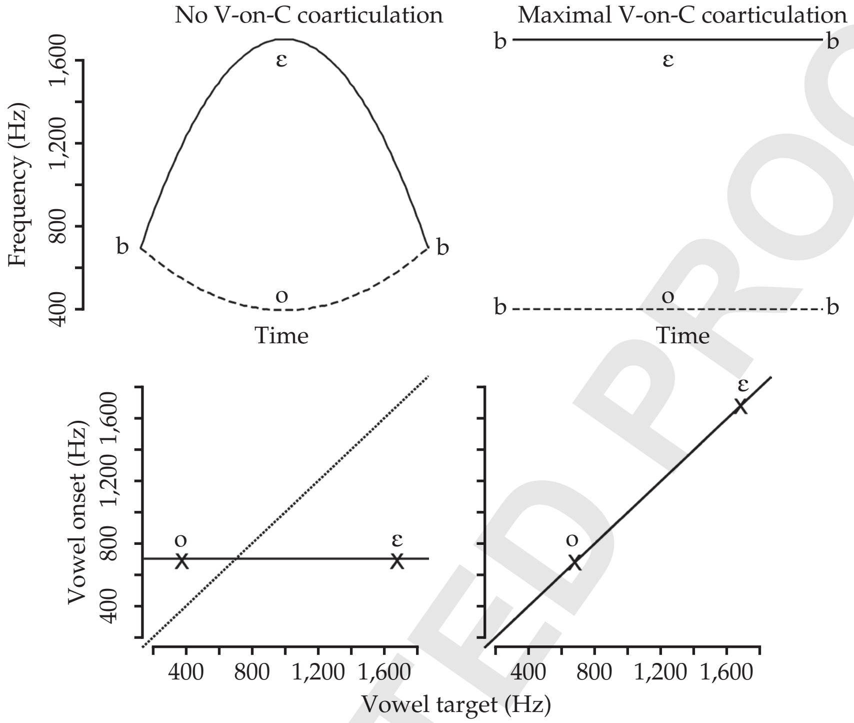

**Figure 3.6** Hypothetical F2 trajectories of [bɛb] (solid) and [bob] (dashed) when there is no V-on-C coarticulation at the vowel onset/offset (left) and when V-on-C coarticulation is maximal (right). Row 1: the trajectories as a function of time. Row 2: a plot of the F2 values in the plane of the vowel target by vowel onset for the data in the first row. The solid line is analogous to the locus equation. The locus frequency can be obtained either from equation (5) or from the point at which the locus equation intersects the dotted line, $F2_{Target} = F2_{Onset}$ (this dotted line overlaps completely with the locus equation on the right meaning that for these data, there is no locus frequency).

onset equals the vowel target as in Figure 3.6 right). More importantly, the fact that the slope varies between 0 and 1 can be used to infer the magnitude of V-on-C coarticulation. This principle is illustrated for some /dVd/ syllables produced by an Australian English male speaker in Figure 3.7.

The V in this case varied over almost all the monophthongs of Australian English and the plots in the first row are F2 as a function of time, showing the same F2 data synchronized firstly at the vowel onset on the left and at the vowel offset on the right. These plots of F2 as a function of time in row 1 of Figure 3.7 show a greater convergence to a common F2 onset frequency for initial compared with final transitions. From this it can be inferred that the size of V-on-C coarticulation is less in initial /dV/ than in final /Vd/ sequences (i.e., /d/ *resists* coarticulatory influences from the vowel to a greater extent in syllable-initial than

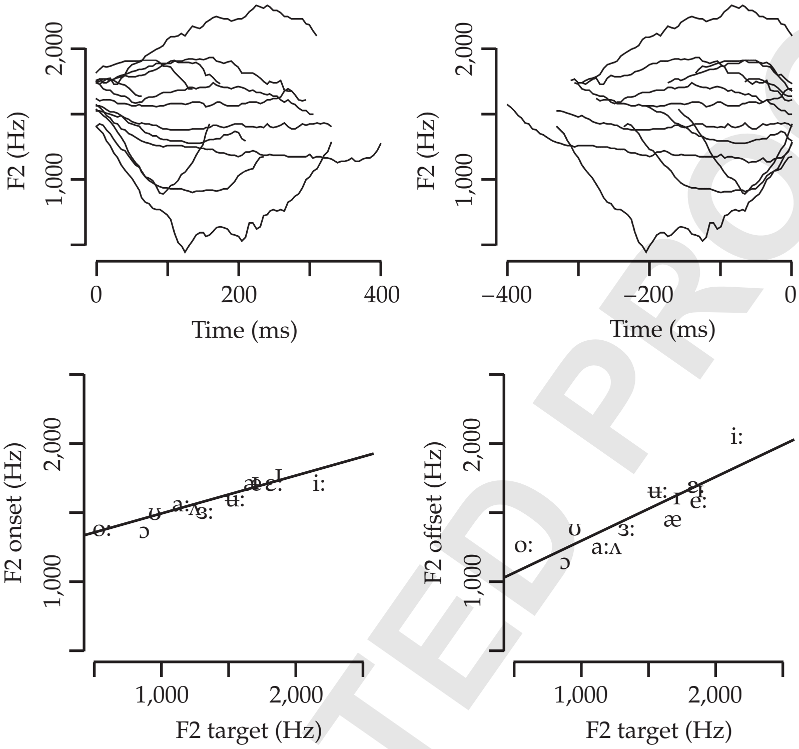

**Figure 3.7** Row 1: F2 trajectories of isolated /dVd/ syllables produced by an adult male speaker of Australian English and synchronized (*t* = 0 ms) at the vowel onset (left) and at the vowel offset (right). There is one trajectory per monophthong (*n* = 14). Row 2: corresponding locus equations with the vowel labels marked at the F2 target × F2 onset positions. The slopes and intercepts of the locus equations are respectively 0.27, 1,220 Hz (initial transitions, left) and 0.46, 829 Hz (final transitions, right).

in syllable-final position). These positional differences are consistent with various other studies showing less coarticulation for initial /d/ compared to final /d/ (Krull, 1989; Sussman et al., 1993).

In Figure 3.7 row 2, F2 at the vowel target has been plotted as a function of the F2 onset and F2 offset respectively and locus equations were calculated by drawing a straight line through each of the two scatters separately. The slope of the regression line (i.e., of the locus equation) is higher for the final /Vd/ than for the initial /dV/ transitions, which is commensurate with the interpretation in this figure that there is greater accommodation of final /d/ than initial /d/ to the vowel.

A locus equation like any straight line in an *x*–*y* plane, has, of course, both a slope and an intercept and various studies (e.g., Fowler, 1994; Sussman, 1994; Chennoukh et al., 1997) have shown how different places of articulation have different values on slopes and intercepts together (the information from both the slope and intercept together is sometimes called a *second-order locus equation*). Whereas the slope says something about the extent of V-on-C coarticulation, the intercept encodes information about the best estimate of the locus frequency weighted by the slope. From (3) and (4) it is evident that the intercept, $c$, locus frequency, $L$, and slope, $\alpha$, are related by $c = L(1 - \alpha)$. Thus the locus frequency can be estimated from the locus equation intercept and slope:

$$L = c/(1 - \alpha) \qquad (5)$$

For the initial /dV/ data (Figure 3.7, row 1, left), the intercept and slope are given by 1,220.3 Hz and 0.27 so the best estimate of the F2 locus is 1,220.3/(1 – 0.27) = 1,671 Hz which is indeed close to the frequency towards which the F2 transitions in row 1 of Figure 3.7 seem to converge.

Some of the main findings to emerge from locus equation (LE) studies in recent years are:

- The data points in the plane of F2 onset × F2 target are tightly clustered about a locus equation and the locus equation parameters (intercept, slope) differ for different places of articulation (Krull, 1989; various studies by Sussman and colleagues referred to earlier).
- Alveolars have the lowest LE slopes which, as discussed earlier, implies that they are least affected by V-on-C coarticulation (e.g., Krull, 1989). They also usually have higher intercepts than bilabials, which is to be expected given the relationship in (5) and the other extensive evidence from perception experiments and acoustic analyses that the F2 locus of alveolars is higher than that of labials.
- It is usually necessary to calculate separate locus equations for velar stops before front and back vowels (Smits et al., 1996a, 1996b) because of the considerable variation in F2 onset frequencies of velars due to the following vowel (or, if velar consonants are pooled across vowels, then they tend to have the highest slopes, as the acoustic and electropalatographic (EPG) data in Tabain, 2000 have shown).
- Subtle place differences involving the same articulator cannot easily be distinguished using LE parameters (Krull et al., 1995; Tabain & Butcher, 1999; Tabain, 2000).
- There is controversy about whether LE parameters vary across manner of articulation (Fowler, 1994) and voicing (Engstrand & Lindblom, 1997; but see Modarresi et al., 2005). For example, Sussman (1994) reports roughly similar slopes for /d, z, n/; however, in an electropalatographic analysis of CV onsets, Tabain (2000) found that LE parameters distinguished poorly within fricatives.
- As already mentioned, Krull (1989) has shown that locus equations can be very useful for analyzing the effects of speaking style: in general, spontaneous speech is likely to have lower slopes because of the greater V-on-C coarticulation than citation-form speech. However in a more recent study, van Son and

Pols (1999) found no difference in intercepts and slopes comparing read with spontaneous speech in Dutch.
- While Chennoukh et al. (1997) relate locus equations to articulatory timing using the distinctive region model (DRM) of area functions (Carré & Mrayati, 1992), none of the temporal phasing measures in VCV sequences using movement data in Löfqvist (1999) showed any support for the assumption that the LE slope serves as an index of the degree of coarticulation between the consonant and the vowel.
- While Sussman et al. (1995) have claimed that "the locus equation metric is attractive as a possible context-independent phonemic class descriptor and a logical alternative to gestural-related invariance notions", the issue concerning the auditory or cognitive status of LEs has been disputed (e.g., Brancazio & Fowler, 1998; Fowler, 1994).

Finally, and this is particularly relevant to the last point above, the claim has been made that it is possible to obtain "perfect classification accuracy (100%) for place of articulation" (Sussman et al., 1991) from LE parameters. However, it is important to recognize that LE parameters themselves are generalizations across multiple data points (Fowler, 1994; Löfqvist, 1999). Therefore, the perfect classification accuracy in distinguishing between three places of articulation is analogous to finding no overlap between three vowel categories that had been averaged by category across each speaker (as in classifying 10 [i], 10 [u], and 10 [a] points in an F1 × F2 space, where each point is an average value per speaker). Seen from this point of view, it is not that entirely surprising that 100 percent classification accuracy could be obtained, especially for citation-form speech data.

## 2.7 *Approximants*

Voiced approximants are similar in acoustic structure to vowels and diphthongs and are periodic with F1–F3 occurring in the 0–4,000 Hz spectral range. As a class, approximants can often be distinguished from vowels by their lower amplitude and from each other by the values of their formant frequencies. Figure 3.8 shows that for the sonorant-rich sentence "Where were you while we were away?" there are usually dips in two energy bands that have been proposed by Espy-Wilson (1992, 1994) for identifying approximants.

Typical characteristics for approximant consonants that have been reported in the literature (and of which some are shown in the spectrogram in Figure 3.8) are as follows:

- [w] has F1 and F2 close together and both low in frequency. The ranges reported for American English are 300–400 Hz for F1 and 600–800 Hz for F2 (e.g., Lehiste, 1964; Mack & Blumstein, 1983). [w], like labials and labial-velars, has a low F2 and this is one of the factors that contributes to sound changes involving these segments (see Ohala & Lorentz, 1977, for further details).
- [j] like [i] has a low F1 and a high F2 – see Figure 3.8.

