3

THE CONTRIBUTION OF PHONETICS TO THE STUDY OF VOWEL DEVELOPMENT AND DISORDERS

Sara Howard and Barry Heselwood

“Now we must pull ourselves together, for we have come to the vowels, and they are very troublesome.” (Rippmann, 1911: 32)

INTRODUCTION

As Ball (1993: 66) observed, “vowels are not as easy to classify as consonants.” Nor are they easy to describe generally in terms of their articulatory and acoustic correlates, and furthermore they pose particular problems for perceptual analysis and transcription. Rippmann, it seems, had a point! This chapter will consider some phonetic approaches to vowel description from both instrumental and perceptual perspectives and will examine evidence from speech development and impaired speech, which helps to illuminate some of the particular challenges and problems posed by vowel description and classification.

VOWELS: A WORKING DEFINITION

First we need to provide a working definition for the scope of the term vowel as used in this chapter. As we shall see, even providing a simple but unambiguous definition is not straightforward. The problem relates in large part to the difference between phonetic and phonological usages of the term. As many authors have noted, from a phonetic perspective
vowels can be described as resonant segments that are produced with an unobstructed laminar airflow escaping centrally through the vocal tract (Ladefoged, 1993; Laver, 1994; Ashby & Maidment, 2005). In contrast, consonants involve some degree of constriction in the vocal tract, which either blocks the airflow momentarily, causes it to become turbulent, or diverts it from a central oral path. Phonologically, on the other hand, vowels are viewed from distributional and functional perspectives, and can be described as those segments that may occupy the center or nucleus of a syllable, as opposed to consonants, which occur in marginal positions in a syllable (onsets and codas).

Ball and Rahilly (1999: 43) point out that “these two approaches to defining consonants and vowels overlap in the majority of cases.” But, as they go on to observe, some sounds that appear to match a phonetic description of the category of vowels (e.g., /j/ and /w/), have patterns of distribution that place them in the consonant category according to a phonological perspective. Catford (1977) provides a long discussion on the ambiguous relationship of approximants and vowels, noting inter alia that /j/ and /w/ are differentiated phonetically from the high vowels /i/ and /u/ on durational grounds, and are very similar in terms of articulatory stricture. Also, Stevens (1998) points out that glides have broader formant bandwidths than their corresponding vowels, due to a slightly greater degree of constriction. Similarly, some sounds we would happily term consonants on the grounds of their phonetic characteristics may occur as syllable nuclei (e.g., [l̩] and [n̩], often termed syllabic consonants). Because of this potential ambiguity, Pike (1947) suggested a two-way categorisation, reserving the terms vowel and consonant to reflect a segment’s phonological behaviour, and using the terms contoid and vocoid to imply phonetic characteristics of segments. In this way, one might distinguish four sets of sounds: contoid consonants (e.g., [p], [s]), vocoid vowels (e.g., [i], [a]), contoid vowels (e.g., [i̯], [n̩]), and vocoid consonants (e.g., [j], [w]). In this chapter, we will restrict our use of the term vowel to those segments that comply with both the phonetic and phonological definitions of vowel given above, in other words, to segments described by Pike (1947) as vocoid vowels.

**STUDYING THE PHONETICS OF VOWELS**

In the discussions later in this chapter references will be made to properties of vowels in the articulatory, acoustic, and auditory domains. Before embarking on those discussions it will be useful to briefly review some of the principal means at our disposal for investigating these properties.
The Articulatory Domain

Clark, Yallop, and Fletcher (2007: 22) observe that “the major challenge in describing the articulation of vocalic sounds is to define the position of the tongue,” and the majority of the following instrumental techniques, when applied to the task of vowel description, aim to provide different types of information on the location and movement of the tongue during vowel description.

Electropalatography (EPG) is a technique that provides information on patterns of contact between the tongue and the palate in the region from the rear of the front teeth or the alveolar ridge to the margins of the hard and soft palate (past and current palatography systems include the Kay Palatometer (Fletcher, McCutcheon, & Wolf, 1975), the Rion Palatograph (Fujimura, Tatsumi, & Kayaga, 1973), the Reading EPG and the WinEPG system; for a review see Gibbon (2008). In our phonetic definition of vowels (above) we said that they require an unobstructed airflow, which means that we would not expect lingualpalatal contact in the central part of the palate. Hardcastle and Gibbon (1997) state that there will be little discernible contact for open vowels or for back vowels. However, particularly for close vowels (e.g., [i], [ɪ], [u]) and for the latter stages of closing diphthongs (e.g., [ɔɪ], /aɪ, [au]), there are clear patterns of contact between the sides of the tongue and the lateral margins of the palate (Byrd, 1995; Gibbon, Lee, & Yuen, 2010). Figure 3.1 shows lingualpalatal contact patterns for the high vowels /i/ and /u/ and the closing diphthong /aɪ/ in Southern Standard British English.

Although EPG appears to be a technique with relatively little to offer in the study of normal vowel production, it is useful in looking at the influence of vowels on consonants and has provided interesting insights into the production of vowels in speakers with impaired speech. Yamashita, Michi, Imai, Suzuki, and Yoshida (1992), for example, describe the speech of a six-year-old child with a history of cleft palate, who maintained strictures of central lingualpalatal contact over extensive areas of the palate during productions perceptually identified as the vowel [i]. Howard (2007) uses EPG to illustrate the case of a 16-year-old girl with a repaired cleft who maintained complete contact across the posterior and central portions of the palate throughout all mid and high vowels (as well as most consonants), releasing airstream laterally or nasally.

Electromagnetic Articulography (EMA) uses electromagnetic transmitter and receiver coils attached to active and passive articulators in the vocal tract to track range and velocity of movement of the active articulators and changing vocal tract configurations over time (Schökle, Grabe, Wenig, Hohne, Schrader, & Conrad, 1987; Perkell, Cohen,
Svirsky, Matthies, Garabieta, & Jackson, 1992; Stone, 2010). Using a system of fixed reference points, up to five coils can be attached to, for example, the upper and lower lips, the tip or blade and dorsum of the tongue, the mandible and even the soft palate to provide a representation of the time course of articulatory movements in the mid-sagittal plane. This is a relatively new technique, but it has already been used to examine aspects of normal and impaired speech production, and has been combined with EPG to give information about lingual activity across different dimensions. Clearly the opportunities it affords for the detailed examination of articulator coordination have enormous potential in the study of normal and atypical vowel production, and the simultaneous monitoring of different articulators means that, for example, the complex interrelationships between lip, tongue, and jaw movements in vowel articulation can be examined. However, its present high cost, and the problems it presents for use with young children, may mean that in practice not a lot of insight into developmental vowel disorders will be gained through EMA for the time being.

Ultrasound, MRI (magnetic resonance imaging) and x-ray imaging are techniques that, for reasons of cost and of safety, respectively, are currently less readily available for the phonetic study of speech production and its clinical application. All, however, have been used to investigate the articulatory properties of vowels as they can provide both static and dynamic images of the movements of the vocal organs during speech.
production, and can sample data from various different spatial perspectives. For the investigation of vowel production, a dynamic, midsagittal view provides the best information on tongue position and movement and on its relationship with other articulators, including the lips, jaw, and velum. Gibbon (2008) and Stone (2005, 2010) both give clear accounts of the technique, outlining methodological problems and pitfalls, which include the difficulty of disentangling overlapping images of different parts of the vocal tract (tongue, teeth, lips, bone, etc.) and identifying precisely which part of an articulator is revealing itself in the image. For example, is the image of the edge of the tongue in a given data sample showing the center of the tongue or one of its lateral margins?

Such questions may, however, be more problematic for the study of consonant articulations than vowels. Certainly x-ray data has provided us with a considerable body of useful information about vowel production. It has been used to identify both the highest point of the tongue and also the point of closest lingual constriction in the vocal tract (not always the same thing) and to relate tongue strictures to spatial adjustments of the mandible and velum (Kent & Moll, 1972; Wood, 1979; Perkell, 1996). Ultrasound imaging has great potential in the description of vowel production because it is a technique that is able to map the cross-sectional shape of the vocal tract in a more complete way than is otherwise possible. Ultrasound makes use of the sound wave reflections that occur at boundaries between tissues or surfaces with different densities (Stone, 2005, 2010), and can thus map data over complex spatial configurations and dimensions. Ultrasound and MRI have been used to explore lingual and mandibular activity in vowel production, consonant–vowel coarticulation, and in relationships between vowel production and velopharyngeal muscle activity (Stone, Shawker, Talbot, & Rich, 1988; Stone & Vatikiotis-Bateson, 1995; Chiang, Lee, Peng & Liu, 2003; Engwall, 2003; Takano & Honda, 2007) and in assessment and intervention for a range of speech impairments (see, for example, Bernhardt, Gick, Backsfalvi, & Ashdown, 2003; Bressmann, Uy, & Irish, 2005; Bressmann, Flowers, Wong, & Irish, 2010).

Electromyography (EMG) detects muscle activation by measuring electrical potential (Gentil & Moore, 1997; Stone, 1997). Small electrodes are attached to the surface of the articulators and detect electrical activity as muscle fibres contract. Increased electrical potential can show that a muscle is being innervated even where there is no observable movement of the organ it relates to, making it possible to infer an intention on the part of the speaker to employ a particular articulator. Because it is possible to place electrodes on a number of discrete sites and collect
simultaneous information from, for example, muscles of the lips, tongue, and jaw, this is a technique, like EMA (above) that has the potential to provide information about the coordination of different articulators in vowel production.

Gentil and Moore (1997) and Stone (1997), however, outline some of the technical problems associated with EMG, which make data collection and interpretation potentially difficult. These include the challenge of accurate positioning of the electrodes to avoid obstructions such as bone or folds of skin, and the difficulty of pinpointing with precision the particular muscle or bundle of muscles to be investigated. Stone (1997: 32) cautions that in using EMG to explore aspects of lingual activity, which is obviously a focus of particular interest in the study of vowel production, because of the complex configuration of the lingual musculature “it is almost impossible to be sure that the signal comes from the muscle of interest.” It is also the case that the need for administration under medical supervision severely inhibits widespread availability and use of EMG. However, studies examining vowel production in normal and impaired speech production have been undertaken using EMG (e.g., Shankweiler, Harris, & Taylor, 1968; Alfonso & Baer, 1982; McGarr & Harris, 1983): Baer, Alfonso, and Honda (1988), Honda, (1996), and Gentil and Moore (1997) argue that the technique has a valuable role to play in the analysis of impaired speech production.

Electrolaryngography (ELG) derives a waveform indicating degrees of vocal fold approximation during the glottal cycle (Hirose, 2010). The efficiency of laryngeal function in voicing has consequences for the quality of resonance in the supralaryngeal chambers (Abberton, Howard, & Fourcin, 1989) that may also affect the perceptual analysis of vowel quality (Lotto, Holt, & Kluender, 1997), particularly at high levels of fundamental frequency (Sundberg & Gauflin, 1979).

The Acoustic Domain

Spectrography is the prime technique for making acoustic measurements of speech production. The speech spectrograph is capable of providing detailed quantitative information on a range of aspects of the speech waveform, including intensity, frequency, duration, and spectral analysis (Kent & Read, 1992). It is not surprising, therefore, that the technique has proved extremely popular for the investigation of normal vowel production, vowel development, and vowel disorders. As Baken (1987: 353) notes in a discussion of vowels, “the sound spectrograph, by making them readily observable, unleashed a flood of research into their characteristics and significance.”
Two aspects of vowel production have proved particularly popular in spectrographic analysis: formant structure and duration. The acoustic structure of vowel productions can normally be adequately specified by measuring the center frequency of the first three or four formant resonances (Fant, 1962) with the first two often being sufficient (Kent & Read, 1992), particularly in languages such as English, which do not contain vowel distinctions that depend solely on contrasts in lip position (Ladefoged, 2001). F1 values relate to tongue position in the vertical domain, and F2 values to the front-back dimension. Iskarous (2010) argues that, at least for steady-state vowel productions, formant structure can provide significant information on location and degree of lingual constriction, as well as presence or absence of lip-rounding. It is important to be aware that formant measurements and comparisons between them are to be seen as relative not absolute values, because they relate to the dimensions of the vocal tract that vary, of course, between individual speakers. There are methods of normalizing formant values to deal with this variation (Adank, Smits, & van Hout, 2004). Vowel duration values, meanwhile, are usually obtained by measuring from the onset of the second formant (F2) to its offset, although Blomgren and Robb (1998) draw attention to the difficulties in determining vowel duration. Farmer (1997) and Kent and Kim (2008) provide useful reviews of the use of spectrography in clinical phonetic analysis, revealing how it has been used to study vowel production across a wide range of speech impairments. In the present chapter we will also make extensive reference to spectrographic studies of vowel development and vowel disorders, as well as the production of vowels by normal speakers.

![Spectrograms of the English f l e e c e, b a t h, g o o s e, and p r i c e vowels produced between glottal stops in a near-RP accent, showing formant structure](image)
Measuring formants and tracking articulator movements are undeniably valuable sources of information that can shed light on why particular vowel productions sound the way they do, and, in the context of disordered speech, don’t sound the way we think they should. However, perceptual analysis is indispensable if we wish to regard vowel sounds, or any speech sounds, as speech *sounds*; that is, as phenomena that impinge on listeners’ consciousness and carry linguistic information. The question as to whether a particular vowel production *sounds* like a realization of English /i/ or /u/, for example, can only be answered by listening to it, not by measuring its formants or obtaining information about the speaker’s tongue and lip movements. For this reason we need to know how the auditory experience of sound compares to its acoustic structure.

Formants are clearly important determinants of our perceptual experience of vowels as shown, for example, by neurograms of the peripheral auditory system’s phase-locked responses to amplitude peaks corresponding to formant frequencies (Delgutte, 1997; Hayward, 2000). However, there is still uncertainty as to how this determination takes place and what other factors may be involved. Neel (2008), for example, cautions that in vowel identification tasks, listeners make use of a complex combination of information from durational values, and formant movement patterns, as well as static F1 and F2 values. Dusan (2007) presents similar evidence about the complexity of the acoustic patterns involved in vowel identification and classification. There is consensus in the psychoacoustic research literature that the auditory system treats the lower end of the frequency spectrum differently to the upper end, devoting approximately two thirds of its resolving power to frequencies below about 3kHz with the remaining one third having to contend with all the frequencies higher than this (for speech that means up to about 10kHz, see Figure 3.3). Perceptual-auditory scales have been developed to reflect this, the most commonly used being the Bark scale (also called the Z scale), which can be related to Hertz to show the non-linear nature of auditory processing (Zwicker & Terhardt, 1980). It is derived from experimental data that suggest that the basilar membrane behaves like a series of discrete but overlapping frequency-band detectors with the bandwidth increasing in size with increased frequency. The ERB-rate scale (equivalent rectangular bandwidth) is in effect very similar but is claimed to be more accurate for lower frequencies (Moore & Glasberg, 1983; see discussion in Hayward, 2000).

The 0–3kHz part of the spectrum contains the ranges in which F1 and F2 vary across vowel qualities (c.200–1000Hz and c.700–2500Hz.

---

**The Auditory Domain**

Measuring formants and tracking articulator movements are undeniably valuable sources of information that can shed light on why particular vowel productions sound the way they do, and, in the context of disordered speech, don’t sound the way we think they should. However, perceptual analysis is indispensable if we wish to regard vowel sounds, or any speech sounds, as speech *sounds*; that is, as phenomena that impinge on listeners’ consciousness and carry linguistic information. The question as to whether a particular vowel production *sounds* like a realization of English /i/ or /u/, for example, can only be answered by listening to it, not by measuring its formants or obtaining information about the speaker’s tongue and lip movements. For this reason we need to know how the auditory experience of sound compares to its acoustic structure.

Formants are clearly important determinants of our perceptual experience of vowels as shown, for example, by neurograms of the peripheral auditory system’s phase-locked responses to amplitude peaks corresponding to formant frequencies (Delgutte, 1997; Hayward, 2000). However, there is still uncertainty as to how this determination takes place and what other factors may be involved. Neel (2008), for example, cautions that in vowel identification tasks, listeners make use of a complex combination of information from durational values, and formant movement patterns, as well as static F1 and F2 values. Dusan (2007) presents similar evidence about the complexity of the acoustic patterns involved in vowel identification and classification. There is consensus in the psychoacoustic research literature that the auditory system treats the lower end of the frequency spectrum differently to the upper end, devoting approximately two thirds of its resolving power to frequencies below about 3kHz with the remaining one third having to contend with all the frequencies higher than this (for speech that means up to about 10kHz, see Figure 3.3). Perceptual-auditory scales have been developed to reflect this, the most commonly used being the Bark scale (also called the Z scale), which can be related to Hertz to show the non-linear nature of auditory processing (Zwicker & Terhardt, 1980). It is derived from experimental data that suggest that the basilar membrane behaves like a series of discrete but overlapping frequency-band detectors with the bandwidth increasing in size with increased frequency. The ERB-rate scale (equivalent rectangular bandwidth) is in effect very similar but is claimed to be more accurate for lower frequencies (Moore & Glasberg, 1983; see discussion in Hayward, 2000).

The 0–3kHz part of the spectrum contains the ranges in which F1 and F2 vary across vowel qualities (c.200–1000Hz and c.700–2500Hz.

---

**The Auditory Domain**

Measuring formants and tracking articulator movements are undeniably valuable sources of information that can shed light on why particular vowel productions sound the way they do, and, in the context of disordered speech, don’t sound the way we think they should. However, perceptual analysis is indispensable if we wish to regard vowel sounds, or any speech sounds, as speech *sounds*; that is, as phenomena that impinge on listeners’ consciousness and carry linguistic information. The question as to whether a particular vowel production *sounds* like a realization of English /i/ or /u/, for example, can only be answered by listening to it, not by measuring its formants or obtaining information about the speaker’s tongue and lip movements. For this reason we need to know how the auditory experience of sound compares to its acoustic structure.

Formants are clearly important determinants of our perceptual experience of vowels as shown, for example, by neurograms of the peripheral auditory system’s phase-locked responses to amplitude peaks corresponding to formant frequencies (Delgutte, 1997; Hayward, 2000). However, there is still uncertainty as to how this determination takes place and what other factors may be involved. Neel (2008), for example, cautions that in vowel identification tasks, listeners make use of a complex combination of information from durational values, and formant movement patterns, as well as static F1 and F2 values. Dusan (2007) presents similar evidence about the complexity of the acoustic patterns involved in vowel identification and classification. There is consensus in the psychoacoustic research literature that the auditory system treats the lower end of the frequency spectrum differently to the upper end, devoting approximately two thirds of its resolving power to frequencies below about 3kHz with the remaining one third having to contend with all the frequencies higher than this (for speech that means up to about 10kHz, see Figure 3.3). Perceptual-auditory scales have been developed to reflect this, the most commonly used being the Bark scale (also called the Z scale), which can be related to Hertz to show the non-linear nature of auditory processing (Zwicker & Terhardt, 1980). It is derived from experimental data that suggest that the basilar membrane behaves like a series of discrete but overlapping frequency-band detectors with the bandwidth increasing in size with increased frequency. The ERB-rate scale (equivalent rectangular bandwidth) is in effect very similar but is claimed to be more accurate for lower frequencies (Moore & Glasberg, 1983; see discussion in Hayward, 2000).

The 0–3kHz part of the spectrum contains the ranges in which F1 and F2 vary across vowel qualities (c.200–1000Hz and c.700–2500Hz.

---

**The Auditory Domain**

Measuring formants and tracking articulator movements are undeniably valuable sources of information that can shed light on why particular vowel productions sound the way they do, and, in the context of disordered speech, don’t sound the way we think they should. However, perceptual analysis is indispensable if we wish to regard vowel sounds, or any speech sounds, as speech *sounds*; that is, as phenomena that impinge on listeners’ consciousness and carry linguistic information. The question as to whether a particular vowel production *sounds* like a realization of English /i/ or /u/, for example, can only be answered by listening to it, not by measuring its formants or obtaining information about the speaker’s tongue and lip movements. For this reason we need to know how the auditory experience of sound compares to its acoustic structure.

Formants are clearly important determinants of our perceptual experience of vowels as shown, for example, by neurograms of the peripheral auditory system’s phase-locked responses to amplitude peaks corresponding to formant frequencies (Delgutte, 1997; Hayward, 2000). However, there is still uncertainty as to how this determination takes place and what other factors may be involved. Neel (2008), for example, cautions that in vowel identification tasks, listeners make use of a complex combination of information from durational values, and formant movement patterns, as well as static F1 and F2 values. Dusan (2007) presents similar evidence about the complexity of the acoustic patterns involved in vowel identification and classification. There is consensus in the psychoacoustic research literature that the auditory system treats the lower end of the frequency spectrum differently to the upper end, devoting approximately two thirds of its resolving power to frequencies below about 3kHz with the remaining one third having to contend with all the frequencies higher than this (for speech that means up to about 10kHz, see Figure 3.3). Perceptual-auditory scales have been developed to reflect this, the most commonly used being the Bark scale (also called the Z scale), which can be related to Hertz to show the non-linear nature of auditory processing (Zwicker & Terhardt, 1980). It is derived from experimental data that suggest that the basilar membrane behaves like a series of discrete but overlapping frequency-band detectors with the bandwidth increasing in size with increased frequency. The ERB-rate scale (equivalent rectangular bandwidth) is in effect very similar but is claimed to be more accurate for lower frequencies (Moore & Glasberg, 1983; see discussion in Hayward, 2000).