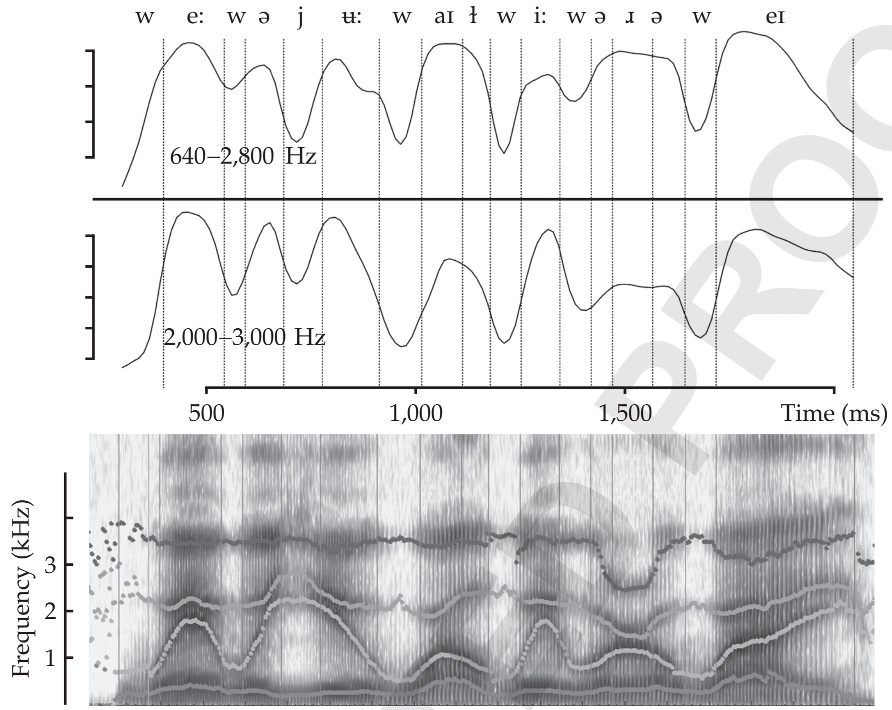

**Figure 3.8** Summed energy values in two frequency bands and the first four formant frequencies superimposed on a spectrogram of the sonorant-rich sentence "Where were you while we were away?" produced by an adult male Australian English speaker. (Adapted from Harrington & Cassidy, 1999)

- American English /r/ and the post-alveolar approximant that is typical in Southern British English have a low F3 typically in the 1,300–1,800 Hz range (Lehiste, 1964; Nolan, 1983), which is likely to be a front cavity resonance (Fant, 1960; Stevens, 1998; Espy-Wilson et al., 2000; Hashi et al., 2003).
- /l/ when realized as a so-called clear [l] in syllable-initial position in many English varieties has F1 in the 250–400 Hz range and a variable F2 that is strongly influenced by the following vowel (Nolan, 1983). F3 in "clear" realizations of /l/ may be completely canceled by an anti-resonance due to the shunting effects of the mouth cavity behind the tongue blade. The so-called dark velarized /l/ that occurs in syllable-final position in many English varieties has quite a different formant structure which, because it is produced with velarization and raising of the back of the tongue, resembles a high back round vowel in many respects: in this case, F2 can be as low as 600–900 Hz (Lehiste, 1964; see also the final /l/ in *while* in Figure 3.8). Bladon and Al-Bamerni (1976) showed that /l/ varies in clarity depending on various prosodic factors, including syllable position; and also that dark realizations of /l/ were

much less prone to coarticulatory influences from adjacent vowels compared with clear /l/.

- Compared with the other approximants, American English /l/ is reported as having longer and faster transition (Polka & Strange, 1985).
- /l/ sometimes has a greater spectral discontinuity with a following vowel that is caused by the complete alveolar closure: that is, there is often an abrupt F1-transition from an /l/ to a following vowel which is not in evidence for the other three approximants (O'Connor et al., 1957).
- In American English, [w] can sometimes be distinguished from [b] because of its slower transition rate into a following vowel (e.g., Mack & Blumstein, 1983).

## 3 Obstruents

Fricatives are produced with a turbulent airstream that is the result of a jet of air being channelled at high speed through a narrow constriction and hitting an obstacle (see Shadle, this volume). For [s] and [ʃ] the obstacles are the upper and lower teeth respectively; for [f] the obstacle is the upper lip and for [x] it is the wall of the vocal tract (Johnson, 2004). The acoustic consequence of the turbulent airstream is aperiodic energy. In Figure 3.9, the distinction between the fricatives and sonorants in the utterance "Is this seesaw safe?" can be seen quite easily from the aperiodic energy in fricatives that is typically above 1,000 Hz. Fricatives are produced with a noise source that is located at or near the place of maximum constriction and their spectral shape is strongly determined by the length of the cavity in front of the constriction – the back cavity makes scarcely any contribution to the spectrum since the coupling between the front and back cavities is weak (Stevens, 1989). Since [s] has a shorter front cavity than [ʃ], and also because [ʃ] but not [s] has a sublingual cavity which effectively lengthens the front cavity (Johnson, 2004), the spectral energy tends to be concentrated at a higher frequency for [s]. Since the length of the front cavity is negligible in [f, θ], their spectra are "diffuse," i.e., there are no major resonances and their overall energy is usually low. In addition, the sibilants [s, ʃ] have more energy at higher frequencies than [f, θ] not just because of the front cavity differences, but also because in the sibilants the airstream hits the teeth producing high-frequency turbulence (Stevens, 1971).

Voiced fricatives are produced with a simultaneous noise and voice sources. In the same spectrogram in Figure 3.9, there is both aperiodic energy in [z̥ð̥] of *is this* above 6,000 Hz and evidence of periodicity, as shown by the weak energy below roughly 500 Hz. The energy due to vocal fold vibration is often weak both in unstressed syllables such as these and more generally in voiced fricatives: this is because the high intraoral air pressure that is required for turbulence tends to cancel the subglottal pressure difference that is necessary to sustain vocal fold vibration. There is sometimes a noticeable continuity in the noise of fricatives with vowel formants (Soli, 1981). This is also apparent in Figure 3.9 as shown

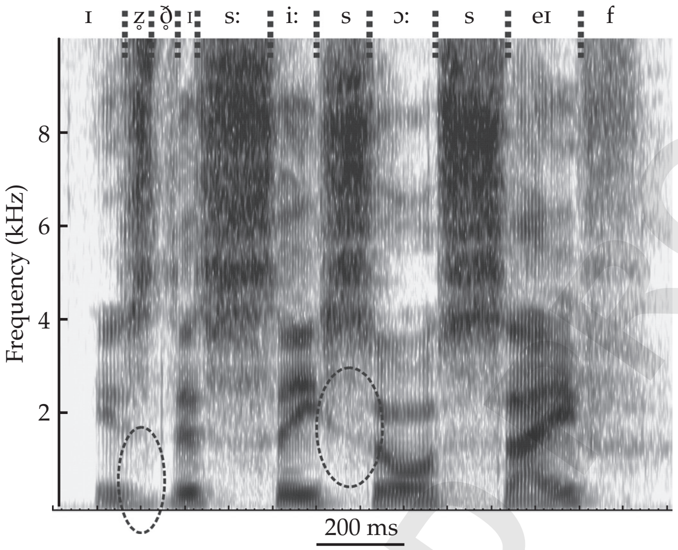

**Figure 3.9** Spectrogram of the sentence "Is this seesaw safe?" produced by an adult male speaker of Australian English. There is evidence of weak periodicity in the devoiced [z̥ð̥] at the boundary of *is this* (ellipse, left) and of an F2 transition in the noise of the second [s] of *seesaw* (ellipse, right). (Adapted from Harrington & Cassidy, 1999)

by the falling F2 transition across the noise in [iso] of *seesaw*. Fricatives especially [s, ʃ] are perceptually salient and they can mask a preceding nasal in vowel-nasal-fricative sequences: Ohala and Busà (1995) reason that this is one of the main factors that contributes to the common loss of nasals before fricatives diachronically (e.g., German *fünf*, but English *five*).

An oral stop is produced with a closure followed by a release which includes *transient*, *frication*, and sometimes *aspiration* stages (Repp & Lin, 1989; Fant, 1973). The transient corresponds to the moment of release and it shows up on a spectrogram as a vertical spike. The acoustics of the frication at stop release are very similar to the corresponding fricative produced at the same place of articulation. Aspiration, if it is present in the release of stops, is the result of a noise source at the glottis that may produce energy below 1 kHz (Figure 3.10). In the acoustic analysis of stops, the *burst* is usually taken to include a section of the oral stop extending for around 20 ms from the transient into the frication and possibly aspiration phases.

## *3.1 Place of articulation: Spectral shape*

From considerations of the acoustic theory of speech production (Fant, 1960; Stevens, 1998), there are place-dependent differences in the spectral shape of stop bursts. Moreover, perception experiments have shown that the burst carries

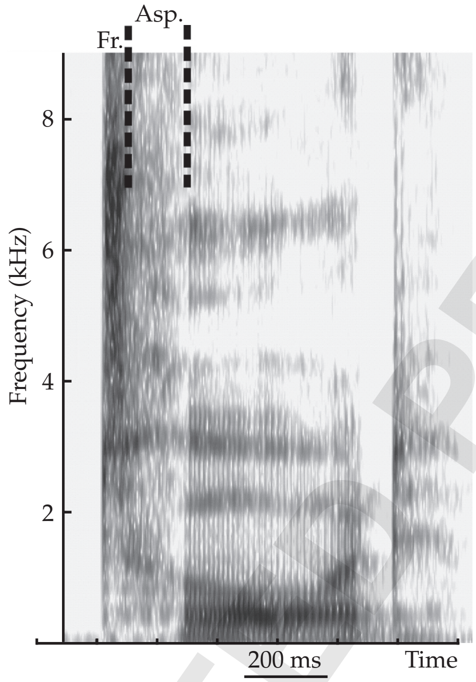

**Figure 3.10** Spectrogram of an isolated production of the nonword [$\text{t}^{\text{h}}$ɔːd] (*tawed*) by a male speaker of Australian English showing the fricated and aspiration stages of the stop.

cues to stop place of articulation (Smits et al., 1996a; Fischer-Jørgensen, 1972). As studies by Blumstein and Stevens (1979, 1980) have shown, labial and alveolar spectra can often be distinguished from each other based on the slope of the spectrum which tends to fall for bilabials, but to rise with increasing frequency above roughly 3,000 Hz for alveolars. The separation of velars from other stops can be more problematic, partly because the vowel-dependent place of articulation variation in velars (fronted before front vowels and backed before back vowels) has such a marked effect on the spectrum. But a prediction from acoustic theory is that velars should have a mid-frequency spectral peak, i.e., a concentration of energy roughly in the 2,000–4,000 Hz range, whereas for the other two places of articulation, energy is more distributed over these frequencies compared with velars. This mid-frequency peak may well be the main factor that distinguishes velar from alveolar bursts before front vowels. Winitz et al. (1972) have shown that velar bursts are often misheard as alveolar before front vowels and this, as well a perceptual reinterpretation of the following aspiration, may be responsible for the diachronic change from /k/ to /tʃ/ in many languages (Chang et al., 2001).

A number of researchers have emphasied that burst cues to place of articulation may not depend on "static" information at a single spectral slice, but instead on the shape of the spectrum as it unfolds in time during the stop release and into the following vowel (e.g., Kewley-Port et al., 1983; Lahiri et al., 1984; Nossair & Zahorian, 1991). Since the burst spectrum of [b] falls with increasing frequency and since vowel spectra also fall with increasing frequency due to the falling glottal spectrum, then the change in spectral slope for [bV] from the burst to the vowel is in general small (Lahiri et al., 1984). As far as velar stops are concerned, these are sometimes distinguished from [b, d] by the presence of mid-frequency peaks that persist between the burst and the vowel onset (Kewley-Port et al., 1983).

Figure 3.11 shows spectra for Australian English [p$^{h}$a, t$^{h}$a, k$^{h}$a] between the burst and vowel onset as a function of normalized time. The displays are averages across five male Australian English speakers and are taken from syllable-initial stressed stops in read speech. The spectral displays were linearly time-normalized prior to averaging so that time point 0.5 is the temporal midpoint between the burst onset and the vowel's periodic onset. One again, the falling, rising, and compact characteristics at the burst are visible for the labial, alveolar, and velar places of articulation respectively. The falling slope is maintained more or less into the vowel for [p$^{h}$a:], whereas for [t$^{h}$a:] the rising spectral slope that is evident at burst onset gives way to a falling slope towards the vowel onset producing a substantial change in energy in roughly the 3–5 kHz range. The same figure shows that the mid-frequency peak visible for [k$^{h}$a] as a concentration of energy at around 2.5 kHz at the burst onset persists through to the onset of the vowel (normalized time point 0.8).

The overall shape of the spectrum has been parameterized with *spectral moments* (e.g., Forrest et al., 1988) which are derived from statistical moments that are sometimes applied to the analysis of the shape of a histogram. Where $x$ is a histogram class interval and $f$ is the count of the number of tokens in a class interval, the $i$th statistical moment, $m_i$, ($i$ = 1, 2, 3, 4) can be calculated as follows:

$$m_1 = \frac{\sum fx}{\sum f} \tag{6}$$

$$m_2 = \frac{\sum f(x-m_1)^2}{\sum f} \tag{7}$$

$$m_3 = \left(\frac{\sum f(x-m_1)^3}{\sum f}\right) m_2^{-1.5} \tag{8}$$

$$m_4 = \left[\left(\frac{\sum f(x-m_1)^4}{\sum f}\right) m_2^{-2}\right] - 3 \tag{9}$$

In *spectral* moments, a spectrum is treated as if it were a histogram so that $x$ becomes the intervals of frequency and $f$ is the dB value at a given frequency. If the

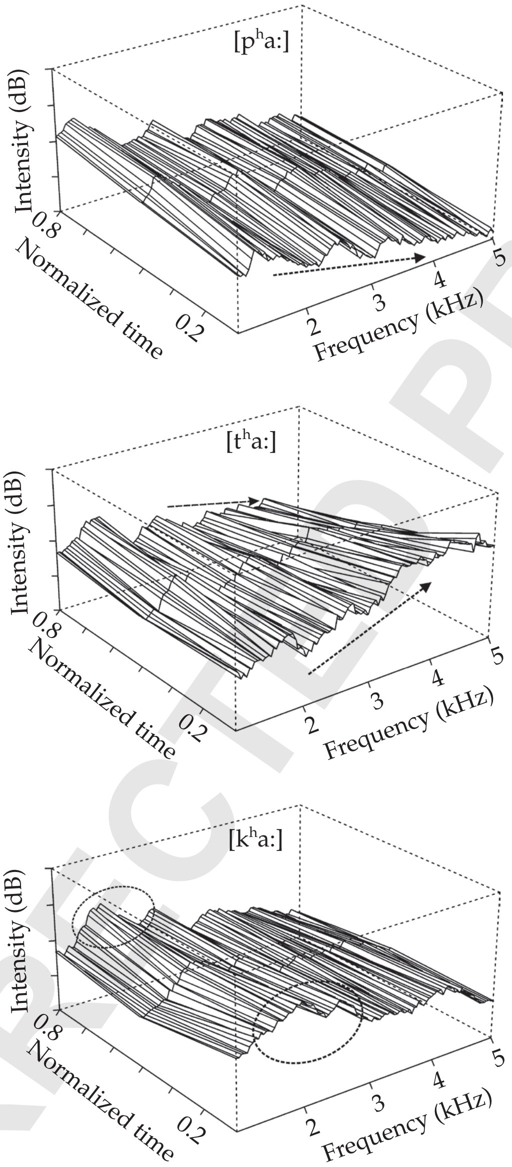

**Figure 3.11** Spectra as a function of normalized time extending from the burst onset (time 0) to the acoustic onset of the vowel (time 0.8) for syllable-initial, stressed bilabial, alveolar, and velar stops preceding [a:] averaged across five male speakers of Australian English. The stops were taken from both isolated words and from read speech and there were roughly 100 tokens per category. The arrows mark the falling and rising slopes of the spectra at burst onset in [p$^{h}$a] and [t$^{h}$a] (and the arrow at time point 0.8 in [t$^{h}$a] marks the falling spectral slope at vowel onset). The ellipses show the mid-frequency peaks that persist in time in [k$^{h}$a]. (Adapted from Harrington & Cassidy, 1999)

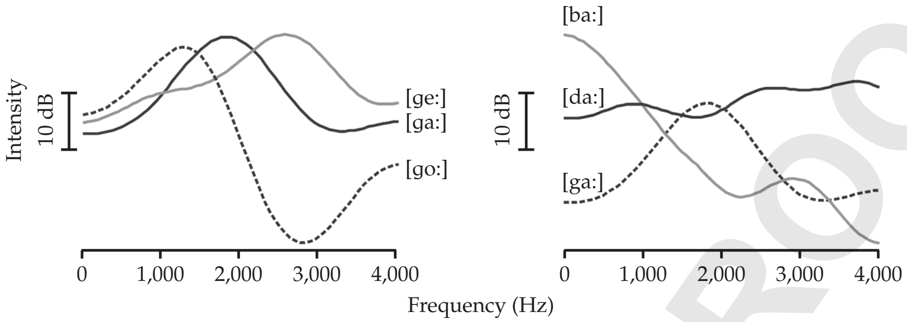

**Figure 3.12** Cepstrally smoothed spectra calculated with a 16 ms window centered at the burst onset in word-initial [b, d, g] stops taken from isolated words produced by an adult male speaker of German. Left: spectra of [ge:, ga:, go:] bursts. Their $m_1$ (spectral center of gravity values) are 2,312 Hz, 1,863 Hz, and 1,429 Hz respectively. Right: spectra of the bursts of [ba:, da:, ga:]. Their $\sqrt{m_2}$ (spectral standard deviation) values are 1,007 Hz, 977 Hz, and 655 Hz respectively.

frequency axis is in Hz, then the units of $m_1$ and $m_2$ are Hz and Hz$^2$ respectively, while the third and fourth moments are dimensionless. It is usual in calculating moments to exclude the DC offset (frequency at 0 Hz) and to rescale the dB values so that the minimum dB value in the spectrum is set to 0 dB.

The first spectral moment $m_1$, gives the frequency at which the spectral energy is predominantly concentrated. Figure 3.12 shows cepstrally smoothed spectra calculated at the burst onset in stop-initial words produced in German. The figure in the left panel shows how $m_1$ decreases across [ge:, ga:, go:], commensurate with the progressive decrease in the frequency location of the energy peak in the spectrum that shifts due to the coarticulatory influence of the backness of the following vowel.

The second spectral moment, $m_2$, or its square root, the *spectral standard deviation*, is a measure of how distributed the energy is along the frequency axis. Thus in the right panel of Figure 3.12, $m_2$ is higher for [ba:, da:] than for [ga:] because, as discussed above, the spectra of the former are relatively more diffuse whereas [g] spectra tend to be more compact with energy concentrated around a particular frequency.

$m_3$ is a measure of asymmetry (see Figure 3.3 where this parameter was applied to F2 of [i:]). Given that spectra are always band-limited, the third spectral moment would seem to be necessarily correlated with $m_1$ (see for example the data in Jongman et al., 2000, table I): that is, $m_3$ is positive or negative if the energy is predominantly concentrated in low- and high-frequency ranges respectively. Finally $m_4$, kurtosis, is an expression of the extent to which the spectral energy is concentrated in a peak relative to the energy distribution in low and high frequencies. In general, $m_4$ is often correlated with $m_2$, although this need not be so (see, e.g., Wuensch, 2006 for some good examples).

Fricative place has been quantified with spectral moments in various studies (e.g., Forrest et al., 1988; Jongman et al., 2000; Tabain, 2001). Across these studies, two of the most important findings to emerge are:

- [s, z] have higher $m_1$ values than [ʃ, ʒ]. This is to be expected given the predictions from articulatory-to-acoustic mapping that the center frequency of the noise is higher for the former. When listeners label tokens from a synthetic /s–ʃ/ continuum, there is a greater probability that the same token is identified as /s/ before rounded compared with unrounded vowels (Mann & Repp, 1980). This comes about firstly because a lowered $m_1$ is a cue both for /ʃ/ and the result of anticipatory lip-rounding caused by rounded vowels; secondly, because listeners compensate for the effects of coarticulation, i.e., they factor out the proportion of $m_1$ lowering that is attributable to the effects of lip-rounding and so bias their responses towards /s/ when tokens are presented before rounded vowels.
- The second spectral moment tends to be higher for nonsibilants than sibilants, which is again predictable given their greater spectral diffuseness (e.g., Shadle & Mair, 1996).

Another way of parameterizing the shape of a spectrum is with the DCT (Nossair & Zahorian, 1991; Watson & Harrington, 1999). This transformation decomposes a signal into a set of half-cycle frequency cosine waves which, if summed, reconstruct the signal to which the DCT was applied. The amplitudes of these cosine waves are the *DCT coefficients* and when the DCT is applied to a spectrum, the DCT coefficients are equivalently *cepstral coefficients* (Nossair & Zahorian, 1991; Milner & Shao, 2006). For an $N$-point signal $x(n)$ extending in time from $n = 0$ to $N - 1$ points, the $m^{\text{th}}$ DCT coefficient, $C_m$, ($m = 0, 1, \ldots N - 1$) can be calculated with:

$$C_m = \frac{2k_m}{N} \sum_{n=0}^{N-1} x(n) \cos\left( \frac{(2n+1)m\pi}{2N} \right) \tag{10}$$

$$k_m = \frac{1}{\sqrt{2}}, \ m = 0; \ k_m = 1, \ m \neq 0$$

It can be shown that the first three DCT coefficients ($C_0$, $C_1$, $C_2$) are proportional to the *mean*, *linear slope*, and *curvature* of the signal respectively (Watson & Harington, 1999).

Figure 3.13 shows some spectral data of three German dorsal fricatives [ç, x, ʃ] taken from 100 read sentences of the Kiel corpus of read speech produced by a male speaker of Standard North German. The spectra were calculated at the fricatives' temporal midpoint with a 256-point discrete Fourier transform (DFT) at a sampling frequency of 16,000 Hz and the frequency axis was transformed to the Bark sale. DCT coefficients were calculated on these Bark spectra over the 500–7,500 Hz range. The fricatives were extracted irrespective of the segmental or prosodic contexts in which they occurred.

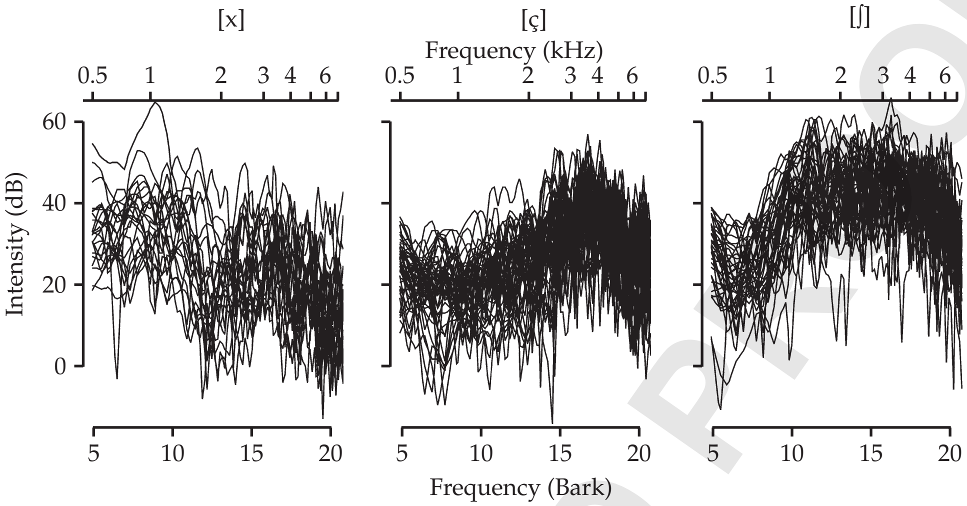

**Figure 3.13** Spectra in the 0–8 kHz range calculated with a 16 ms DFT at the temporal midpoint of the German fricatives [x] (left, *n* = 25), [ç] (center, *n* = 50), and [ʃ] (right, *n* = 39) and plotted with the frequency axis proportional to the Bark scale. The data are from read sentences produced by one male speaker of Standard German.

As is well known, [ç] and [x] are allophones of one phoneme in German that are predictable from the frontness of the preceding vowel, but they also have very different spectral characteristics. As discussed in Johnson (2004), the energy in back fricatives like [x] tracks F2 of the following vowel, whereas in palatal fricatives like [ç], the energy is concentrated at a higher frequency and is continuous with the flanking vowel's F3. As Figure 3.13 shows, the palatal [ç] patterns more closely with [ʃ] because [x] has a predominantly falling spectrum whereas the spectra of [ç] and [ʃ], which show a concentration of energy in the 2–5 kHz range, are rising. The distinction between [ʃ] and [ç] could be based on curvature: [ʃ], there is a greater concentration of energy around 2–3 kHz so that the [ʃ] spectra have a greater resemblance to an inverted U-shape than those of [ç].

Figure 3.14 shows the distribution of the same spectra on the DCT coefficients $C_1$ and $C_2$. Compatibly with these predictions from Figure 3.13, [x] is separated from the other fricatives primarily on $C_1$ (spectral slope) whereas the [ʃ]–[ç] distinction depends on $C_2$ (spectral curvature). Thus together $C_1$ and $C_2$ provide quite an effective separation between these three dorsal fricative classes, at least for this single speaker.