The 0–3kHz part of the spectrum contains the ranges in which F1 and F2 vary across vowel qualities (c.200–1000Hz and c.700–2500Hz.
respectively in adult males) and indicates that these formants will be resolved more clearly than the higher formants, which are known to play less of a role in vowel perception. However, things are not as straightforward as we might wish them to be in this respect. Frequency resolution decreases continuously, although not at a uniform rate, as frequency increases. One result of this is that the first three or four harmonics of a vowel are probably resolved separately, i.e., those that contribute their energy to F1. The question then arises whether F1 is a real percept, and if so, how are the harmonics integrated after resolution. A further complication is that although F2, F3, and F4 are resolved separately in peripheral processing, in central auditory processing they become integrated when F2 is high (Bladon, 1983; Hayward, 2000). What may actually be perceived in this case is referred to as F2' (F2 prime) and is explained as the weighted integration of F2, F3, and F4 (Chistovich, Sheikin, & Lublinskaja, 1979). By applying Z and ERB-rate scales to spectrographic data it is possible to derive cochleagrams that purport to represent the spectral components of vowels in a more perceptually realistic way (Johnson, 2003).

When relating spectral and spectrographic data to vowel qualities it is as well to bear these factors in mind.
Perceptual Analysis

Perceptual analysis is no doubt the most commonly used source of information about speech sounds, particularly in clinical settings, and Stoel-Gammon and Pollock (2008) note that most studies of vowel development and disorder employ phonetic transcription. However, while listening may seem an easy, well-practiced and low-budget procedure it can be attended by a number of pitfalls awaiting the unwary analyst. If developmental or disordered vowel data collected by perceptual analysis are to be valid and reliable, some of these difficulties need to be taken into account. For example, even at the level of deciding how to sample and/or record speech data, we need to make informed decisions about the advantages and disadvantages of different approaches. We need to be aware that in recorded speech the fidelity of the acoustic signal isn’t always high—this is often the case where recordings have been made in SLT clinics or schools due to poor equipment, problems with positioning of microphones, excessive background noise, etc. Should transcription, then, be carried out live in situ or should use be made of audio and/or video recordings? For speech data generally and for the disambiguation of aspects of vowel production such as lip posture and jaw position we would be well-advised to take note of Kelly and Local (1989: 35), who caution that “in doing phonetic transcription it is important to pay attention to at least part of what a speaker can be seen to be doing” (our italics), so we would wish to have some recourse to visual data in our collection of the speech sample. But we also need to be aware that one condition that provides visual information, real-time speech production, simply happens too quickly for us to be able to make reliable perceptual judgments about it. Indeed, Amorosa, von Benda, Wagner, and Keck (1985) argue that live transcription unsupplemented by subsequent reference to audio or video recordings cannot provide a reliable basis for even the most superficial of phonological analyses of developmental speech disorders. They observe that in their data the time pressure of live transcription resulted in a significant “normalisation” of the data towards the expected forms.

This, in turn, brings us to another problem with transcription; the effect of the listener’s expectations of the speech output on the accuracy of their transcription. Laver (1994: 556) remarks that the most challenging type of speech output for listeners to transcribe is material from their own native language, because of the strength of the effects of the native phonological system: “listeners tend to force the new material through the perceptual grid of the phonological categories of their own language.” This will apply, of course, to all speech data, including speech
that differs from normal native language categories by virtue of being immature, impaired, or merely from a different social or regional accent from that of the listener (for further discussion of transcription of clinical data, see Heselwood & Howard, 2008). This not only encourages the phenomenon of phonemic false evaluation (Buckingham & Yule, 1987), where speech material that does not readily conform to native segmental categories may nonetheless be wrongly assigned to one of those categories and thus inaccurately transcribed, but it also means that where the listener knows the target word or utterance, they are likely to “normalize” or transcribe towards that form (Oller & Eilers, 1975). Ingrisano, Klee, and Binger (1996: 46) note, rather alarmingly, that “transcribers in contextual conditions don’t seem to recognise utterances as unintelligible, rather, they loosely transcribe ‘what it is they think they heard,’” and Oller and Eilers (1975) suggest that the effect of knowing the target utterance can lead listeners to add elements to, or omit elements from, their transcription when there is no supporting evidence in the acoustic data. This effect can be seen in the perceptual analysis of vowel data, where, for example, Pye, Wilcox, and Siren (1988) note that the vowels [ɪ] and [ə] when they occur in unstressed syllables are frequently added or omitted in the transcription of children’s speech.

The literature on the perceptual analysis of vowels is not extensive, reflecting in part an apparent imbalance between transcription practices for consonants and vowels. A number of authors have observed that it is frequently the case that consonants are subject to narrow transcription where vowels, even in the same data, are transcribed broadly (Crystal, 1982; Grunwell, 1987; Ball, 1988; Ball, 1991). Butcher (1989) suggests that vowels have not customarily been transcribed in detail because of the technical difficulties of transcription, in comparison with consonants, and Vieregge and Maassen (1999) support this position, noting that in atypical developmental data, vowels are harder to transcribe than consonants. In a detailed investigation of transcription reliability for developmental speech disorders, Shriberg et al. (2010) report higher agreement for vowels than consonants where broad transcription is concerned: conversely, in narrow transcription, the situation is reversed, with lower inter- and intratranscriber agreement being reported for vowels, even with experienced transcribers. The general avoidance of narrow vowel transcription may be in part responsible for the widespread belief that vowel impairments are rare. Ball (1991) argues that the routine narrow transcription of vowels may prove extremely important in the analysis of disordered sound systems in order to capture clinically significant consonant-vowel interactions and coarticulations. To this end, Ball, Müller, Rutter, and Klopfenstein (2010) provide extensive
advice on the transcription of vowels for clinical purposes, while Teoh and Chin (2009) include vowels in their discussion of the importance of using narrow phonetic transcription for the speech production of individuals with a cochlear implant. Local (1983), furthermore, in a single case study of vowel development in a child from the north-east of Britain, demonstrates convincingly how the use of narrow transcription can illustrate subtle but significant variation in vowel productions which, as he suggests, is “something to be accounted for and not something troublesome to be got rid of at any cost.”

In terms of the level of perceptual difficulty of different types of vowel, Pye et al. (1988) suggest that diphthongs are relatively easier to transcribe than monophthongs, an observation that may relate in part to the comment by Norris, Harden, and Bell (1980) that in general segments with longer durations are easier than their shorter counterparts. Maassen, Offeringa, Vieregge, and Thoonen (1996) mention that in their studies of the transcription of developmentally disordered speech production in Dutch, low vowels were easier to transcribe than central vowels, an observation that has intriguing parallels with the normal developmental order of emergence of vowels in speech development (Stoel-Gammon, 1985; Kent & Miolo, 1995) and also with the frequency of occurrence of peripheral versus central vowels in the languages of the world (Ladefoged & Maddieson, 1996; Schwartz, Boë, Valée, & Abry, 1997). Shriberg et al. (2010) make the important observation that vowels are more challenging to transcribe in conversational speech, where they are subject to various types of reduction, than when elicited in isolation or in single words in the context of formal clinical assessments. An interesting study by Lotto, Holt, and Kluender (1997) suggests that voice quality can affect the perceptual analysis of vowels, with listeners in their study identifying significantly more tokens of synthesized vowels with breathy voice quality as high vowels, in comparison with vowels having modal voice quality. It is worth noting that non-phonetic factors may also influence the perception of vowels, with studies demonstrating that in vowel identification tasks listeners are influenced by their beliefs about the gender, age, and/or social class of the speaker (Johnson, Strand, & D’Imperio, 1999; Hay, Warren, & Drager, 2006).

Despite the problems outlined above, we have to recognize that perceptual analysis is important for two reasons. First, it completes the bridge between the speaker and the hearer in the sense that without perceptual judgements we are dealing with phenomena devoid of communicative value—we don’t speak palatograms or hear spectrograms, nor is a vowel simply the sum of the measurements we can make in the various domains. Second, it engages us more fully with the data so we
are less likely to miss significant details and more likely to detect possible patterns that we can then go on to investigate instrumentally if we think it might prove fruitful to do so. To turn one’s back on perceptual analysis because of its methodological imperfections is tantamount not only to deciding not to listen to the data and to rely entirely on instruments, but also to deciding that what the speech sounds like is unimportant (Heselwood & Howard, 2008; Heselwood, 2009). We would do well always to remember that the role of instruments is to fill in the details in certain parts of the whole picture, not to define what that whole picture is.

DESCRIPTION, CLASSIFICATION, AND TRANSCRIPTION OF VOWELS IN NORMAL AND DISORDERED SPEECH

When we wish to denote a vowel in transcription, there are four basic questions to consider.

1. What taxonomic framework should we use for classifying vowels?
2. What does a vowel symbol denote in terms of articulatory, acoustic, and auditory properties?
3. Should we use vowel symbols with language-independent values or the values they have in the accent of the speaker’s speech community?
4. Are the available transcriptional conventions adequate for normal and clinical vowel data?

**Description and Classification**

“What are phoneticians really doing when they describe a vowel sound by allocating it to a certain box in their scheme of categories, or a certain point on their vowel diagrams?” (Ladefoged, 1967: 53)

To begin with, it may be useful to distinguish between description and classification, a distinction that isn’t always made yet has some important implications. We can describe sounds in as little or as much detail as we like, and we can use any property as a descriptor. For example, [a] can be described just as “sonorant,” or has having an F2 value of around 1720Hz in adult males (Peterson & Barney, 1952). Alternatively, we can attempt to give a near-exhaustive description of the whole vocal tract during its production, including such observations as pharyngeal width, position of tongue tip, convexity of the tongue surface, etc.; description can also include dynamic accounts of vocal tract activity during speech.
production. Classification, on the other hand, takes an essentially static view. It requires a parsimonious set of categories to which vowels and consonants are assigned, and those categories have to be set up according to certain principles. For speech sounds, categories are based on judgements as to what the significant aspects of their production are: pharyngeal width, for example, is not regarded as significant in the production of \( [t] \) but position of the tongue tip is, so it is incorporated into the classification.

It should be clear that as far as description is concerned there is no reason to treat consonants and vowels any differently: the same vocal organs are involved in both and their actions can be described in the same terms. However, they have traditionally been assigned to categories in different classificatory frameworks because of particularities in their production. Consonants involve identifiable strictures, and both the location and the degree of that stricture have been set up as classificational criteria with a nomenclature derived largely from the superior speech organs—alveolar, palatal, velar, etc. By contrast, vowels have been classified according to the location of the highest point of the tongue in a two-dimensional space defined by the axes close-open (or high-low) and front-back. A third classificatory dimension is provided by the rounded-spread axis of lip posture. The reason for this difference is probably twofold: the highest point of the tongue for many vowels is too far away from any of the superior organs for its approximation to them to be readily conceptualized as a stricture, and the quality of the vowel cannot be attributable simply to that approximation—rather it is attributable to the distribution of volume throughout the whole supralaryngeal vocal tract. In fact, this last point applies to sonorant consonants too, and also in a modified sense to obstruents insofar as the vocal tract dimensions in front of the stricture contribute to the quality of aperiodic sound in stop-bursts and fricatives (Stevens, 1998).

An account of attempts at vowel classification in European, and especially British, phonetics dating back to the early 17th century is given by Ladefoged (1967), showing how early descriptions were limited in the number of descriptive parameters which they used. It was not until the late 19th century that Bell (cited in Ladefoged, 1967) devised a method of description that included categories for tongue position along both the vertical and horizontal parameters and also made reference to the position of the lips. The most well-known framework for the classification of vowels, however, is probably the Cardinal Vowel system. Although the system has attracted much criticism over the years, not least from Butcher (1982: 50) who has described it as “theoretically inadequate and scientifically redundant,” it is still widely...
used today, and Ladefoged (1993) observes that it has aided in a more precise description of accents and languages of the world than any other method currently available. The Cardinal Vowel system was developed in the early 20th century by Daniel Jones to replace the system of three-term articulatory labels—close, back, rounded, etc.—which had proved too imprecise for adequate classification (Abercrombie, 1967). Jones was motivated by the observation that the consonants of a foreign language “are as a rule best acquired by directing attention to tactile and muscular sensations, whereas in learning vowels it is necessary to direct attention more particularly to the acoustic qualities of the sounds” (Jones, 1972: 26, original italics). He began by defining two articulatory positions at opposite ends of the vocal tract’s vowel space and noting the vowel qualities associated with them (see Table 3.1).

From these he derived six further vowel qualities giving a front series of four and a back series of four, which he termed the primary Cardinal Vowels (Table 3.2), and which he plotted onto a vowel quadrilateral (see Figure 3.4).

Both the front and back routes from [i] to [a] involved what Jones described as equal degrees of “acoustic separation,” i.e., the steps from one vowel to the next should be such as to be judged auditorily equal. Further Cardinal Vowels were derived by changing the lip shape from spread to rounded and vice versa, or from neutral to rounded, and by specifying vowels in the central areas of the vowel space, all these together being called the Secondary Cardinal Vowels.

There is, however, as Ladefoged (1967) has noted, an ambivalence in Jones’ account. On the one hand he stresses the acoustic relations as the defining ones and is adamant that “The values of Cardinal Vowels cannot be learnt from written descriptions; they should be learnt by oral instruction from a teacher who knows them” (p. 34), but on the other he is drawn into specifying for each vowel in the series what he terms an

<table>
<thead>
<tr>
<th>Articulatory Position</th>
<th>Vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>tongue: close front; lips: spread</td>
<td>[i], CV no. 1</td>
</tr>
<tr>
<td>tongue: open back; lips: not rounded</td>
<td>[a], CV no.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Articulatory Position</th>
<th>Vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>tongue: close front; lips: spread</td>
<td>[i], CV no. 1</td>
</tr>
<tr>
<td>tongue: open back; lips: not rounded</td>
<td>[a], CV no. 5</td>
</tr>
</tbody>
</table>
“approximate” tongue and lip position but which is given with rather more exactness than “approximate” suggests. He also compares the degree of tongue movement required in the front and back series to attain acoustic equidistance, and goes so far as to refer to the Cardinal Vowels as “a set of fixed vowel-sounds having known acoustic qualities and known tongue and lip positions” (p. 1972: 28, italics added). Ladefoged and Maddieson (1996) remind us, however, that tongue height is a difficult parameter to use with confidence, both because for the back vowels it is not only the height of the tongue, but also the relative height of the soft palate that must be taken into account, and also because, as Stevens and House (1955, 1961) have pointed out, the location of the most significant tongue stricture in many back vowels is not at the point of maximum lingual elevation, but occurs in the pharyngeal cavity. Wood (1979) uses x-ray evidence to confirm that these constrictions lie outside the “vowel space” originally proposed by Jones. More recently, Esling (2005) has proposed a three-way categorization of vowels into front, raised, and retracted, in which those vowels traditionally categorized as low back are reclassified as retracted.

It might also be argued that the lip positions for the Cardinal Vowels are underspecified in Jones’ original accounts. He describes Cardinal Vowels 1 to 5 as having spread or neutral lips, and Cardinal Vowels 6, 7, and 8 as having open, indeterminate, and close lip-rounding respectively. This corresponds to the observation that the degree of lip-rounding in the Cardinal Vowel system correlates quite closely with the height of the tongue (Catford, 1988; Ladefoged & Maddieson, 1996). However, Catford goes on to point out that it is useful to distinguish between

![Figure 3.4 Traditional vowel quadrilateral, showing primary and secondary Cardinal Vowels (primary to left, secondary to right, rounded vowels in brackets)](image-url)
different types of lip-rounding: he associates endolabial rounding (lip-pouting, with the rounding formed by the inner surfaces of the lips) with back rounded vowels, and exolabial rounding (lip-pursing, with the rounding formed by the outer surfaces of the lips) with front rounded vowels. These different types of rounding are useful to note in the description of vowels in different accents and languages. For example, Iivonen (1994) describes significant differences in the types of rounding found in Swedish [yː] and [uː].

Given these ambivalences and complications, it is hardly surprising that the practice has remained of interpreting the Cardinal Vowels according to precisely those same three articulatory dimensions that were originally deemed inadequate. In fact, the current IPA vowel quadrilateral is generally seen as an articulatory space rather than an auditory one, and the arrangement into rounded and unrounded series rather than primary and secondary, although in practice only affecting the open back vowels, is a shift towards a more explicitly articulatory framework.

What is common to the traditional classificatory schemes for consonants and vowels, then, is their articulatory basis even where for vowels acoustic–auditory principles have been introduced. It would therefore seem an advantage, in principle at least, if they could be combined into one framework. Catford (1977) shows how this could be done using a system of “polar coordinates” in which Cardinal Vowels 1, 8, 7, 6, and 5 form a series of “narrow approximants” corresponding with consonantal places of articulation (see Table 3.3).

Further vowels are derived by progressively opening the approximation, as in the Cardinal system, from close through half-close and half-open to open (see Figure 3.5). An important difference between Catford’s scheme and the Cardinal Vowel scheme of Jones, however, is that these terms do not have the same interpretation. For example, [ɑ] is classed as open by Jones but as closed by Catford. The difference lies in how the

---

**Table 3.3** Catford’s Vowel System

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Place of Articulation</th>
<th>Homorganic Voiced Fricative</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>palatal</td>
<td>ʝ</td>
</tr>
<tr>
<td>u</td>
<td>advanced velar</td>
<td>ɣ</td>
</tr>
<tr>
<td>o</td>
<td>velar</td>
<td>ɣ</td>
</tr>
<tr>
<td>œ</td>
<td>uvular</td>
<td>ʁ</td>
</tr>
<tr>
<td>a</td>
<td>pharyngeal</td>
<td>ʕ</td>
</tr>
</tbody>
</table>
tongue is seen to occupy the vowel space. For Jones, the significant fact is that in the vertical plane the highest point of the tongue is low down compared, say, to [u] and the jaw is open, but Catford's proposal invites us to regard the proximity of the tongue root to the rear pharyngeal wall as the salient feature, and this is similar to the tongue dorsum’s proximity to the velum in the production of [u].

However, Catford advances acoustic reasons concerning the near-universality of the [i, u, a] point vowel triangle, and physiological reasons concerning the proprioceptive discreteness of tongue-raising and tongue-retracting muscles, to explain why this system is not as advantageous as it might appear. He concludes by saying that “we must continue to treat vowels differently from consonants for purposes of practical classification. It is equally clear, however, that from a purely theoretical point of view vowels can be well fitted into the normal taxonomic parameters of location and stricture type if we wish to treat them this way” (1977: 186–187). Catford’s scheme does not address the relationship of articulatory behaviors to the acoustic or auditory domains and it is the difficulty of establishing such relationships, which to some extent justifies regarding vowels as importantly different from consonants.

The difficulty in addressing auditory–articulatory relationships is exemplified in the use and interpretation of vowels symbols. For example, a vowel symbol such as [a] represents a particular auditory quality but may have rather different articulatorily-motivated labels, depending
on the classification scheme one adopts. In the IPA scheme, [ɑ] is described as “open back”, in Catford’s scheme it is “close pharyngeal” and in Esling’s it is “retracted.”

**Articulatory, Acoustic or Auditory?**

“Phoneticians are thinking in terms of acoustic fact and using physiological fantasy to express the idea.” (Russell, 1928, q. in Ladefoged, 1967: 72)

It is important to be clear about the information value of a vowel symbol. Is it to be understood as denoting the speaker’s vocal tract configuration, its acoustic output in terms of formant resonance patterns, or the phonetically-trained listener’s auditory impression? Several writers have commented that vowels are auditory qualities to which we attach articulatory labels (Ladefoged, 1975; Catford, 1977; Ball, 1993) and we have seen how even Daniel Jones was unable to really free himself from this practice. Most of the time the inference from auditory impression to tongue and lip positions in normal speech is likely to be accurate enough for our purposes and can to a significant but limited extent be objectively validated by formant frequency measurements: There is a robust inverse relation between the height of F1 and jaw height (Keating, 1983), although this may be less true in the case of back vowels (Stone, 2010), and F2 frequency is indicative of the position of the bulk of the tongue along the front-back axis of the vowel space (Stevens, 1998). It is these relationships that have made it possible to construct formant charts that map onto the traditional vowel quadrilateral (Ladefoged, 1975), but there are often practical problems in identifying formants accurately, particularly among close and back vowels (Ladefoged, 1967), i.e., the set of vowels that fall into Catford’s “narrow approximant” class (Catford, 1977). Ladefoged (2001) notes that for British and American English F3 is fairly predictable from the values for F1 and F2 for any given vowel, but that in languages that, unlike English, produce vowel contrasts that have the same tongue position but which are distinguished by different lip positions (for example Swedish: Iivonen, 1994), F3 assumes a greater importance in looking at the acoustic identity of vowels.

It has, in addition, to be noted that the role of F2 as a determinant of auditory quality differs between front and back vowels. If F3 is within 3.5 Bark of F2—i.e., if the responses on the basilar membrane to the two formants are within three and a half critical bands (about 4.2mm in a mature adult) of each other—then what is perceived in this frequency region is a single F2’ (see above). F3 is only within 3.5 Bark of F2 in front
vowels. In back vowels F3 is further away from F2 and does not contribute to the F2 percept (Stevens, 1997, 1998).

Insofar as the articulatory–acoustic–auditory relationships are constant it doesn’t much matter if auditory impressions are expressed in articulatory terms, or indeed if the articulatory terms are, in Russell’s words, a fantasy. But how constant are they, and how safely can we infer from one domain to another? An experiment on the auditory judgement of lip posture in vowels (Lisker, 1989) revealed significant inaccuracies with, for example, unrounded [ɯ] (Cardinal 16) judged rounded in 63% of responses, and rounded [y] (Cardinal 9) judged to be rounded in only 44%, a result that reinforces fears regarding intersubjective disagreements and the reliability of vowel transcriptions. The problem is compounded by the fact that a given set of formant frequencies can be produced by more than one vocal tract shape (Ladefoged, Harshman, Goldstein, & Rice, 1978), and Stevens (1989) offers evidence to show that articulatory tolerance is not monotonic with respect to different zones of the vowel space. The point (or quantal) vowels—that is, [i], [ɑ], [u]—according to Stevens, remain acoustically stable across a wider range of articulatory variation than non-point vowels.