## 3.2 Place of articulation in obstruents: Other cues

Beyond these considerations of gross spectral shape discussed in the preceding section and F2 locus cues in formant transitions discussed in 2.6, place of articulation within obstruents is cued by various other acoustic attributes, in particular:

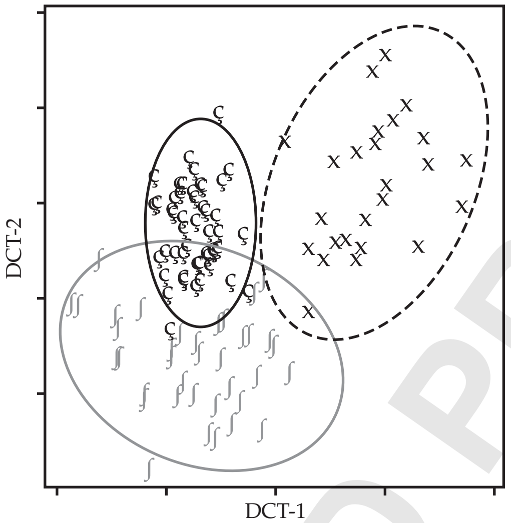

**Figure 3.14** 95 percent confidence ellipses for three fricatives in the plane of DCT-1 and DCT-2 obtained by applying a DCT to the Bark-scaled spectra in Figure 3.13.

- The bursts of labials tend to be weak in energy (Fischer-Jørgensen, 1954; Fant, 1973) since they lack a front cavity and perceptual studies have shown that this energy difference in the burst can be used by listeners for distinguishing labials from alveolars (e.g., Ohde & Stevens, 1983). The overall intensity of the burst relative to that of the vowel has also been used by Jongman et al. (1985) for place of articulation distinctions in voiceless coronal stops produced by three adult male talkers of Malayalam.
- The duration of the stop release up to the periodic onset of the vowel in CV syllables, i.e., voice onset time (VOT), can also provide information about the stop's place of articulation: in carefully controlled citation-form stimuli, within either voicing category, velar stops have longer VOTs than alveolar stops, whose VOTs are longer than those of bilabial stops (e.g., Kewley-Port, 1982).
- The *amplitude of the frication noise* has been shown to distinguish perceptually the sibilant fricatives [s, ʃ] from nonsibilants like [f, θ] (Heinz & Stevens, 1961). Ali et al. (2001) found an asymmetry in perception such that decreasing the amplitude of sibilants leads them to be perceived as nonsibilants (whereas increasing the amplitude of nonsibilants does not cause them to be perceived as sibilants).
- Studies by Harris (1958) and Heinz and Stevens (1961) showed that, whereas the noise carried more information for place distinctions than formant transitions, F2 and F3 may be important in distinguishing [f] from [θ] given that labiodentals and dentals have very similar noise spectra (see Tabain, 1998 for an analysis of spectral information above 10 kHz for the labiodental/dental

fricative distinction). More recently, Nittrouer (2002) found that in comparison with children, adults tended to be more reliant on noise cues than formant transition cues in distinguishing [f] from [θ]. F2 transitions in noise have been shown to be relevant for distinguishing [s] from [ʃ] acoustically and perceptually (Soli, 1981).

## 3.3 *Obstruent voicing*

VOT is the duration from the stop release to the acoustic periodic onset of the vowel and it is perhaps the most salient acoustic cue for distinguishing domain-initial voiced from voiceless stops in English and in many languages (Lisker & Abramson, 1964, 1967). If voicing begins during the closure (as in the example in Figure 3.5), then VOT is negative. The duration of the noise in fricatives is analogous to VOT in stops and it has been shown to be an important cue for the voicing distinction within syllable-initial fricatives (Cole & Cooper, 1975) although noise duration is not always consistently less in voiced than in voiceless fricatives (Jongman, 1989).

VOT differences can be related to differences in the onset frequency and transition of F1. When the vocal tract is completely occluded, F1 is at its theoretically lowest value. Then, with the release of the stop, F1 rises (Stevens & House, 1956; Fant, 1960). The F1-transition rises in both voiced and voiceless CV stops, but since periodiocity starts much earlier in voiced stops (in languages that use VOT for the voicing distinction), much more of the transition is periodic and the onset of voiced F1 is often considerably lower (Fischer-Jørgensen, 1954).

In a set of synthesis and perception experiments, Liberman et al. (1958) showed that delaying the onset of F1 relative to the burst and to F2 and F3 was a primary cue for the voiced/voiceless distinction (see also Darwin & Seton, 1983). Subsequent experiments in speech perception have shown that a rising periodic F1-transition (e.g., Stevens & Klatt, 1974) and a lower F1-onset frequency (e.g, Lisker, 1975) cue voiced stops and that there may be a trading relationship between VOT and F1-onset frequency (Summerfield & Haggard, 1977). Thus as is evident in comparing [k$^h$] with [g] in Figure 3.5, both F2 and F3 converge back towards a common onset frequency near the burst, but the first part of these transitions are aperiodic in the voiceless stop. Also, although F1 rises in both cases, the rising part of the transition is aperiodic in [k$^h$] resulting in a higher F1-onset frequency at the beginning of the voiced vowel.

In many languages, voiceless stops are produced with greater articulatory force and as a result the burst amplitude (Lisker & Abramson, 1964) and the rate at which the energy increases is sometimes greater in voiceless stops (Slis & Cohen, 1969). In various perception experiments, Repp (1979) showed that increasing the amplitude of aspiration relative to that of the following vowel led to greater voiceless stop percepts. The comparison of burst amplitude across stop voicing categories is one example in which *first-differencing* the signal can be important. When a signal is differenced, i.e., samples at time points $n$ and $n - 1$ are subtracted from each other, there is just under a 6 dB rise per octave or doubling of frequency

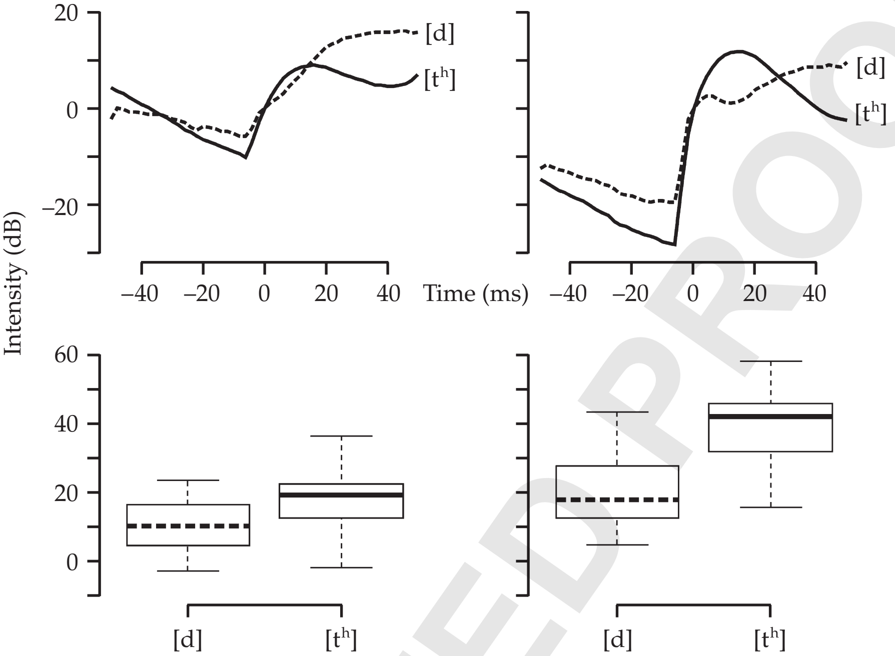

**Figure 3.15** Row 1: averaged dB-RMS trajectories of [d] ($n$ = 22) and [t$^h$] (n = 69) calculated with a 10 ms rectangular window on sampled speech data without (left) and with (right) first-differencing. 0 ms marks the burst onset. The averaging was done after rescaling the amplitude of each token relative to 0 dB at the burst onset. The stops are from two male speakers of Australian English and were extracted from prevocalic stressed syllable-initial position from 100 read sentences per speaker irrespective of vowel context. Row 2: boxplots showing the corresponding distribution of [d, t$^h$] on the parameter $b$–$a$, where $b$ and $a$ are respectively the dB values 10 ms after, and 10 ms before the burst onset. (The height of the rectangle marks the interquartile range).

in the spectrum, so that the energy at high frequencies is boosted (see Ellis, this volume). Given that at stop release there may well be greater energy in the upper part of the spectrum in voiceless stops, the effect of first-differencing is likely to magnify any energy differences across voiced and voiceless stops. In Figure 3.15, the root-mean-square (RMS) energy has been calculated in voiced and voiceless stops: in the left panels, there was no differencing of the sampled speech data, whereas in the right panels the speech waveform was first differenced before the RMS energy calculation was applied. As the boxplots show, there is only a negligible difference in burst amplitude across the voicing categories on the left; but with the application of first differencing, the rise in amplitude of the stop burst is much steeper and the difference in energy 10 ms before and after the release is noticeably greater in the voiceless stop.

The fundamental frequency is higher after voiceless than voiced obstruents (House & Fairbanks, 1953; Lehiste and Peterson, 1961; Hombert et al., 1979) and this has been shown to be a relevant cue for the voicing distinction both in stops (e.g., Whalen et al., 1993) and in fricatives (Massaro & Cohen, 1976). Löfqvist et al. (1989) have shown that these voicing-dependent differences in $f_0$ are the result of increased longitudinal tension in the vocal folds (but see Hombert et al., 1979, for an aerodynamic interpretation).

Several studies have concerned themselves with the acoustic and perceptual cues that underlie final (e.g., *duck*/*dug*) and intervocalic (*rapid*/*rabid*) voicing distinction. Denes (1955) showed that the distinction between /ju:s/ (*use*, noun) and /ju:z/ (*use*, verb) was based primarily on the vowel duration acoustically and perceptually. The acoustic cues that signal the final voicing in pairs have also been shown to include the F1-offset frequency and rate of F1-offset transition (e.g., Wardrip-Fruin, 1982).

Lisker (1978, 1986) showed that voicing during the closure is one of the main cues distinguishing *rapid* and *rabid* in English. Kohler (1979) demonstrated that the cues for the same phonological contrast have different perceptual rankings in different languages. He showed that, whereas voicing during the closure is a more salient cue than vowel : consonant duration ratios in French, it is the other way round in German. Another important variable in the post-vocalic voicing distinction in German can be the drop of the fundamental frequency contour in the vowel which is greater preceding voiced stops (Kohler, 1985).

## 4 Nasal Consonants and Nasalized Vowels

Nasal consonants are detectable on spectrograms by the presence of a *nasal murmur* corresponding to the phase of nasal consonant production in which the oral tract is closed and air passes through the nasal cavity. The overall amplitude of the nasal murmur is low and the energy is concentrated predominantly in a low frequency range. The beginning and end of the nasal murmur can often be quite easily detected by abrupt spectral discontinuities that are associated with the combined lowering/raising of the velum and closing/opening of the oral tract at the onset/offset of the nasal consonant. Such discontinuities are considered by Stevens (1985, 2002) to carry some of the main cues to the place of articulation in nasal consonants – this point is discussed again more fully below. In English and in many languages, this abruptness, i.e., syntagmatic distinction between the vowel and nasal, is a good deal more marked in syllable-initial nasal-vowel than in syllable-final vowel-nasal transitions (e.g., Repp & Svastikula, 1988; Redford and Diehl, 1999). These syllable-position differences can, in turn, be related to studies of sound change showing a greater propensity for the vowel and nasal to merge when the nasal is syllable-final (e.g., Hajek, 1997).

The spectrum of the nasal murmur is characterized by a set of nasal formants (N1, N2, . . .) that are the result of excitation of the combined nasal-pharyngeal tube. N1 has been calculated from vocal tract models to occur in the 300–400 Hz region and higher nasal formants occur for an adult male tract at intervals of about 800 Hz (Fant, 1960; Flanagan, 1972). Various studies also concur that that nasal formant bandwidths are broad and that N1 is high in amplitude (e.g., Fujimura, 1962).

In addition, the oral cavity acts as a side-branching resonator to the main nasal-pharyngeal tube and this results in *oral anti-formants* that absorb energy from the main nasal-pharyngeal tube. The presence of anti-formants is one of the reasons why the overall amplitude of nasal consonants is low. (Another is that, because the mouth cavity is sealed, the amount of acoustic energy leaving the vocal tract is much less than for vowels.) The spectral effect of introducing an anti-formant is both to produce a spectral dip at the anti-formant frequency and to alter the spectral balance or spectral tilt (Atal, 1985) and it is this change of spectral balance that may be as important a cue for nasalization as the frequency at which the anti-formant occurs.

The center frequency of the first oral anti-formant (FZ1) in nasal consonants is predicted to vary inversely with the length of the mouth cavity and is lowest for [m], higher for [n], highest for [ŋ] (Fant, 1960; Fujimura, 1962). A uvular nasal [N] has no anti-formants since, as the tongue constriction is so far back in the mouth, there is no oral side-branching resonator. Since FZ1 for [n] tends to coincide with N3 in roughly the 1,800 Hz range and since FZ1 for [m] occurs at a lower frequency, [n] nasal murmurs are expected to have less energy in the 1,500–2,000 Hz range than those of [m]. These FZ1-dependent spectral differences between [m] and [n] were incorporated into a metric for distinguishing between these two places of articulation in Kurowski and Blumstein (1987).

The F2 locus theory should also be applicable to nasal consonants and Liberman et al. (1954) were the first to show place of articulation in nasal consonants could be cued by formant transitions that pointed to different locus frequencies; more recently, locus equations have been applied to place of articulation distinctions in nasal consonants (Sussman, 1994). There have been several studies in which nasal murmurs and transitions have been cross-spliced, in which for example an [m]-murmur is combined with transitions appropriate for [n] (see Recasens, 1983 for a review). One of the first of these was by Malécot (1956), who showed that listeners' judgments were predominantly guided by the transitions and not the murmur. On the other hand, Kurowski and Blumstein (1984) found that nasal place was more accurately identified from a section of the waveform spanning the murmur–vowel boundary than from waveforms containing either only the murmur or only the vowel transition. This finding is consistent with studies in speech perception (Repp, 1986, 1987; Repp & Svastikula, 1988; Ohde et al., 2006) and with various acoustic studies (Kurowski & Blumstein, 1987; Seitz et al., 1990; Harrington, 1994) showing that the salient cues to nasal place of articulation are at the murmur–vowel boundary.

The spectrograms in Figure 3.16 of five phonemically contrastive nasal consonants in the Central Australian language Warlpiri recorded from one female speaker by Andrew Butcher in 2005 show evidence of differences in both the murmur and the transitions. In particular:

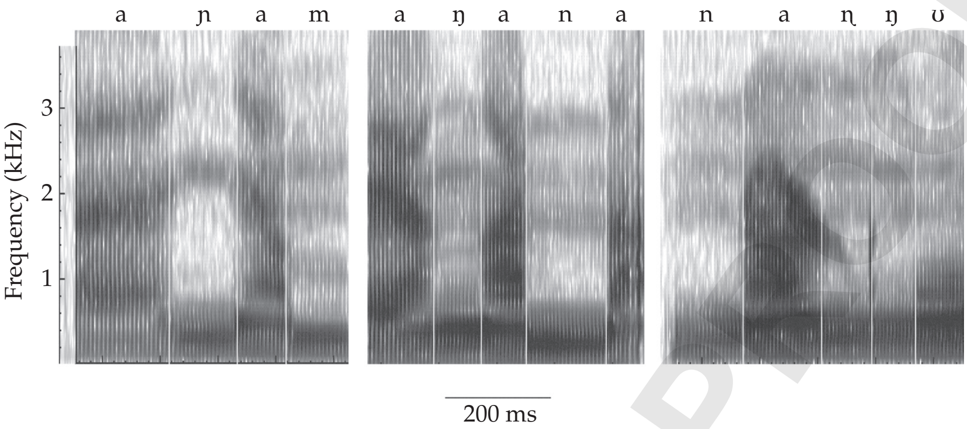

**Figure 3.16.** Spectrograms of /a#ɲampu/ (left), /a#ŋana/ (center), /naɳŋʊ/ (right) produced by a female speaker of the Central Australian language Warlpiri (# is a word boundary).

- Compatibly with some of the studies reviewed above, [n] lacks very much energy in the 1–1.5 kHz range because of the presence of an anti-formant in this frequency region.
- The lack of energy is also evident for [ɲ] but it occurs over a wider range from 1–2 kHz possibly because, since the mouth cavity is shorter for the palatal nasal, NZ1 for [ɲ] is at a higher frequency than for [n]. Also, [ɲ] has an intense formant at around 2,200 Hz that is reminiscent of F2 of the palatal vowels [i] or [ɪ].
- [ŋ] has quite a flat spectrum up to 3,000 Hz, i.e., no marked dips or peaks. The absence of any marked energy dips is expected given that the lowest anti-formant is predicted to occur above 3,000 Hz (Fant, 1960).
- It is evident that some of the distinguishing characteristics between these nasals are in the formant transitions. For example, F2 rises in the vowel of [aɲ], F2 falls in the vowel of [aŋ], and there is a steeply falling F2 (or possibly F3) in the vowel of [aɳ].

In the acoustic analysis of nasal consonants, some researchers have tried to parameterize place differences using formant variables (e.g., Chen, 1995, 1997). Although such an approach has the advantage of linking acoustic structure to vocal tract activity, it is, in practice, extremely difficult to identify with any certainty both whether a particular resonance is a nasal or an oral formant (and if so which formant number) and also whether a dip that can be seen in a spectrogram or spectral slice really is due to an anti-formant or else to a trough between formant peaks. Then there is the added complexity that vowels adjacent to nasal consonants are often nasalized which again makes the identification of vowel oral formant frequencies problematic. For this reason, a whole-spectrum approach is often favored in the acoustic (e.g., Qi & Fox, 1992) and perceptual (e.g., Kataoka et al., 2001) analysis of nasal place of articulation, which does nevertheless often allow relationships to be established to formant frequencies. For example, Kurowski and Blumstein (1987) reasoned that the energy change from the murmur into the following vowel should be greater for [n] than [m] in the 11–14 Bark (approx. 1,450–2,300 Hz) range because this is the frequency region in which [n], but not [m] has an anti-formant (see also Qi & Fox, 1992 for compatible evidence).

Energy change at the murmur–vowel boundary can be parameterized with a *difference spectrum* (i.e., by subtracting a spectrum close to the vowel onset from a spectrum close to the murmur offset) and both Kurowski and Blumstein (1987) and Seitz et al. (1990) showed high classification scores for the [m–n] distinction using metrics based on difference spectra. The analysis in Harrington (1994) also included information from the murmur and from the vowel, but the classification was based on *combined*, rather than differenced, spectral information across the murmur–nasal boundary. The idea that the combination of separately processed murmur and vowel spectra provides salient cues to place of articulation in nasals has also been shown in perception experiments of adult and child speech (Ohde et al., 2006).

Nasal vowels can occur phonetically due to the effects of context and in many languages they contrast phonemically with oral vowels. There is a correlation between phonetic vowel height and velum height: when high vowels are nasalized, the velum is not lowered to the same extent as when low vowels are nasalized. The reasons for this may be either physiologically determined by a muscular connection between the velum and the tongue (e.g., Moll, 1962; but see Lubker, 1968) or based on auditory factors that require a certain ratio of oral to nasal impedance for nasalization to be perceptible (House & Stevens, 1956; and see the discussion in Abramson et al., 1981).

When a vowel is nasalized, the mouth aperture is bigger than the nose aperture and as a result, the nasal cavity becomes a side-branching resonator to the oral-pharyngeal tube, which introduces an additional set of *nasal formants* and *nasal anti-formants* into the spectrum (House & Stevens, 1956; Fant, 1960; Fujimura, 1962). Some of the main acoustic consequences that result from coupling of the oral and nasal tubes in the production of nasalized vowels are as follows:

- There are changes to the oral formants. In particular, F1 moves up in frequency, is lowered in intensity, and has a broader bandwidth (House & Stevens, 1956).
- Compared with oral vowels, nasal vowels often have a greater density of formants in the 0–3,000 Hz range due to the presence of both oral and nasal formants. In the spectrogram on the left in Figure 3.16, the word-medial /a/ in /a#ɲampu/ is evidently more nasalized than the preboundary /a#/ in the same word: there are at least three resonance peaks for the former compared with two for the latter in the 500–2,500 Hz range. Similarly, the nasalized realization of /iː/ in *meaning* in Figure 3.17 has an additional nasal resonance at around 1,000 Hz compared with the oral production of /iː/ in *deeper* produced in the same prosodic phrase by the same speaker. An increase in

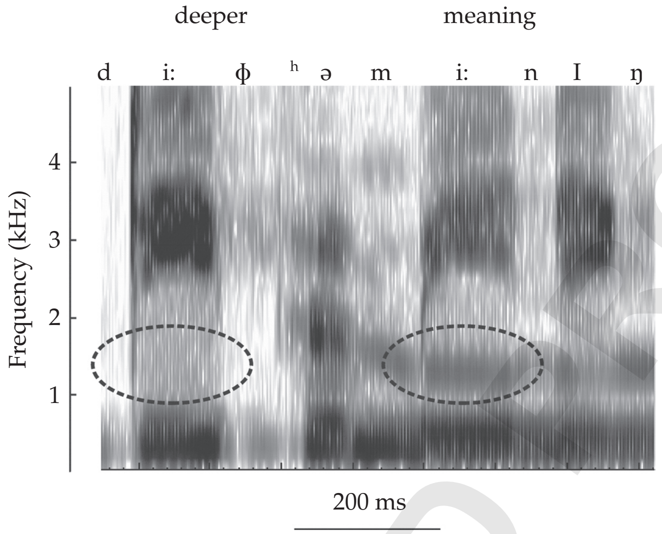

**Figure 3.17** Spectrogram of *deeper meaning* from the 2004 Queen Elizabeth II Christmas broadcast data (Harrington, 2006). The ellipses extend over the interval of [i:] in the two words showing the absence and presence respectively of a nasal formant at just under 1.5 kHz. (/p/ in *deeper* has been transcribed with a bilabial fricative since, as the aperiodic energy over this interval shows, the closure of the stop is evidently not complete).

the amplitude between F1 and F2 is common when high vowels are nasalized and recently Kataoka et al. (2001) have shown that this amplitude increase is correlated with perceived hypernasality in children.

- In mid vowels, i.e., vowels that have F1 roughly in the 300–800 Hz region, F1 is replaced with a triplet of an oral formant, nasal formant, and nasal anti-formant, i.e. an F1–N1–NZ1 combination (e.g., Hawkins & Stevens, 1985). Often F1 and N1 are not distinct, so that the overall effect in comparing oral and nasal mid vowels is that the bandwidth of the first formant (merged F1 and N1) is considerably broader than F1 of the corresponding oral vowel (Hattori et al., 1958). The peak amplitude of the first peak in nasalized mid vowels is also likely to be lower, both because of the broader bandwidth, and because of the presence of NZ1.

There is a loss of perceptual contrast especially along the height dimension when vowels are nasalized i.e., high vowels tend to be perceived to be lower and low vowels are perceived to be phonetically higher (Wright, 1975, 1986). The acoustic basis for this loss of distinction is likely to be that in high vowels, the mouth-cavity dependent F1 is raised, while in low vowels, the entire spectral center of gravity in the region of F1 is lowered due to the presence of N1 that is lower in frequency than F1 (Krakow et al., 1988). The perceptual lowering effect of nasalization was demonstrated by Beddor et al. (1986). They synthesized two continua from *bad* to *bed* by lowering F1. In one continuum, the vowels were oral, /bæd–bɛd/, and in the other they were nasal, /bæ̃d–bɛ̃d/. They found more tokens from the nasal continuum were labelled as *bad* than from the oral continuum (see Ohala, 1993 for a number of sound changes that are consistent with this effect). In a related experiment, Krakow et al. (1988) additionally synthesized a continuum from *bend* to *band* using nasal vowels (thus /bæ̃nd–bɛ̃nd/). They found that the responses from /bæ̃nd–bɛ̃nd/ patterned with the *oral* continuum from *bad–bed* /bæd–bɛd/ rather than with nasal /bæ̃d–bɛ̃d/. They reasoned that this is because listeners attribute the acoustic nasalization in /bæ̃nd–bɛ̃nd/ not to the vowel (as they necessarily must do in /bæ̃d–bɛ̃d/ since vowel nasalization has no coarticulatory raison d'être in this case) but to the contextual effects of the following /n/ consonant.