Perkell attributes the many-to-one relation between vocal tract shape and acoustic output to the phenomenon of “motor equivalence” (1997: 366) meaning that deflections away from a target position for an articulator can be corrected for by compensatory controlled adjustments of other speech organs. A study of normal vowel production in German by Maurer, Grone, Landis, Hoch, and Schönle (1993) using electromagnetic articulography (EMA) confirmed that significantly different vocal tract configurations can produce perceptually identical timbres. They also found that very similar configurations can be responsible for significantly different formant patterns as a function of a varying Fo (thus agreeing with some previous research findings, e.g., Carrell, Smith, & Pisoni, 1981), and that the degree of articulatory movement is not directly or easily related to degree of perceived change of vowel quality. These findings lead them to state that “there is no evidence of an imperative relationship between the vocal tract shape and the vowel identity in real vocalizations” (p. 141).

Some caution must be exercised, however, before accepting the validity of Maurer et al.’s (1993) conclusion that the source-filter theory of vowel production (Fant, 1960; Stevens & House, 1961) is contra-indicated by their results. It was in sustained isolated vowel productions that they claim to have found a lack of evidence for predictable relationships between vocal tract shape and vowel quality. The instruction to the two subjects in these productions was to move the articulators around while
trying to keep vowel quality constant. In a different task, where subjects repeated lexical items in response to the experimenter’s productions (rather misleadingly described as spontaneous speech by the authors), a much more robust relationship was found. From this we might conclude rather that there are a number of different configurations that will produce similar output involving adjustments not measured or considered by Maurer et al., (1993) e.g., altering larynx height, lip and tongue posture in the lateral dimension, changes of shape and volume in the oro- and laryngo-pharynx, etc. (Laver, 1980), but that most of these configurations are not habitually used for producing vowels in spoken German because in speech much shorter time-slots are available than in deliberately sustained isolated vowels and configurations must accommodate to consonantal context. That is to say, the temporal constraints operating in normally-spoken syllables may favour the use of articulatory-acoustic relationships that are more straightforwardly predictable by the source-filter model. Interesting in this respect is Strange’s (1989) conclusion that steady states are not crucial for vowel identification and that vowels may be dynamically specified. In fact vowels rarely attain steady states in speech due to the formant transitions caused by neighbouring consonants. We therefore need to consider whether, from an acoustic point of view at least, the term “monophthong” is strictly applicable to most vowel data. It does seem nonetheless to retain its validity in the subjective realm of perception, suggesting that transitions may not affect the stability of vowel percepts in central auditory processing. As a complication, however, it has been noted that monophthongs are harder to transcribe than diphthongs (Pye et al., 1988).

Studies of electromyographic (EMG) vowel data by Honda and associates, which were produced in a [əp_p] context, have led to the claim that “Vowels have analogous motor and sensory representations in the anterior and posterior cortical areas” (Honda, 1996: 49), and that the relations between them are robust (Maeda & Honda, 1994). Robustness of this relation, far from being compromised by many-to-one mappings, may in fact rely on them. Perkell states that at least for [i] and [ɑ] “acoustic stability is achieved by virtue of a non-linear relation between constriction location and vowel formants, in combination with a physiologically-based non-linear relation between motor commands and the degree of constriction” (1996:20, original italics). In cases of congenital and acquired physical abnormalities, speakers may explore these non-linear relations to discover ways of producing the right acoustic/auditory results. Thus, although Whitehill, Ciocca, Chan, and Samman (2006) provide detailed acoustic evidence of limited tongue movements in the front-back dimension in speakers following a glossectomy, Ferrier,
Johnston, and Bashir (1991) describe auditorily-acceptable realizations of /i/ by a child with insufficient tongue length to attain a high front articulation in the region of the hard palate, and Morrish (1984, 1988), reporting on adult speakers following glossectomy, describes subtle adjustments of the jaw and pharynx that facilitate auditory vowel distinctions. Laaksonen, Riger, Haaponen, Harris, and Seikaly (2010), meanwhile, also identify compensatory and adaptive behaviors in vowel production in individuals following reconstructive surgery for the tongue after oral cancer. Temporary changes to the vocal tract have also been shown to bring about compensatory articulatory changes aimed, apparently, at maintaining the auditory qualities of vowels. Hamlet and Stone (1976), for example, identified patterns of apparently unconscious compensatory behavior in normal adult speakers after the insertion of dental prostheses. In comparison, a study by Daniloff, Bishop, and Ringel (1977) demonstrated a lack of compensatory features in children’s vowel articulations under conditions of oral anaesthesia. Vowels became less well-differentiated and typically were more centralized, a pattern that has also been reported in vowel articulations in a number of speech disorders, including speech associated with hearing impairment (Ange-locci, Kopp, & Holbrook, 1964; Dagenais, & Critz-Crosby, 1992), stammering (Blomgren, Robb, & Chen, 1998), and long-term tracheostomy (Kamen & Watson, 1991).

In the light, then, of research into articulatory–acoustic–auditory relationships, and the fact that vowels only have linguistic value when heard by a listener, it would seem prudent and appropriate to use vowel symbols primarily to denote auditory qualities. In normal vocal tracts the articulatory correlates can be recovered approximately through application of the acoustic theory of speech production, and, if necessary, more precisely through the use of instrumentation. Where we have good reason to believe that a speaker’s vocal tract is structurally or functionally incapable of assuming typical articulatory configurations for certain vowels, then instrumental investigation of the articulatory domain is essential if we are to have any useful information about his/her vowel productions. That is to say, we need to recognize those conditions under which inferences about vocal tract configurations based on auditory impressions and acoustic measurements are less reliable. If there is little or no tongue then any symbol used to denote a vowel quality obviously cannot be interpreted in terms of tongue position (Barry & Timmerman, 1985).

Taking a developmental perspective, there will be an impact on the acquisition of vowels and vowel oppositions if the articulatory–acoustic–auditory relationships do not remain balanced over the
normal course of anatomical growth from birth through the language acquisition period. In abnormal patterns of growth any effect of this kind may be amplified making it more difficult for stable representations of vowel sounds to become established. Any model of the development of an infant’s articulatory–auditory feedback loop (Locke & Pearson, 1992) has to take account of the relationship between the auditory and vocal motor systems in terms of anatomy, physiology, and neurology. We know that vocal tract architecture changes quite extensively over the first four or five years of life (Buhr, 1980; Kent, 1992; Vorperian & Kent, 2007; Beck, 2010). Indeed, Lieberman, Crelin, and Klatt (1972) anticipate many later researchers in observing that the vocal tracts of human neonates more closely resemble those of other primates, such as apes and chimpanzees, than those of adult humans. Buhr (1980) lists the main differences between the vocal tracts of new-borns and adults as follows: The human neonate has a relatively high larynx (so high, in fact, that for the first six months of life the epiglottis makes contact with the soft palate; Beck, 2010), and consequently a reduced pharyngeal space; typically the tongue is also relatively large and thus fills the oral cavity quite extensively; and, significantly, the “right-angle” configuration of the adult vocal tract, where the lip-to-velum area is on a horizontal plane relative to the vertical velum-to-larynx area, is not present in the neonate or infant vocal tract where there is a much more continuous, horizontal configuration of the tract from lips to larynx. These observations lead Buhr to suggest that acoustic measurements of F1 and F2 are more useful in describing infant vowel productions than articulatory categories such as high, mid, low, or front and back. As well as these anatomical differences, which resolve towards the adult geometry over the first four or five years of life, there are neuromuscular immaturities that limit the range and accuracy of independent movements of the tongue, lips, and mandible (Fletcher, 1973; Beck, 2010). The infant’s task, which may have profound implications for vowel production, is to achieve independent but coordinated movement of the different articulators over a period when the vocal tract is undergoing continual changes in shape and dimensions (Vorperian & Kent, 2007). Lieberman (1980) is among those who have postulated an innate mechanism for “normalization” during this period, to allow the infant to accommodate to vocal tract changes and to master normal vowel articulations during this period of anatomical, physiological and perceptual development.

A number of authors suggest that the patterns of typical vowel articulations in early pre-speech and speech behaviour can be linked to a lack of articulatory independence, particularly of the tongue and mandible. Buhr (1980) notes that the front vowels, [i, e, ɛ], which seem to occur
early in speech development, can be produced without independent lingual movement, merely by modifications of the position of the mandible, in comparison, for example, with the later developing vowel [u], which requires integration of lingual, labial and mandibular activity. Fletcher (1973), Kent and Murray (1982), and Davis and MacNeilage (1990) also underline the important role of the mandible in early vowel production and to the vowels found in early babbling, and Steeve (2010), in a single case study of an infant followed from eight to 22 months, observes that jaw kinematics for vowel babble differ from that found in non-speech activity such as chewing. We must bear in mind, however, that this mandibular activity may not be intentionally producing differentiated vowel sounds in early output, a possibility supported by the observation of Meier, McGarvin, Zakia, and Willerman (1997) that interspersed in early babbling sequences are silent rhythmic mandibular movements or “jaw wags,” which seem to correspond motorically with audible babbled syllables, but have no auditory consequences. Fletcher (1973) suggests that the infant tongue is limited to the “thrusting and rocking” movements associated with sucking and swallowing and that neuromuscular immaturity may prevent the early development of other lingual movements necessary for the production of different vowels. Similarly, Kent (1992) suggests that early tongue movements are likely to be in a horizontal (anterior-posterior) plane, rather than in the vertical plane that would involve raising and lowering of the tongue body. Given the large body of literature that emphasizes the significant role of vocal tract development and change on the production and development of vowels in infancy, it is useful to also weigh conflicting evidence. Clement and Wijnen (1994), studying Dutch speech development, found very similar patterns in the limitations on vowel productions in normal two-year-olds and in four-year-olds with developmentally delayed sound systems, which in turn were different from both normal four-year-olds and from adults. As the authors surmised that significant vocal tract maturation would have taken place by four years old, they argued that the immature speech output patterns in these children could not be accounted for by anatomical factors. Nittrouer, Studdert-Kennedy, and Neely (1996) also argue that the acoustic patterns of consonant–vowel interaction observed in children of different ages cannot be explained solely by differences in vocal tract anatomy or dimensions and a number of studies emphasize the influence of the ambient language on the emergence of early vowel productions (see, for example, Stokes and Wong, 2002; Rvachew, Alhaidary, Mattock, and Polka, 2008; Lee, Davis, and MacNeilage, 2010). We can therefore conclude that the significant anatomical differences between the infant and adult vocal tracts pose a real
problem for the perceptual analysis of early vocalization and early vowel productions (see below), but that anatomical factors are not the only story.