## 5 Concluding Comment

The three areas that have made substantial contributions to acoustic phonetics that were outlined at the beginning of this chapter are all certain to continue to be important for progress in the field in the future. As a result of the advances in speech physiology, in particular using techniques such as MRI and ultrasound, it is now possible to draw upon a much wider range of vocal tract cross-sectional data allowing more realistic articulatory-to-acoustic models to be developed. Making use of the extensive speech corpora that have become available in the last 15–20 years due largely to the needs of speech technology will be important for expanding our understanding of variability due to different speaking styles, age groups, language varieties, and many other factors. With the availability of larger amounts of training data that are now available from speech corpora, it should be possible in the future to incorporate into acoustic phonetic studies more sophisticated probabilistic and, above all, time-dependent models of the speech signal. Just this kind of information is becoming increasingly important in both phonology and linguistics (Bod et al., 2003). Analyzing large speech corpora will also be essential to ensure a greater convergence in the future between basic speech research and speech technology, so that more of the knowledge that has been described in this chapter can be incorporated more explicitly into the design of human–machine communication systems.

## ACKNOWLEDGMENTS

My thanks to David Broad and Frantz Clermont for a helpful discussion of Broad and Clermont (1987), to Janine Lilienthal and Nils Ülzmann for proofreading an earlier version of the manuscript, and to John Laver and Sallyanne Palethorpe for many helpful comments.

## REFERENCES

Abramson, A., Nye, P., Henderson, J., & Marshall, C. (1981) Vowel height and the perception of consonantal nasality. *Journal of the Acoustical Society of America*, 70, 329–39.

Ali, A., Van der Spiegel, J., & Mueller, P. (2001) Acoustic-phonetic features for the automatic classification of fricatives. *Journal of the Acoustical Society of America*, 109, 2217–35.

Assmann, P. (1991) The perception of back vowels: Centre of gravity hypothesis, *Quarterly Journal of Experimental Psychology*, 43, 423–8.

Atal, B. S. (1985) Linear predictive coding of speech. In F. Fallside & W. A. Woods (eds.), *Computer Speech Processing* (pp. 81–124). Englewood Cliffs, NJ: Prentice-Hall.

Atal, B. S. & Hanauer, S. (1971) Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50, 637–55.

Aylett, M. & Turk, A. (2006) Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America*, 119, 3048–58.

Beckman, M. E., Edwards, J., & Fletcher, J. (1992) Prosodic structure and tempo in a sonority model of articulatory dynamic. In G. Docherty & D. R. Ladd (eds.), *Papers in Laboratory Phonology II: Gesture, Segment, and Prosody* (pp. 68–86). Cambridge: Cambridge University Press.

Beddor, P. S., Krakow, R. A., & Goldstein, L. M. (1986) Perceptual constraints and phonological change: A study of nasal vowel height, *Phonology Yearbook* 3, 197–217.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003) Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America*, 113, 1001–24.

Bird, S. & Harrington, J. (2001) Speech annotation and corpus tools. *Speech Communication*, 33, 1–4.

Bladon, R. A. W. (1982) Arguments against formants in the auditory representation of speech. In R. Carlson and B. Granström (eds.), *The Representation of Speech in the Peripheral Auditory System* (pp. 95–102). Amsterdam: Elsevier Biomedical.

Bladon, R. A. W. & Al-Bamerni, A. (1976) Coarticulation resistance in English /l/, *Journal of Phonetics*, 4, 137–50.

Bladon, R. A. W. & Lindblom, B. (1981) Modeling the judgment of vowel quality differences, *Journal of the Acoustical Society of America* 69, 1414–22.

Blumstein, S. E. & Stevens, K. N. (1979) Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America* 66, 1001–17.

Blumstein, S. E. & Stevens, K. N. (1980) Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America* 67, 648–62.

Bod, R., Hay, J., & Jannedy, S. (2003) *Probabilistic Linguistics*. Cambridge, MA: MIT Press.

Brancazio, L. & Fowler, C. (1998) The relevance of locus equations for production and perception of stop consonants. *Perception and Psychophysics*, 60, 24–50.

Broad, D. & Clermont, F. (1987) A methodology for modeling vowel formant contours in CVC context.

*Journal of the Acoustical Society of America*, 81, 155–65.

Broad, D. & Fertig, R. H. (1970) Formant-frequency trajectories in selected CVC utterances. *Journal of the Acoustical Society of America* 47, 1572–82.

Broad, D. J. & Wakita, H. (1977) Piecewise-planar representation of vowel formant frequencies. *Journal of the Acoustical Society of America*, 62, 1467–73.

Bybee, J. (2000) Lexicalization of sound change and alternating environments. In M. Broe & J. Pierrehumbert (eds.), *Papers in Laboratory Phonology V: Acquisition and the Lexicon* (pp. 250–68). Cambridge: Cambridge University Press.

Carlson, R., Fant, G., & Granström, B. (1975) Two-formant models, pitch and vowel perception. In G. Fant & M. A. A. Tatham (eds.), *Auditory Analysis and Perception of Speech* (pp. 55–82). New York: Academic Press.

Carré, R. & Mrayati, M. (1992) Distinctive regions in acoustic tubes: Speech production modeling. *Journal d'Acoustique*, 5, 141–59.

Cassidy, S. & Harrington, J. (1995) The place of articulation distinction in voiced oral stops: Evidence from burst spectra and formant transitions. *Phonetica*, 52, 263–84.

Chang, S., Plauché, M. C., & Ohala, J. J. (2001) Markedness and consonant confusion asymmetries. In E. Hume & K. Johnson (eds.), *The Role of Speech Perception in Phonology* (pp. 79–101). San Diego CA: Academic Press.

Chen, M. Y. (1995) Acoustic parameters of nasalized vowels in hearing impaired and normal-hearing speakers. *Journal of the Acoustical Society of America*, 98, 2443–53.

Chen, M. Y. (1997) Acoustic correlates of English and French nasalized vowels. *Journal of the Acoustical Society of America*, 102, 2360–70.

Chennoukh, S., Carré, R., & Lindblom, B. (1997) Locus equations in the light of articulatory modeling. *Journal of the Acoustical Society of America*, 102, 2380–9.

Chistovich, L. A. (1985) Central auditory processing of peripheral vowel spectra. *Journal of the Acoustical Society of America*, 77, 789–805.

Chistovich, L. A. & Lublinskaya, V. V. (1979) The "center of gravity" effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research*, 1, 185–95.

Cole, R. A. & Cooper, W. E. (1975) Perception of voicing in English affricates and fricatives. *Journal of the Acoustical Society of America*, 58, 1280–7.

Cooper, F. S., Liberman, A. M., & Borst, J. M. (1951) The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proceedings of the National Academy Sciences of the United States of America*, 37, 318–25.

Cox., F. (1998) The Bernard data revisited. *Australian Journal of Linguistics*, 18, 29–55.

Darwin, C. & Seton, J. (1983) Perceptual cues to the onset of voiced excitations in aspirated initial stops. *Journal of the Acoustical Society of America*, 73, 1126–35.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955) Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27, 769–73.

Delattre, P., Liberman, A. M., Cooper, F. S., & Gerstman, F. J. (1952) An experimental study of the acoustic determinants of vowel color: Observations on one- and two-formant vowels synthesised from spectrographic patterns. *Word*, 8, 195–210.

Denes, P. (1955) Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America*, 27, 761–4.

Dunn, H. (1950) The calculation of vowel resonances in an electrical vocal tract. *Journal of the Acoustical Society of America*, 22, 740–53.

Edwards, J., Beckman, M. E., & Fletcher, J. (1991) The articulatory kinematics of final lengthening, *Journal of the Acoustical Society of America*, 89, 369–82.

Engstrand, O. & Lindblom, B. (1997) The locus line: Does aspiration affect its steepness? *Reports from the Department of Linguistics: Umea University (PHONUM)*, 4, 101–4.

Espy-Wilson, C. Y. (1992) Acoustic measures for linguistic features distinguishing the semivowels /w j r l/ in American English. *Journal of the Acoustical Society of America*, 92, 736–57.

Espy-Wilson, C. Y. (1994) A feature-based semivowel recognition system. *Journal of the Acoustical Society of America*, 96, 65–72.

Espy-Wilson, C. Y., Boyce, S., Jackson, M., Narayanan, S., & Alwan, A. (2000) Acoustic modeling of American English /r/. *Journal of the Acoustical Society of America*, 108, 343–56.

Essner, C. (1947) Recherche sur la structure des voyelles orales. *Archives Néerlandaises de Phonétique Expérimentale*, 20, 40–77.

Fant, G. (1960) *Acoustic Theory of Speech Production*. The Hague Mouton.

Fant, G. (1973) *Speech Sounds and Features*. Cambridge, MA: MIT Press.

Fischer-Jørgensen, E. (1954) Acoustic analysis of stop consonants, *Miscellenea Phonetica*, 2, 42–59.

Fischer-Jørgensen, E. (1972) Tape-cutting experiments with Danish stop consonants in initial position, *Annual Report, Institute of Phonetics, University of Copenhagen*, 6, 104–68.

Flanagan, J. L. (1972) *Speech Synthesis, Analysis, and Perception*. New York: Springer-Verlag.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988) Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America*, 84, 115–24.

Fourakis, M. (1991) Tempo, stress, and vowel reduction in American English. *Journal of the Acoustical Society of America*, 90, 1816–27.

Fowler, C. A. (1994) Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception and Psychophysics*, 55, 597–611.

Fowler, C. A. & Housum, J. (1987) Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction, *Journal of Memory and Language*, 26, 489–504.

Fry, D. B. (1965) The dependence of stress judgements on vowel formant structure. In E. Zwirner & W. Bethge (eds.), *Proceedings of the Fifth International Conference of Phonetic Sciences* (pp. 306-3–1). Basel: Karger.

Fujimura, O. (1962) Analysis of nasal consonants. *Journal of the Acoustical Society of America*, 34, 1865–75.

Gay, T. (1968) Effect of speaking rate on diphthong formant movements. *Journal of the Acoustical Society of America*, 44, 1570–73.

Gay, T. (1970) A perceptual study of American English diphthongs. *Language and Speech*, 13, 65–88.

Gottfried, M., Miller, J. D., & Meyer, D. J. (1993) Three approaches to the classification of American English diphthongs. *Journal of Phonetics*, 21, 205–29.

Hajek, J. (1997) *Universals of Sound Change in Nasalization*. Oxford: Blackwell.

Harrington, J. (1994) The contribution of the murmur and vowel to the place of articulation distinction in nasal consonants. *Journal of the Acoustical Society of America*, 96, 19–32.

Harrington, J. (2006) An acoustic analysis of "happy-tensing" in the Queen's Christmas Broadcasts. *Journal of Phonetics*, 34, 439–57.

Harrington, J. & Cassidy, S. (1994) Dynamic and target theories of vowel classification: Evidence from monophthongs and diphthongs in Australian English. *Language and Speech*, 37, 357–73.

Harrington, J. & Cassidy, S. (1999) *Techniques in Acoustic Phonetics*. Dordrecht: Kluwer.

Harrington, J., Fletcher, J., & Beckman, M. E. (2000) Manner and place conflicts in the articulation of accent in Australian English. In M. Broe (ed.), *Papers in Laboratory Phonology V* (pp. 40–55). Cambridge: Cambridge University Press.

Harris, K. S. (1958) Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech*, 1, 1–7.

Hashi, M., Honda, K., & Westbury, J. (2003) Time-varying acoustic and articulatory characteristics of American English [ɹ]: A cross-speaker study. *Journal of Phonetics*, 31, 3–22.

Hattori, S., Yamamoto, K., & Fujimura, O. (1958) Nasalization of vowels in relation to nasals. *Journal of the Acoustical Society of America*, 30, 267–74.

Hawkins, S. & Stevens, K. N. (1985) Acoustic and perceptual correlates of the non-nasal-nasal distinction for vowels. *Journal of the Acoustical Society of America*, 77, 1560–75.

Hay, J., Sato, M., Coren, A., Moran, C., & Diehl, R. (2006) Enhanced contrast for vowels in utterance focus: A cross-language study. *Journal of the Acoustical Society of America*, 119, 3022–33.

Heinz, J. M. & Stevens, K. N. (1961) On the properties of voiceless fricative consonants. *Journal of the Acoustical Society of America*, 33, 589–93.

Hillenbrand, J., Clark, M., & Nearey, T. (2001) Effects of consonant environment on vowel formant patterns. *Journal of the Acoustical Society of America*, 109, 748–63.

Hillenbrand, J., Houde, R., & Gayvert, R. (2006) Speech perception based on spectral peaks versus spectral shape. *Journal of the Acoustical Society of America*, 119, 4041–54.

Hillenbrand, J. & Nearey, T. (1999) Identification of resynthesized /hVd/ utterances: Effects of formant contour. *Journal of the Acoustical Society of America*, 105, 3509–23.

Holbrook, A. & Fairbanks, G. (1962) Diphthong formants and their movements. *Journal of Speech and Hearing Research*, 5, 38–58.

Hombert, J-M., Ohala, J., & Ewan, W. (1979) Phonetic explanations for the development of tones. *Language*, 55, 37–58.

House, A. S. & Fairbanks, G. (1953) The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America*, 25, 105–13.

House, A. S. & Stevens, K. N. (1956) Analog studies of the nasalization of vowels. *Journal of Speech and Hearing Disorders*, 21, 218–32.

Huang, C. B. (1986) The effect of formant trajectory and spectral shape on the tense/lax distinction in American vowels. *IEEE International Conference on Acoustics Speech and Signal Processing*, 893–6.

Huang, C. B. (1992) Modeling human vowel identificatin using aspects of format trajectory and context. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (eds.), *Speech Perception, Production and Linguistic Structure* (pp. 43–61). Amsterdam: IOS Press.

Hunnicutt, S. (1985) Intelligibility versus redundancy: Conditions of dependency. *Language and Speech*, 28, 47–56.

Hunnicutt, S. (1987) Acoustic correlates of redundancy and intelligibility. *Speech Transmission Laboratory, Quarterly Status Progress Report*, 2–3, 7–14.

Ito, M., Tsuchida, J., & Yano, M. (2001) On the effectiveness of whole spectral shape for vowel perception. *Journal of the Acoustical Society of America*, 110, 1141–9.

Johnson, K. (2004) *Acoustic and Auditory Phonetics*. Oxford: Blackwell.

Johnson, K. (2005) Speaker normalization in speech perception. In D. Pisoni & R. Remez (eds.), *The Handbook of Speech*

*Perception* (pp. 363–89). Malden, MA: Blackwell.

Jongman, A. (1989) Duration of frication noise required for identification of English fricatives. *Journal of the Acoustical Society of America*, 85, 1718–25.

Jongman, A., Blumstein, S. E., & Lahiri, A. (1985) Acoustic properties for dental and alveolar stop consonants: A cross-language study. *Journal of Phonetics*, 13, 235–51.

Jongman, A. R., Wayland, S., & Wong, S. (2000) Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America*, 108, 1252–63.

Joos, M. (1948) Acoustic phonetics. *Language*, 24, 1–136.

Jurafsky, D., Bell, A., & Girand, C. (2003) The role of the lemma in form variation. In N. Warner and C. Gussenhoven (eds.), *Laboratory Phonology VII* (pp. 3–34). Mouton Berlin: de Gruyter.

Kataoka, R., Warren, D., Zajac, D., Mayo, R., and Lutz, R. (2001) The relationship between spectral characteristics and perceived hypernasality in children. *Journal of the Acoustical Society of America*, 109, 2181–9.

Kelso, J. A. S., Vatikiotis-Bateson, E., Saltzman, E., & Kay, B. (1985) A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling. *Journal of the Acoustical Society of America*, 77, 266–80.

Kewley-Port, D. (1982) Measurement of formant transitions in naturally produced stop consonant-vowel syllables, *Journal of the Acoustical Society of America*, 72, 379–89.

Kewley-Port, D., Pisoni, D. B., and Studdert-Kennedy, M. (1983) Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, 73, 1779–93.

Kiefte, M. & Kluender, K. (2005) The relative importance of spectral tilt in monophthongs and diphthongs. *Journal of the Acoustical Society of America*, 117, 1395–1404.

Klatt, D. H. (1982) Speech processing strategies based on auditory models. In R. Carlson and B. Granström (eds.), *The Representation of Speech in the Peripheral Auditory System* (pp. 181–96). Amsterdam: Elsevier Biomedical.

Klein, W., Plomp, R., & Pols, L. C. W. (1970) Vowel spectra, vowel spaces and vowel identification. *Journal of the Acoustical Society of America*, 48, 999–1009.

Koenig, W., Dunn, H. K., & Lacy, L. Y. (1946) The sound spectrograph. *Journal of the Acoustical Society of America*, 18, 19–49.

Kohler, K. J. (1979) Parameters in the production and the perception of plosives in German and French. *Arbeitsberichte, Institute of Phonetics, University of Kiel*, 12, 261–92.

Kohler, K. J. (1985) F0 in the perception of lenis and fortis plosives. *Journal of the Acoustical Society of America*, 78, 21–32.

Krakow, R., Beddor, P., Goldstein, L., & Fowler, C. (1988) Coarticulatory influences on the perceived height of nasal vowels. *Journal of the Acoustical Society of America*, 83, 1146–58.

Krull, D. (1989) Second formant locus patterns and consonant vowel coarticulation in spontaneous speech. *Phonetic Experimental Research at the Institute of Linguistics, University of Stockholm*, 10, 87–108.

Krull, D., Lindblom, B., Shia, B. E., & Fruchter, D. (1995) Cross-linguistic aspects of coarticulation: An acoustic and electropalatographic study of dental and retroflex consonants. *Proceedings of the 13th International Congress of Phonetic Sciences*, Stockholm, Sweden, 3, 436–9.

Kuehn, D. P. & Moll, K. L. (1976) A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics*, 4, 303–20.

Kurowski, K. & Blumstein, S. E. (1984) Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants. *Journal of the Acoustical Society of America*, 76, 383–90.

Kurowski, K. & Blumstein, S. E. (1987) Acoustic properties for place of articulation in nasal consonants. *Journal of the Acoustical Society of America*, 81, 1917–27.

Labov, W. (2001) *Principles of Linguistic Change*, vol. 2: *Social Factors*. Oxford: Blackwell.

Ladefoged, P. (1971) *Preliminaries to Linguistic Phonetics*. Chicago: University of Chicago Press.

Ladefoged, P. (1985) The phonetic basis for computer speech processing. In F. Fallside & W. A. Woods (eds.), *Computer Speech Processing* (pp. 3–27). Englewood Cliffs, NJ: Prentice-Hall.

Ladefoged, P. & Bladon, R. A. W. (1982) Attempts by human speakers to reproduce Fant's nomograms. *Speech Communication*, 9, 231–98.

Lahiri, A., Gewirth, L., & Blumstein, S. E. (1984) A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. *Journal of the Acoustical Society of America*, 76, 391–404.

Lawrence, W. (1953) The synthesis of speech from signals which have a low information rate. In W. Jackson (ed.), *Communication Theory* (pp. 460–9). London: Butterworth.

Lehiste, I. (1964) *Acoustical Characteristics of Selected English Consonants*. Bloomington: Indiana University Press.

Lehiste, I. & Peterson, G. (1961) Transitions, glides, and diphthongs. *Journal of the Acoustical Society of America*, 33, 268–77.

Liberman, A. M., Delattre, P. C. and Cooper, F. S. (1958) The role of selected stimulus variables in the perception of voiced and voiceless stops in initial position. *Language and Speech*, 1, 153–67.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954) The role of consonant–vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs*, 68, 1–13.

Lieberman, P. (1963) Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, 6, 172–87.

Liljencrants, J. & Lindblom, B. (1972) Numerical simulations of vowel quality systems: The role of perceptual contrast. *Language*, 48, 839–62.

Lindblom, B. (1963) Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 35, 1773–81.

Lindblom, B. (1990) Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle and A. Marchal (eds.), *Speech Production and Speech Modeling* (pp. 403–39). Dordrecht: Kluwer Academic.

Lindblom, B. (1996) Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America*, 99, 1683–92.

Lindblom, B. & Studdert-Kennedy, M. (1967) On the role of formant transitions in vowel recognition. *Journal of the Acoustical Society of America*, 42, 830–43.

Lindblom, B. E. F. & Sundberg, J. E. F. (1971) Acoustical consequences of lip, tongue, jaw, and larynx movement. *Journal of the Acoustical Society of America*, 50, 1166–79.

Lisker, L. (1975) Is it VOT or a first-formant transition detector? *Journal of the Acoustical Society of America*, 57, 1547–51.

Lisker, L. (1978) Rapid vs. rabid: A catalogue of acoustic features that may cue the distinction. *Haskins Laboratories Status Reports on Speech Research*, 54, 127–32.

Lisker, L. (1986) Voicing in English: A catalogue of acoustic features signaling /b/ vs. /p/ in trochees. *Language and Speech*, 29, 3–11.

Lisker, L. & Abramson, A. S. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20, 384–422.

Lisker, L. & Abramson, A. S. (1967) Some effects of context on voice onset time in English stops. *Language and Speech*, 10, 1–28.

Löfqvist, A. (1999) Interarticulator phasing, locus equations, and degree of coarticulation. *Journal of the Acoustical Society of America*, 106, 2022–30.

Löfqvist, A., Baer, T., McGarr, N., & Story, R. (1989) The cricothyroid muscle in voicing control. *Journal of the Acoustical Society of America*, 85, 1314–21.

Lubker, J. (1968) An electromyographic-cinefluorographic investigation of velar function during normal speech production. *Cleft Palate Journal*, 5, 1–18.

Mack, M. & Blumstein, S. E. (1983) Further evidence of acoustic invariance in speech production: The stop–glide contrast. *Journal of the Acoustical Society of America*, 73, 1739–50.

Malécot, A. (1956) Acoustic cues for nasal consonants. *Language*, 32, 274–84.

Mann, V. A. & Repp, B. H. (1980) Influence of vocalic context on perception of the [ʃ]–[s] distinction. *Perception and Psychophysics*, 28, 213–28.

Massaro, D. W. & Cohen, M. M. (1976) The contribution of fundamental frequency and voice onset time to the /zi/–/si/ distinction. *Journal of the Acoustical Society of America*, 60, 704–17.

Milner, B. & Shao, X. (2006) Clean speech reconstruction from MFCC vectors and fundamental frequency using an integrated front-end. *Speech Communication*, 48, 697–715.

Modarresi, G., Sussman, H., Lindblom, B., & Burlingame, E. (2005) Locus equation encoding of stop place: Revisiting the voicing/VOT issue. *Journal of Phonetics*, 33, 101–13.

Molis, M. (2005) Evaluating models of vowel perception. *Journal of the Acoustical Society of America*, 118, 1062–71.

Moll, K. L. (1962) Velopharyngeal closure on vowels. *Journal of Speech and Hearing Research*, 5, 30–7.

Moon, S.-J. & Lindblom, B. (1994) Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America*, 96, 40–55.

Munson, B. & Soloman, N. (2004) The effect of phonological neighborhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research*, 47, 1048–58.

Nearey, T. M. & Assmann, P. (1986) Modeling the role of vowel inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, 80, 1297–1308.

Nittrouer, S. (2002) Learning to perceive speech: How fricative perception changes, and how it stays the same. *Journal of the Acoustical Society of America*, 112, 711–19.

Nolan, F. (1983) *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press.

Nossair, Z. B. & Zahorian, S. A. (1991) Dynamical spectral features as acoustic correlates for initial stop consonants. *Journal of the Acoustical Society of America*, 89, 2978–91.

O'Connor, J., Gerstman, L., Liberman, A. M., Delattre, P., & Cooper, F. S. (1957) Acoustic cues for the perception of initial /w, j, r, l/ in English. *Word*, 13, 24–43.

Ohala, J. J. (1993) The phonetics of sound change. In Charles Jones (ed.), *Historical Linguistics: Problems and Perspectives* (pp. 237–78). London: Longman.

Ohala, J. J. & Busà, M. G. (1995) Nasal loss before voiceless fricatives: A perceptually-based sound change. *Rivista di Linguistica*, 7, 125–44.

Ohala, J. J. & Lorentz, J. (1977) The story of [w]: An exercise in the phonetic explanation for sound patterns. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, 3, 577–99.

Ohde, R. N., Haley, K., & Barne, C. (2006) Perception of the [m]–[n] distinction in consonant-vowel (CV) and vowel-consonant (VC) syllables produced by child and adult talkers. *Journal of the Acoustical Society of America*, 119, 1697–1711.

Ohde, R. N. & Stevens, K. N. (1983) Effect of burst amplitude on the perception of stop consonant place of articulation. *Journal of the Acoustical Society of America*, 74, 706–14.

Öhman, S. E. G. (1966) Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, 39, 151–68.

Öhman, S. E. G. (1967) Numerical model of coarticulation. *Journal of the Acoustical Society of America*, 41, 310–20.

Palethorpe, S., Wales, R., Clark, J. E., & Senserrick, T. (1996) Vowel classification in children. *Journal of the Acoustical Society of America*, 100, 3843–51.