At the other end of the loop, the ear is affected by changes during infancy and childhood. In the outer ear for example, compared to adults, infants have a much shorter external canal leading to a smaller tympanic membrane set at a more oblique angle (Lowry, 1978); the shorter canal has a higher natural resonance frequency and will impose a different transfer function on incoming acoustic stimuli (Keefe, Burns, Bulen, & Campbell, 1994). It isn’t known exactly when the transfer function becomes fully mature but it may be as late as seven or eight years (Werner & Marean, 1996). Similarly, regarding the middle ear, it isn’t known when the impedance characteristics of the ossicular chain reach mature levels but research shows differences of up to 8dB in conduction in parts of the frequency spectrum at two years of age (Werner & Marean, 1996). In the inner ear the third of the three rows of outer hair cells in the organ of Corti is not present at birth (Atkinson, Barlow, & Braddick, 1982); the function of outer hair cells is to increase the sensitivity of the inner hair cells (Stach, 1998). Norton and Widen (1990) report age-related differences in otoacoustic emissions persisting to 13 years. Myelination of the auditory nerve fibers is not complete until four years of age (Tanner, 1989), which may mean that transmission of information to the auditory cortex is relatively less efficient. Although these immaturities of the auditory system are minor by comparison with those of the vocal tract and motor abilities, they do introduce variables into the articulatory–auditory relationship from the auditory end such that we perhaps cannot regard an infant or child’s auditory capacity as homeostatic, nor regard as constants the constraints it exerts on the less consistent behaviours of the articulators. On the other hand, perhaps we should not make too much of these auditory immaturities. A study by Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy, and Mehler (1988) indicated that new-born infants can extract information from certain vowel stimuli accurately enough to recognize them on subsequent presentations. Findings of this nature corroborate other research evidence that infants are particularly tuned in to vocalic contrasts (Kuhl & Miller, 1982).

The fact that there are developmental variables at both ends of the articulatory–auditory loop raises the question as to what extent there might be a significant mismatch between, say, an infant’s auditory impressions of his/her own vowel sounds at 12 months and at 24 months. Of course this will not be the only mismatch the infant has to contend with. There will be gross acoustic differences between the infant’s vowels
and those of adults due to differences in vocal tract dimensions with formant values being up to twice those of adult vowels (Kent & Murray, 1982). This particular difficulty may be avoided by a normalization process (Kent, 1992). However, normalization would not so easily account for differences in the infant’s own productions over time, and in fact the concept of normalization as a necessary part of speech processing has been questioned (Pisoni, 1997).

In addition to the effects of physical growth and development there is the problem of knowing when we can attribute phonological intentionality to early vocalizations and start to relate them in a meaningful way to the “target” vowels in real lexical items. One problem here is that early words are produced as gestalts, which cannot justifiably be analyzed at a segmental level (Bates, Dale, & Thal, 1995) meaning that the vowel is not separable from the rest of the syllable and so cannot be seen as the intentional realization of a vowel phoneme. Furthermore, how can we reliably distinguish between babbled syllables and gestalt words, and what problems might children have moving through this phase of development? Locke (1993: 89) has an interesting perspective on this latter point. He argues that infants may have difficulty in moving from “vowel-as-voice to vowel-as-voice-and-linguistic-unit” because of the prevalence of vowel-like articulations to express emotion and affect in early vocalizations. Indeed a number of authors have posited a discontinuity between babbling and speech for vowels as opposed to consonants (Davis & MacNeilage, 1990), but not all evidence supports this view. For example, Boysson-Bardies, Halle, Sagart, and Durand (1989) point to the early phonological differentiation of vowels in their cross-linguistic study showing how ten-month-old infants from different language backgrounds were already showing consistent language-specific patterns in their use of vowels, a finding supported by Rvachew et al. (2008). In addition, Blake and Fink (1987) used narrowly transcribed phonetic data to argue that the relation between sound and meaning emerges gradually during the transition period from babbling to speech. However, as Morris (2010: 227) notes, “[p]honetic transcription of young children’s speech is difficult,” and researchers have questioned the methodological validity of applying IPA transcriptional conventions to early infant vocalizations and babbling (Oller, 2001; Morris, 2010) and other approaches are sometimes adopted. McGowan, Nittrouer, and Chenausky (2008), for example, used broad categories (voiced, nasal, egressive, whispered) for vowel classification in their study of the early vocalizations of children with a hearing impairment, reporting a reasonable level of listener agreement, with most disagreements related to the presence or absence of nasalization.
Cardinal Qualities or Accent-Specific Qualities?

Besides the problem of what exactly is the information value of a vowel symbol, there is the question of the relationship of the symbol to the vowel space and its quadrilateral projection. For example, the symbols for the English vowels are also used for Cardinal and IPA vowels but the vowel qualities are quite noticeably different. The symbol for Cardinal 8 and the English goose vowel is [u] but the latter is less peripheral and in many accents is slightly diphthongized (Wells, 1982). If [u] is to be used with its cardinal value then a transcription of a normal goose vowel would have to have diacritics to denote these facts, and a transcription of a disordered goose would generally require additional diacritics to show in what way it differed from the expected norm, e.g., too open, less rounded, etc. The principle is represented in Table 3.4.

The advantage of using Cardinal qualities as reference points in clinical transcription is that they can be interpreted by anyone trained in the system without knowledge of the accent, or even the language, of the speaker. The disadvantage is the cumbersome nature of the diacritics. If symbols are used with their accent-specific values there is a twofold advantage: generally fewer diacritics are needed so transcription is easier to make and to read, and one can see more readily whether a vowel was produced normally or not. Interpretation of the accent value of the symbol needs, however, to be facilitated by a short account of the accent's vowel qualities using the Cardinal Vowels as reference points (Table 3.5).

Of course, it won't always be the case that use of diacritics is the most appropriate way to represent a disordered vowel. Where we judge the vowel quality to be normal for a different target (so-called substitutions)

### Table 3.4 Using Vowel Symbols with their Cardinal/IPA Values

<table>
<thead>
<tr>
<th>Cardinal Vowel</th>
<th>Accent-Specific Normal</th>
<th>Accent-Specific Disordered</th>
</tr>
</thead>
<tbody>
<tr>
<td>symbol without diacritics</td>
<td>symbol with diacritic/s</td>
<td>symbol with additional diacritic/s</td>
</tr>
<tr>
<td>e.g., CV1 [u]</td>
<td>e.g., [ü]</td>
<td>e.g., [ü̞]</td>
</tr>
</tbody>
</table>

### Table 3.5 Using Vowel Symbols with their Accent-Specific Values

<table>
<thead>
<tr>
<th>Cardinal Vowel</th>
<th>Accent-Specific Normal</th>
<th>Accent-Specific Disordered</th>
</tr>
</thead>
<tbody>
<tr>
<td>reference point for accent-specific vowel quality</td>
<td>symbol without diacritics</td>
<td>symbol with diacritic/s</td>
</tr>
<tr>
<td>e.g., CV8 [u]</td>
<td>e.g., [u]</td>
<td>e.g., [u̞]</td>
</tr>
</tbody>
</table>
we would simply use the symbol for the error vowel, in which case our judgment as to whether an error was *phonetically* deviant or not in terms of the set of normal vowel qualities of the accent would be reflected in the presence or absence respectively of diacritics (Table 3.6). For example, use of a normal [a] TRAP vowel quality where the target is a DRESS vowel would be transcribed with [a], but use of a quality between TRAP and DRESS where this is a quality not typically found in the accent, could be transcribed with [ɛ̞] or [a̝].

There will, however, be cases where this approach is not appropriate. Vowel symbols for qualities present in a speech sample but not present in the normal inventory for the accent will frequently be needed. For example, in a language with only the three vowels /i/, /a/, /u/, a realization of one of them as a mid-central vowel could not really be represented accurately with the target symbol plus diacritics; or the use of a back unrounded vowel for English GOOSE where the simplest recourse might be to [u] rather than [u]. Note here the implication that [u] represents an unrounded equivalent of GOOSE, not of Cardinal Vowel 8: symbols representing opposite lip postures are used with reference to accent-specific vowels, not Cardinal or IPA vowels.

The basic question, then, is whether it is more useful for a clinical transcription to relate to the norms of the speaker’s speech community, or to a set of fixed, absolute universal phonetic qualities, e.g., those of the cardinal or IPA systems. This question arises most acutely when considering vowels, but in fact applies to consonants too—does IPA [ʃ] mean exactly the same to an English-speaking phonetician as it does to a French- or Japanese-speaking one? Is it true that “for the phonetician there is no universal truth independent of the observer” (Ladefoged, 1990: 335), i.e., is our interpretation of IPA and Cardinal Vowel symbols inevitably influenced by our own speech experience? If we accept Ladefoged’s claim, it perhaps suggests that our reference point for clinical transcription should be the norms of the relevant speech community. It also reinforces the widely-held view in clinical phonetics and phonology that familiarity with those norms is essential if the aim

Table 3.6 Using Vowel Symbols with their Accent-Specific Values, Including the Possibility of Cross-Category Realizations

<table>
<thead>
<tr>
<th>Normal Target Realization</th>
<th>Normal Non-Target Realization</th>
<th>Phonetically Deviant Realization</th>
</tr>
</thead>
<tbody>
<tr>
<td>target symbol without diacritics</td>
<td>non-target symbol without diacritics</td>
<td>symbol (target or non-target) with diacritics</td>
</tr>
<tr>
<td>[ɛ̞]</td>
<td>[a]</td>
<td>[ɛ]</td>
</tr>
</tbody>
</table>

"Downloaded By: 10.3.98.104 At: 00:15 15 Oct 2020; For: 978020303103890, chapter3, 10.4324/978020303103890; Sara Howard and Barry Heselwood"
is to phonetically resettle a client’s speech into its sociolinguistic context (Docherty & Khatabb, 2008). Cox (2008) provides a useful discussion of these issues in the clinical context, exploring the tension between abstract, phonemic transcription systems and detailed, narrow phonetic transcription, in the context of Australian English. She notes (Cox, 2008: 332) that “transcription systems are inherently based on compromise between phonemic clarity and phonetic accuracy,” but demonstrates that where a long-established reference system no longer reflects current usage by the speech community, it can be of little relevance for clinical assessment.

The Cardinal Vowel system is still, however, valuable for clinical phoneticians for it enables the vowel qualities of a particular accent—our points of clinical transcriptional reference—to be specified with a descriptive precision that allows others without familiarity with the accent to get a good idea of what they are. The English GOOSE vowel for RP can be described as a slightly mid-centralized Cardinal Vowel 8, and any realization judged normal for an RP speaker can be transcribed as [u]; any production judged to deviate from the norm can be transcribed in such a way as to show the nature of the deviation—[uː] if fully unrounded, [u] if partly unrounded, [u] if raised, etc. The possibility that different people may have slightly different interpretations of cardinal qualities is something we simply have to live with for there is currently no alternative: “to abandon the Cardinal Vowel system is to abandon the only internationally known method of specifying vowels at all accurately” (Ladefoged, 1967: 142).

Plotting the normal vowel qualities of a client’s accent on a vowel quadrilateral is an important first step so that they can be situated in a vowel space whose limits are defined by Cardinal or IPA vowel qualities, and the client’s disordered or immature productions can be compared to them. Ascertaining these norms, however, is problematic because any study of the spoken form of a language is complicated by the fact that across speech communities there are sociolinguistic differences that include differences of pronunciation. Vowels are particularly affected, making it imperative that before diagnosing or remediating any vowel disorder, SLTs should be well acquainted with the vowel qualities and the vowel system of the client’s speech community. Of course this is no easy task, although sources are available containing useful information in this regard; for example, Wells (1982), Foulkes and Docherty (1999), and Ball (1992) have proposed the construction of “clinical socio-linguistic checklists” for use by SLTs, which could include vowel data.

But it is not only across speech communities that vowel differences are found. Sociolinguistic influences cut through speech communities
along lines of gender, class, age, and ethnicity, resulting in a number of vowel variants co-existing not only in the community at large but also in the speech of individuals. Often one variant predominates, but over a period of time another may take its place and become the new “norm,” and new variants may begin to appear (Watt & Tillotson, 2001). Foulkes and Docherty (2007) provide a review of some current changes in vowel qualities in England from a sociolinguistic perspective, and Docherty and Khattab (2008) discuss the clinical implications of sociophonetic differences and the importance of acknowledging and accounting for them in assessment and intervention.

One therefore has to try to distinguish between a speaker using a new variant, and one using a pathologically deviant variant. Confronted by these in a paediatric SLT clinic the temptation might be to intervene without taking the time to distinguish the sociolinguistic from the pathological.

Transcriptional Conventions

Ball (1991: 61), warning that “It is not possible to use a phonemic transcription when you do not know the phonology,” advocates the use of narrow phonetic transcription in the analysis of speech development and speech disorder, a view shared by Kelly and Local (1989: 26) who argue that “it is not possible to have too much phonetic detail . . . we must attend to and reflect everything that we can discriminate.” Indeed, to use the broad vowel phoneme symbols of the target phonology for vowels that are clearly atypical and auditorily distant from the norms, is actually providing no greater detail than the kind of attempts to describe the auditory quality of vowels using orthographic symbols typified by the early effort of Hasluck and Hasluck (1898: 8). They stated that “In ‘Cockney’ pronunciation . . . the tendency is to transfer the sounds from one vowel to another. Thus the a in ‘day’ is unconsciously converted into a sound resembling the i in ‘die’ and when the word ‘die’ is intended it comes out more like doy.” What then, does the present literature offer us to aid in the transcription of vowels? Over the last 10 years or so considerable work has gone into developing extensions to the IPA specifically for the representation of disordered speech (Duckworth, Allen, Hardcastle, & Ball, 1990; Ball & Local, 1996). A number of symbols and diacritics have been agreed and are known as the extensions to the IPA (extIPA) (these are reproduced in Clinical Linguistics and Phonetics 8/3, 263). There are several new consonant symbols for sounds that, as far as is known, only occur in disordered speech, and Ball, Rahilly, and Tench (1996) provide a detailed account of the use of IPA
and extIPA symbols to capture different aspects of vowel articulation. They provide a check-list of features that they argue should all be considered in the narrow transcription of vowels: airstream mechanism and direction of airflow; phonation and voice quality; vertical and horizontal tongue positions; lip shape; secondary articulations; length; and what they term vowel stability, which distinguishes a diphthong, e.g., [ai̯] from a sequence of two consecutive vowel articulations [a:i]. Recent developments to the IPA and extIPA systems have not, however, provided any new vowel symbols; the Cardinal Vowel/IPA system is assumed to define the totality of available vowel space in both articulatory and acoustic/auditory terms (Catford, 1977; Ball, 1993). No doubt this is true for all mature structurally-normal vocal tracts, but it may not always be true in the case of immature or abnormal ones.

As we have already noted, there may be significant differences in the anatomy of the vocal tracts of infants and speakers with, for example, a cleft palate or a glossectomy, and these differences may make it difficult for us to use phonetic symbols with their normally assumed articulatory implications. Very young infants, for example, do not have a developed pharynx (Beck, 2010) and therefore their vocalic sounds do not have the same kind of formant structure as those of older children and adults—they have been termed “quasi” resonant as opposed to fully resonant by some writers (e.g., Oller, 1980) and “vocants” by others (e.g., Martin, 1981) for this reason. Oller and Lynch (1992) discuss the inappropriateness of IPA conventions for early vocalizations and draw attention to a problem that attended early studies; namely, how to interpret acoustic displays without a set of concepts to relate them to. Later studies benefited from a greater understanding of the stages of vocal development (Proctor, 1989), which provided researchers with such concepts, e.g., “squeal,” “coo,” etc. (Oller, 1980; Stark, 1980). Although these approaches have also been criticized (Bauer & Robb, 1992), the descriptors used, or their conventional abbreviations (SQ, etc.) can be incorporated into transcriptions and interpreted narrowly enough to function much like symbols. Regarding IPA conventions, whereas some researchers use IPA transcription conventions for vowel data in children as young as 15 months (e.g., Selby, Robb, & Gilbert, 2000), many researchers have chosen to avoid phonetic transcriptions based on linear strings of segments for vowel production during the early stages of speech development. Instead they have attempted to categorize sounds in terms of broader physiological parameters, often linked to the mandibular cycle of opening and closing that features so strongly in early output and relates to the gradual emergence of vowel-like and consonant-like sounds (Oller, 1980; Koopmans-van Beinum, & van der Stelt, 1986; Bauer & Robb, 1992).
A further challenge for the accurate transcription of vowels is noted by Goldstein and Pollock (2000), who observe that the number of vowels in the sound system of a particular language may affect transcription. They note, for example, that in comparison with the relatively large and complex vowel system of English, Spanish has a small and simple vowel inventory. Goldstein and Pollock (2000: 231) question whether this “less crowded vowel space” may cause small but potentially significant variations in the production of individual vowels in speech development and speech disorder to go unnoticed in perceptual analysis because they do not cross the relatively large vowel spaces occupied by each of the small number of vowels in the system.

Certain peculiarities of vowel resonance can be transcribed using the Voice Quality Symbols (VoQS—Ball, Esling, & Dickson, 1995), and nuances of timbre that characterize individual voice qualities can be captured—i.e., a speaker’s habitual “articulatory setting” (Laver, 1980) such as palatalized or pharyngealized voice. Conventions are also provided for such things as electrolarynx speech, oesophageal speech, diplophonia, harsh voice, and other phenomena affecting the phonatory component of vowels, which as we have seen may have consequences for resonance definition and the perception of vowel quality.

Speakers with anatomically highly deviant vocal tracts, e.g., those who have had large amounts of tongue removed, may produce resonances that even VoQS conventions cannot adequately represent. This highlights a methodological problem that might easily be overlooked. Articulatory settings are judged against a speaker-specific “neutral” setting in which the vocal tract approximates to uniform cross-section and the tongue has “a regularly curved convex shape” (Laver, 1994: 403). Deviant anatomy, whether congenital or resulting from surgical intervention, may prevent approximation to these conditions: if a large part of the anterior section of the tongue has been removed then it cannot assume a curved convex shape. It therefore becomes difficult to heed Wirz and Beck’s otherwise justifiable warning that “neutral should not be confused with any notion of normality” (1995: 49).

The possibility of a neutral setting as defined by Laver may not be a reality for speakers with grossly atypical vocal tract architecture. If a speaker has an abnormally high larynx, for example, or a tongue that protrudes because it is oversized, then to use the VoQS symbols $\L$ (raised larynx), or $\Theta$ (protruded tongue voice), is not really describing a deviation from that speaker’s neutral setting but rather his/her neutral setting itself. The reason they need to be represented is that they deviate from some notional norms, however imprecisely defined, of larynx height or tongue size; otherwise we wouldn’t be drawing attention to
them. At present there are no conventions for distinguishing between abnormal neutral settings and deviations from normal neutral settings. What, for example, should we do in the case of a speaker with an abnormal neutral setting deviating from it by raising his/her larynx yet higher, or pushing the tongue even further out? This kind of problem highlights the need to make a clear decision about whether one’s transcription is aiming to represent the listener’s auditory experience or the speaker’s articulatory behavior (Hewlett, 1985).

The availability of instrumental information to supplement perceptual analysis introduces greater delicacy into our potential representations of speech so that finer distinctions are now possible in a number of phonetic parameters than was the case before. Transcriptional conventions such as those offered by the IPA, extIPA, and VoQS are not always able to represent such fine distinctions. For example, the retracted tongue root diacritic [˫], often used with vowel symbols, cannot indicate the degree of retraction and therefore measurements of that parameter from x-ray or other sources cannot be represented in transcriptions using that convention, neither can dynamic data such as articulator velocity. Another example of instrumentation usefully supplementing transcription is in the area of vowel duration. We know that young children take time to learn to control the durational aspects of vowel production (Stoel-Gammon & Herrington, 1990; Clement & Wijnen, 1994) and that vowel durations are vulnerable across a range of speech impairments, but identifying subtle durational differences may be difficult using perceptual analysis alone. Ball, Rahilly, and Tench (1996) demonstrate how vowel length can be differentiated using phonetic symbols and diacritics: shorter than normal vowel length [ā]; normal length [a]; half-long vowel [aˑ]; long vowel [aː], where these distinctions are to be understood as relative rather than absolute. For precise quantitative information, however, spectrographic analysis has proved a very fruitful method of investigation, although as Blomgren and Robb (1998: 406) note, “determining the duration of a vowel segment is no simple matter.” There is perhaps a widening gulf between what a segmentally-conceived transcription can represent and what increasingly sophisticated instrumentation can reveal. Historically this gulf can be traced to the fact that IPA-type transcription conventions are based on classification of sounds whereas instrumental data are by nature more richly descriptive, i.e., they include much more than just the location and degree of those strictures that phonetic theory identifies as the salient ones and on which classifications are based.

There have been attempts to develop forms of transcription with a more descriptive rather than classificatory orientation by analyzing
speech “horizontally” into separate parameters instead of “vertically” into separate segments (Abercrombie, 1965). Dynamic properties can be represented by lines that move up and down as one follows them from left to right, and degrees of a property can also be iconically represented (Tench, 1978; Ball & Rahilly, 1999). A major disadvantage, however, is that parametric transcriptions are cumbersome to make and difficult to read, although the latter objection can be overcome by combining parametric and IPA-type transcriptions together. A further reason why parametric transcription is not widely employed is probably that, where no instrumental information is available, they would have to be derived from segmental transcriptions anyway by applying phonetic theory. In the case of vowels under these conditions, a parametric transcription showing the movements of the tongue, lips, and velum through time is only going to show what we infer those movements to have been, not necessarily what they actually were. For the parametric approach to be useful it has to use reliable information about the vocal tract’s movements, and this has to come from instrumental sources. An alternative to converting instrumental data into parametric form is simply to show them in graphic form. For example, an EPG print, an EMA trace, a spectrum envelope, etc., or combinations of these, could be indexed to the relevant vowel symbol in a transcription of a particular utterance.