Palethorpe, S., Watson, C. I., & Barker, R. (2003) Acoustic analysis of monophthong and diphthong production in acquired severe to profound hearing loss. *Journal of the Acoustical Society of America*, 114, 1055–68.

Peterson, G. & Barney, H. L. (1952) Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24, 175–84.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986) Speaking clearly for the hard of hearing, II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, 29, 434–46.

Polka, L. & Strange, W. (1985) Perceptual equivalence of acoustic cues that differentiates /r/ and /l/. *Journal of the Acoustical Society of America*, 78, 1187–97.

Pols, L. C. W., Tromp, H. R. C., & Plomp, R. (1973) Frequency analysis of Dutch vowels from 50 male speakers. *Journal of the Acoustical Society of America*, 53, 1093–1101.

Potter, R. K., Kopp, G., & Green, H. (1947) *Visible Speech*. New York: Dover Publications.

Qi, Y. & Fox, R. A. (1992) Analysis of nasal consonants using perceptual linear prediction. *Journal of the Acoustical Society of America*, 91, 1718–26.

Recasens, D. (1983) Place cues for nasal consonants with special reference to Catalan. *Journal of the Acoustical Society of America*, 73, 1346–53.

Redford, M. & Diehl, R. (1999) The relative perceptual distinctiveness of initial and final consonants in CVC syllables. *Journal of the Acoustical Society of America*, 106, 1555–65.

Repp, B. H. (1979) Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. *Language and Speech*, 27, 173–89.

Repp, B. H. (1986) Perception of the m–n distinction in CV syllables. *Journal of the Acoustical Society of America*, 79, 1987–99.

Repp, B. H. (1987) On the possible role of auditory short-term adaptation in perception of the prevocalic m–n contrast. *Journal of the Acoustical Society of America*, 82, 1525–38.

Repp, B. H. & Lin, H.-B. (1989) Acoustic properties and perception of stop consonant release transients. *Journal of the Acoustical Society of America*, 85, 379–96.

Repp, B. H. & Svastikula, K. (1988) Perception of the [m]–[n] distinction in VC syllables. *Journal of the Acoustical Society of America*, 83, 237–47.

Schouten, M. E. H. & Pols, L. C. W. (1979a) Vowel segments in consonantal context: A spectral study of coarticulation, Part I. *Journal of Phonetics*, 7, 1–23.

Schouten, M. E. H. & Pols, L. C. W. (1979b) CV- and VC-transitions: A spectral study of coarticulation, Part II. *Journal of Phonetics*, 7, 205–24.

Seitz, P. F., McCormick, M. M., Watson, I. M. C., & Bladon, R. A. (1990) Relational spectral features for place of articulation in nasal consonants. *Journal of the Acoustical Society of America*, 87, 351–8.

Shadle, C. H. & Mair, S. J. (1996) Quantifying spectral characteristics of fricatives. In H. Bunnell & W. Idsari (eds.), *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP '96)* (pp. 1517–20). New Castle, DE: Citation Delaware.

Shepard, R. N. (1972) Psychological representation of speech sounds. In E. David & D. P. Denes (eds.), *Human Communication: A Unified View* (pp. 67–113). New York: McGraw Hill.

Slis, I. & Cohen, A. (1969) On the complex regulating the voiced–voiceless distinction. *Language and Speech*, 12, 80–102.

Sluijter, A. M. C. & Heuven, V. J. van (1996) Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100, 2471–85.

Sluijter, A. M. C., Heuven, V. J. van, & Pacilly, J. J. A. (1997) Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America*, 101, 503–13.

Smiljanić, R. & Bradlow, A. (2005) Production and perception of clear speech in Croatian and English. *Journal of the Acoustical Society of America*, 118, 1677–88.

Smits, R. ten Bosch, L., & Collier, R. (1996a) Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants, I: Perception experiment. *Journal of the Acoustical Society of America*, 100, 3852–64.

Smits, R. ten Bosch, L., & Collier, R. (1996b) Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants, II: Modeling and evaluation. *Journal of the Acoustical Society of America*, 100, 3865–81.

Soli, S. D. (1981) Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation. *Journal of the Acoustical Society of America*, 70, 976–84.

Stack, J., Strange, W., Jenkins, J., Clarke, W., & Trent, S. (2006) Perceptual invariance of coarticulated vowels over variations in speaking rate. *Journal of the Acoustical Society of America*, 119, 2394–405.

Stevens, K. N. (1971) Airflow and turbulence for noise for fricative and stop consonants: Static considerations, *Journal of the Acoustical Society of America*, 50, 1180–92.

Stevens, K. N. (1972) The quantal nature of speech: Evidence from articulatory-acoustic data. In E. Davis & D. P. Denes (eds.), *Human Communication: A Unified View* (pp. 51–66). New York: McGraw Hill.

Stevens, K. N. (1985) Evidence for the role of acoustic boundaries in the perception of speech sounds. In V. A. Fromkin (ed.), *Phonetic Linguistics* (pp. 243–55). New York: Academic Press.

Stevens, K. N. (1989) On the quantal nature of speech. *Journal of Phonetics*, 17, 3–46.

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

Stevens, K. N. (2002) Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, 111, 1872–91.

Stevens, K. N. & House, A. S. (1955) Development of a quantitative description of vowel articulation. *Journal of the Acoustical Society of America*, 27, 484–93.

Stevens, K. N. & House, A. S. (1956) Studies of formant transitions using a vocal tract analog. *Journal of the Acoustical Society of America*, 28, 578–85.

Stevens, K. N. & House, A. S. (1963) Perturbation of vowel articulations by consnantal context: An acoustical study. *Journal of Speech and Hearing Research*, 6, 111–28.

Stevens, K. N. & Klatt, D. H. (1974) Role of formant transitions in the voiced–voiceless distinction of stops.

*Journal of the Acoustical Society of America*, 55, 653–9.

Strange, W. (1999) Perception of vowels: Dynamic constancy. In J. Pickett (ed.), *The Acoustics of Speech Communication* (pp. 153–65). Boston: Allyn & Bacon.

Strange, W. & Bohn, O.-S. (1998) Dynamic specification of coarticulated German vowels: Perceptual and acoustical studies. *Journal of the Acoustical Society of America*, 104, 488–504.

Strange, W., Jenkins, J. J., & Johnson, T. L. (1983) Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, 74, 695–705.

Strange, W., Verbrugge, R. R., Shankweiler, D. P., & Edman, T. R. (1976) Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, 60, 213–24.

Summerfield, A. Q. & Haggard, M. P. (1977) On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America*, 62, 435–48.

Sussman, H. M. (1994) The phonological reality of locus equations across manner class distinctions: Preliminary observations, *Phonetica*, 51, 119–31.

Sussman, H. M., Fruchter, D., & Cable, A. (1995) Locus equations derived from compensatory articulation. *Journal of the Acoustical Society of America*, 97, 3112–24.

Sussman, H. M., Hoemeke, K. A., & Ahmed, F. S. (1993) A crosslinguistic investigation of locus equations as a phonetic descriptor of articulation. *Journal of the Acoustical Society of America*, 94, 1256–68.

Sussman, H. M., McCaffrey, H., & Matthews, S. A. (1991) An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, 90, 1309–25.

Syrdal, A. K. (1985) Aspects of a model of the auditory representation of American English vowels. *Speech Communication*, 4, 121–35.

Syrdal, A. K. & Gopal, H. S. (1986) A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, 79, 1086–100.

Tabain, M. (1998) Nonsibilant fricatives in English: Spectral information above 10 kHz. *Phonetica*, 55, 107–30.

Tabain, M. (2000) Coarticulation in CV syllables: A comparison of Locus Equation and EPG data. *Journal of Phonetics*, 28, 137–59.

Tabain, M. (2001) Variability in fricative production and spectra: Implications for the hyper- and hypo- and quantal theories of speech production. *Language and Speech*, 44, 57–94.

Tabain, M. & Butcher, A. (1999) Stop consonants in Yanyuwa and Yindjibarndi: A locus equation perspective. *Journal of Phonetics*, 27, 333–58.

Taylor, H. C. (1933) The fundamental pitch of English vowels. *Journal of Experimental Psychology*, 16, 565–82.

Terbeek, D. (1977) Cross-language multidimensional scaling study of vowel perception. *UCLA Working Papers in Phonetics, University of California, Los Angeles*, 37.

Traunmüller, H. (1981) Perceptual dimension of openness in vowels. *Journal of the Acoustical Society of America*, 69, 1465–75.

Traunmüller, H. (1984) Articulatory and perceptual factors controlling the age- and sex-conditioned variability in formant frequencies of vowels. *Speech Communication*, 3, 49–61.

Tsao, T-C., Weismer, G., & Iqbal, K. (2006) The effect of intertalker speech rate variation on acoustic vowel space. *Journal of the Acoustical Society of America*, 119, 1074–82.

Turner, G. S., Tjaden, K., & Weismer, G. (1995) The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. *Journal of Speech and Hearing Research*, 38, 1001–3.

Vaissière, J. (2007) Area functions and articulatory modeling as a tool for investigating the articulatory, acoustic and perceptual properties of sounds across languages. In M. J. Solé, P. Beddor & M. Ohala (eds.), *Experimental Approaches to Phonology* (pp. 54–72). Oxford: Oxford University Press.

van Bergem, D. R. (1993) Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication*, 12, 1–23.

van Son, R. J. J. H. (1993) Spectro-temporal features of vowel segments: Studies in language and language use. Ph.D. thesis, University of Amsterdam.

van Son, R. J. J. H. & Pols, L. C. W. (1990) Formant frequencies of Dutch vowels in a text, read at normal and fast rate. *Journal of the Acoustical Society of America*, 88, 1683–93.

van Son, R. J. J. H. & Pols, L. C. W. (1992) Formant movements of Dutch vowels in a text, read at normal and fast rate. *Journal of the Acoustical Society of America*, 92, 121–7.

van Son, R. J. J. H. & Pols, L. C. W. (1999) An acoustic description of consonant reduction. *Speech Communication*, 28, 125–40.

van Son, R. J. J. H. & Pols, L. C. W. (2003) How efficient is speech? *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam*, 25, 171–84.

Wardrip-Fruin, C. (1982) On the status of temporal cues to phonetic categories: Preceding vowel duration as a cue to voicing in final stop consonants. *Journal of the Acoustical Society of America*, 71, 187–95.

Watson, C. I. & Harrington, J. (1999) Acoustic evidence for dynamic formant trajectories in Australian English vowels. *Journal of the Acoustical Society of America*, 106, 458–68.

Weismer, G., Laures, J. S., Jeng, J.-Y., Kent, R. D., & Kent, J. F. (2000) Effect of speaking rate manipulations of acoustic and perceptual aspects of the dysarthria in Amyotrophic Lateral Sclerosis. *Folia Phoniatrica et Logopaedica*, 52, 201–19.

Whalen, D. H., Abramson, A., Lisker, L., & Mody, M. (1993) F0 gives voicing information even with unambiguous voice onset times. *Journal of the Acoustical Society of America*, 47, 36–49.

Winitz, H., Scheib, M. E., & Reeds, J. A. (1972) Identification of stops and vowels for the burst portion of /p, t, k/ isolated from conversational speech. *Journal of the Acoustical Society of America*, 51, 1309–17.

Wood, S. (1986) The acoustical significance of tongue, lip, and larynx maneuvers in rounded palatal vowels. *Journal of the Acoustical Society of America*, 80, 391–401.

Wright, J. (1975) Nasal-stop assimilation: Testing the psychological reality of an English MSC. In C. A. Ferguson, L. M. Hyman, & J. J. Ohala (eds.), *Nasalfest* (pp. 389–97). Stanford: Language Universals Project, Dept. of Linguistics, Stanford University.

Wright, J. (1986) The behavior of nasalized vowels in the perceptual vowel space. In J. J. Ohala and J. J. Jaeger (eds.), *Experimental Phonology* (pp. 45–67). Orlando: Academic Press.

Wright, R. (2003) Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (eds.), *Papers in Laboratory Phonology VI: Phonetic Interpretation* (pp. 75–87). Cambridge: Cambridge University Press.

Wuensch, K. (2006) Skewness, kurtosis, and the normal curve. http://core.ecu.edu/psyc/wuenschk/StatsLessons.htm.

Zahorian, S. & Jagharghi, A. (1993) Spectral shape features versus formants as acoustic correlates for vowels. *Journal of the Acoustical Society of America*, 94, 1966–82.