VOWELS AND PROSODY

So far attention has been restricted to vowels without reference to the larger units in which vowels occur. The minimum quantum of speech is the syllable and the structural role of a vowel is to function as the nucleus of a syllable. It is useful to approach this function from two angles: the relations vowels have in syllables with consonants, and their role in larger multisyllabic rhythmic structures.

Relations with Consonants

As syllable nuclei, vowels have intimate spectral and temporal relations with tautosyllabic consonants to the extent that they encode acoustic information about them in formant transitions and duration values (Fry, 1979; Kent & Read, 1992). Acquisition of vowels must entail some sorting of this information and the recognition of which properties are intrinsic to the vowel and which are coarticulatory. At least, this must be the case if the notion “acquisition of vowels” is to be interpreted as acquiring a stock of vowel segments abstracted away from contextual influences (Kuhl, 1980). There is some research showing that some
children may find this kind of processing of vocalic information difficult (Tallal & Stark, 1980).

Many phonologists view vowels as having a key organizational role in syllables, and explain the sequencing of consonants in syllable margins with reference to their resemblance to vowels in terms of sonority. They claim that syllables can be conceptualized as a sonority profile, with the “unmarked” syllable types exhibiting low sonority onsets rising to a high plateau. Vowels provide this high plateau. The predominance of CV structures in immature speech has been noted for a long time (Jakobson, 1968; Kent & Bauer, 1985), while in mature speech sonorant consonants occur adjacent to vowels with the extreme margins being occupied by obstruents. The sonority profile view of the syllable is based on the principle of syntagmatic contrast, that consonants with low sonority followed by vowels with high sonority provide a perceptual aid to the listener who can more easily keep track of events through time if successive elements are maximally different. It also rather neatly correlates with the “frame and content” view of syllable production based on the mandibular cycle of an alternating closed and open buccal chamber (MacNeilage & Davis, 1990). Lindblom’s conception of the syllable as “a gestalt trajectory coursing through the phonetic (articulatory/acoustic/perceptual) space” (1986: 502) accommodates both the sonority and frame-and-content characterizations.

It may be, however, that sonority itself is not the heart of the matter. Ohala (1984) has suggested that signal modulation in general may be the important factor, in which case obstruent-vowel sequences, with their rapid modulation of several acoustic parameters at once, can be seen as particularly good examples. But other contrasts occur that are not explainable in terms of sonority, e.g., the sequences [ju] and [wi] that Ohala points out are much more common than [ji] and [wu] (see the discussion of this in Christman, 1992). Furthermore, Heselwood (1999) has discussed problems with defining sonority adequately, and has argued that some oral stops seem sometimes to function both as sonorants and as obstruents (Heselwood, 1998), which obscures the match between sonority and the openness of the vocal tract.

We have seen earlier how for purposes of general phonetic taxonomy different classificatory frameworks have been used for consonants and vowels, but more recently there has been some discussion in non-linear approaches to phonology as to whether a single set of features can account for the internal structure of both classes (Gierut, Cho, & Dinnsen, 1993). This would be particularly useful for investigating and describing cases of consonant–vowel interaction (Bates & Watson, 1996; also this volume). A number of reports of individual cases have noted
consistent patterns of consonant-vowel interaction (Oller, 1973; Braine, 1974; Camarata & Gandour, 1984; Wolfe & Blocker, 1990), with the co-occurrence of front vowels with coronal consonants and round vowels with labial consonants being repeatedly observed (Gierut et al., 1993; Davis & MacNeilage, 1995), although in a study of 23 young children reported in Vihman (1992) the statistical significance of consonant-vowel associations of this kind were not clear for the sample as a whole, with a large amount of intersubject variability. Similar results for a group of nine children studied by Tyler and Langsdale (1996) suggest that these interactions may not be common across children’s sound systems generally.

Some more striking examples of C-V interaction have been noted where long vowels and diphthongs have been realized as short vowel plus consonant (Harris, Watson, & Bates, 1999), and vice versa (Reynolds, 1990; Song & Demuth, 2008). The locus for these realizations that cross the vowel–consonant divide may be the second mora of the syllable, precisely the one syllabic position in English with a paradigm comprising almost all the vowel and consonant phonemes of the language. The content of the second mora in English varies across the whole sonority hierarchy (Zec, 1995) and across the whole range of articulatory stricture, although interspeaker variation in the study by Song and Demuth calls the moraic explanation at least partially into question. These behaviours may indicate a problem sorting out the phonotactics that some children simplify by disallowing certain consonantal features, e.g., [stop], or [obstruent], while others disallow vowels in that position. It would be interesting to see if such developmental behaviours occur where the ambient language does not have post-vocalic consonants.

Bernhardt and Stemberger (1998) cite research by Öhman indicating that in speech production vowels and consonants are programmed separately, yet they also acknowledge that “vowels can interact with consonants” (p. 101), which suggests that any separation in programming cannot be total. There may even be a developmental progression from integration in infancy towards separation in maturity, which could account for high levels of C-V interactional constraints in babbling and early speech in at least some children (Davis & MacNeilage, 1990,1995; Tyler & Langsdale, 1996; MacNeilage, 1997) that are not evident later.

**Rhythmic Structures**

Natural speech is inherently rhythmic. An important aspect of the rhythm of English is the succession of syllables of contrasting degrees of stress and duration, which can be seen in the same terms as the contrasts
between successive segments within syllables, i.e., as another instance of signal modulation that may aid the listener’s perception. The perceptually most prominent syllables typically occur in words of high semantic content and these can be expected also to often have high imageability, particularly in child-directed speech. Insofar as this is the case, the occurrence of maximally differentiated vowel qualities in stressed syllables correlates with the visual prominence of the word’s referent, e.g., *It’s on the table by the window* produced as [ɪts ɒn ðə tɛibl baɪ ðə wɪndəʊ] or in a more reduced form as something like [ts̚ ɒ ̃ðə tɛibl ba ðə wɪndə] where the reductions affect those items that already have low perceptual prominence. Such reduced forms are intrinsic to the prosodic patterns of typical adult English connected speech production, and contrast with vowel production in words in isolation (Johnson, 2004).

In early speech development vowel production is more accurate in target stressed syllables and children often appear uncertain as to the phonetic content of unstressed ones (Peters, 1995). This affects weak forms of items such as modal and auxiliary verbs, prepositions, pronouns, conjunctions, and copulas, and may also affect children’s ability to recognize stems that take part in vowel alternations (Clark, 1995). Speech that is more towards the syllable-timed end of the syllable-stress timing continuum is reported for two-year-olds by Allen and Hawkins (1980), who also observed deletion of unstressed syllables when word-initial or when adjacent to another unstressed syllable. Grabe, Post, and Watson (1999) present evidence suggesting that rhythmic patterns resulting from the wide variation in vowel duration associated with stress-timing are harder for children to acquire than patterns where variation is less.

It is important therefore when assessing a vowel disorder to take account of rhythm, as it may sometimes be the rhythm that is the problem, rather than the vowels. Acquisition of speech rhythm has a developmental path (Young, 1991) and poor non-linguistic rhythmic skills have been noted in children with speech disorders (Henry, 1990). Furthermore, for connected speech production in multi-word utterances, vowel production may be influenced by the tension between achieving articulatory accuracy for individual gestures, while producing utterances with acceptable rhythm, stress, and rate (Howard, 2007).

Processing capacity may also be relevant in assessing vowel disorders. Vowels are the principal carriers of perceptually prominent pitch movements and durational values associated with stress and tonicity, which encode information about grammatical class, information focus, discourse structure, and speaker affect (Cruttenden, 1997). Attention to these factors may be at the expense of target vowel quality in speakers with restricted linguistic processing capacity (Crystal, 1987).
CONCLUSION

Our examination of the range of instrumental and perceptual approaches to vowel description has shown what a rich range of different kinds of information they can provide and has highlighted the care with which they must be used and interpreted. We have seen how anatomically atypical vocal tracts offer particular challenges for vowel description, both in perceptual and instrumental analysis; and in terms of phonetic transcription, we have seen how the IPA, extIPA, and VoQS systems provide a coherent and flexible resource for the detailed description of normal and atypical vowel productions. We have seen how important it is to be clear about the implications of vowel symbols in phonetic transcription. Is a particular symbol to be interpreted as having specific articulatory values in relation to a speaker’s tongue and jaw movements, or is it meant to have a purely auditory value which, given Perkell’s theory of “motor equivalence” and the anatomical differences found in some speakers with speech impairments, could have been produced by somewhat different vocal tract configurations. We would suggest that it is important to adopt a listener-perspective on vowels where phonetic symbols imply particular auditory qualities that are not, in turn, associated with invariant articulatory values. Aspects of articulation may, of course, be recovered via different types of instrumental analysis if required. Our examination of vowel description and classification has also shown how the Cardinal Vowel system provides us with an important framework for plotting an individual speaker’s vowel system against a known set of reference vowels, but at the same time we have argued that our reference point for the transcription of vowels, whether normal or atypical, should be the accent of the speech community of the individual speaker.

Although it is unlikely that we would ever actually combine all the different perceptual and instrumental methods available in the analysis of any single set of vowel data, it is important that we are aware of them and have an understanding of their relative strengths and weaknesses and of the potential pitfalls associated with their use. In this way we can be aware not only of what our chosen analytical methods show, but also what they do not show; for example, as we have noted, a perceptual analysis of vowel durations does not provide the quantitative accuracy of a spectrographic analysis, and a spectrographic analysis, meanwhile, is unable to provide the detailed qualitative information about types of lip-rounding that visual perceptual analysis is capable of. An understanding of the opportunities and the limitations of different analytic techniques allows us to select the most suitable ones for investigating
our data, depending on our specific aims, which may range from clinical assessment and the assessment of developmental appropriateness to the investigation of intra- or interspeaker variability or the disentangling of sociolinguistic from pathological variation. What is clear above all is the importance of spending time on selecting appropriate methods and carrying them out with care, in order to gain detailed qualitative and quantitative insights into the rich domain of vowel production.

NOTES

1. In light of the discussion in the preceding section concerning the articulatory basis of vowel classification, the terms “Cardinal Vowels” and “IPA vowels” will be used more or less interchangeably: they have in common that they are presented as fixed universal qualities independent of particular languages, although the IPA set contains more non-peripheral vowels; and the International Phonetic Association fully endorsed the principles of the Cardinal Vowel system (Pullum & Ladusaw, 1986).

2. Ladefoged’s suggestion that Cardinal Vowels be identified by underlining (1975) is perhaps not a good one to follow given the standard IPA retracted articulation diacritic.

3. Unfortunately Maurer et al. (1993) do not provide information on the durations of the sustained vowels in their data, and state that “There were no restrictions on the duration of the vocalizations” (Maurer et al., 1993: 130).


REFERENCES


The Contribution of Phonetics


